What measure of effect size when comparing two groups based on their means?

Marie Delacre[1], Christophe Ley[2], Daniel Lakens[3], & Christophe Leys[1]

[1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

[2] Universiteit Gent, Department of Applied Mathematics, Computer Science and Statistics, Gent, Belgium

[3] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, the Netherlands

Author Note

Correspondence concerning this article should be addressed to Marie Delacre, CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Abstract

Keywords: keywords

Word count: X

What measure of effect size when comparing two groups based on their means?

Intro

During decades, researchers in social science (Henson & Smith, 2000) and education (Fan, 2001) have overestimated the ability of the null hypothesis (H0) testing to determine the importance of their results. The standard for researchers in social science is to define H0 as the absence of effect (Meehl, 1990). For example, when comparing the mean of two groups, researchers commonly test the H0 that there is no mean difference between groups (Steyn, 2000). Any effect that is significantly different from zero will be seen as sole support for a theory.

Such an approach has faced many criticisms among which the most relevant to our concern is that the null hypothesis testing highly depends on sample size: for a given alpha level and a given difference between groups, the larger the sample size, the higher the probability of rejecting the null hypothesis (Fan, 2001; Kirk, 2009; Olejnik & Algina, 2000; Sullivan & Feinn, 2012). It implies that even tiny differences could be detected as statistically significant with very large sample sizes (McBride, Loftis, & Adkins, 1993)[1].

Facing this argument, it has become an adviced practice to report the *p*-value assorted by a measure of the effect size, that is, a quantitative measure of the magnitude of the experimental effect (Cohen, 1965; Fan, 2001; Hays, 1963). This practice is also highly endorsed by the American Psychological Association (APA) and the American Educational Research Association (AERA) (American Educational Research Association, 2006; American Psychological Association, 2010). However, only a limited number of studies have properly

[1] Tiny differences might be due to sampling error, or to other factors than the one of interest: even under the assumption of random assignent (which is a necessary but not sufficient condition), it is almost impossible to be sure that the only difference between two conditions is the one defined by the factor of interest. Other tiny factors of no theoretical interest might slighly influence results, making the probability of getting an actual zero effect very low. This is what Meehl (1990) calls 'systematic noise'.

reported effect size in the last decades.

Generally, there is a high confusion between the effect size and other related concepts such as the Clinical significance. Moreover, there are several situations that call for effect size measures and, in the current litterature, it is not always easy to know which measure to use in which context. We will therefore begin this paper with 3 sections in which we will:

1. Clearly define what is a measure of effect size;
2. List the different situations that call for effect sizes measures;
3. Define required properties of the effect size estimators depending on the specific situation.

Moreover, it is highly recommended to compute a confidence interval around the point effect size. In a fourth section, we will therefore summarize in how far it is an added value to mention the confidence interval around the effect size.

After these general adjustments, we will focus our attention on "between-subject" designs where individuals are randomly assigned into one of two independent groups and group scores are compared based on their means[2]. Because it has been widely argued that there are many fields in psychology where the assumption of equal variances between two populations is ecologically unlikely (Delacre, Lakens, & Leys, 2017; Erceg-Hurn & Mirosevich, 2008; Grissom, 2000), it is becoming more common in statistical software to present a t-test that does not hold under this assumption by default, namely the Welch's t-test (e.g., R, Minitab). However, similar issues for the measures of effect sizes have received less attention (Shieh, 2013), and Cohen's d_s remains persistent[3]. One possible reason is that researchers cannot find a consensus on which alternative should be used (Shieh, 2013). We will limit our study to the standardized mean difference, called the d-family, because it is the

[2] We made this choice because *t*-tests are still the most commonly used tests in the field of Psychology.

[3] For example, in Jamovi, Cohen's ds is provided, independently of whether one performs Student's or Welch's t-test.

dominant family of estimators of effect size when comparing two groups based on their means (Peng, Chen, Chiang, & Chiang, 2013; Shieh, 2013), and we will see that even in this very specific context, there is little agreement between researchers as to which is the most suitable estimator. According to us, the main reason is that it is difficult, based on currently existing measures, to optimally serve all the purposes of an effect size measure. Throughout this section, we will:

1. Present the main measures of the *d*-family that are proposed in the literature,related to the purpose they serve, and introduce a new one, namely the "transformed Shieh's *d*" that should help at reaching all the purposes simultaneously;

2. Present and discuss the results of simulations we performed, in order to compare existing measures and our newly introduced one;

3. Summarize our conclusions in practical recommendations. In this section, we will provide useful tools (i.e., an R package) to compute relevant measures of effect sizes and related information.

Measure of effect size: what it is, what it is not

The effect size is commonly referred to as the practical significance of a test. Grissom & Kim (2005) define the effect size as the extent to which results differ from what is implied by the null hypothesis. In the context of the comparison of two groups based on their means, depending on the defined null hypothesis (considering the absence of effect as the null hypothesis), we could define the effect size either as the magnitude of differences between parameters of two populations groups are extracted from (e.g. the mean; Peng & Chen, 2014) or as the magnitude of the relation between one dichotomous factor and one dependent variable (American Educational Research Association, 2006). Both definitions refer to the most famous families of measures of effect sizes (Rosenthal, 1994): the *d*-family and the *r*-family.

Very often, the contribution of the measures of effect size is overestimated. First, benchmarks about what should be a small, medium or large effect size might have contributed to viewing the effect size as a measure of the importance or the relevance of an effect in real life, but it is not (Stout & Ruble, 1995). The effect size is only a mathematical indicator of the magnitude of a difference, which depends on the way a variable is converted into numerical indicator. In order to assess the meaningfulness of an effect, we should be able to relate this effect with behaviors/meaningful consequences in the real world (Andersen, McCullagh, & Wilson, 2007). For example, let us imagine a sample of students in serious school failure who are randomly divided into two groups: an experimental group following a training program and a control group. At the end of the training, students in the experimental group have on average significantly higher scores on a test than students in the control group, and the difference is large (e.g. 30 percents). Does it automatically mean that students in the experimental condition will be able to pass to the next grade and to continue normal schooling? Whether the computed magnitude of difference is an important, meaningful change in everyday life refers to the interpretation of treatment outcomes and is neither a statistical nor mathematical concept, but is related to the underlying theory that posits an empirical hypothesis. This concept is sometimes called *Clinical significance* (Grissom & Kim, 2012; Thompson, 2002) or *Social significance* (Tyler, 1931) in the current literature. However, in our conception, we should use a more general term and we propose to rename this concept to *Applied significance*[4].

Second, in the context of the comparison of two groups based on their means, the effect size should not replace the null hypothesis testing. Statistical testing allows the researcher to determine whether the oberved departure from H0 occured by chance or not (Stout & Ruble, 1995), while effect size estimators allow to assess the practical signficance of an effect, and as reminds Fan (2001): *"a practically meaningful outcome may also have occured by chance,*

[4] In our conception Applied significance encompasses all what refers to the relevance of an effect in real life, such as for instance clinical, personnal, social, professionnal relevance

and consequently, is not trustworthy" (p.278). For this reason, the use of confidence intervals around the effect size estimate is highly recommended (Bothe & Richardson, 2011).

Different purposes of effect size measures

Effect size measures can be used in an *inferential* perspective:

- The effect sizes from previous studies can be used in a prior power analysis when planning a new study (Lakens, 2013; Prentice & Miller, 1990; Stout & Ruble, 1995; Sullivan & Feinn, 2012; Wilkinson & the Task Force on Statistical Inference, 1999);

- We can compute confidence limits around the point estimator (Shieh, 2013) in order to replace conventional hypothesis testing : if the null hypothesis area is out of the confidence interval, we can conclude that the null hypothesis is false.

Measures of effect size can also be used in a *comparative* perspective, that is, to assess the stability of results across designs, analysis, samples sizes (Wilkinson & the Task Force on Statistical Inference, 1999). This includes

- the comparison of results from 2 or more studies (Prentice & Miller, 1990);

- the incorporation of results in meta-analysis (Lakens, 2013; Li, 2016; Nakagawa & Cuthill, 2007; Stout & Ruble, 1995; Wilkinson & the Task Force on Statistical Inference, 1999).

Finally, effect size measures can be used for *interpretative* purposes, namely to assess the practical significance of a result (beyond statistical significance; Lakens, 2013; American Psychological Association, 2010; Prentice & Miller, 1990).

Properties of a good effect size estimator

The empirical value of an estimator (called estimate) depends on the sampling, in other words, different samples extracted from the same population will of course lead to different estimates for a same estimator. The *sampling distribution* of the estimator is the distribution of all estimates, based on all possible samples of size n extracted from one population. Studying the sampling distribution is very useful, as it allows us to assess the qualities of estimator. More specifically, three desirable properties a good estimator should possess for inferential purposes are: **unbiasedness**, **consistency** and **efficiency** (Wackerly, Mendenhall, & Scheaffer, 2008).

An estimator is unbiased if the distribution of estimates is centered around the true population parameter. On the other hand, an estimator is positively (or negatively) biased if the distribution is centered around a value that is higher (or smaller) than the true population parameter (see Figure 1). In other words, the bias tells us if estimates are good, on average. The *bias* of a point estimator $\hat{\delta}$ can be computed as

$$\hat{\delta}_{bias} = E(\hat{\delta}) - \delta \tag{1}$$

where $E(\hat{\delta})$ is the expectation of the sampling distribution of the estimator (i.e. the population average) and δ is the true (population) parameter.

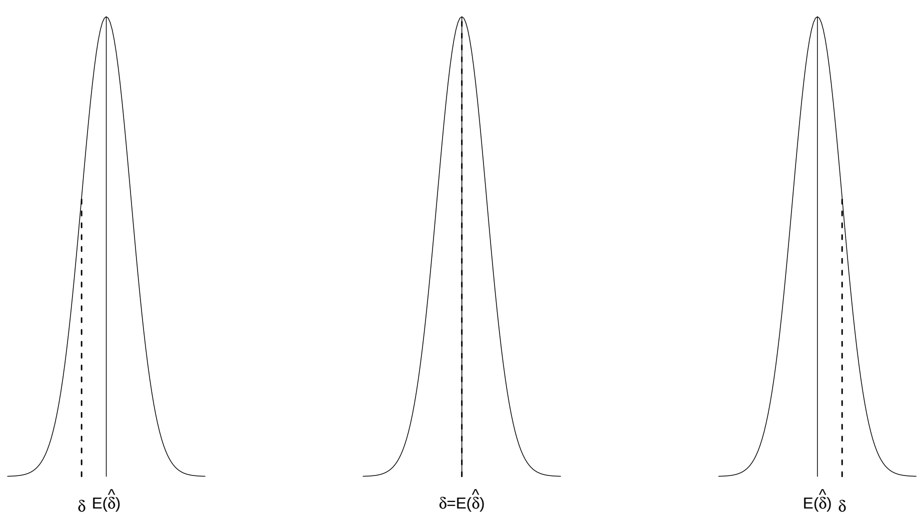

Figure 1. Samplig distribution for a positively biased (left), an unbiased (center) and a negatively biased estimator (right)

Moreover, since there is a strong relationship between the bias and the size of any estimator (the larger an estimator, the larger the bias), it might be interesting to also define the *relative bias* as the ratio between the bias and the population parameter:

$$\hat{\delta}_{relative\ bias} = \frac{E(\hat{\delta}) - \delta}{\delta} \tag{2}$$

While the bias informs us about the quality of estimates on average, in particular their capacity of lying close to the true value, it says nothing about individual estimates. Imagine a situation where the distribution of estimates is centered around the real parameter but with such a large variance that some point estimates are very far from the center. This would be problematic, since we then do not know if this estimate, based on the sample at hand, is close to the truth or far off. Therefore it is not only essential for an estimator to be unbiased, but the variability of its sampling distribution should also ideally be small. Put simply, we hope that *all* possible estimates are close enough of the true population parameter, in order to be sure that for *any* estimate, one has a correct estimation of the real parameter. Among two unbiased estimators $\hat{\delta}_1$ and $\hat{\delta}_2$, we therefore say that $\hat{\delta}_1$ is **more efficient** than $\hat{\delta}_2$ if

$$Var(\hat{\delta}_1) \leq Var(\hat{\delta}_2) \tag{3}$$

Where $Var(\hat{\delta})$ is the variance of the sampling distribution of the estimator $\hat{\delta}$. Among all unbiased estimators, the more efficient will be the one with the smallest variance [5]. Again, the variance of an estimator $\hat{\delta}$ is a function of its size (the larger the estimator, the larger the variance) and therefore, we might be interested in computing the *relative variance* as the ratio between the variance and the square of the population estimator:

[5] The famous Crámer-Rao inequality provides a theoretical lower bound for the variance of unbiased estimators. An estimator reaching this bound is therefore most efficient.

$$\hat{\delta}_{relative\ variance} = \frac{Var(\hat{\delta})}{\delta^2} \tag{4}$$

Note that both unbiasedness and efficiency are very important. An unbiased estimator with such a large variance that somes estimates are extremely far from the real parameter is as undesirable as a parameter which is highly biased. In some situations, it is better to have a slightly biased estimator with a tight shape around the biased value (so that each estimate remains relatively close to the true parameter and one can apply bias correction techniques) rather than an unbiased estimator with a large variance (Raviv, 2014).

Finally, the last property of a good point estimator is **consistency**: consistency means that the bigger the sample size, the closer the estimate is to the population parameter. In other words, the estimates *converge* to the true population parameter.

Beyond the inferential properties, Cumming (2013) reminds that an effect size estimator needs to have a constant value across designs in order to be easily interpretable and to be included in meta-analysis. In other words, it should achieve the property of **generality**.

Confidence interval around a point estimator

We already mentioned that confidence interval around a point estimate could replace conventional hypothesis testing. A confidence interval contains all the information that a p-value of a test based on the same estimator does: if the area of the null hypothesis is out of the $(1-\alpha)$-confidence interval, then the hypothesis test would also result in a p-value below the nominal alpha level. Hypothesis tests and confidence intervals based on the same statistical quantity (this is an essential requirement) are thus directly related. At the same time, the intervals provide extra information about the precision of the sample estimate for inferential purposes, and therefore on how confident we can be in the observed results (Altman, 2005; Ellis, 2015): the narrower the interval, the higher the precision. On the other hand, the wider the confidence interval, the more the data lacks precision (for example, because the sample size is too small).

Different measures of effect sizes

The d-family effect sizes are commonly used with "between-subject" designs where individuals are randomly assigned into one of two independent groups and groups scores means are compared. The population effect size is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{5}$$

where both populations follow a normal distribution with mean μ_j in the j^{th} population (j=1,2) and common standard deviation σ. They exist different estimators of this effect size measure varying as a function of the chosen standardizer (σ). For all estimators, the mean difference is estimated by the difference $\bar{X}_1 - \bar{X}_2$ of both sample means. When used for inferential purposes, some estimators require both the assumptions of normally distributed data and the equality of variances, while others rely solely on the assumption of normality.

Alternatives when variances are equal between groups

When we have good reasons to assume equality of variances between groups, then the most common estimator of δ is Cohen's d_s where the sample mean difference is divided by a pooled error term (Cohen, 1965):

$$Cohen's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)\times SD_1+(n_2-1)\times SD_2}{n_1+n_2-2}}} \tag{6}$$

Where SD_j is the standard deviation and n_j the sample size of the j^{th} sample (j=1,2). The reasoning behind this measure is to make use of the fact that both samples share the same population variance (n.d.), hence we achieve a more accurate estimation of the population variance by pooling both estimates of this parameter (i.e SD_1 and SD_2). Since the larger the sample size, the more accurate the estimate, we give more weight to the estimate based on the larger sample size.

Cohen's d_s is directly related with Student's t-test:

$$Cohen's\ d_s = t_{student} \times \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \leftrightarrow t_{student} = cohen's\ d_s \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{7}$$

The relationship described in equation 7 between Cohen's d_s and the Student's t statistic, whose distribution is well known, allows us to theoretically determine the sampling distribution of Cohen's d_s, and therefore, its theoretical expectency, bias and variance under the assumption of normality and equality of variances. As one can deduce from Table 1, with $n_1 + n_2 \geq 4$, the bias is always positive (IF N=2, NEGATIVE BIAS, AND IF N=3, INFINITE BIAS). While the bias tends to zero with large sample size, it is substantial with small sample size, even under the assumptions of normality and equal variances (Lakens, 2013) and, for this reason, Hedges & Olkin (1985) has defined a bias-corrected version:

$$Hedge's\ g_s = Cohen's\ d_s \times (1 - \frac{3}{4 \times (n_1 + n_2) - 9}) \tag{8}$$

Not only is this corrected estimator unbiased, but also has a smaller variance than Cohen's d_s, especially with small sample size. While the pooled error term is the best choice when variances are equal between groups (Grissom & Kim, 2001), it may not be well advised for use with data that violate this assumption (Cumming, 2013; Grissom & Kim, 2001, 2005; Kelley, 2005, 2005; Shieh, 2013). When variances are unequal between groups, the expression in equation 5 is no longer valid because both groups don't share a common population variance. If we pool the estimates of two unequal population variances, the estimator of effect size will be lower as it should be in case of positive pairing (i.e. the group with the larger sample size is extracted from the population with the larger variance) and larger as it should be in case of negative pairing (i.e. the group with the larger sample size is extracted from the population with the smaller variance). Because the assumption of equal variances across populations is very rare in practice (Cain, Zhang, & Yuan, 2017; Delacre et al., 2017; Delacre, Leys, Mora, & Lakens, 2019; Erceg-Hurn & Mirosevich, 2008; Glass, Peckham, & Sanders, 1972; Grissom, 2000; Micceri, 1989; Yuan, Bentler, & Chan, 2004), both Cohen's d_s and Hedge's g_s should be abandoned in favor of a robust alternative to unequal population variances.

Table 1. *Expentency, bias and variance of different estimators*

Estimator	Expectency	Bias	Variance
$Cohen's\ d_s$	$\delta_{cohen} \times \frac{\sqrt{\frac{N-2}{2}} \times \Gamma(\frac{N-3}{2})}{\Gamma(\frac{N-2}{2})}$	$\delta_{cohen} \times \left(\frac{\sqrt{\frac{N-2}{2}} \times \Gamma(\frac{N-3}{2})}{\Gamma(\frac{N-2}{2})} - 1\right)$	$\frac{N-2}{(N-4) \times \frac{n_1 n_2}{N}} \times \left(1 + \frac{n_1 n_2}{N} \times \delta^2_{cohen}\right) - \delta^2_{cohen} \times \left[\frac{\sqrt{\frac{N-2}{2}} \times \Gamma(\frac{N-3}{2})}{\Gamma(\frac{N-2}{2})}\right]^2$
	$\approx \frac{\delta_{Cohen}}{\left(1 - \frac{3}{4(n_1+n_2)-9}\right)}$	$\approx \delta_{Cohen} \left[\frac{1}{\left(1 - \frac{3}{4(n_1+n_2)-9}\right)} - 1\right]$	$\frac{N-2}{(N-4) \times \frac{n_1 n_2}{N}} \times \left(1 + \frac{n_1 n_2}{N} \times \delta^2_{cohen}\right) - \delta^2_{cohen} \times \left[\frac{1}{\left(1 - \frac{3}{4(n_1+n_2)-9}\right)}\right]^2$
$Hedges's\ d_s$	δ_{cohen}	/	$Var(Cohen's\ d_s) \times \left[\frac{\Gamma(\frac{N-2}{2})}{\sqrt{\frac{N-2}{2}} \times \Gamma(\frac{N-3}{2})}\right]^2$
			$Var(Cohen's\ d_s) \times \left[1 - \frac{3}{4(n_1+n_2)-9}\right]^2$
$Glass's\ d_s$	$\delta_{glass} \times \frac{\sqrt{\frac{n_c-1}{2}} \times \Gamma(\frac{n_c-2}{2})}{\Gamma(\frac{n_c-1}{2})}$	$\delta_{glass} \times \left(\frac{\sqrt{\frac{n_c-1}{2}} \times \Gamma(\frac{n_c-2}{2})}{\Gamma(\frac{n_c-1}{2})} - 1\right)$	$\frac{n_c-1}{(n_c-3) \times \frac{n_1 n_2}{N}} \times \left(1 + \frac{n_1 n_2}{N} \times \delta^2_{glass}\right) - \delta^2_{glass} \times \left[\frac{\sqrt{\frac{n_c-1}{2}} \times \Gamma(\frac{n_c-2}{2})}{\Gamma(\frac{n_c-1}{2})}\right]^2$
$Shieh's\ d_s$	$\delta_{Shieh} \times \frac{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}{\Gamma(\frac{df}{2})}$	$\delta_{Shieh} \times \left(\frac{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}{\Gamma(\frac{df}{2})} - 1\right)$	
		$with\ df = \frac{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)^2}{\frac{(SD_1^2/n_1)^2}{n_1-1} + \frac{(SD_2^2/n_2)^2}{n_2-1}}$	

Note. N is the total sample size (n1+n2). Equations to compute expectency, bias and variances of Cohen's ds, Hedge's gs and Glass's ds require the assumptions of normality and equal variances.

Alternatives when variances are unequal between populations

In his review, Shieh (2013) mentions three options available in the literature to deal with the case of unequal variances: the sample mean difference divided by (A) the non pooled average of both variance estimates, (B) the Glass's d_s and (C) the Shieh's d_s.

The sample mean difference, divided by the non pooled average of both variance estimates was suggested by Cohen (1988). We immediately exclude this alternative because it suffers from several limitations:

- it results in a variance term of an artificial population and is therefore very difficult to interpret (Grissom & Kim, 2001);

- unless both sample sizes are equal, the variance term does not correspond to the variance of the mean difference (Shieh, 2013);

- unless the mean difference is null, the measure is biased. Moreover, the bigger the sample size, the larger the variance around the estimate.

When comparing one control group with one experimental group, Glass, McGav, & Smith (2005) recommend using the standard deviation SD of the control group as standardizer. It is also advocated by Cumming (2013), because, according to him, it is what makes the most sense, conceptually speaking. This yields

$$Glass's\ d_s = \frac{\bar{X}_{experimental} - \bar{X}_{control}}{SD_{control}} \tag{9}$$

One argument in favour of using the SD of the control group as standardizer is the fact that it is not affected by the experimental treatment. When it is easy to identify which group is the "control" one, it is therefore convenient to compare the effect size estimation of different designs studying the same effect. However, defining this group is not always obvious (Coe, 2002). This could induce large ambiguity because depending of the chosen SD as standardizer, measures could be substantially different (Shieh, 2013). The Glass d_s also has limitations when used for inference. The standardizer is estimated from only a part of the sample (since only one group is taken into account in variance estimation), which might potentially reduce accuracy (think of the desirable property of consistency). While being a consistant measure, Glass's d_s can be shown to be highly positively biased when there are less than 300 participants (Hedges, 1981; Olejnik & Hess, 2001), especially for small effect sizes.

Kulinskaya & Staudte (2007) were the first to advice the use of a standardizer that take the sample sizes allocation ratios into account, in addition to the variance of both samples. Shieh (2013), following Kulinskaya & Staudte (2007), proposed a modification of the exact *SD* of the sample mean difference:

$$Shieh's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SD_1^2/q_1 + SD_2^2/q_2}}; \quad q_j = \frac{n_j}{N}(j = 1, 2) \tag{10}$$

where $N = n_1 + n_2$. Shieh's d_s is directly related with Welch's t-test:

$$Shieh's\ d_s = \frac{t_{welch}}{\sqrt{n_1 + n_2}} \leftrightarrow t_{welch} = Shieh's\ d_s \times \sqrt{n_1 + n_2} \tag{11}$$

The exact distribution of Welch's t statistic is more complicated than the exact distribution of Student's t statistic, but it can be approximated (Shieh, 2013; Welch, 1938). Again, it allows us to theoretically determine the sampling distribution of Shieh's d_s, and therefore, its theoretical expectency, bias and variance under the assumption of normality (see Table 1).

As one can deduce from Table 1, with $n_1 + n_2 \geq 4$, the bias is always positive (IF N=2, NEGATIVE BIAS, AND IF N=3, INFINITE BIAS). While the bias tends to zero with large sample size, it is substantial with small sample size, even under the assumptions of normality and equal variances (Lakens, 2013) and, for this reason, Hedges & Olkin (1985) has defined a bias-corrected version. Not only is this corrected estimator unbiased, but also has a smaller variance than Cohen's d_s, especially with small sample sizes.

According to the statistical properties of Welch's statistic under heteroscedasticity, it does not appear possible to define a proper standardised effect size without accounting for the relative group size of subpopulations in a sampling scheme. At the same time, the lack of generality caused by taking this specificity of the design into account has led Cumming (2013) to question its usefulness in terms of interpretability: when keeping constant the mean difference ($\bar{X}_1 - \bar{X}_2$) as well as SD_1 and SD_2, Shieh's d_s will vary as a function of the sample sizes allocation ratio (the dependency of Shieh's d_s value on the sample sizes allocation ratio is detailed and illustrated in Appendix 1, and also in the following shiny application: https://mdelacre.shinyapps.io/improve-the-interpretability-of-shieh-s-d-shiny-app/).

Fortunately, this apparent paradox can be resolved. It is possible to find a modified measure of Shieh's d_s that does not depend on sample sizes ratio, namely by answering the following question: "whatever the real sample sizes ratio, what value of Shieh's d_s would have been computed if design were balanced (i.e. $n_1 = n_2$), keeping all other parameters constant?"

It can be shown that the relationship between Shieh's δ when samples sizes are equal between groups and Shieh's δ for any other sample sizes allocation ratios can be expressed as follows:

$$Shieh's\ \delta_{n_1=n_2} = Shieh's\ \delta \times \frac{(nratio+1) \times \sigma_{n_1 \neq n_2}}{2 \times \sigma_{n_1=n_2} \times \sqrt{nratio}} \tag{12}$$

with

$$nratio = \frac{n_1}{n_2}$$

$$\sigma_{n_1=n_2} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

$$\sigma_{n_1 \neq n_2} = \sqrt{(1-\frac{n_1}{N}) \times \sigma_1^2 + (1-\frac{n_2}{N}) \times \sigma_2^2}$$

$Shieh's\ \delta_{n_1=n_2}$ can thefore be estimated using this equation:

$$Shieh's\ d_s^* = Shieh's\ d_s \times \frac{(nratio+1) \times SD_{n_1 \neq n_2}}{2 \times SD_{n_1=n_2} \times \sqrt{nratio}} \tag{13}$$

with

$$SD_{n_1=n_2} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

and

$$SD_{n_1 \neq n_2} = \sqrt{(1-\frac{n_1}{N}) \times SD_1^2 + (1-\frac{n_2}{N}) \times SD_2^2}$$

$Shieh's\ d_s^*$ can be compared across two different studies using different sample sizes allocation ratio and could thus be included in meta-analysis.

Monte Carlo Simulations

Simulation 1: assessing the bias, efficiency and consistency of 5 estimators.

Method. We performed Monte Carlo simulations using R (version 3.5.0) to assess the bias, efficiency and consistency of Cohen's d_s, Hedge's g_s, Glass's d_s (using respectively the sample SD of the first or second group as a standardizer), Shieh's d_s and our transformed measure of Shieh's d_s, that we will note later d_s^*.

A set of 100,000 datasets were generated for 1,008 scenarios as a function of different criterions that will be explained below. In 252 scenarios, samples were extracted from a normally distributed population and in 756 scenarios, samples were extracted from non normal population distributions. In order to assess the quality of estimators under realistic deviations from the normality assumption, we referred to the review of Cain et al. (2017). Cain et al. (2017) investigated 1,567 univariate distributions from 194 studies published by authors in Psychological Science (from January 2013 to June 2014) and the American Education Research Journal (from January 2010 to June 2014). For each distribution, they computed the Fisher's skewness (G1) and kurtosis (G2):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{\sqrt{(m_2)^3}} \tag{14}$$

with s = standard deviation, n = sample size, m_2 = second centered moment and m_3 = third centered moment.

$$G_2 = \frac{n-1}{(n-2)(n-3)} \times [(n+1)(\frac{m_4}{(m_2)^2} - 3) + 6] \tag{15}$$

with s = standard deviation, n = sample size and m_3=third centered moment. They found values of kurtosis from G2 = -2.20 to 1,093.48. According to their suggestions, throughout our simulations, we kept constant the population kurtosis value at the 99th percentile of their distribution of kurtosis, i.e. G2=95.75. Regarding skewness, we simulated population parameter values which correspond to the 1st and 99th percentile of their distribution of skewness, i.e. respectively G1 = -2.08 and G1 = 6.32. We also simulated samples extracted from population where G1 = 0, in order to assess the main effect of high kurtosis on the quality of estimators. All possible combinations of population skewness and kurtosis and the number of scenarios for each combination are summarized in Table 2.

Table 2. *Number of Combinations of skewness and kurtosis in our simulations*

	Kurtosis		
	0	95.75	**TOTAL**
0	252	252	**504**

			Kurtosis	
Skewness	-2.08	/	252	**252**
	6.32	/	252	**252**
	TOTAL	**252**	**756**	**1008**

Note. Fisher's skewness (G1) and kurtosis (G2) are presented in Table 2. The 252 combinations where both G1 and G2 equal 0 correspond to the normal case.

For the 4 resulting combinations of skewness and kurtosis (see Table 2), all other parameter values were chosen in order to illustrate the consequences of factors known to play a key role on quality of estimators. We manipulated the population mean difference $(\mu_1 - \mu_2)$, the sample sizes (n), the sample size ratio (n-ratio $= \frac{n_1}{n_2}$), the population *SD*-ratio (i.e. $\frac{\sigma_1}{\sigma_2}$), and the sample size and population variance pairing. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in step of 1 (so does $\mu_1 - \mu_2$)[6]. Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10 (so does $\frac{\sigma_1}{\sigma_2}$). The simulations for which both σ_1 and σ_2 equal 1 are the particular case of homoscedasticity (i.e. equal population variances across groups). Sample size of both groups (n_1 and n_2) were 20, 50 or 100. When sample sizes of both groups are equal, the n-ratio equals 1 (it is known as a balanced design). All possible combinations of n-ratio and population *SD*-ratio were performed in order to distinguish positive pairings (the group with the largest sample size is extracted from the population with the largest *SD*), negative pairings (the group with the smallest sample size

[6] In the original plan, we had added 252 simulations in which mu1 and mu2 were both null. We decided to not present the results of these simulations, because the relative bias and the relative variance appeared to us to be very useful to fully understand the estimators comparison, and computing them is impossible when the real mean difference is zero.

is extracted from the population with the smallest *SD*), and no pairing (sample sizes and/or population *SD* are equal across all groups). In sum, the simulations grouped over different sample sizes yield 5 conditions based on the n-ratio, population *SD*-ratio, and sample size and population variance pairing, as summarized in Table 3. Table 3. *5 conditions based on the n-ratio, SD-ratio, and sample size and variance pairing*

		***n*-ratio**		
		1	**>1**	**<1**
	1	a	b1	b2
***SD*-ratio**	**>1**	c1	d1	e1
	<1	c2	e2	d2

Note. The n-ratio is the sample size of the first group (n_1) divided by the sample size of the second group (n_2). When all sample sizes are equal across groups, the n-ratio equals 1. When $n_1 > n_2$, n-ratio > 1, and when $n_1 < n_2$, n-ratio < 1. *SD*-ratio is the population *SD* of the first group (σ_1) divided by the population *SD* of the second group (σ_2). When $\sigma_1 = \sigma_2$, *SD*-ratio $= 1$. When $\sigma_1 > \sigma_2$, *SD*-ratio > 1. Finally, when $\sigma_1 < \sigma_2$, *SD*-ratio < 1. QUESTION: EST-CE-VRAIMENT NECESSAIRE DE DIRE CECI? REP MARIE: JE PENSE QUE oui, POUR NE PAS ETRE OBLIGE DE RETOURNER VOIR DANS LE TEXTE CE QUI CORRESPOND A QUOI.

Results. Before detailing estimators comparison for each condition, it might be interesting to make some general comments.

1) When the normality assumption is met (i.e. when G1 and G2 = 0, left in Figures 3 to 7), bias and variance of all estimators is so small that any detected differences are

marginal. However, the further from the normality assumption (i.e. when moving from left to right in Figures 3 to 7), the larger the value of all envisaged indicators of quality (i.e. bias, relative biaS, efficiency and relative efficiency). Note that in a purpose of readability, the ordinate axis is not on the same scale depending on the combination G1/G2. However, if the distribution shape influences all our indicators of quality, most of the time, there is no appearant interaction effect between estimators and distribution shape: the general appearance of barplots is almost always the same for all combinations of skewness and kurtosis (the only is exception is for the Glass's d_s when population distributions are skewed, as it will be described later). As a conclusion, the further from the normality assumption, the larger the below mentioned differences between estimators.

2) The fact that the bias of all estimators is very small when the normality assumption is met does not mean that all estimators are relevant in any conditions when the normality assumption is met. Because of the pooled error term, Cohen's d_s and Hedge's d_g should be avoided when population variances and sample sizes are unequal across groups, as reminded in the section "Different mesures of effect size". [However, because the pooled standard deviation will be poorly estimated, both at sample and population levels, this cannot be seen throughout the size of the bias (i.e. bias $= E(\hat{\delta}) - \delta$ and both $E(\hat{\delta})$ and δ are badly estimated).]. CF. LIGNE 395: IT'S A BIT WEIRD TO SAY THAT THE COMMON STD DEVIATION IS POORLY ESTIMATED AS SUCH QUANTITY DOES NOT EXIST (BOTH GROUPS DON'T SHARE A COMMON POPULATION VARIANCE. . . .) SO MAYBE WE SHOULD REPHRASE?

3) Throughout this section, we will **compare** the quality of different estimators. We chose very extreme (although realistic) conditions, and we know that none of the parametric measures of effect size will be robust against such extreme conditions. Our goal is therefore to study the robustness of the estimators against normality violations

only in comparison with the robustness of other indicators, but not in absolute terms.

After these general remarks, we will analyze each condition separately. In all Figures presented below, averaged results for each sub-condition are presented under five different configurations of distributions, using the legend described in Figure 2.

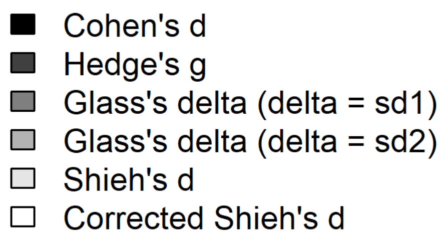

Figure 2. Legend

Figures 3 and 4 show that for all configurations where sample sizes are equal between groups (conditions a and c), estimator bias tends to decrease and precision is also improved with increasing sample sizes, meaning that all estimators are consistent. Moreover, Shieh's d_s and Shieh's d_s^* are identical, because our transformation is operant only when the sample sizes ratio differs from 1. We can demonstrate that the bias of Shieh's d_s and Shieh's d_s^* is exactly half the size of the bias of Cohen's d_s, and that their variance is exactly four times smaller than the variance of Cohen's d_s. Due to the relation described in equation 16 when sample sizes are equal between groups, such proportions mean that relative to their respective true effect size, Cohen's d_s, Shieh's d_s and Shieh's d_s^* perform all as well, as we can see in the second and fourth rows in Figures 3 and 4. These two rows also reveal that the relative bias and variance of Hedges's g_s is also identical to the three prementioned ones.

$$Shieh's\ \delta_{n_1=n_2} = \frac{Cohen's\ \delta_{n_1=n_2}}{2} \tag{16}$$

When samples are extracted from symmetric ditributions (the two first columns in Figures 3 and 4), both glass's d_s estimates (i.e. using SD_1 and SD_2) show least precision and highest bias rates, in comparison with all other measures, which is not surprising, as the standardizer is estimated based on half the sample size. In terms of relative bias, Glass's d_s shows similar qualitys when using either SD_1 or SD_2 as standardizer[7], however, when population variances differ across groups, the relative variances are unequal (Figure 4). It is well known that in equation 9, the variance of the numerator estimate (i.e. $\bar{X}_1 - \bar{X}_2$) depends on both σ_1 and σ_2, as reminded in equation 17:

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{17}$$

On the other side, σ is estimated based on only one sample SD (either SD_1 or SD_2). When computing the Glass's d_s with the estimate of the smaller population SD as standardizer (i.e. SD_2 in Figure 5, and SD_1 in Figure 6), the estimate of the larger

[7] When looking at the raw bias in Figure 4, one could believe that the bias is always more important when choosing SD2 as a standardizer. It is only an artefact of simulations. The bias is always more important when choosing the sample extracted from the smaller population SD as standardiser, because it results in a larger effect size estimate, and the larger the effect size estimate, the larger the raw bias. In our simulations, while the population SD of the first group always equals 1, in half of the simulations in condition c, the population SD of the second group is lower than 1 (meaning that the more biased Glass's estimate will occure when choosing SD2 < 1 as standardiser), and in the other half, the population SD of the second group is larger than 1 (meaning that the more biased Glass's estimate will occure when choosing SD1 as standardizer). Of course, for X, a constant mean difference and z, the standardizer, X/z will always result in a larger effect size measure when $z < 1$. This is confirmed by the identical average relative bias for both measures of Glass's ds.

population SD will increase the variance of the numerator in equation 9 but will not impact the variance of the denominator, resulting in a more variable estimator. When choosing the estimate of the larger population SD as standardizer, the estimate of the smaller population SD will decrease the variance of the numerator in equation 9 but will not impact the variance of the denominator, resulting in a less variable estimator. CA SEMBLE ETRE LE CONTRAIRE QD DISTRIBUTION NORMALE…BIZARRE?[8]

When samples are extracted from skewed distributions, both glass's d_s estimators (i.e. using SD_1 and SD_2) remain the more biased and variable when variances are equal across groups. When variances are unequal across groups, glass's d_s is sometimes more biased and variable than all other estimators, and sometimes less biased and variable. The explanation lies in two concepts:

- a non null correlation between the sample means and standard deviations;
- the number of observations based on which the standardizer is computed.

When samples are extracted from skewed distributions, there is a non null correlation between the sample means and sample standard deviations, resulting in a non null correlation of opposite sign between $\bar{X}_1 - \bar{X}_2$ and respectively SD_1 and SD_2.[9]). When the population mean difference $\mu_1 - \mu_2$ is positive (like in our simulations), Glass's d_s is always

[8] Remind that in our simulations, the first population SD always equal 1, and the second population SD varies from .1 to 10. The difference in terms of relative variance is more important when comparing SD1=1 and SD2=10 as standardizer than when comparing SD1=1 and SD2=.1 as standardizer, explaining why the difference between both Glass's estimators seems more important in FIgure 6 than in Figure 5

[9] when distributions are right-skewed, the correlation between sample means and standard deviation is positive. Because SD1 (SD2) is positively (negatively) correlated with the mean difference estimates, it results in a positive (negative) correlation between SD1 (SD2) and the mean difference. when distributions are left-skewed, the correlation between sample means and standard deviation is negative. Because SD1 (SD2) is positively (negatively) correlated with the mean difference estimates, it results in a negative (positive) correlation between SD1 (SD2) and the mean difference.

more biased and variable when choosing as standardizer the SD that is negatively correlated with $\bar{X_1} - \bar{X_2}$ (i.e. SD_2 when distributions are right-skewed and SD_1 when distributions are left-skewed). When the population mean difference is negative, the reverse is true. For interested reader, this is detailed and explained in Appendix 3.

When the standardizer takes both SD_1 and SD_2 into account (i.e. all estimators but Glass's d_s), the correlation between standardizer and mean difference will be null, as long as population variances are equal across groups. However, when population variances are unequal across groups, the sign of the correlation between the standardizer and the mean difference will be the same as the one of the correlation between the mean difference and the estimates of the larger population variance (e.g. if samples are extracted from right-skewed distributions and $\sigma_2 > \sigma_1$, there will be a negative correlation between the sample mean differences and standardizer). And so.... CONCLURE CE TRUC.

Figure 7 shows that when population variances are equal but sample sizes are unequal between groups, as when sample sizes were equal, while estimators are consistent, glass's d_s is generally more biased and variable that all other estimators. Again, this is due to the fact that the standardizer is estimated based on part of the total sample and unsurprisingly, the bias is even larger when choosing the SD of the smallest group as a standardizer. As long as samples are extracted from symmetric distributions, the bias and variance of glass's d_s is only a function of the sample size of the group from which standardizer is computed (because $\sigma_1 = \sigma_2$). However, when samples are extracted from skewed distribution, because of the correlation between sample mean and sample SD, glass's d_s becomes even more biased and variable when the chosen standardizer is negatively correlated with the mean difference and associated with the smaller sample size (i.e. when choosing SD_2 as standardizer, with $n_1 > n_2$ when distributions are right-skewed; and when choosing SD_1 as standardizer, with $n_1 < n_2$ when distributions are left-skewed). [10].

[10] Again, we should remind that in all our simulations, the population mean difference is positive. If mean

As previously, the bias of Shieh's d_s is smaller than the Cohen's d_s one (as well as the Hedge's g_s one). However, the difference is smaller than previously. Remember that when sample sizes differ between groups, Shieh's d_s is always more than twice smaller than Cohen's d_s (see Appendix 1 for more details). As a consequence, if both Cohen's d_s and Shieh's d_s performed as well, the bias of Shieh's d_s should be more than twice smaller than Cohen's d_s bias (and the variance of Shieh's d_s should be more than four time smaller than Cohen's d_s bias), but it's not. It's confirmed by the second and fourth rows in Figure 7 where we can see that the relative bias and variance of Shieh's d_s are larger than the relative bias and variance of Cohen's d_s, that remains the best indicator in terms of bias. However, it is very interesting to note that our transformed Shieh's d_s^* is on average less biased and variable than original Shieh's d_s, both if raw and relative terms. This measure seems to perform almost as well as Cohen's d_s.

Figure 8 and 9 refer to conditions where there is a pairing between population variances and sample sizes. We know that in these configurations, the pooled variance will be poorly estimated (see the second remark at the beginning of the result section), and therefore, we will not discuss the Cohen's d_s and Hedge's g_s.We will only compare the quality of Glass's d_s, Shieh's d_s and Shieh's d_s^*.

Figure 8 shows that when variances are unequal, and the largest group is associated with largest variance, the more biased and variable estimator is Glass's d_s when choosing the standard deviation of the smallest group as standardizer. REM: AGAIN ONE OBSERVE THE SAME INTERACTION EFFECT BETWEEN STANDARDISER IN GLASS MEASURE AND SENSE OF ASYMMETRY AS OBSERVED FOR FIGURE 3 (IN SAME DIRECTION: WITH NEGATIVE SKEWNESS, WORST WHEN CHOOSING SD1 AND WHEN POSITIVE SKEWNESS, WORST WHEN CHOOSING SD2). Glass's d_s when

difference were negative, glass's d_s would be more biased and variable when the chosen standardizer is positively correlated with the mean difference and associated with the smaller sample size.

choosing the standard deviation of the largest group as standardizer, Shieh's d_s and transformed Shieh's d_s^* perform very similarly, both in terms of bias and efficiency.

Figure 9 shows that when variances are unequal, and the largest group is associated with smallest variance, as in all other configurations, the more biased and variable estimator is Glass's d_s when choosing the standard deviation of the smallest group as standardizer (sauf quand asymetrie négative… not true anymore when there is asymmetry… explain it).

In summary, Cohen's d_s and Hedge's d_s remain the best measure when the asssumption of equal variances is met. When variances are unequal across populations, Cohen's d_s and Hedge's g_s perform exactly as well as Shieh's d_s and transformed Shieh's d_s^*, as long as sample sizes are equal across groups. However, when variances and sample sizes are both unequal across groups, Cohen's d_s and Hedge's g_s become irrelevant. Glass's d_s is most of the time the more biased and variable measure. We presume this could be explained by the estimation of the SD based on a subsample, because the bias is larger when standardizer is estimated based on the smallest group. Only under very specific conditions (when there is a negative correlation between sample sizes and variances and the sample size of the control group is larger than the sample size of the experimental group), Glass's d_s performs the best in comparison with all other estimators.

Conclusion. SUMMARY GLASS: il y a plusieurs facteurs "aggravants":

Pour le biais: (1) SD calculé sur base du plus petit n (–> mesure plus variable et biaisée car distributions plus asymétrique) –> vrai pour toute distribution (2) SD négativement corrélé avec la différence de moyenne quand mu1-mu2 > 0 (= choix de SD2 quand asymétrie positive, et de SD1 quand asymétrie négative, vu qu'on calcule m1-m2) OU SD positivement corrélé avec la différence de moyenne quand mu1-mu2 < 0 (= choix de SD1 quand asymétrie positive, et de SD2 quand asymétrie négative, vu qu'on calcule m1-m2). –> vrai seulement quand distributions asymétriques

Pour la variance: (1) et (2) jouent, mais il y a en plus: (3) SD calculé sur base du plus petit sigma

Shieh's d_s and our transformed Shieh's d_S^* are the only measure that have an acceptable bias and variance in all configurations. Considering the fact that our transformed Shieh's d_s^* is much more generalizable (and therefore interpretable) than Shieh d_s, we would recommend the use of this measure in all situations, unless we have very good reason to believe that variances are the same across populations.

Simulation 2: confidence intervals. TO DO #### Method ### Results #### Conclusion

Altman, G. D. (2005). Why we need confidence intervals. *World Journal of Surgery*, *29*, 554–556. doi:10.1007/s00268-005-7911-0

American Educational Research Association. (2006). Standards for reporting on empirical social science research in aera publications. *Educational Researcher*, *35*, 33–40. doi:10.3102/0013189X035006033

American Psychological Association. (2010). *Publication manual of the american psychological association [apa] (6 ed.)* (American Psychological Association.). Washington, DC:

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology*, *29*, 664–672.

Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *American Journal of Speech-Language Pathology*, *20*, 233–242.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior*

Research Methods, *49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important.* Retrieved from https://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1965). Some statistical issues in psychological research. In *Handbook of clinical psychology* (B. B. Wolman., pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Routledge Academic.). New York, NY.

Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on shieh (2013). *Behavior Research Methods*, *45*, 968–971. doi:10.3758/s13428-013-0392-4

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi:10.5334/irsp.82

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of welch's f-test instead of the classical f-test in one-way anova. *International Review of Social Psychology*, *32*(1), 1–12. doi:http://doi.org/10.5334/irsp.198

Ellis, P. D. (2015). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* (Cambridge University Press.). Cambridge, UK.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*(5), 275–282. doi:10.1080/00220670109598763

Glass, G. V., McGav, B., & Smith, M. L. (2005). *Meta-analysis in social research* (Sage.). Beverly Hills, CA.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. doi:10.1037//0022-006x.68.1.155

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135–146. doi:10.1037/1082-989X.6.2.135

Grissom, R. J., & Kim, J. J. (2012). *Effect size for research* (Routledges.). New York, NY.

Grissom, R. R., & Kim, J. J. (2005). *Effect size for research: A broad practical approach.* (Lawrence Erlbaum Associates, Mahwah, N.J.). London.

Hays, W. L. (1963). *Statistics for psychologists* (Holt, Rinehart & Winston.). New York.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (Academic Press.). Cambridge, Massachusetts. doi:10.1016/C2009-0-03396-0

Henson, R. I., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA task force report and current trends. *Journal of Research and Development in Education*, *33*(4), 285–296.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervales around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69. doi:10.1177/0013164404264850

Kirk, R. E. (2009). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. doi:10.1177/0013164496056005002

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. doi:10.3389/fpsyg.2013.00863

Li, J. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, *48*(4), 1560–1574. doi:10.3758/s13428-015-0667-z

McBride, G. B., Loftis, J. C., & Adkins, N. C. (1993). What do significance tests really tell us about the environment? *Environmental Management*, *17*(4), 423–432.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi:10.1037/0033-2909.105.1.156

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*, 591–605. doi:10.1111/j.1469-185X.2007.00027.x

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286. doi:10.1006/ceps.2000.1040

Olejnik, S., & Hess, B. (2001). *Revisiting the efficacy of glass's estimator of effect size for program impact analysis*. Retrieved from https://eric.ed.gov/?id=ED452210

Peng, C.-Y., & Chen, L.-T. (2014). Beyond cohen's d: Alternative effect size measures for between-subject designs. *THE JOURNAL OF EXPERIMENTAL EDUCATION*, *82*(1), 22–50. doi:10.1080/00220973.2012.745471

Peng, C.-Y., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The Impact of APA and AERA Guidelines on Effect size Reporting. *Contemporary Educational Psychology*, *82*(1), 22–50. doi:10.1080/00220973.2012.745471

Prentice, D., & Miller, D. T. (1990). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164.

Raviv, E. (2014). *Bias vs. Consistency*. Retrieved March 25, 2020, from https://eranraviv.com/bias-vs-consistency/

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The hand-book of research synthesis* (pp. 231–244). New-York: Sage.

Shieh, G. (2013). Confidence intervals and sample size calculations for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVA. *Behavior Research Methods*, *45*(1), 2–37. doi:10.3758/s13428-012-0228-7

Steyn, H. S. (2000). Practical significance of the difference in means. *Journal of*

Industrial Psychology, *26*(3), 1–3.

Stout, D. D., & Ruble, T. L. (1995). Assessing the practical signficance of empirical results in accounting education research: The use of effect size information. *Journal of Accounting Education*, *13*(3), 281–298.

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

Thompson, B. (2002). "Statistical","Practical", and "Clinical": How Many Kinds of Significance Do Counselors Need to Consider? *Journal of Counseling & Development*, *80*, 64–71.

Tyler, R. W. (1931). What is Statistical Significance? *Educational Research Bulletin*, *X*(5), 115–142.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications (7th edition)* (Brooks/Cole, Cengage Learning.). Belmont, USA.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. doi:10.1007/bf02295644

(n.d.).

Altman, G. D. (2005). Why we need confidence intervals. *World Journal of Surgery*, *29*, 554–556. doi:10.1007/s00268-005-7911-0

American Educational Research Association. (2006). Standards for reporting on empirical social science research in aera publications. *Educational Researcher*, *35*, 33–40. doi:10.3102/0013189X035006033

American Psychological Association. (2010). *Publication manual of the american psychological association [apa] (6 ed.)* (American Psychological Association.). Washington, DC:

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology*, *29*, 664–672.

Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *American Journal of Speech-Language Pathology*, *20*, 233–242.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1

Coe, R. (2002). *It's the effect size, stupid. What effect size is and why it is important.* Retrieved from https://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1965). Some statistical issues in psychological research. In *Handbook of clinical psychology* (B. B. Wolman., pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Routledge Academic.). New York, NY.

Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on shieh (2013). *Behavior Research Methods*, *45*, 968–971. doi:10.3758/s13428-013-0392-4

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi:10.5334/irsp.82

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of welch's f-test instead of the classical f-test in one-way anova. *International Review of Social Psychology*, *32*(1), 1–12. doi:http://doi.org/10.5334/irsp.198

Ellis, P. D. (2015). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* (Cambridge University Press.). Cambridge, UK.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*(5), 275–282. doi:10.1080/00220670109598763

Glass, G. V., McGav, B., & Smith, M. L. (2005). *Meta-analysis in social research* (Sage.). Beverly Hills, CA.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. doi:10.1037//0022-006x.68.1.155

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135–146. doi:10.1037/1082-989X.6.2.135

Grissom, R. J., & Kim, J. J. (2012). *Effect size for research* (Routledges.). New York, NY.

Grissom, R. R., & Kim, J. J. (2005). *Effect size for research: A broad practical approach.* (Lawrence Erlbaum Associates, Mahwah, N.J.). London.

Hays, W. L. (1963). *Statistics for psychologists* (Holt, Rinehart & Winston.). New York.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (Academic Press.). Cambridge, Massachusetts. doi:10.1016/C2009-0-03396-0

Henson, R. I., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA task force report and current trends. *Journal of Research and Development in Education*, *33*(4), 285–296.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervales around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69. doi:10.1177/0013164404264850

Kirk, R. E. (2009). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. doi:10.1177/0013164496056005002

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. doi:10.3389/fpsyg.2013.00863

Li, J. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, *48*(4), 1560–1574. doi:10.3758/s13428-015-0667-z

McBride, G. B., Loftis, J. C., & Adkins, N. C. (1993). What do significance tests really tell us about the environment? *Environmental Management*, *17*(4), 423–432.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi:10.1037/0033-2909.105.1.156

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*, 591–605. doi:10.1111/j.1469-185X.2007.00027.x

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286. doi:10.1006/ceps.2000.1040

Olejnik, S., & Hess, B. (2001). *Revisiting the efficacy of glass's estimator of effect size for program impact analysis*. Retrieved from https://eric.ed.gov/?id=ED452210

Peng, C.-Y., & Chen, L.-T. (2014). Beyond cohen's d: Alternative effect size measures for between-subject designs. *THE JOURNAL OF EXPERIMENTAL EDUCATION*, *82*(1), 22–50. doi:10.1080/00220973.2012.745471

Peng, C.-Y., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The Impact of APA and AERA Guidelines on Effect size Reporting. *Contemporary Educational Psychology*, *82*(1), 22–50. doi:10.1080/00220973.2012.745471

Prentice, D., & Miller, D. T. (1990). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164.

Raviv, E. (2014). *Bias vs. Consistency*. Retrieved March 25, 2020, from https://eranraviv.com/bias-vs-consistency/

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The hand-book of research synthesis* (pp. 231–244). New-York: Sage.

Shieh, G. (2013). Confidence intervals and sample size calculations for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVA. *Behavior Research Methods*, *45*(1), 2–37. doi:10.3758/s13428-012-0228-7

Steyn, H. S. (2000). Practical significance of the difference in means. *Journal of Industrial Psychology*, *26*(3), 1–3.

Stout, D. D., & Ruble, T. L. (1995). Assessing the practical signficance of empirical results in accounting education research: The use of effect size information. *Journal of Accounting Education*, *13*(3), 281–298.

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

Thompson, B. (2002). "Statistical","Practical", and "Clinical": How Many Kinds of Significance Do Counselors Need to Consider? *Journal of Counseling & Development*, *80*, 64–71.

Tyler, R. W. (1931). What is Statistical Significance? *Educational Research Bulletin*, *X*(5), 115–142.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications (7th edition)* (Brooks/Cole, Cengage Learning.). Belmont, USA.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. doi:10.1007/bf02295644

(n.d.).

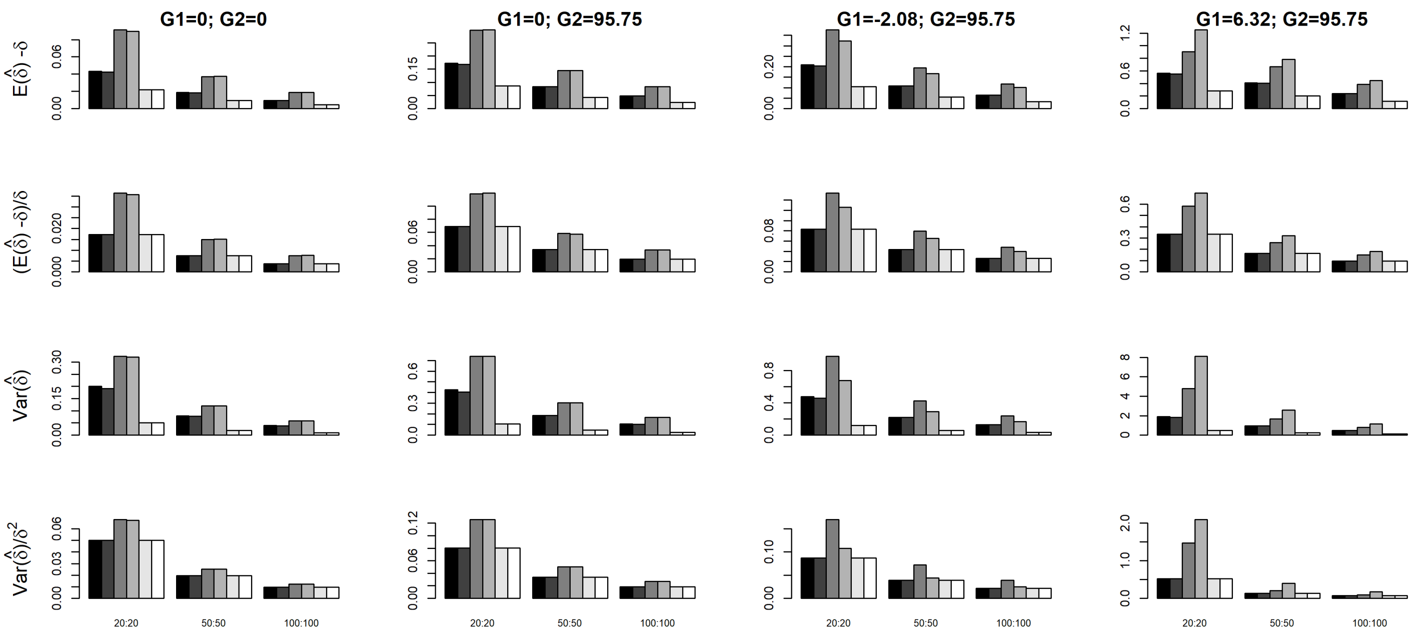

Figure 3. Bias and efficiency of five estimator of standardized mean difference, when variances and sample sizes are equal across groups (condition a)

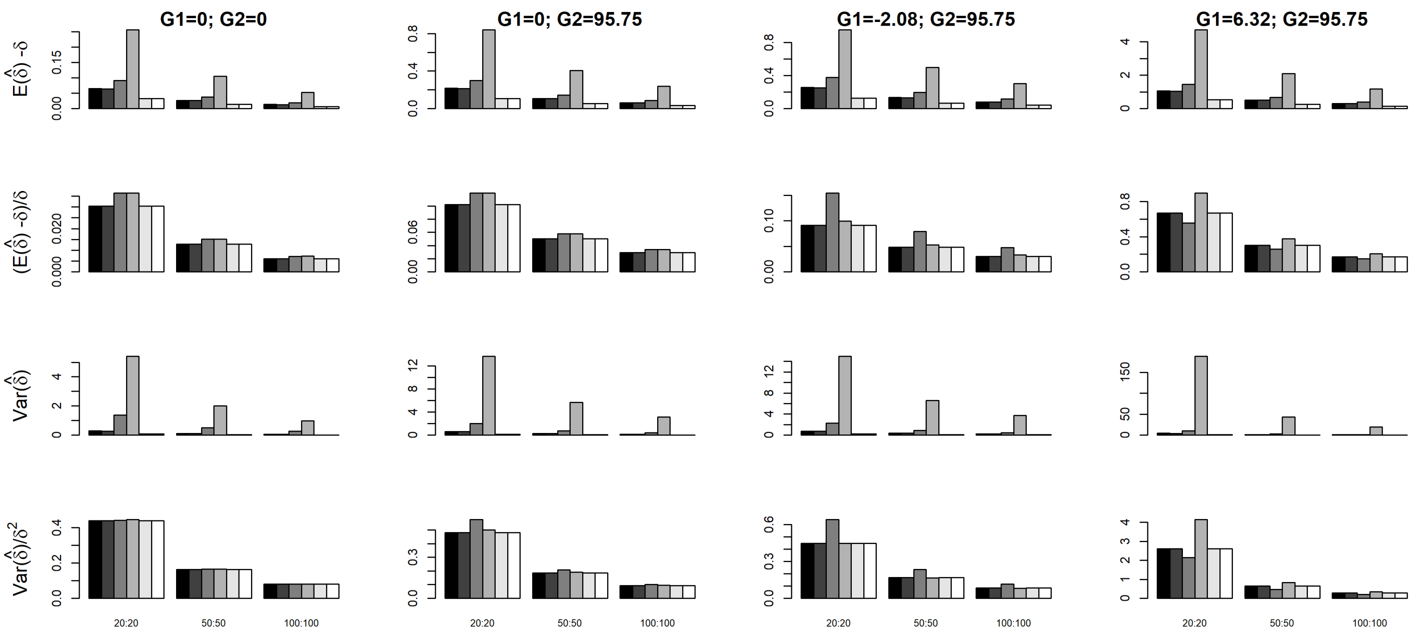

Figure 4. Bias and efficiency of five estimator of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c)

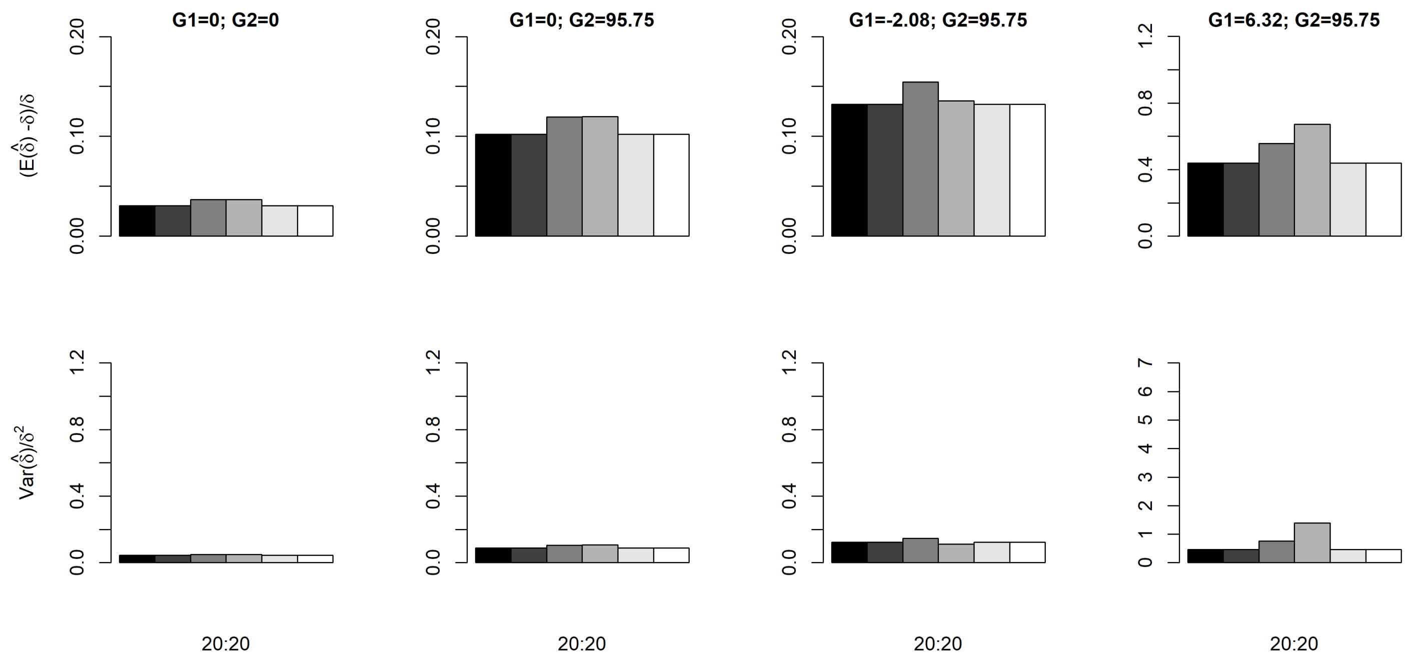

Figure 5. Bias and efficiency of Glass's ds, when sample sizes are equal across groups and population variances are unequal, when SD1 is larger than SD2 (condition c1)

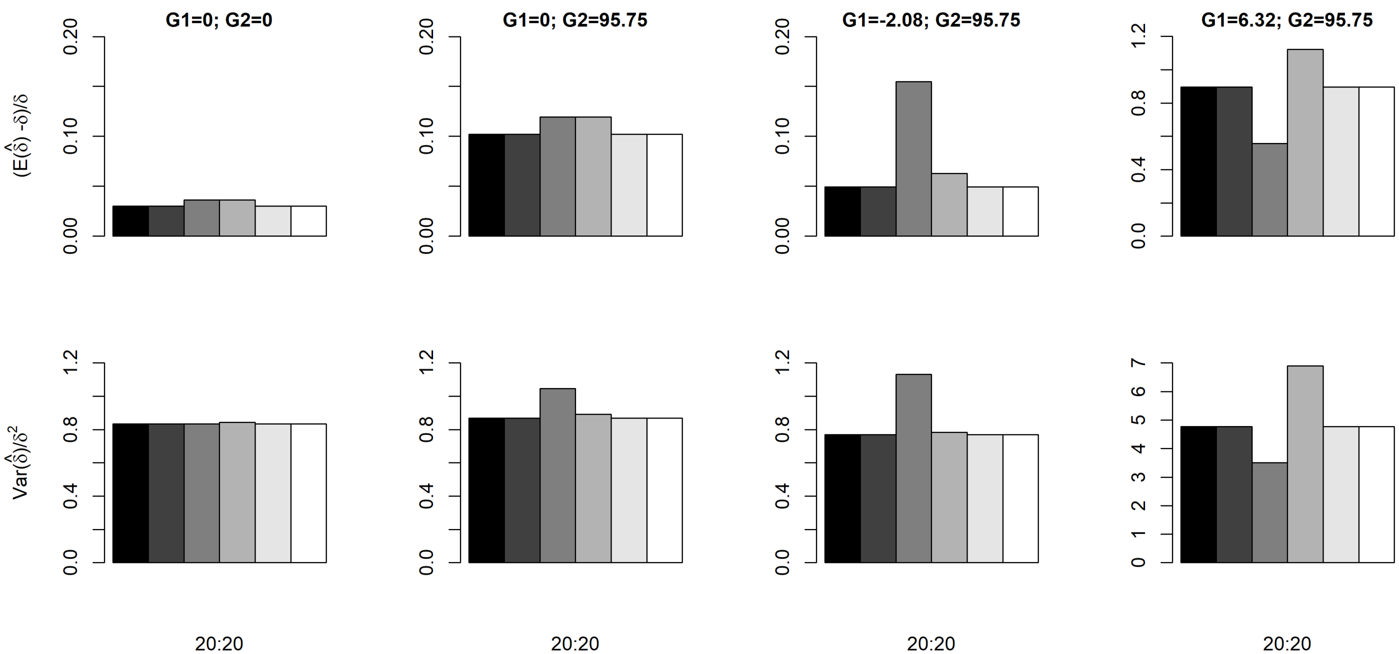

Figure 6. Bias and efficiency of Glass's ds, when sample sizes are equal across groups and population variances are unequal, when SD1 is lower than SD2 (condition c2)

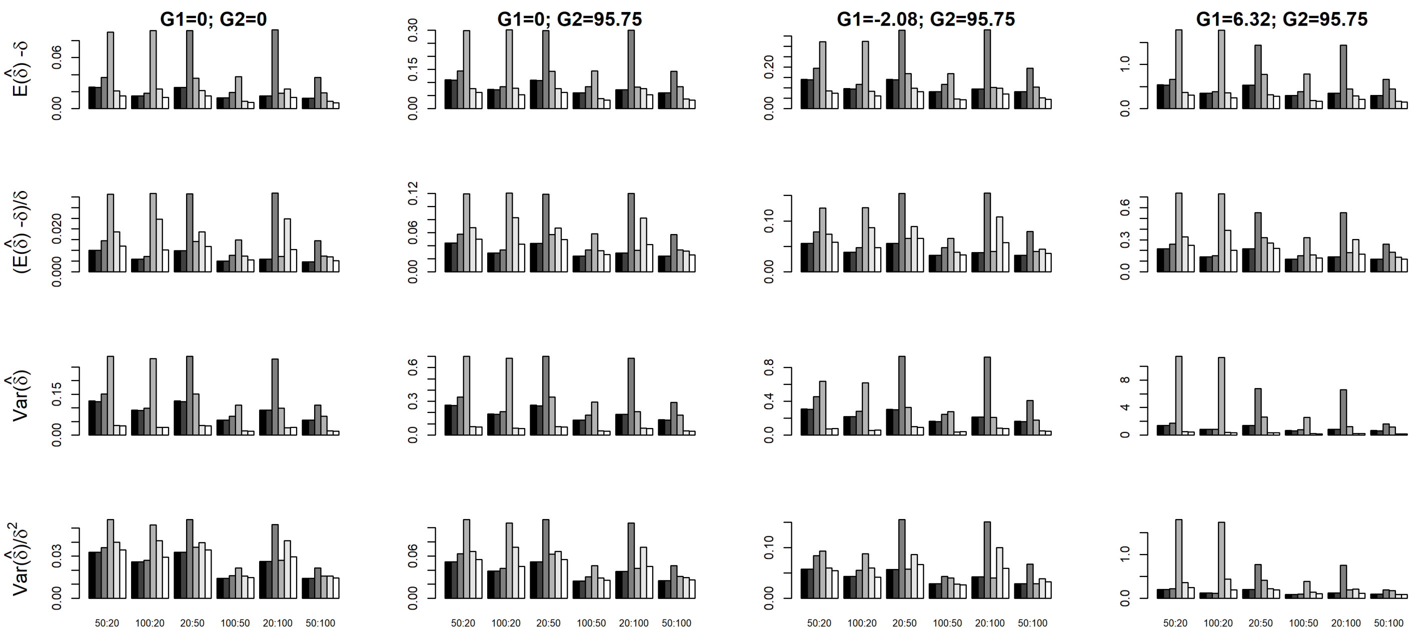

Figure 7. Bias and efficiency of five estimator of standardized mean difference, when variances are equal across groups and sample sizes are unequal

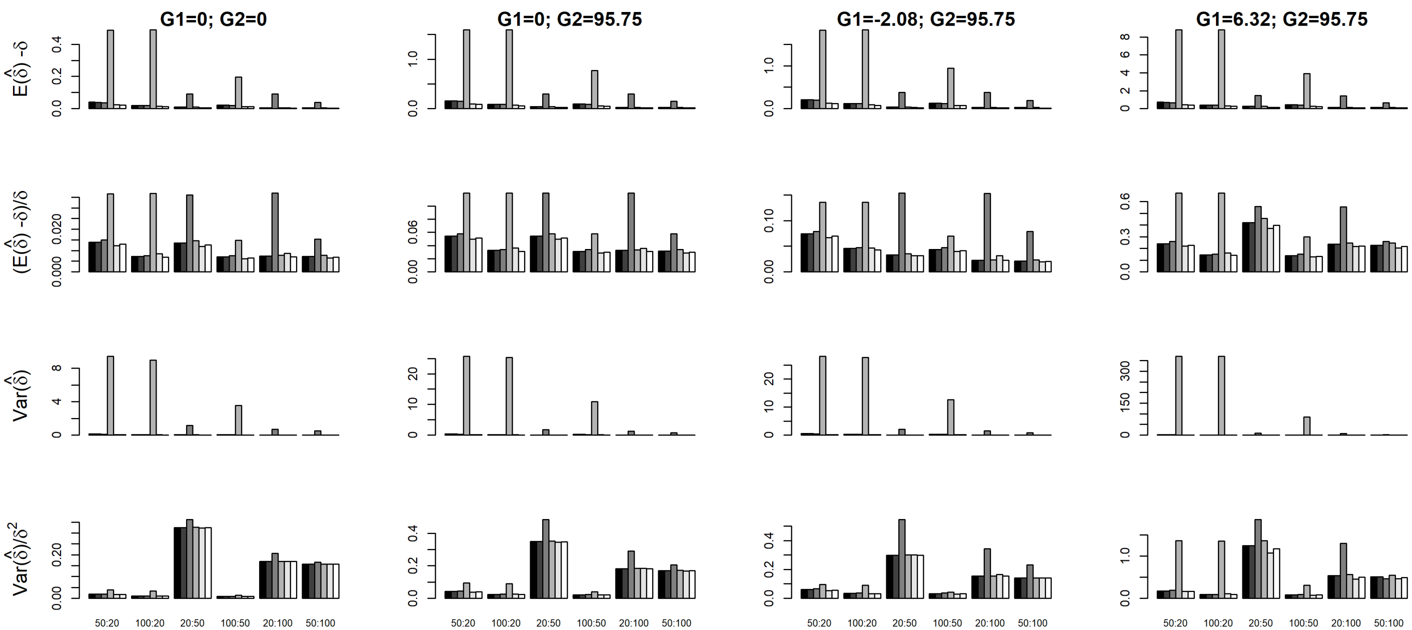

Figure 8. Bias and efficiency of five estimator of standardized mean difference, when variances and sample sizes are unequal across groups, with positive correlation between them

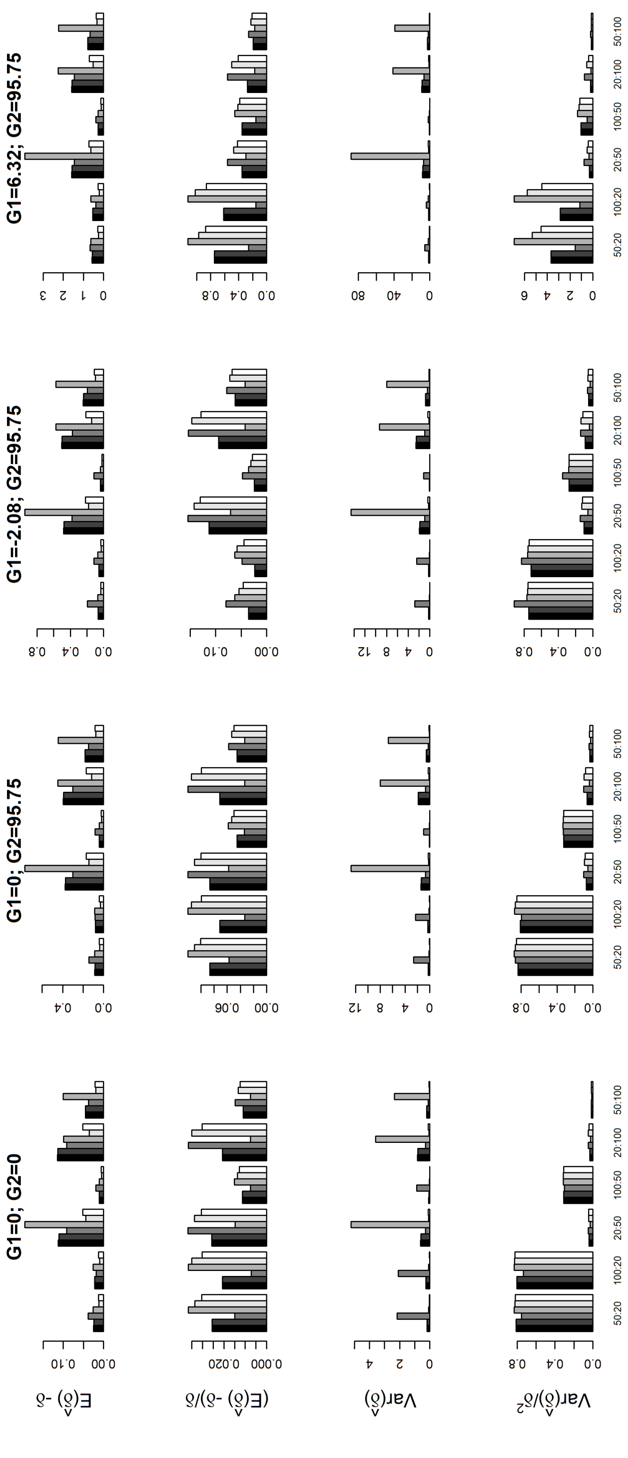

Figure 9. Bias and efficiency of five estimator of standardized mean difference, when variances and sample sizes are unequal across groups, with negative correlation between them

Appendix

Appendix 1: The mathematical study of Shieh's δ

Paste Appendix 1 when it will be finished

Appendix 2: Confidence intervals

Paste Appendix 2 when it will be finished

Appendix 3: a priori power analyses

Paste Appendix 3 when it will be finished (Cumming & Finch, 2001)

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