Practical Assessment, Research, and Evaluation

Volume 16 *Volume 16, 2011* Article 10

2011

Cohen's d vs Alternative Standardized Mean Group Difference Measures

Sorel Cahan

Eyal Gamliel

Recommended Citation
Cahan, Sorel and Gamliel, Eyal (2011) "Cohen's d vs Alternative Standardized Mean Group Difference Measures," *Practical Assessment, Research, and Evaluation*: Vol. 16 , Article 10.
DOI: https://doi.org/10.7275/t1wf-5r27
Available at: https://scholarworks.umass.edu/pare/vol16/iss1/10

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 10, June 2011 ISSN 1531-7714

First Among Others? Cohen's *d* vs. Alternative Standardized Mean Group Difference Measures

Sorel Cahan, *Hebrew University of Jerusalem, Israel* and
Eyal Gamliel, *Ruppin Academic Center, Israel*

Standardized effect size measures typically employed in behavioral and social sciences research in the multi-group case (e.g., η^2, f^2) evaluate between-group variability in terms of either total or within-group variability, such as variance or standard deviation – that is, measures of dispersion about the mean. In contrast, the definition of Cohen's *d*, the effect size measure typically computed in the two-group case, is incongruent due to a conceptual difference between the numerator – which measures between-group variability by the intuitive and straightforward raw difference between the two group means – and the denominator - which measures within-group variability in terms of the difference between all observations and the group mean (i.e., the pooled within-groups standard deviation, S_W). Two congruent alternatives to *d,* in which the root square or absolute mean difference between all observation pairs is substituted for S_W as the variability measure in the denominator of *d,* are suggested and their conceptual and statistical advantages and disadvantages are discussed.

A frequent research design in behavioral and social sciences research involves comparison between the conditional means of a quantitative variable *Y* (e.g., a test score) in groups defined by a categorical treatment variable *X* (e.g., method of instruction). In field studies, the between-group raw mean *Y* variability is typically interpreted as reflecting the inequality between-groups (e.g., schools, countries, or groups defined by socio-demographic characteristics, such as age, gender, SES). In experimental studies, the between-group mean *Y* variability is indicative of the treatment's raw effect size (ES), expressed in *Y*'s metric. In order to remove the original measurement unit and obtain a "pure" number, the raw ES measure (i.e., the between-group means variance or standard deviation) must be standardized (Cohen, 1988).

The use of such standardized ES measures, which are comparable across populations and variables, is mandated by the APA (American Psychological Association, 2009). Moreover, it is highly recommended as a safeguard against the misinterpretation of statistical significance or insignificance as indicators of the scientific or real-life importance or lack thereof, respectively (e.g., Cohen, 1994; Kline, 2004; Thompson, 2002; Wilkinson, 1999). The use of effect size measures is particularly frequent in meta-analyses of studies in behavioral and social sciences (Bowers, Kirby, & Deacon, 2010; Hunter & Schmidt, 2004; Roth, Bevier, Bobko, Switzer, & Tyler, 2001).

Standardized ES Measures: Evaluating Between-group Variability in Terms of Within-group or Across-groups (Total) Variability

Typically, standardized ES measures evaluate *between-group* variability in terms of either (a) *total* (i.e., across groups) or (b) *within-group* variability, where variability is defined in terms of variances or standard deviations.[1] The first type of standardized ES measure is best illustrated by the widely known *eta squared* (η^2), defined as the ratio of between-group variance (S^2_B) and total variance (S^2_T):

[1] A variability measure in the denominator is not required in ES measures defined in terms of the difference between two statistics that are unit free (i.e., "pure"), such as *q* and *h*, which express the difference between two proportions and two correlation coefficients, respectively (Cohen, 1988).

$$\eta^2 = \frac{S^2_B}{S^2_T} = \frac{S^2_B}{S^2_B + S^2_W} \quad (1)$$

where S^2_W is the pooled within-group variance (e.g., Algina, Keselman, & Penfield, 2005; Cohen, 1988). Due to the inclusion of the numerator in the denominator, η^2 ranges between 0 and 1 and can be meaningfully interpreted in terms of the proportion of the total variance lying between groups (e.g., Algina et al., 2005).

The squared root of eta squared (η) is also used as a standardized ES measure:

$$\eta = \sqrt{\eta^2} = \frac{S_B}{S_T} \quad (2)$$

Because $\eta > \eta^2$, η apparently gives a more generous impression of the effect size (Cohen, 1988).

The second type of standardized ES measure is best illustrated by

$$f^2 = \frac{S^2_B}{S^2_W} \quad (3)$$

In contrast to η^2, f^2 evaluates the between-group variance in terms of the *within-group*, rather than total, variance. Hence, its scale has no upper bound. The corresponding standard deviation version is

$$f = \sqrt{f^2} = \frac{S_B}{S_W} \quad (4)$$

f has the advantage of being easily and intuitively interpreted in terms of the "standard deviation of the standardized means" (Cohen, 1988, p. 275).

The Special Case of Two Groups

The two-group case is unique in terms of the possible definitions of standardized ES measures. In this case, the between-group variability can be expressed, in addition to S^2_B or S_B, by the intuitive and straightforward mean difference ($\overline{Y}_1 - \overline{Y}_2$), which, by definition, equals the mean of the *between-group* raw differences between all observation pairs:

$$\overline{Y}_1 - \overline{Y}_2 = E(Y_{1i} - Y_{2j}) \quad (5)$$

where 1 and 2 indicate groups, and *i* and *j* stand for the members of the two groups. In order to obtain a 'pure' number, free of the original measurement unit, the mean difference has to be standardized, that is, divided by a measure of variability (Cohen, 1988, p. 20). Cohen's *d*, which divides the mean difference by the pooled within-groups standard deviation, is a prime example of such a standardized mean difference (SMD) measure (Kelly & Rausch, 2006; McGrath & Meyer, 2006)[2]. Formally, the *population d* is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \quad (6)$$

where μ_1 and μ_2 are the means of the two populations represented in the sample by the two groups, and σ is the standard deviation of either population under the assumption of equal variances (Cohen, 1988, p. 20). The corresponding formula for the sample *d* is:

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_W} \quad (7)$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are the means of the two groups and S_W is the *pooled within-groups* standard deviation (Cohen, 1988, p. 66).[3] *d* is the most frequently used statistic in the context of meta-analysis for experimental and intervention studies in social and behavioral sciences (Hunter & Schmidt, 2004) as well as in educational research (e.g., Bowers et al, 2010; Roth et al., 2001).

Recently, statistical properties of Cohen's *d* – e.g., the quantitative robustness of the two parameters involved in its computation (Algina et al., 2005; Hogarty & Kromrey, 2001; Wilcox & Keselman, 2003) – and *d*'s relative merits and disadvantages over the closely related point-biserial correlation coefficient (r_{pb}; McGrath & Meyer, 2006) have been the focus of renewed interest. The purpose of this paper is to contribute to the critical evaluation of *d* as an SMD measure by (a) pointing to the conceptual incongruence between its numerator and its denominator and examining its origins; and (b) suggesting alternative, more congruent, SMD measures and comparing them to *d* and to one another.

The Conceptual Incongruence of Cohen's d

Common to the various standardized effect size measures in the multi-group case (e.g., Eq.1- Eq.4 above) is the conceptual identity between the expressions in the nominator and the denominator. Both are measures of dispersion about the mean, expressed either as variances or as

[2] Cohen's *d* is an adaptation of the *f* measure to the two-group case, whereby the mean difference is substituted for S_B in the numerator.

[3] It is interesting to note that the intuitive appeal of Cohen's *d* numerator has led Cohen (1977, p. 276) to suggest a generalized version of *d* that applies to any number of groups (namely, the *standardized range of population means*) in which the range $\overline{Y}_{\max} - \overline{Y}_{\min}$ was substituted for the mean difference in the two group case. However, due to interpretation difficulties, this generalized measure has not been adopted by the scientific community.

standard deviations. In contrast, *d* is defined in terms of the ratio between two conceptually different entities (Eq. 2): (a) the *signed raw* difference *between two* (mean) *values* ($\overline{Y}_1 - \overline{Y}_2$); and (b) the (within-group) *positive* root mean square difference (RSMD) *between all observations and the group mean* (the pooled within-groups standard deviation – S_W).

Clearly, this definition is inconsistent with the definition of the other standardized measures of effect size. More importantly, the definition entails a clear incongruence between the discrepancies in the nominator (the *signed* raw between-groups mean difference) and the denominator (a *positive* RMSD measure of dispersion about the mean).

The incongruity relates to two different and independent aspects: (1) the nature of the discrepancy: signed in the numerator vs. positive in the denominator; and (2) the essence of the numerical values involved in its computation: a discrepancy between two (mean) values in the numerator vs. a discrepancy between all observations and their respective group mean in the denominator.

Importantly, these two dimensions differ in terms of their inevitability. (1) above is endemic to the definition of a standardized mean difference, and, therefore, inevitable and common to all possible alternative SMD measures. The rationale underlying such measures (Cohen, 1988) requires that the (signed) raw difference between the two group means (the numerator) be standardized through division by a (necessarily positive) measure of within-group variability (the denominator), such as the RMSD between all observations and the mean (i.e., the Standard Deviation, SD):

$$SD = \sqrt{\sum (X_i - \overline{X})^2 / n} \qquad (8)$$

This aspect of the incongruity between the numerator and the denominator in *d*'s definition can thus be thought of as the unavoidable cost of expressing the between-group variability in the two group case in terms of the intuitively preferable and more informative (signed) raw difference between the two group means, rather than in terms of their variance or standard deviation. The only way to avoid it is to disregard this possibility and express between-group variability in the two group case as well in terms of the standard deviation or variance of the two group means (i.e., the between-group standard deviation S_B or the between-group variance S^2_B, respectively), that is, by substituting *f* (Eq.4) or f^2 (Eq.3) for *d*.

The second aspect of the incongruity in *d*'s definition is the lack of identity between the definitions of the discrepancies in its numerator and denominator – between two (mean) values vs. between all observations and their respective group mean, respectively((2) above). Unlike the first aspect, discussed above, this incongruity is not an inevitable characteristic of standardized mean difference measures, per se. Rather, it is specific to the expression of within-group variability through the arbitrary use of the pooled within-groups *SD* - and other measures of dispersion about the mean, such as the Mean Absolute Deviation from the Mean, *MAD* (Gorard, 2005; Itzhaki, 2003):

$$MAD = \sum |X_i - \overline{X}| / n . \qquad (9)$$

Therefore, instead of using dispersion about the mean measures, alternative variability measures can be defined in terms of root square or absolute mean differences *between all observation pairs* - measures which are both more conceptually congruent and possibly more intuitively meaningful. Yet, like *d*, they are also unbounded.

Two such variability measures suggest themselves:

1. The RMSD between all possible pairs of observations, *E* (Kendall & Stuart, 1977):

$$E = \sqrt{\sum (X_i - X_j)^2 / n} \quad i \neq j, \qquad (10)$$

Like the *SD*, *E* is a RMSD measure. However, it differs from the *SD* by averaging differences between all pairs of observations rather than between the observations and the mean. The two variability expressions are proportional: $E = \sqrt{2}\ SD$ (Kendall & Stuart, 1977).

2. Gini's Mean Absolute Difference between each observation and every other observation, *GMD* (Gini, 1912; Yntema, 1933):

$$GMD = \sum |X_i - X_j| / n \quad i \neq j. \qquad (11)$$

Like the *MAD* (Eq.9 above), *GMD* is defined in terms of absolute deviations. However, it differs from the *MAD* by averaging differences between all pairs of observations rather than between the observations and the mean. Hence, it is conceptually, however not mathematically, equivalent to *E*.

The two (sample) congruent SMD measures resulting from the standardization of the raw mean difference (i.e., *d*'s numerator) by using *E* and *GMD* (d_1 and d_2, respectively) are detailed below, followed by a critical examination of their advantages and disadvantages relative to *d* and to one another.

Two Congruent SMD Measures: d_1 and d_2

1. d_1.

d_1 results from the substitution of E_W for S_W in the denominator of *d*:

$$d_1 = \frac{\overline{Y_1} - \overline{Y_2}}{E_W} \quad (12)$$

where E_W is the pooled within-groups RMSD *between all possible pairs of observations* (Assuming equal group size, E_W equals the average of the two groups' E values). Thus, d_1 expresses the between-group mean difference in terms of the RMS difference between all pairs of observations, rather than between all observations and their mean. Because $E_W = \sqrt{2} * S_W$ (e.g., Kendall & Stuart, 1977), d_1 is smaller than d by a factor of $\sqrt{2}$ (i.e., $d_1 = d/\sqrt{2} \approx 0.7*d$). The proportional relation between d_1 and d implies perfect correlation between them ($r_{d_1 d} = 1$) and identity of their statistical properties as estimators of the respective population parameters.

2. d_2

d_2 results from the substitution of GMD_W for S_W in the denominator of d:

$$d_2 = \frac{\overline{Y}_1 - \overline{Y}_2}{GMD_W} \quad (13)$$

where GMD_W is the pooled within-groups mean of absolute differences *between all pairs of observations* (assuming equal group size, GMD_W equals the average of the two groups' GMD values). Even though GMD and SD are not functionally related, a strong statistical relation exists between them (correlation coefficient of about 0.95; Gorard, 2005). Hence, similar correlation exists between d_2 and d: $r_{d_2 d} = .95$. Furthermore, because $r_{d_1 d} = 1$ (see above), $r_{d_1 d_2} = .95$ as well.

The expected magnitude relations between d and d_2 are an inverse function of the magnitude relations between the pooled within-groups standard deviation (S_W) and the pooled within-groups mean difference (GMD_w) and they depend on the distributional characteristics of X. Assuming normal or rectangular distribution of X, $GMD_w = 1.15 * S_W$ (Greselin & Maffenini, 2005). Consequently, in this case d_2 will be smaller than d by about 13% ($d_2 \approx 0.87*d$) and larger than d_1 by about 23% ($d_2 \approx 1.23*d_1$). Note, however, that d, d_1, and d_2 are expressed on different scales (i.e., units of measurement) and, therefore, their values are not directly comparable.

To conclude, the main argument of this paper is that d_1 and d_2, both of which express within-group variability in terms of the mean difference between all observation pairs, rather than between all observations and the group mean – are conceptually more congruent than Cohen's d and, therefore, preferable to d from this perspective as standardized mean difference measures. What then are the possible arguments against the substitution of d_1 or d_2 for d as the preferred SMD measure? Two types of argument can be raised. The first type is based on considerations of "familiarity":

a. The SD is the most common variability measure in the behavioral and social sciences research. Thus, expressing mean differences in SD units is easily understood and facilitates communication among researchers and professionals.

b. Due to the extensive use of Cohen's d in behavioral and social sciences research, professionals and researchers are familiar with the meaning of its values, including the admittedly arbitrary (McGrath & Meyer, 2006) benchmarks offered by Cohen (1988) for "small", "medium" or "large" effect sizes.

c. Much of the work of meta-analysis is based on d. If another measure were to be used, a great deal of theory would need to be developed to support its use.

Clearly, however, such technical arguments cannot and do not provide justification for adhering to a conceptually inferior SMD measure. Even though familiarity, ease of communication and previous use are important characteristics of statistical measures, they do not exhaust the arguments for choosing between alternative measures. Nor do they figure among the most important considerations. Conceptual considerations and statistical features of the various alternatives are much more critical. Furthermore, the functional relation between d' and d, which allows for their direct calibration, is likely to greatly facilitate the transition from d to d'.

The second type of argument against the substitution of the more congruent and conceptually preferable SMD measures suggested above is statistical. Note, however, that, because d_1 is proportional to d ($d = \sqrt{2} * d_1$), it shares all of d's statistical characteristics. Hence, from the statistical perspective, d can only be preferable to d_2:

1. If several conditions are met (i.e., normal distribution of the population, random sample in each of the two groups, lack of measurement error), d is a more efficient estimator of the respective population parameter than d_2, due to the higher efficiency of the SD relative to GMD as estimator of the population variability (Gorard, 2005; Itzhaki, 2003). However, this advantage of the SD and of d as parameter estimators disappears in real life situations, where the normality assumptions do not hold true (Yitzhaki, 2003), and where there are measurement errors and sampling is non-random (Barnett & Lewis, 1978; Gorard, 2005; Huber, 1981).

2. Given several assumptions, d (and d_1, which is proportional to it; see above) is known to be mathematically related to other group difference measures (e.g., U and the point bi-serial correlation; Cohen, 1988). This functional relation does not hold

with regard to d_2, which is only statistically related to *d*.

3. Because its computation involves the *SD*, *d* (and the mathematically related d_1) has the advantage of ease of mathematical manipulation relative to d_2.

Conclusions and Implications

The main conclusion of our analysis is that, even though Cohen's *d* is a widely used SMD measure, it has several drawbacks. Notable among them is the incongruence between the numerator and denominator, illustrated in this paper for the first time (to the best of our knowledge). Each of the two alternative SMD measures suggested in this paper is clearly preferable to *d* in this respect. Both express the (signed) raw difference between the two group means in terms of the expected within-group absolute or root square mean difference between two randomly sampled observations. For example, d_2= 0.5 indicates that the (signed) difference between the two group means is half the within-group average (absolute) difference between observation pairs.

This conceptual advantage of d_1 and d_2 relative to Cohen's *d* becomes more visible and impressive if the mean difference between the two groups (i.e., the numerator of *d*, d_1 and d_2) is conceived as the mean of the between-group differences between all observation pairs (which, by definition, equals the difference between the two group means; Eq. 5). Clearly, evaluating the *between-group* difference between all observations pairs in terms of the corresponding *within-group* difference between them (d_1 and d_2's denominators) makes more sense than evaluating it in terms of the difference between all observations *and their group mean* (the denominator of *d*).

Furthermore, the conceptual superiority of the suggested SMD measures is not offset by statistical or computational considerations. First, d_1 is identical to *d* from this perspective. Hence, at most, these considerations should lead to preferring d_1 to d_2, rather than preferring *d* to d_2. Secondly, *d*'s statistical superiority relative to d_2 is only hypothetical: In real life situations d_2 has higher estimation efficiency than Cohen's *d*. Furthermore, ease of computation can no longer be advocated as a critical consideration in the choice of statistical measures, certainly not as a valid counterargument to conceptual considerations.

Whether the conceptual and statistical advantages of d_1 or d_2 presented in this paper justify their substitution for Cohen's *d* as the SMD of choice is an open question. This kind of dilemma necessarily involves value judgments, has no unequivocal correct answer and can be resolved only by future debate in the methodological literature. Such a discourse is closely related to another debate, namely the pros and cons of various measures of variability, in general, and the universal use of the *SD*, in spite of its well documented shortcomings (Gorard, 2005; Yitzhaki, 2003), in particular. We hope that this paper will contribute to the initiation of such constructive discussions.

References

Algina, J., Keselman, H.J., & Penfield, R.D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317-328.

American Psychological Association. (2009). *Publication manual of the American Psychological Association (6th ed.)*. Washington, DC: American Psychological Association.

Bowers, P.N., Kirby, J.R., Deacon, S.H. (2010). The Effects of Morphological Instruction on Literacy Skills: A Systematic Review of the Literature. *Review of Educational Research, 80*, 144-179.

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

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist, 49*(12), 997-1003.

Gini, C. (1912). *Variabilità e mutabilità*. Bologna: Tipographia di Paolo Cuppini.

Gorard, S. (2005). Revisiting a 90-year-old debate: The advantages of the mean deviation. *British Journal of Educational Studies, 53*, 417–430.

Greselin, F., & Maffenini, W. (2005). *Confidence intervals for Gini's mean difference: Simulation results*. Paper presented in International Conference in Memory of C. Gini and M.O. Lorenz, Sienna.

Hogarty, K.Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Huber, P. (1981). *Robust statistics*. New York: Wiley.

Hunter, J. E., & Schmidt, F. C. (2001). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363-385.

Kendall, M.G. & Stuart, A., (1977). *Advanced theory of statistics*. London: Griffin.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

McGrath, R.E., & Meyer, G.J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386-401.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*(2), 297-330.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.

Wilcox, R.R., & Keselman, A. J. (2003). Modern robust data analysis methods. Measures of central tendency. *Psychological Methods, 8,* 254-274.

Wilkinson, L., and The task force on statistical inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

Barnett, V., & Lewis, T. (1978) *Outliers in statistical data.* Chichester: Wiley.

Yitzhaki, S. (2003). Gini's mean difference: A superior measure of variability for non-normal distributions, *Metron, LXI*, 285-316.

Yntema, D.B. (1933). Measures of the inequality in the personal distribution of wealth or income. *Journal of the American Statistical Association, 28,* 423-433.

Citation:

Cahan, Sorel & Eyal Gamliel (2011) First Among Others? Cohen's *d* vs. Alternative Standardized Mean Group Difference Measures. *Practical Assessment, Research & Evaluation*, 16(10). Available online: http://pareonline.net/getvn.asp?v=16&n=10.

Notes

A previous version of this paper was presented at the Annual Meeting of the American Educational Research Association (AERA). New-York, March, 2008.

Authors:

Sorel Cahan
School of Education
Hebrew University of Jerusalem
Israel
sorelc@mscc.huji.ac.il

Eyal Gamliel
Behavioral Sciences Department
Ruppin Academic Center
Israel
eyalg@ruppin.ac.il