What measure of effect size using when performing a Welch's t-test?

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

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

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

[3] Universiteit Gent, Department of Applied Mathematics, Computer Science and Statistics,4 Gent, Belgium

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 using when performing a Welch's t-test?

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 slightly 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.

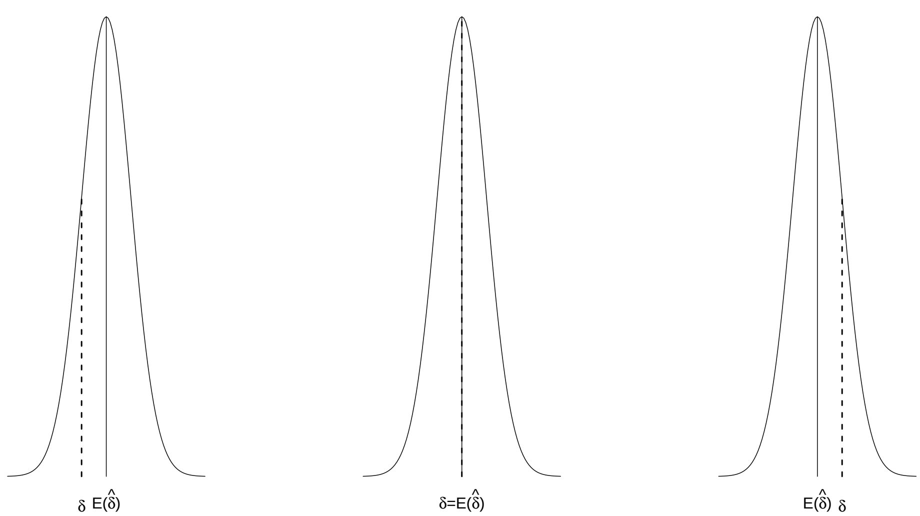

Figure 1. Samplig distribution for a positively biased (left), an unbiased (center) and a negatively biased estimator (right)

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 reducing the effect size impact in computing the *relative variance* as the ratio between the variance and the square of the population estimator:

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

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

This doesn't entirely remove the impact of the effect size, but greatly reduces it. SEE TABLES 1 –> 4: ONLY THE SECOND TERM IN THE ADDITION CONTAINS δ^2, SO WHEN DIVIDING THE VARIANCE BY δ^2, THE FIRST TERM IN THE ADDITION DEPENDS ON THE EFFECT SIZE. 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. For all of them, the mean difference is estimated by the difference $\bar{X}_1 - \bar{X}_2$ of both sample means. When the equality of variances assumption is assumed, σ is estimated by pooling both samples standard deviations (S_1 and S_2). When the equality of variances assumption cannot be assumed, alternatives to the common standard deviation are available. Throughout this section, we will present some of these estimators, separately depending on whether they rely on the assumption of equality of variances or not. For each of them, we will provide information about their theoretical bias, variance and consistency.

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):

$$Cohen's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)\times S_1^2+(n_2-1)\times S_2^2}{n_1+n_2-2}}} \tag{6}$$

Where S_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 (Keselman, Algina, Lix, Deering, & Wilcox, 2008), hence we achieve a more accurate estimation of the population variance by pooling both estimates of this parameter (i.e S_1 and S_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-statistic:

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

Under the assumption of normality and equal variances between groups, Student's t-statistic follows a t-distribution with known degrees of freedom and noncentrality parameter [6]:

$$df_{student} = n_1 + n_2 - 2 \tag{8}$$

$$ncp_{student} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} \times \sqrt{\frac{n_1 n_2}{n_1+n_2}}, \quad where\ \sigma_{pooled} = \sqrt{\frac{(n_1-1)\times\sigma_1^2+(n_2-1)\times\sigma_2^2}{n_1+n_2-2}} \tag{9}$$

The relationship described in equation 7 and the theoretical distribution of Student's t-statistic allow us to theoretically determine the sampling distribution of Cohen's d_s, and

[6] Under the null hypothesis of no differences between sample means, Student's t-statistic will follow a central t-distribution with $n_1 + n_2 - 2$ degrees of freedom. However, when the null hypothesis is false, the distribution of this quantity will not be centered, and noncentral t-distribution will arise.

therefore, its theoretical expectancy and variance when the assumptions of normality and equal variances are met. All these equations are provided in Table 1. For interested readers, bias and variance of Cohen's d_s have been thoroughly studied, based on Table 1, so as to determine the way different parameters influence them, and results are detailed and available on Github (see Appendix 1 in https://github.com/mdelacre/Effect-sizes/; it will be the same for all estimators described later).

While Cohen's d_s is a consistent estimator, its bias and variance are substantial with small sample sizes, even under the assumptions of normality and equal variances (Lakens, 2013). In order to compensate for Cohen's d_s bias with small sample sizes, Hedges & Olkin (1985) has defined a bias-corrected version:

$$Hedges'\ g_s = Cohen's\ d_s \times \frac{\Gamma(\frac{df_{Student}}{2})}{\sqrt{\frac{df_{Student}}{2}} \times \Gamma(\frac{df_{Student}-1}{2})} \quad (10)$$

Where $df_{Student}$ has been defined in equation 8,and $\Gamma()$ is the gamma function. This equation can be approximated as follows:

$$Hedges'\ g_s = Cohen's\ d_s \times \left(1 - \frac{3}{4N-9}\right) \quad (11)$$

Hedges' g_s is theoretically unbiased when the assumptions of normality and equal variances are met (see Table 1), and it has a smaller variance than Cohen's d_s, especially with small sample sizes [7]. As Cohen's d_s, its variance depends on the total sample size (N), the sample sizes ratio $\left(\frac{n_1}{n_2}\right)$ and the population effect size (δ_{Cohen}). How these parameters influence the variance of Hedges' g_s will be summarized in a later section in which we will

[7] $.52 \leq \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}\right]^2 < 1$ for $3 \leq df < \infty$. The larger the total sample size, the smaller the difference between the variance of Cohen's d_s and Hedges' g_s.

compare different estimators through Monte Carlo simulations (see the section "Monte Carlo Simulations").

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 Hedges' g_s should be abandoned in favor of an alternative robust to unequal population variances.

Table 1

Expentency, bias and variance of Cohen's d_s and Hedges' g_s under the assumptions that independent residuals are normally distributed with equal variances across groups.

	df	Expectancy	Variance
$Cohen's\ d_s$	$N-2$	$\delta_{cohen} \times \frac{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}{\Gamma(\frac{df}{2})}$	$\frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen} \left[\frac{df}{df-2} - \left(\frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)} \right)^2 \right]$
		$\approx \frac{\delta_{Cohen}}{\left(1-\frac{3}{4N-9}\right)}$	$\approx \frac{N \times df}{n_1 n_2 \times (df-2)} + \delta^2_{Cohen} \left[\frac{df}{df-2} - \left(\frac{1}{1-\frac{3}{4N-9}} \right)^2 \right]$
$Hedges's\ g_s$	$N-2$	δ_{Cohen}	$Var(Cohen's\ d_s) \times \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})} \right]^2$
			$\approx Var(Cohen's\ d_s) \times \left[1 - \frac{3}{4N-9}\right]^2$

Note. Cohen's d_s is a biased estimator, because its expectation differ from the population effect size. On the other hand, Hedges' g_s is an unbiased estimator, because its expectation equals the population effect size; equations in Table 1 require $df \geq 3$ (i.e. $N \geq 5$).

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 Glass's d_s, (B) the Shieh's d_s and (C) the non pooled average of both variance estimates.

Glass's d_s. 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

$$Glass's\ d_s = \frac{\bar{X}_e - \bar{X}_c}{S_c} \tag{12}$$

Where $\bar{X}_e$ *and* $\bar{X}_c$ are respectively the sample means of the experimental and control groups, and S_c is the sample SD of the control group. 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 distribution of Glass's d_s is defined as following (Algina, Keselman, & Penfield, 2006):

$$Glass's\ d_s \sim \sqrt{\frac{1}{n_c} + \frac{S_e^2}{n_e \times S_c^2}} \times t_{df,ncp} \tag{13}$$

Where n_c and n_e are respectively the sample sizes of the control and experimental groups, and df and ncp are defined as follows:

$$df = n_c - 1 \tag{14}$$

$$ncp = \frac{\mu_c - \mu_e}{\sigma_c \times \sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}}} \tag{15}$$

Where μ_c and μ_e are respectively the mean of the populations control and experimental groups are extracted from. Thanks to equation 13, we can compute its theoretical expectancy variance when the assumptions of normality is met (See Table 2), and therefore determine which factors influence bias and variance, and how they do so (see Appendix 1).

Table 2

Expentency, bias and variance of Glass's d_s and Cohen's d'_s and Shieh's d_s under the assumptions that independent residuals are normally distributed.

	df	Expectancy	Variance
$Glass's\ d_s$	$n_c - 1$	$\delta_{glass} \times c_f$	$\frac{df}{df-2} \times \left(\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \sigma_c^2}\right) + \delta_{Glass}^2 \left(\frac{df}{df-2} - c_f^2\right)$
$Cohen's\ d'_s$	$\frac{(n_1-1)(n_2-1)(s_1^2+s_2^2)^2}{(n_2-1)s_1^4+(n_1-1)s_2^4}$	$\delta'_{Cohen} \times c_f$	$\frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta'_{Cohen})^2 \left(\frac{df}{df-2} - c_f^2\right)$
		$\approx \delta'_{Cohen} \times \frac{4df-1}{4(df-1)}$	$\approx \frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta'_{Cohen})^2 \left[\frac{df}{df-2} - \left(\frac{4\,df-1}{4(df-1)}\right)^2\right]$
$Shieh's\ d_s$	$\approx \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	$\delta_{Shieh} \times c_f$	$\frac{df}{(df-2)N} + \delta_{Shieh}^2 \left(\frac{df}{df-2} - c_f^2\right)$

Note. $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; all estimators are biased estimators, because their expectations differ from the population effect size δ; equations require $df \geq 3$ and at least 2 subjects per group.

Shieh's d_s. Kulinskaya & Staudte (2007) were the first to advice the use of a standardizer that takes 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:

$$Shieh's\ d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/q_1 + S_2^2/q_2}}; \quad q_j = \frac{n_j}{N}(j = 1, 2) \tag{16}$$

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

$$t_{welch} = Shieh's\ d_s \times \sqrt{N} \tag{17}$$

Where $N = n_1 + n_2$. The exact distribution of Welch's t-statistic is more complicated than the exact distribution of Student's t-statistic, but it follows a t-distribution with degrees of freedom and noncentrality parameters that can be approximated as follows, under the assumption of normality (Shieh, 2013; Welch, 1938):

$$df_{Welch} \approx \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}} \tag{18}$$

$$ncp_{Welch} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}} \times \sqrt{N} \tag{19}$$

The relationship described in equation 17 and the theoretical distribution of Welch's t-statistic allow us to theoretically approximate the sampling distribution of Shieh's d_s. Based on the sampling distribution of Shieh's d_s, we can estimate its theoretical expectancy and variance under the assumption of normality (see Table 2), and therefore and therefore determine which factors influence bias and variance, and how they do so (see Appendix 1).

It can be demonstrated that when variances and sample sizes are equal across groups, the biases and variances of Shieh's d_s and Cohen's d_s are identical except for a constant, as shown in equations 20 and 21:

$$Shieh's\ d_{s,bias} = 2 \times Cohen's\ d_{s,bias} \quad (considering\ \sigma_1 = \sigma_2\ and\ n_1 = n_2) \tag{20}$$

$$Shieh's\ d_{s,variance} = 4 \times Cohen's\ d_{s,variance} \quad (considering\ \sigma_1 = \sigma_2\ and\ n_1 = n_2) \tag{21}$$

Due to the relation described in equation 22 when sample sizes are equal between groups (as explained in Appendix 2), such proportions mean that relative to their respective true effect size, Cohen's d_s and Shieh's d_s are equally good. This is a good illustration of the fact that biases and variances should always be studied relative to the population effect size (and not in absolute terms), as we will do later.

$$Shieh's\ \delta_{n_1=n_2} = \frac{Cohen's\ \delta_{n_1=n_2}}{2} \tag{22}$$

Except for this very specific situation, according to the statistical properties of Welch's statistic under heteroscedasticity, Shieh's d_s accounts for the sample sizes allocation ratio. 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 illustrated in the following shiny application: https://mdelacre.shinyapps.io/ShiehvsCohen/).

Cohen's d'_s. The sample mean difference, divided by the non pooled average of both variance estimates was suggested by Welch (1938). This yields:

$$Cohen's\ d'_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(S_1^2+S_2^2)}{2}}} \tag{23}$$

Where $\bar{X}_j$ is the mean and S_j is the standard deviation of the j^{th} sample (j = 1,2). We know the distribution of Cohen's d'_s (Huynh, 1989):

$$cohen's\ d'_s \sim \sqrt{\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)(\sigma_1^2+\sigma_2^2)}{2}} \times t_{df^*,ncp^*} \tag{24}$$

Where df^* and ncp^* are defined as follows:

$$df^* = \frac{(n_1-1)(n_2-1)(s_1^2+s_2^2)^2}{(n_2-1)s_1^4+(n_1-1)s_2^4} \tag{25}$$

$$ncp^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}} \times \sqrt{\frac{n_1 n_2(\sigma_1^2+\sigma_2^2)}{2(n_2\sigma_1^2+n_1\sigma_2^2)}} \tag{26}$$

Thanks to equation 26, we can compute its theoretical expectancy variance when the assumptions of normality is met (See Table 2), and therefore determine which factors influence bias and variance, and how they do so (see Appendix 1). This estimator has been widely criticized, because:

- 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).

However, we will show throughout the simulation section that this estimator show very good inferential properties.

Glass's g_s, Shieh's g_s and Hedges' g'_s. As for Cohen's d_s, an Hedges' correction can be applied in order to compensate for the bias of Glass's d_s, Shieh's d_s and Cohen's d'_s with small sample sizes (see Table 2). This correction has the following general form:

$$g_s = d_s \times \frac{\Gamma(\frac{\nu}{2})}{\sqrt{\frac{\nu}{2}} \times \Gamma(\frac{\nu-1}{2})} \tag{27}$$

Where ν are provided in equation 15 for Glass's g_s, in equation 25 for Hedges' g'_s and in equation 18 for Shieh's g_s. The three corrected estimators are theoretically unbiased when the assumptions of normality is met. Their variance are a function of the same factors as their biased equivalent, however, due to the correction, they have a smaller variance, especially with small sample size, as shown in Table 3. In summary:

- The variances of Hedges' g'_s and Shieh's g_s depend on the total sample size (N), their respective population effect size (δ), and the interaction between the sample sizes ratio and the SD-ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$.
- The variance of Glass's g_s also depends on N, δ and $\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}$. In addition, there is also a main effect of the SD-ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$ on its variance.

How these parameters influence the variance of the estimators will be summarized and illustrated in the section dedicated to the Monte Carlo simulations.

Table 3

Expentency, bias and variance of Glass's d_s and Cohen's d'_s and Shieh's d_s under the assumptions that independent residuals are normally distributed.

	df	Expectancy	Variance
$Glass's\ g_s$	$n_c - 1$	δ_{glass}	$Var(Glass's\ d_s) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}\right)^2$
$Cohen's\ g'_s$	$\frac{(n_1-1)(n_2-1)(s_1^2+s_2^2)^2}{(n_2-1)s_1^4+(n_1-1)s_2^4}$	δ'_{Cohen}	$Var(Cohen's\ d'_s) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}\right)^2$
$Shieh's\ g_s$	$\approx \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	δ_{Shieh}	$Var(Shieh's\ d_s) \times \left(\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}\right)^2$

Note. $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; all estimators are unbiased estimators, because their expectations equal the population effect size δ; equations require $df \geq 3$ and at least 2 subjects per group.

Monte Carlo Simulations: 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 g_s, Glass's g_s (using respectively the sample SD of the first or second group as a standardizer), Hedges' g'_s and Shieh's g_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 (in order to insure the reliability of our calculation method) 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):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{\sqrt{(m_2)^3}} \tag{28}$$

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

$$G_2 = \frac{n-1}{(n-2)(n-3)} \times [(n+1)(\frac{m_4}{(m_2)^2} - 3) + 6] \tag{29}$$

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 5.

Table 4

Number of Combinations of skewness and kurtosis in our simulations.

		Kurtosis		
		0	95.75	**TOTAL**
	0	252	252	**504**
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 5. 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 5), all other parameter values were chosen in order to illustrate the consequences of factors identified as playing a key role on the variance of unbiased 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 $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in step of 1 (so does $\mu_1 - \mu_2$)[8]. 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

[8] In the original plan, we had added 252 simulations in which μ_1 and μ_2 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.

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 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 6.

Table 5

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.

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

1) We previously introduced the fact that raw bias and variances are sometimes misleading. They can give the illusion of huge differences between two estimators, even if these differences only reflect a change of unit (i.e. different population effect sizes). To better understand this, imagine a sample of 15 people for whom we know the height (in meters) and we compute a sample variance of 0.06838. If we convert sizes to centimeters and compute the sample variance again, we find a measure of 683.8 (i.e. 10^3 larger). Both measures represent the same amount of variability, but they are expressed in different units. Similar things occur when comparing the estimates of different population measures. To avoid this possible confusion, we will only present the relative bias and relative variance in all Figures (and anytime we will mention the biases and variances in the results section, we will be referring to relative bias and variance). For interested reader, the raw bias and variance are available on Github.

2) In a purpose of readability, we used different scales for the ordinate axis of our plots, for both relative bias and relative variance, when $G_1 = 6.32$ and $G_2 = 95.75$, in comparisons with all other conditions. Indeed, relative biases and variances are much larger for this specific combination G_1/G_2.

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 relative bias and relative variance for each sub-condition are presented under five different configurations of distributions, using the legend described in Figure 2. When describing the Glass's g_s estimators, we will systematically call "control group" the group the standardizer is computed from (i.e. the first group when using SD_1 as standardizer, the second group when using SD_2 as standardizer). The other group will be called "experimental group".

Figure 2. Legend

When variances are equal across groups. Figures 3 and 4 represent configurations where the equality of variances assumption is met. According to our expectations, one observes that the bias of all estimators is approximately zero as long as the normality assumption is met (first column in both Figures)[9]. However, the further from the normality assumption (i.e. when moving from left to right in Figures), the larger the bias.

Figure 3 illustrates scenarios where both population variances and sample sizes are equal across groups (condition a). One can first notice that all estimators are consistent, as their bias and variance decrease when the total sample size increase. For any departure from the normality assumption, the biases of Hedges' g_s, Hedges' g'_s and Shieh's d_s are similar[10] and smaller than the bias rate of both glass's g_s estimates (i.e. using S_1 and S_2), because the three firsts depend on the total sample size (N; the larger the total sample size, the lower

[9] When looking at the relative bias for all estimators, the maximum departure from zero is 0.0064 when sample sizes are equal across groups, and 0.0065 with unequal sample sizes.

[10] While the biases of Cohen's d_s, Cohen's d'_s and Shieh's d_s are identical, the bias of Hedges' g_s is marginally different than the biases of Hedges' g'_s and Shieh's g_s (these last two being identical). Indeed, because of the sampling error, difference remain between sample variances, even when population variances are equal groups. Because the Hedges' correction applied to Cohen's d_s does not imply the sample variances (unlike the one applied on both other estimators), the bias of Hedges' g_s is slighly different than the bias of Hedges' g'_s and Shieh's g_s

the bias), while Glass's g_s only depends on the control group size (n_c; the larger the control group size, the lower the bias). Moreover, when samples are extracted from skewed distributions, Glass's g_s will show unequal biases as a function of the chosen standardizer (S_1 or S_2), even if both S_1 and S_2 are estimates of the same population variance, based on the same sample size. This is due to non-null correlations of opposite sign between the mean difference ($\bar{X}_1 - \bar{X}_2$) and respectively S_1 and S_2. For interested reader, when a non nul correlation occur between the sample means difference ($\bar{X}_1 - \bar{X}_2$) and the estimators standardizers as well as the way they impact the biases and variances of estimators is detailed in Appendix 3.

VARIANCE : TEXTE THEORIQUE, DE QUOI DEPEND CHAQUE ESTIMATEUR: voir annexe 1. Pour le reste, en discuter avec Christophe. Ensuite, dire ce qui diffère quand on s'écarte de la normalité.

Figure 4 illustrates scenarios where population variances are equal across groups and sample sizes are unequal (condition b). For any departures from the normality assumptions, Hedges' g_s shows the smallest bias. Hedges' g_s and Hedges' g'_s are consistent estimators (i.e. the larger the sample sizes, the lower the bias), unlike Shieh's g_s and Glass's g_s. The bias of Glass's g_s does not depend either on the size of the experimental group or on the total sample size. The only way to decrease the bias of Glass's d_s is therefore to add subjects in the control group[11]. Regarding Shieh's g_s, while increasing sample sizes will decrease the bias for a given sample size ratio, there is a large impact of the sample sizes ratio in order that when the sample sizes ratio moves away from 1 by adding subjects, bias increases. On the other side, when the sample sizes ratio moves closer to 1 by adding subjects, the bias will decrease[12].

[11] e.g. when there are 50 subjects in the control group, the bias will remain identical when there are respectively 20 and 100 subjects in the experimental group

[12] e.g. when there are 20 subjects in one group, the bias of Shieh's d_s is larger when there are 100 subjects in

When samples are extracted from skewed distributions and have unequal sizes (i.e. $n_1 \neq n_2$, the two last columns in Figure 4), for a constant total sample size, $Glass' s g_s$, Shieh's g_s and Hedge's g'_s will show unequal biases depending on which group is the largest one (e.g. when distributions are right-skewed, the bias of all estimator when n_1 and n_2 are respectively 50 and 20 are not the same as their bias when n_1 and n_2 are respectively 20 and 50). This is due to non-null correlations of opposite sign between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and their respective standardizers depending on which group is the largest one, as detailed in Appendix 3.

VARIANCE: the variance of Glass's d_s depends on the sample size of both control and experimental groups, however, Glass's d_s will be less variable when there are more subjects in the control group than when there are more subjects in the experimental group. Finally, as previously observed in Figure 3, the bias and variances of Glass's estimates also depends on the correlation between $\bar{X}_1 - \bar{X}_2$ and the standardizer, when distribution are skewed. [13]

In conclusion, Glass's g_s should always be avoided when the equality of variance assumption is met. Hedge's g_s, Hedges' g'_s and Shieh's g_s are equally performant as long as the sample size ratio is close to 1. However, when designs are highly unbalanced, Shieh's g_s

---

the other group than when there are only 50 subjects in the other group, because despite the increasing total sample size, the sample size ratio increases. On the other side, when there are 100 subjects in one group, the bias and variance of Shieh's d_s are smaller when there are 50 subjects in the other gruop than when there are only 20 subjects in the other group, because the sample size ratio decreases

[13] In our simulations, the worst configuration will occure when the standardizer is computed with the smaller group and is negatively correlated with the sample means difference (i.e. when choosing SD2 when distributions are right-skewed, and when choosing SD1 when distributions are left-skewed). Again, we should remind that in all our simulations, the population mean difference is positive. If mean 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.

is not consistent anymore. While Hedge's g'_s is consistent, Hedges's g_s remains a better estimator.

When variances are unequal across groups. Figures 5 to 10 represent configurations where the equality of variances assumption is not met. According to our expectations, one observes that the bias of all estimators is approximately zero as long as the normality assumption is met (first column in the three Figures), and the further from the normality assumption (i.e. when moving from left to right in Figures), the larger the bias[14]. You might find it surprising that the bias of Hedges' g_s remain very small troughout these conditions. As reminded in the section "Different measures of effect size", Hedges' g_s should be avoided when population variances and sample sizes are unequal across groups, because of the pooled error term. When pooling the estimates of two unequal population variances, the resulting estimator will be lower (in case of positive pairing) or larger (in case of negative pairing) as it should be. At the same time, when pooling two unequal population variances, the population effect size will also be lower (in case of positive pairing) or larger (in case of negative pairing) as it should be. As a consequence, the distorsion cannot be seen through the difference between the expected estimator and the population effect size measure. For this reason, the bias and variance of Hedges' g_s will not be taken into account in the following comparisons.

Figures 5 and 6 are dedicated to scenarios where population variances are unequal between groups and sample sizes are equal (condition c). In Figure 5, scenarios are subdivided as a function of the sample sizes and one can notice that all estimators are consistent, as their bias and variance decrease when the total sample size increases. In Figure 6, scenarios are subdivided as a function of the SD-ratio. Because the comparison pattern remains very similar for all sample sizes, we present only scenarios when sample sizes equal 20. One observes then that the biases of Shieh's g_s and Hedges' g'_s are identical, for

[14] When looking at the relative bias for all estimators, the maximum departure from zero is 0.0173 when sample sizes are equal across groups, 0.0088 when there is a positive correlation between sample sizes and sample variances, and 0.0274 when there is a negative correlation between sample sizes and sample variances.

any departures from the normality assumption. When distributions are symmetric (second column in Figure 6), the bias of both Shieh's g_s and Hedges' g'_s depend on the SD-ratio in order that the larger the difference between σ_1 and σ_2, the larger the bias. On the other side, the bias of Glass's g_s does not depend on the SD-ratio, but it remains larger than the bias of Shieh's g_s and Hedges' g'_s, even when the SD-ratio is very large $\left(i.e.\ \frac{\sigma_1}{\sigma_2} = 10\ or\ 0.1\right)$. However, when samples are extracted from skewed distributions, the bias of Glass's g_s is sometimes smaller and sometimes larger than the bias of Shieh's g_s and Hedge's g'_s. This is mainly due to the fact that when two samples of same sizes are extracted from two skewed distributions with unequal variances (i.e. $\sigma_1 \neq \sigma_2$, the two last columns in Figure 6), there will be non-null correlations of opposite sign between the mean difference ($\bar{X}_1 - \bar{X}_2$) and the standardizer of *all* estimators, depending on which population variance is the largest one. [15].

VARIANCE: Glass's more variable than Shieh's g_s and Hedge's g'_s. Moreover, the relative variance of Glass's d_s is larger when computing the standardizer based on the sample extracted from the less variable population, as shown in Figures 6 and **??**.

Figures 7 to 10 are dedicated to scenarios where both samples sizes and population variances differ across groups. Because of a high number of combinations between the sample sizes-ratio and the variances-ratio in our simulations, we decided to present only conditions where the total sample sizes equals either 150 (with $n_1 = 100$ and $n_2 = 50$, or $n_1 = 50$ and $n_2 = 100$), or 120 (with $n_1 = 100$ and $n_2 = 20$, or $n_1 = 20$ and $n_2 = 100$)). VERIFIER SI CA VAUT LA PEINE DE LAISSER LES 2. JE M ATTENDS A AVOIR UN PATTERN DIFFERENT DANS LES DEUX CAS DU A L INTERACTION ENTRE NRATIO ET SD RATIO EN DEFAVEUR DU SHIEH.

---

[15] When population variances are unequal, a non-null correlation occurs between standardizers estimates and $\bar{X}_1 - \bar{X}_2$. The sign of the correlation between the standardizer and the means difference will be the same as the sign of the correlation between the mean difference and the estimate of the larger population variance. For interested readers, this is detailed in Appendix 3.

Figures 7 and 8 refer to scenarios where both population variances sample sizes are unequal between groups, with a positive correlation between them (i.e. the larger variance is associated with the larger sample size; condition d).

Figures 9 and 10 refer to scenarios where both population variances sample sizes are unequal between groups, with a negative correlation between them (i.e. the larger variance is associated with the larger sample size; condition d).

Figure **??** and **??** refer to conditions where there is a pairing between population variances and sample sizes. We know that in these configurations, computing a pooled variance term does not make sense (see the second remark at the beginning of the result section), and therefore, we will not discuss the Cohen's d_s.We will only compare the performances of Glass's d_s, Shieh's d_s and Shieh's d_s^*.

Figure **??** 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 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 **??** 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 remains the best measure when the asssumption of equal variances is met. When variances are unequal across populations, Cohen's d_s performs 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 becomes irrelevant. Glass's d_s is most of the time the more biased and variable measure. 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. Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods*, *5*(1), 1–13. doi:10.22237/jmasm/1146456060

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.

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., & 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.

Huynh, C.-L. (1989). A unified approach to the estimation of effect size in meta-analysis. San Francisco: Paper presented at the Annual Meeting of the American Educational Research Association.

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

Keselman, H. J., Algina, J., Lix, L. M., Deering, K. N., & Wilcox, R. R. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*(2), 110–129. doi:10.1037/1082-989X.13.2.110

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

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

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods*, *5*(1), 1–13. doi:10.22237/jmasm/1146456060

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.

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., & 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.

Huynh, C.-L. (1989). A unified approach to the estimation of effect size in meta-analysis. San Francisco: Paper presented at the Annual Meeting of the American Educational Research Association.

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

Keselman, H. J., Algina, J., Lix, L. M., Deering, K. N., & Wilcox, R. R. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*(2), 110–129. doi:10.1037/1082-989X.13.2.110

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

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

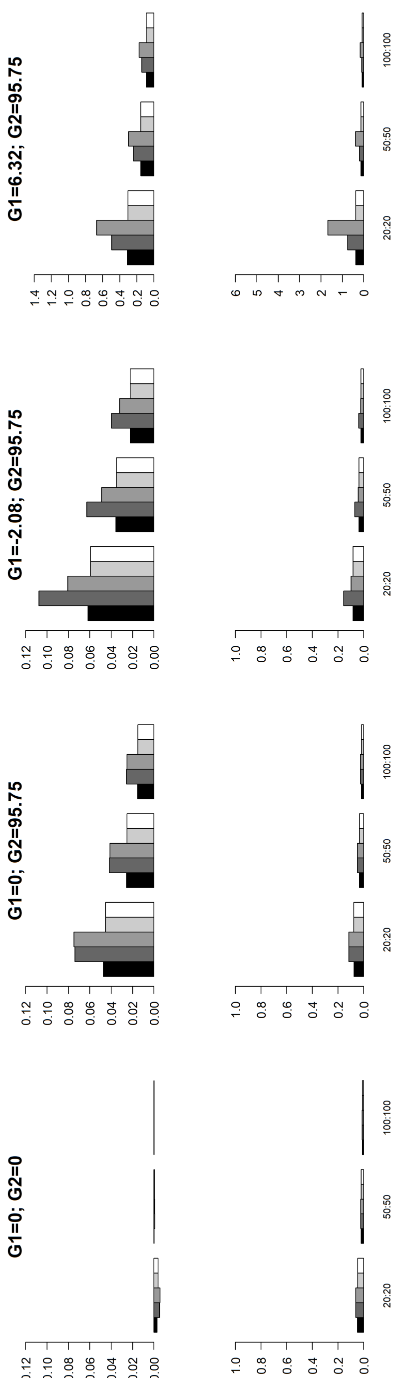

Figure 3. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)

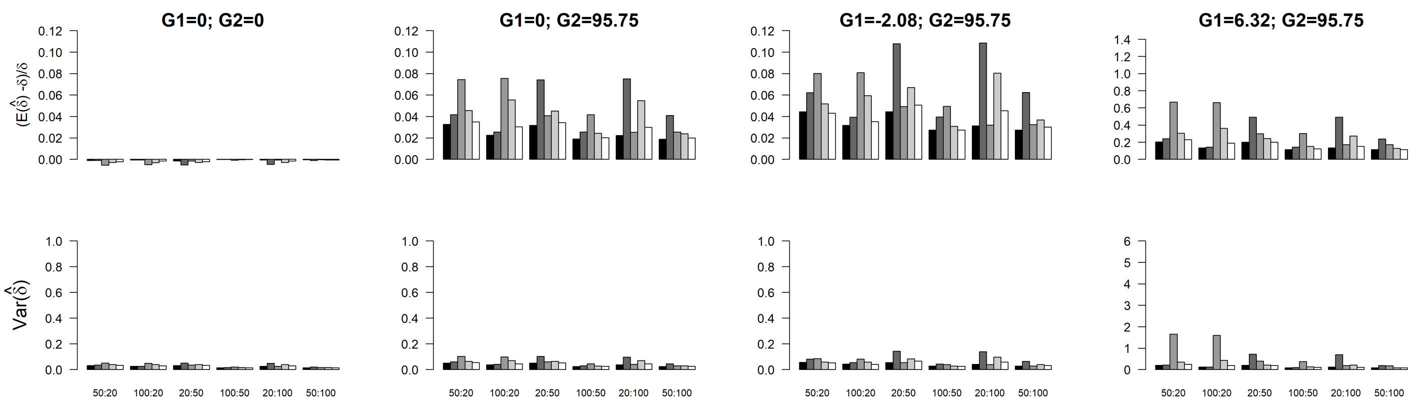

Figure 4. Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)

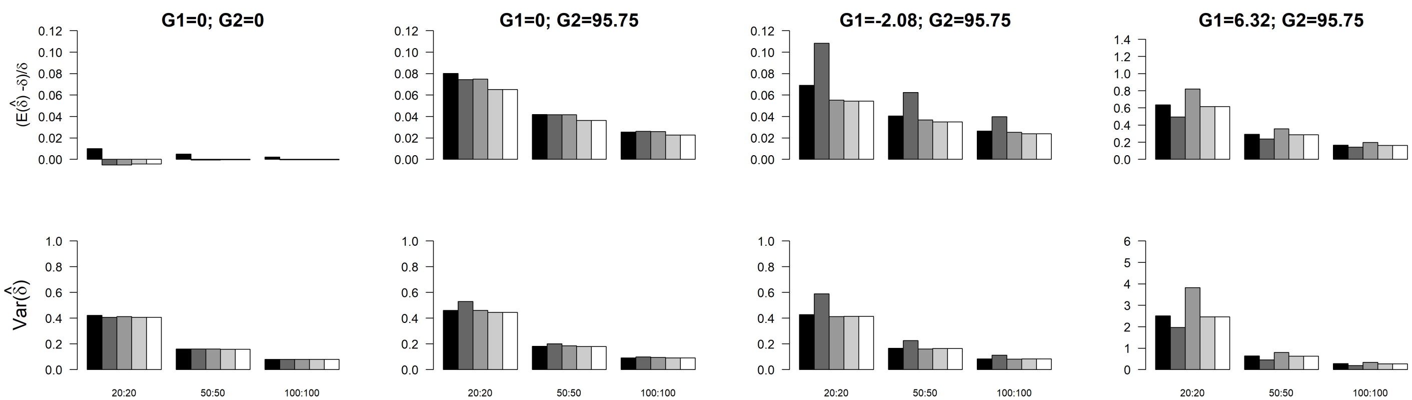

Figure 5. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c), as a function of n-ratio

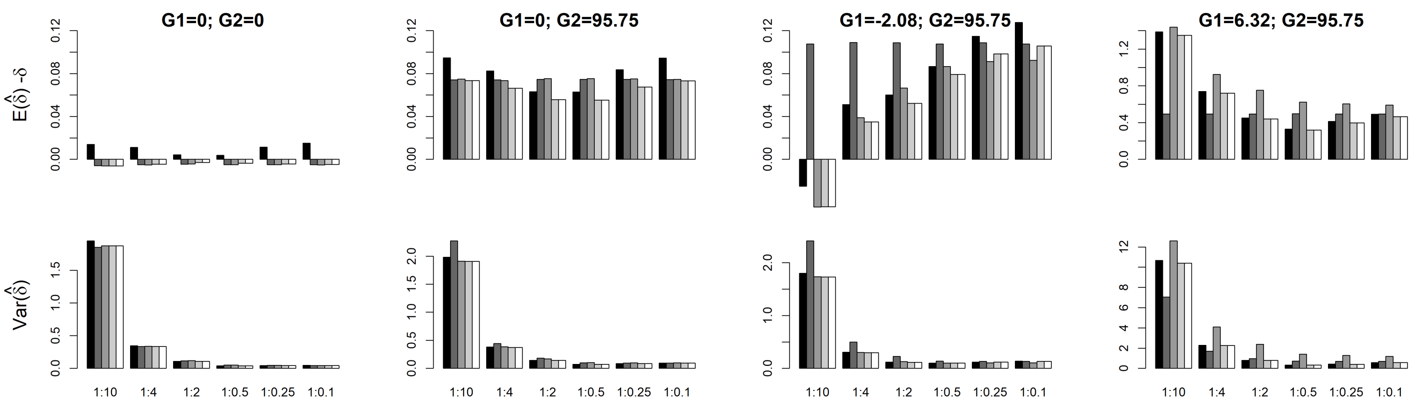

Figure 6. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the SD-ratio (when $n_1 = n_2 = 20$)

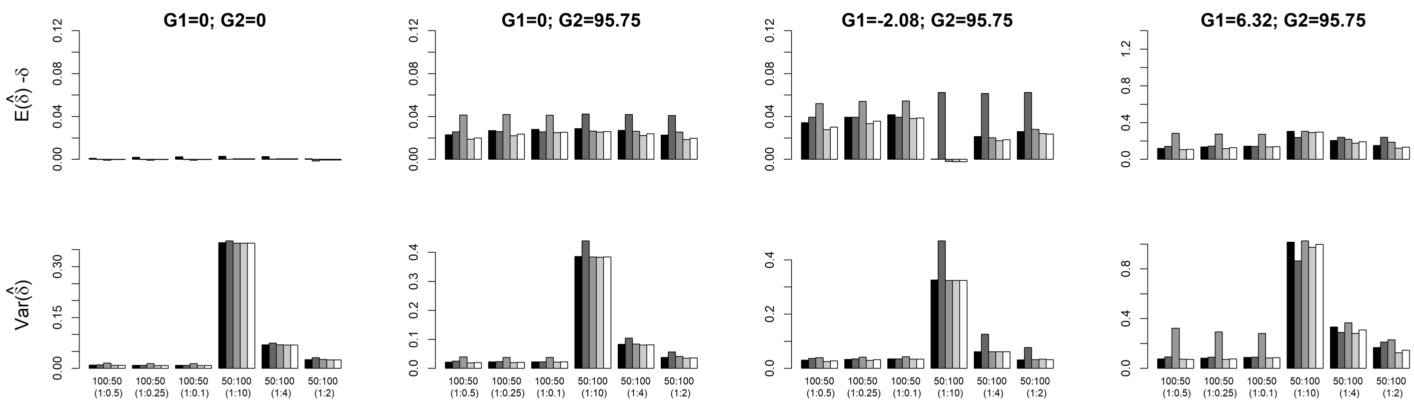

Figure 7. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups and total sample (N) equals 150, with positive correlation between them (condition d)

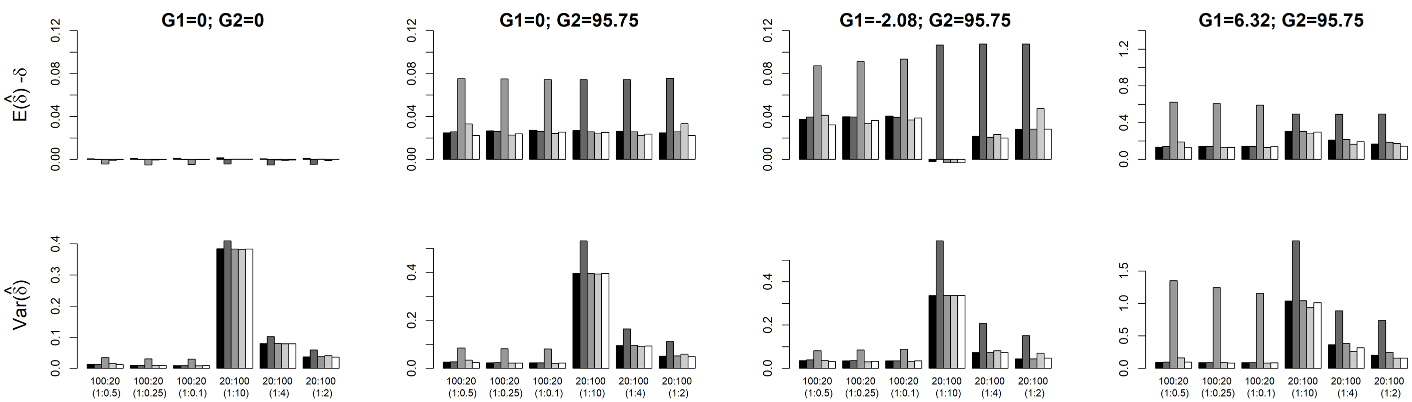

Figure 8. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups and total sample (N) equals 120, with positive correlation between them (condition d)

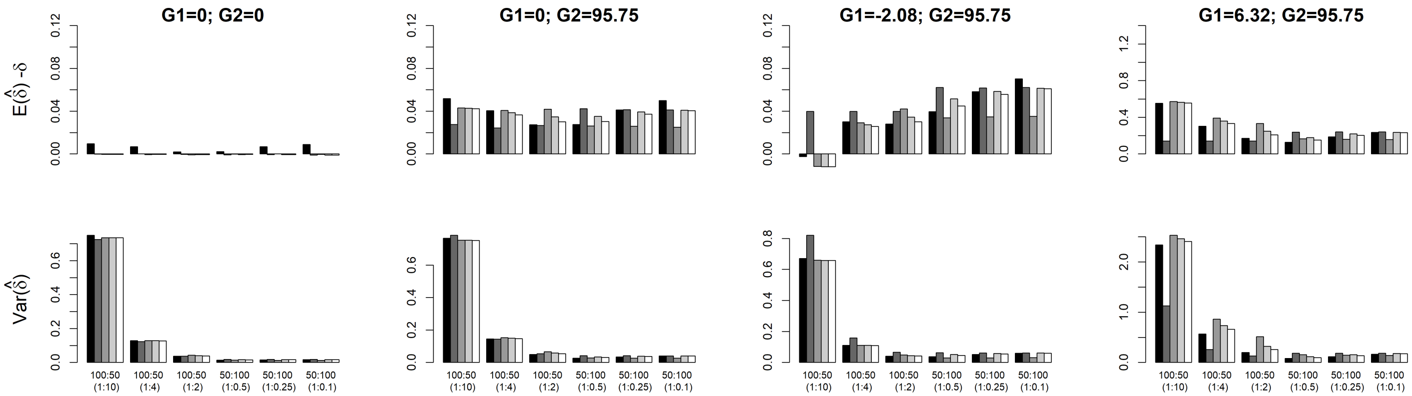

Figure 9. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups and total sample (N) equals 150, with negative correlation between them (condition e)

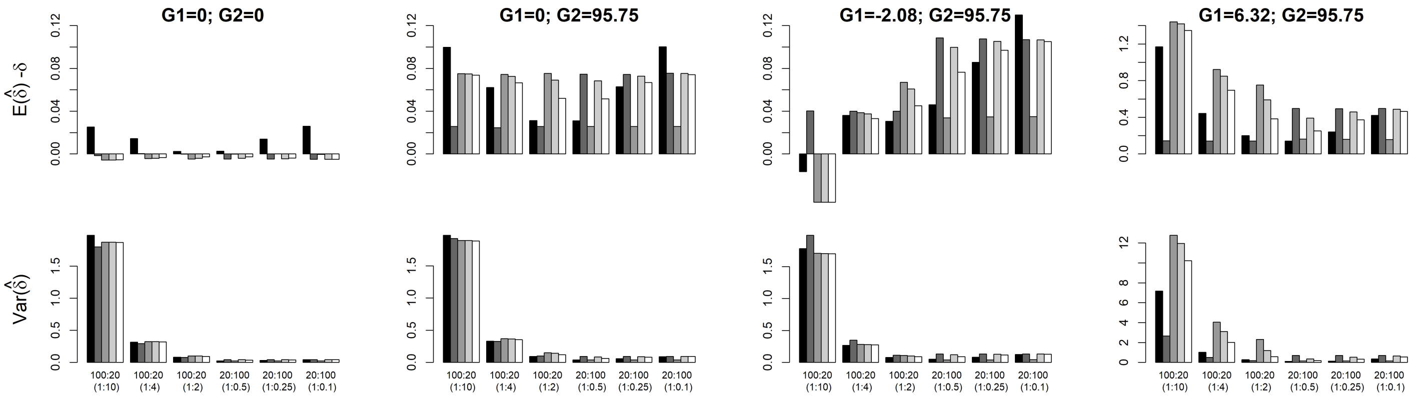

Figure 10. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups and total sample (N) equals 120, with negative correlation between them (condition e)

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.