1   What measure of effect size when comparing two independent means?

2   What measure of effect size when comparing two independent means?

3   Effect sizes are important issues in applied research. There is a strong call for

4 researchers to report and interpret effect sizes and associated confidence intervals. This

5 practice is also highly endorsed by the American Psychological Association (APA) and the

6 American Educational Research Association (AERA; American Psychological Association,

7 2010; American Educational Research Association, 2006).

8   In "between-subject" designs where individuals are randomly assigned into one of two

9 independent groups and group scores are compared based on their means, the dominant

10 estimator of effect size is Cohen's d_s, where the sample mean difference is divided by the

11 pooled sample standard deviation (Peng, Chen, Chiang, & Chiang, 2013; Shieh, 2013). This

12 estimator is available in many statistical softwares, such as SPSS or Stata. However,

13 computing the pooled sample standard deviation implies that both sample variances are

14 estimates of a common population variance, and it has been widely argued that there are

15 many fields in psychology where the assumption of equal variances between two populations

16 is ecologically unlikely (Delacre, Lakens, & Leys, 2017; Erceg-Hurn & Mirosevich, 2008;

17 Grissom, 2000). The debate surrounding the assumption of equal variances has been widely

18 explored in the context of hypothesis testing and it is becoming more common in statistical

19 software to present a t-test that does not hold under this assumption by default when

20 performing hypothesis tests, namely the Welch's t-test (e.g., R, Minitab). However, similar

21 issues for the measures of effect sizes have received less attention. One possible reason is that

22 researchers cannot find a consensus on which alternative should be used (Shieh, 2013). Even

23 within the very specific context of standardized sample mean difference estimates, there is

24 little agreement between researchers as to which is the most suitable estimator. In this

25 article, we will review the main measures of d family effect sizes that are proposed in the

26 literature, namely Cohen's d_s, Glass's d_s, Shieh's d_s and Cohen's d_s^*. We will compare them

27 through simulations and suggest that Cohen's d_s^* is a better default than Cohen's d_s. We will

₂₈ also provide practical recommendations and useful tools (i.e. R package, Shiny app) to
₂₉ compute relevant measures of effect sizes and confidence intervals around them.

₃₀   Before reviewing the main measures of the *d*-family, we will first list the different
₃₁ situations that call for effect sizes measures (and the link between effect size and both
₃₂ statistical and practical significance). We will also describe the properties that a good effect
₃₃ size estimator ought to possess, properties that we will take into account in our Monte Carlo
₃₄ simulations in order to compare different estimators.

Effect size: definition and purposes

₃₆   The effect size is a measure of the magnitude of an effect. In the context of the
₃₇ comparison of two groups based on their means, when the null hypothesis is the absence of
₃₈ effect, *d*-family effect size estimators estimate the magnitude of the differences between
₃₉ parameters of two populations groups are extracted from (e.g. the mean; Peng & Chen,
₄₀ 2014). Such a measure can be used in three different perspectives.

₄₁   First, effect size measures can be used for *interpretative* purposes, namely to assess the
₄₂ practical significance of a result (i.e. all what refers to the relevance of an effect in real life,
₄₃ such as clinical, personal, social, professionnal relevance). Note, however, that in this context
₄₄ it is important not to overestimate the contribution of the measures of effect size: effect size
₄₅ is **not** a measure of the importance or the relevance of an effect in real life (even if
₄₆ benchmarks about what should be a small, medium or large effect size might have
₄₇ contributed to viewing the effect size as so; Stout & Ruble, 1995). It is only a mathematical
₄₈ indicator of the magnitude of a difference, which depends on the way a variable is converted
₄₉ into numerical indicator. In order to assess the meaningfulness of an effect, we should be
₅₀ able to relate this effect size estimate with behaviors/meaningful consequences in the real
₅₁ world (Andersen, McCullagh, & Wilson, 2007). For example, let us imagine a sample of
₅₂ students in serious school failure who are randomly divided into two groups: an experimental

₅₃ group following a training program and a control group. At the end of the training, students

₅₄ in the experimental group have on average significantly higher scores on a test than students

₅₅ in the control group, and the difference is large (e.g. 30 percents). Does it automatically

₅₆ mean that students in the experimental condition will be able to pass to the next grade and

₅₇ to continue normal schooling? Whether the computed magnitude of difference is an

₅₈ important, meaningful change in everyday life refers to the interpretation of treatment

₅₉ outcomes and is neither a statistical nor mathematical concept, but is related to the

₆₀ underlying theory that posits an empirical hypothesis.

₆₁   Second, effect size measures can be used for *comparative* purposes, that is, to assess the

₆₂ stability of results across designs, analyses, sample sizes. This includes the comparison of

₆₃ results from 2 or more studies and the incorporation of results in meta-analysis.

₆₄   Third, effect size measures can be used for *inferential* purposes. The effect sizes from

₆₅ previous studies can be used in a prior power analysis when planning a new study (Lakens,

₆₆ 2013; Prentice & Miller, 1990; Stout & Ruble, 1995; Sullivan & Feinn, 2012; Wilkinson & the

₆₇ Task Force on Statistical Inference, 1999). Moreover, while point effect size estimators should

₆₈ not replace the null hypothesis testing by themselves (Fan, 2001) [1], conventional hypothesis

₆₉ testing and confidence interval around a point estimate are equivalent decision procedures. A

₇₀ confidence interval contains all the information that a *p*-value of a test based on the same

₇₁ estimator does: if the area of the null hypothesis is out of the $(1 - \alpha)$-confidence interval,

₇₂ then the hypothesis test would also result in a *p*-value below the nominal alpha level.

₇₃ Hypothesis tests and confidence intervals based on the same statistical quantity (this is an

₇₄ essential requirement) are thus directly related. At the same time, the intervals provide extra

₇₅ information about the precision of the sample estimate for inferential purposes, and therefore

———

[1] Statistical testing and point effect size estimators do not answer the same question, even if they are related. Statistical testing allows the researcher to determine whether data are surprisingly different from what is expected under the null.

76 on how confident we can be in the observed results (Altman, 2005; Ellis, 2015): the narrower

77 the interval, the higher the precision. On the other hand, the wider the confidence interval,

78 the more the data lacks precision (for example, because the sample size is too small).

Properties of a good effect size estimator

80   The empirical value of an estimator (called estimate) depends on the sampling, in

81 other words, different samples extracted from the same population will of course lead to

82 different estimates for a same estimator. The *sampling distribution* of the estimator is the

83 distribution of all estimates, based on all possible samples of size n extracted from one

84 population. Studying the sampling distribution is very useful, as it allows us to assess the

85 qualities of an estimator. More specifically, three desirable properties a good estimator

86 should possess for inferential purposes are: **unbiasedness**, **consistency** and **efficiency**

87 (Wackerly, Mendenhall, & Scheaffer, 2008).

88   An estimator is unbiased if the distribution of estimates is centered around the true

89 population parameter. On the other hand, an estimator is positively (or negatively) biased if

90 the distribution is centered around a value that is higher (or lower) than the true population

91 parameter (see Figure 1). In other words, the bias tells us if estimates are accurate, on

92 average. The *bias* of a point estimator $\hat{\delta}$ can be computed as

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

93 where $E(\hat{\delta})$ is the expectation of the sampling distribution of the estimator and δ is the true

94 (population) parameter.

95   As we can see in Tables 1 and 2, the bias is directly related with the population effect

96 size: the larger the population effect size, the larger the bias. It might therefore be interesting

97 to also define the *relative bias* as the ratio between the bias and the population effect size:

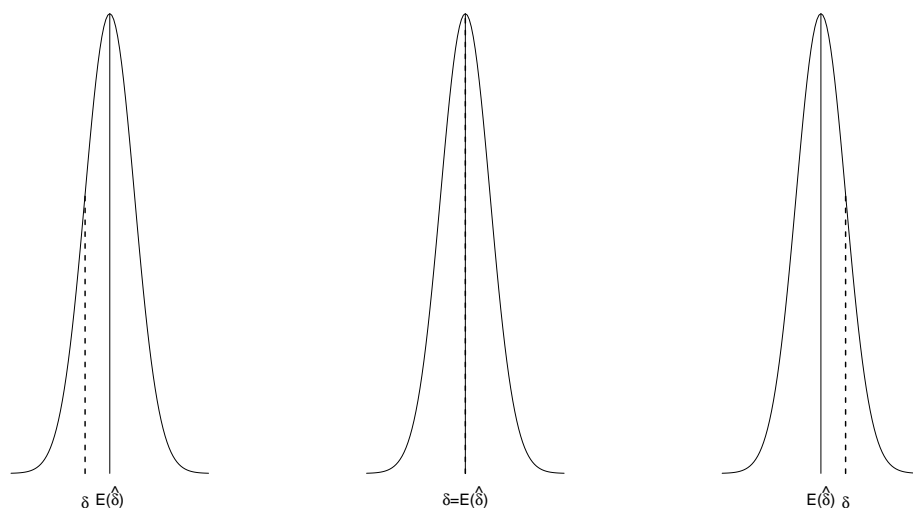$$\delta_{relative\ bias} = \frac{E(\hat{\delta}) - \delta}{\delta} \tag{2}$$

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

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 to 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 [2]. Again,

---

[2] The Cramér-Rao inequality provides a theoretical lower bound for the variance of unbiased estimators. An

110 the variance of an estimator $\hat{\delta}$ is a function of its size (the larger the estimator, the larger the

111 variance) and, therefore, we might be interested in reducing the effect size impact by

112 computing the *relative variance* as the ratio between the variance and the square of the

113 population estimator:

$$relative\ var(\hat{\delta}_1) = \frac{Var(\hat{\delta})}{\delta^2} \tag{4}$$

114 Note that both unbiasedness and efficiency are very important. An unbiased estimator with

115 such a large variance that some estimates are extremely far from the real parameter is as

116 undesirable as a parameter which is highly biased. In some situations, it is better to have a

117 slightly biased estimator with a tight shape around the biased value (so that each estimate

118 remains relatively close to the true parameter and one can apply bias correction techniques)

119 rather than an unbiased estimator with a large variance (Raviv, 2014).

120   Finally, the last property of a good point estimator is **consistency**: consistency means

121 that the bigger the sample size, the closer the estimate is to the population parameter. In

122 other words, the estimates *converge* to the true population parameter.

123   Beyond these inferential properties, Cumming (2013) reminds that an effect size

124 estimator needs to have a constant value across designs, in order to be easily interpretable

125 and to be included in meta-analysis. In other words, it should achieve the property of

126 **generality**. To our concern, while we focus on between-subject designs, we expect to

127 highlight an estimate of the population effect size with a constant value for any sample sizes

128 ratio.

<div align="center">

Different measures of effect sizes

</div>

      The *d*-family effect sizes are commonly used for mean differences between groups or

conditions. The population effect size is defined as

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

———

estimator reaching this bound is therefore most efficient.

130 where both populations follow a normal distribution with mean μ_j in the jth population

131 $(j = 1, 2)$ and common standard deviation σ. There exist different estimators of this effect

132 size measure. For all of them, the mean difference is estimated by the difference $\bar{X}_1 - \bar{X}_2$ of

133 both sample means. When the equality of variances assumption is assumed, σ is estimated

134 by pooling both sample standard deviations (S_1 and S_2). When the equality of variances

135 assumption cannot be assumed, alternatives to the common standard deviation are available.

136 Throughout this section, we will present some of these estimators, separately depending on

137 whether they rely on the assumption of equality of variances or not. For each of them, we

138 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):

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

140 where S_j is the standard deviation and n_j the sample size of the jth sample $(j = 1, 2)$. The

141 reasoning behind this measure is to make use of the fact that both samples share the same

142 population variance (Keselman, Algina, Lix, Deering, & Wilcox, 2008), hence we achieve a

143 more accurate estimation of the population variance by pooling both estimates of this

144 parameter (i.e. S_1 and S_2). Since the larger the sample size, the more accurate the estimate,

145 we give more weight to the estimate based on the larger sample size. Cohen's d_s is directly

146 related with Student's t-statistic:

$$t_{Student} = \frac{Cohen's\ d_s}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leftrightarrow Cohen's\ d_s = t_{Student} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5}$$

147 Under the assumption of normality and equal variances between groups, Student's t-statistic

148 follows a t-distribution with known degrees of freedom

$$df_{Student} = n_1 + n_2 - 2 \tag{6}$$

149 and noncentrality parameter [3]

$$ncp_{Student} = \frac{\delta_{Cohen}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

150 where $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ and $\sigma_{pooled} = \sqrt{\frac{(n_1-1) \times \sigma_1^2 + (n_2-1) \times \sigma_2^2}{n_1 + n_2 - 2}}$. The relationship described in

151 equation 5 and the theoretical distribution of Student's t-statistic allow us to theoretically

152 determine the sampling distribution of Cohen's d_s, and therefore, its theoretical expectation

153 and variance when the assumptions of normality and equal variances are met. All these

154 equations are provided in Table 1. For interested readers, we have theoretically studied the

155 bias and variance of Cohen's d_s based on Table 1 (as well as the bias and variance of all

156 other estimators described later, based on Tables 2 and 3) in Supplemental Material 1 [4], so

157 as to determine the way different parameters influence them. Main results will be illustrated

158 in a later section called "Monte Carlo Simulations: assessing the bias, efficiency and

159 consistency of 5 estimators" .

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 and Olkin

(1985) has defined a bias-corrected version:

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

where $df_{Student}$ has been defined in equation 6 and $\Gamma()$ is the gamma function (for integers,

$\Gamma(x)$ is the factorial of x minus 1: $\Gamma(x) = (x-1)!$; Goulet-Pelletier & Cousineau, 2018).

This equation can be approximated as follows:

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

———

[3] 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 a noncentral t-distribution will arise.

[4] All supplemental Materials are available on Github: https://github.com/mdelacre/Effect-sizes/

160 where N is the total sample size. Hedges' g_s is theoretically unbiased when the assumptions

161 of normality and equal variances are met (see Table 1). Moreover, while the variance of both

162 Cohen's d_s and Hedges' g_s depend on the same parameters (i.e. the total sample size (N) and

163 the sample sizes ratio $\left(\frac{n_1}{n_2}\right)$), Hedges' g_s is less variable, especially with small sample sizes [5].

164   While the pooled error term is the best choice when variances are equal between

165 groups (Grissom & Kim, 2001), it may not be well advised for use with data that violate this

166 assumption (Cumming, 2013; Grissom & Kim, 2001, 2005; Kelley, 2005, 2005; Shieh, 2013).

167 When variances are unequal between groups, the expression in equation A is no longer valid

168 because both groups do not share a common population variance. If we pool the estimates of

169 two unequal population variances, the estimator of effect size will be lower as it should be in

170 case of positive pairing (i.e. the group with the larger sample size is extracted from the

171 population with the larger variance) and larger as it should be in case of negative pairing

172 (i.e. the group with the larger sample size is extracted from the population with the smaller

173 variance). Because the assumption of equal variances across populations is very rare in

174 practice (Cain, Zhang, & Yuan, 2017; Delacre et al., 2017; Delacre, Leys, Mora, & Lakens,

175 2019; Erceg-Hurn & Mirosevich, 2008; Glass, Peckham, & Sanders, 1972; Grissom, 2000;

176 Micceri, 1989; Yuan, Bentler, & Chan, 2004), both Cohen's d_s and Hedges' g_s should be

177 abandoned in favor of an alternative robust to unequal population variances.

---

[5] $.52 \leq \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}\right]^2 < 1$ 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.

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

[178] *Note.* $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ and $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}{\Gamma(\frac{df}{2})}$; Cohen's d_s is a biased estimator, because its expectation differs from the

[179] population effect size. Moreover, the larger the population estimator (δ_{Cohen}), the larger the bias. Indeed, the bias is the

[180] difference between the expectation and δ_{Cohen}: $\delta_{bias} = \delta_{Cohen} \times (c_f - 1)$. On the other hand, Hedges' g_s is an unbiased estimator,

[181] because its expectation equals δ_{Cohen}; 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, and Smith (2005) recommend using the standard deviation 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

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

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 S_c 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 on the chosen SD as standardizer, measures could be substantially different (Shieh, 2013). The distribution of Glass's d_s is defined as (Algina, Keselman, & Penfield, 2006):

$$Glass's \ d_s \sim \sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}} \times t_{df,ncp} \tag{7}$$

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:

$$df = n_c - 1 \tag{8}$$

$$ncp = \frac{\delta_{Glass}}{\sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}}}$$

where $\delta_{Glass} = \frac{\mu_c - \mu_e}{\sigma_c}$ and μ_c and μ_e are respectively the mean of the populations control and experimental groups are extracted from. Thanks to equation 7, we can compute its

theoretical expectation and variance when the assumption of normality is met (see Table 2),

and therefore determine which factors influence bias and variance, and how they do so (see

Supplemental Material 1).

Table 2

Expentency, bias and variance of Glass's d_s and Cohen's d_s^* and Shieh's d_s under the assumption that independent residuals are normally distributed.

	df	Expectation	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^*	$\frac{(n_1-1)(n_2-1)(\sigma_1^2+\sigma_2^2)^2}{(n_2-1)\sigma_1^4+(n_1-1)\sigma_2^4}$	$\delta_{Cohen}^* \times c_f$	$\frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta_{Cohen}^*)^2 \left(\frac{df}{df-2} - c_f^2\right)$
		$\approx \delta_{Cohen}^* \times \frac{4df-1}{4(df-1)}$	$\approx \frac{df}{df-2} \times \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{\sigma_1^2+\sigma_2^2} + (\delta_{Cohen}^*)^2 \left[\frac{df}{df-2} - \left(\frac{4df-1}{4(df-1)}\right)^2\right]$
Shieh's d_s	$\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	$\delta_{Shieh} \times c_f$	$\frac{df}{(df-2)N} + \delta_{Shieh}^2 \left(\frac{df}{df-2} - c_f^2\right)$

Note. $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \frac{\mu_c-\mu_e}{\sigma_c}$, $\delta_{Shieh} = \frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N}+\frac{\sigma_2^2}{n_2/N}}}$ and $\delta_{Cohen}^* = \frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}$; all estimators are biased estimators, because their expectations differ from the population effect size δ. Moreover, the larger the population estimator (δ), the larger the bias. Indeed, the bias is the difference between the expectation and δ: $\delta_{bias} = \delta \times (c_f - 1)$. equations require $df \geq 3$ and at least 2 subjects per group.

Shieh's d_s. Kulinskaya and 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 and Staudte (2007), proposed a modification of the exact *SD* of the sample mean difference:

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

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

$$Shieh's\ d_s = \frac{t_{Welch}}{\sqrt{N}} \leftrightarrow t_{welch} = Shieh's\ d_s \times \sqrt{N} \tag{9}$$

206 The exact distribution of Welch's *t*-statistic is more complicated than the exact distribution
207 of Student's *t*-statistic, but it can be approximated, under the assumption of normality, by a
208 *t*-distribution with degrees of freedom and noncentrality parameters (Shieh, 2013; Welch,
209 1938):

$$df_{Welch} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1} + \frac{(\sigma_2^2/n_2)^2}{n_2-1}} \tag{10}$$

$$ncp_{Welch} = \delta_{Shieh} \times \sqrt{N} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n_2}}}$$

210 where $\delta_{Shieh} = \frac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N}+\frac{\sigma_2^2}{n_2/N}}}$. The relationship described in equation 9 and the theoretical
211 distribution of Welch's *t*-statistic allow us to theoretically approximate the sampling
212 distribution of Shieh's d_s. Based on the sampling distribution of Shieh's d_s, we can estimate
213 its theoretical expectation and variance under the assumption of normality (see Table 2),
214 and this way determine which factors influence bias and variance, and how they do so (see
215 Supplemental Material 1).

216   As demonstrated in the Appendix, when variances and sample sizes are equal across
217 groups, the biases and variances of Cohen's d_s and Shieh's d_s are identical except for a
218 constant, as well as δ_{Cohen} and δ_{Shieh}:

$$\delta_{Cohen} = 2 \times \delta_{Shieh} \quad (\text{considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \tag{11}$$

219

$$Bias_{Cohen's\ d_s} = 2 \times Bias_{Shieh's\ d_s} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2 \text{)} \tag{12}$$

220

$$Var_{Cohen's\ d_s} = 4 \times Var_{Shieh's\ d_s} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2 \text{)} \tag{13}$$

221 We can deduce from equations 11, 12 and 13 that relative to their respective population

222 effect size, Cohen's d_s and Shieh's d_s are equally accurate. This is a good illustration of our

223 motivation to favor relative bias and variance (previously defined in equations 2 and 4) over

224 the most commonly used raw bias and variance (previously defined in equations 1 and 3).

225   Except for this very specific situation, according to the statistical properties of Welch's

226 statistic under heteroscedasticity, Shieh's d_s accounts for the sample sizes allocation ratio.

227 The lack of generality caused by taking this specificity of the design into account has led

228 Cumming (2013) to question its usefulness in terms of interpretability: when keeping

229 constant the mean difference $(\bar{X}_1 - \bar{X}_2)$ as well as S_1 and S_2, Shieh's d_s will vary as a

230 function of the sample sizes allocation ratio (unlike Cohen's d_s^* that we will define right

231 below). At the population level, δ_{Shieh} also depends on the sample sizes allocation ratio, as

232 illustrated in the following shiny application: https://mdelacre.shinyapps.io/shiehvsCohen/.

Cohen's d_s^*. The sample mean difference divided by the square root of the non

pooled average of both variance estimates was suggested by Welch (1938). This yields:

$$Cohen's\ d_s^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(S_1^2 + S_2^2)}{2}}}$$

233 where $\bar{X}_j$ is the mean and S_j is the standard deviation of the jth sample (j = 1,2). We know

234 the distribution of Cohen's d_s^* (Huynh, 1989):

$$Cohen's\ d_s^* \sim \sqrt{\frac{2(n_2 \times \sigma_1^2 + n_1 \times \sigma_2^2)}{n_1 n_2 (\sigma_1^2 + \sigma_2^2)}} \times t_{df^*, ncp^*} \tag{14}$$

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

$$df^* = \frac{(n_1 - 1)(n_2 - 1)(\sigma_1^2 + \sigma_2^2)^2}{(n_2 - 1)\sigma_1^4 + (n_1 - 1)\sigma_2^4} \tag{15}$$

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

where $\delta^*_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$. Thanks to equation 14, we can compute its theoretical expectation and variance when the assumption of normality is met (see Table 2), and therefore determine which factors influence bias and variance, and how they do so (see Supplemental Material 1). This estimator has been widely criticized, because:

- it results in a variance term of an artificial population (i.e. since the variance term does not estimate the variance of one or the other group, the composite variance is an estimation of the variance 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 exhibits very good inferential properties. Moreover, it has a constant value across sample sizes ratios, as shown in the previously introduced Shiny App.

Glass's g_s, Shieh's g_s and Hedges' g_s^*. As for Cohen's d_s, a 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:

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

where the distinct values of ν are provided in equation 8 for Glass's g_s, in equation 15 for Hedges' g_s^* and in equation 10 for Shieh's g_s. The three corrected estimators are theoretically unbiased when the assumption of normality is met. Their variance is a function of the same parameters 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

255   respective population effect size (δ), and the interaction between the sample sizes ratio

256   and the SD-ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2} \right)$.

257   •   The variance of Glass's g_s also depends on N, δ and $\frac{n_c}{n_e} \times \frac{\sigma_c}{\sigma_e}$. In addition, there is also

258   a main effect of the SD-ratio $\left(\frac{\sigma_c}{\sigma_e} \right)$ on its variance.

259   How these parameters influence the variance of the estimators will be summarized and

260   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 assumption that independent residuals are normally distributed.

	df	Expectation	Variance
Glass's g_s	$n_c - 1$	δ_{glass}	$Var(Glass's\ d_s) \times \left(\dfrac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$
Cohen's g_s^*	$\dfrac{(n_1-1)(n_2-1)(\sigma_1^2+\sigma_2^2)^2}{(n_2-1)\sigma_1^4+(n_1-1)\sigma_2^4}$	δ_{Cohen}^*	$Var(Cohen's\ d_s^*) \times \left(\dfrac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$
Shieh's g_s	$\approx \dfrac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1}+\frac{(\sigma_2^2/n_2)^2}{n_2-1}}$	δ_{Shieh}	$Var(Shieh's\ d_s) \times \left(\dfrac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \right)^2$

Note. $c_f = \dfrac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \frac{\mu_c-\mu_e}{\sigma_c}$, $\delta_{Shieh} = \dfrac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N}+\frac{\sigma_2^2}{n_2/N}}}$ and $\delta_{Cohen}^* = \dfrac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}$; 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 was generated for 1,008 scenarios as a function of different criteria that will be explained below. In 252 scenarios, samples were extracted from a normally distributed population (in order to ensure 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 Fisher's skewness

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

and kurtosis

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

where n is the sample size and m_2, m_3 and m_4 are respectively the second, third and fourth centered moments. 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 4.

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**

[287]   *Note.* Fisher's skewness (G1) and kurtosis (G2) are presented in Table 4. The 252

[288] combinations where both G1 and G2 equal 0 correspond to the normal case.

[289]   For the 4 resulting combinations of skewness and kurtosis (see Table 4), all other

[290] parameter values were chosen in order to illustrate the consequences of factors identified as

[291] playing a key role on the variance of unbiased estimators. We manipulated the population

[292] mean difference ($\mu_1 - \mu_2$), the sample sizes (n), the sample size ratio (n-ratio $= \frac{n_1}{n_2}$), the

[293] population *SD*-ratio (i.e. $\frac{\sigma_1}{\sigma_2}$), and the sample size and population variance pairing $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2} \right)$.

[294] In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in steps of 1 (so does $\mu_1 - \mu_2$)[6].

––––––

[6] In the original plan, we had added 252 simulations in which μ_1 and μ_2 were both null. We decided not to

present the results of these simulations in the main article, because the relative bias and the relative variance

appeared to us to be very useful to fully understand the comparison of the estimators, and computing them

is impossible when the real mean difference is zero. Indeed, for these specific configurations, both relative

bias and relative variance would have infinite values due to the presence of the population effect size term in

295 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

296 simulations for which both σ_1 and σ_2 equal 1 are the particular case of homoscedasticity

297 (i.e. equal population variances across groups). The sample sizes of both groups (n_1 and n_2)

298 were 20, 50 or 100. When sample sizes of both groups are equal, the n-ratio equals 1 (this is

299 known as a balanced design). All possible combinations of n-ratio and population SD-ratio

300 were performed in order to distinguish scenarios where both sample sizes and population

301 variances are unequal across groups (with positive pairing when the group with the largest

302 sample size is extracted from the population with the largest SD, and negative pairing when

303 the group with the smallest sample size is extracted from the population with the smallest

304 SD) and scenarios with no pairing between sample sizes and variances (sample sizes and/or

305 population SD are equal across all groups). In sum, the simulations grouped over different

306 sample sizes yield 4 conditions (a, b, c and d) based on the n-ratio, population SD-ratio, and

307 sample size and population variance pairing, as summarized in Table 5. We chose to divide

308 scenarios into these 4 conditions because analyses in Supplemental Material 1 revealed main

309 and interaction effects of sample sizes ratio and SD-ratio on the bias and variance of some

310 estimators.

---

their denominator. However, these extra simulations were included in the simulation checks, in Supplemental

Material 2.

Table 5

4 conditions based on the n-ratio and the SD-ratio.

		n-ratio		
		1	**>1**	**<1**
	1	a	b	b
SD-ratio	**>1**	c	d	d
	<1	c	d	d

Results. Before detailing the comparison of the estimators 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. 10^4 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 readers, illustrations of the raw bias and variance are available on Github.

2) For the sake of readability, for each condition, the scale we used for the ordinate axis of our plots when comparing the relative bias of all estimators and $G_1 = 6.32$ and $G_2 = 95.75$ is different from the scale we used for all other combinations G_1/G_2. In the same way of thinking, we also used different scales for the ordinate axis of our plots that compare the relative variance of all estimators, as a function of the condition.

3) Throughout this section, we will **compare** the relative bias and variance 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, for different sub-conditions, the averaged relative bias and relative variance of five estimators are presented. 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 S_1 as standardizer, the second group when using S_2 as standardizer). The other group will be called "experimental group".

When variances are equal across groups. Figures 2 and 3 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)[7]. However, the further from the normality assumption (i.e. when moving from left to right in Figures), the larger the bias.

We will observe that 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 performing as long as

---

[7] When looking at 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.

350 the sample size ratio is close to 1 (condition a; see Figure 2). However, when designs are

351 highly unbalanced (condition b; see Figure 3), Shieh's g_s is not consistent anymore, and

352 while Hedge's g_s^* remain consistent, Hedges's g_s is a better estimator. For interested readers,

353 these findings are detailed in the three paragraphs below.

354   Figure 2 illustrates scenarios where both population variances and sample sizes are

355 equal across groups (condition a). One can first notice that all estimators are consistent, as

356 their bias and variance decrease when the total sample size increases. For any departure

357 from the normality assumption, both bias and variance of Hedges' g_s, Shieh's g_s and Hedges'

358 g_s^* are similar[8] and smaller than the bias and variance of Glass's g_s estimates using either S_1

359 or S_2 as standardizer. Moreover, when samples are extracted from skewed distributions,

360 Glass's g_s will show different bias and variance as a function of the chosen standardizer (S_1

361 or S_2), even if both S_1 and S_2 are estimates of the same population variance, based on the

362 same sample size. This is due to non-null correlations of opposite sign between the mean

363 difference ($\bar{X}_1 - \bar{X}_2$) and respectively S_1 and S_2. In Supplemental Material 3, we detailed in

364 which situation a non nul correlation occurs between the sample mean difference ($\bar{X}_1 - \bar{X}_2$)

365 and the standardizer of compared estimators as well as the way this correlation impacts the

366 bias and variance of estimators.

367   Figure 3 illustrates scenarios where population variances are equal across groups and

368 sample sizes are unequal (condition b). For any departures from the normality assumptions,

369 Hedges' g_s shows the smallest bias and variance. Hedges' g_s and Hedges' g_s^* are consistent

370 estimators (i.e. the larger the sample sizes, the lower the bias and the variance), unlike

---

[8] While the bias and variance of Cohen's d_s, Cohen's d_s^* and Shieh's d_s are identical, the bias and variance of Hedges' g_s are marginally different from the bias and variance of Hedges' g_s^* and Shieh's g_s (these last two having identical bias and variance). Indeed,because of the sampling error, differences remain between sample variances, even when population variances are equal between groups. Since the Hedges' correction applied to Cohen's d_s does not contain the sample variances (unlike the one applied on both other estimators), the bias and variance of Hedges' g_s are slighly different from the bias and variance of Hedges' g_s^* and Shieh's g_s.

371 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

372 experimental group or on the total sample size. The only way to decrease the bias of Glass's

373 g_s is therefore to add subjects in the control group. On the other hand, the variance of

374 Glass's g_s depends on both sample sizes, but not in an equivalent way: in order to reduce the

375 variance, it is much more efficient to add subjects in the control group and when the size of

376 the experimental group decreases so does the variance, even when the total sample size is

377 increased. Regarding Shieh's g_s, for a given sample size ratio, the bias and variance will

378 decrease when sample sizes increase. However, there is a large effect of the sample sizes ratio

379 such that when the sample sizes ratio moves away from 1 by adding subjects, bias and

380 variance might increase.[9] On the other hand, when the sample sizes ratio moves closer to 1

381 by adding subjects, the bias will decrease.

382   When samples are extracted from skewed distributions and have unequal sizes (the two

383 last columns in Figure 3), for a constant total sample size, Glass's g_s, Shieh's g_s and Hedges'

384 g_s^* will show different bias and variance depending on which group is the largest one

385 (e.g. when distributions are right-skewed, the bias and variance of all these estimators when

386 n_1 and n_2 are respectively 50 and 20 are not the same as their bias and variance when n_1

387 and n_2 are respectively 20 and 50). This is due to a non-null correlation of opposite sign

388 between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and their respective standardizers depending on

389 which group is the largest one, as detailed in Supplemental Material 3. One observes that

390 under these configurations, the bias and variance of Glass's g_s are sometimes a bit smaller

---

[9] Regarding variance, in Supplemental Material 1, we mentioned that when the population effect size is nul,

the larger the total sample size, the lower the variance, whatever the sample sizes ratio is constant or not.

We also mentioned that this is no longer true when the population effect size is not zero and in our

simulations, the effect size is never zero. The effect size effect is partially visible in Figure 3 because we do

not entirely remove the effect size effect when we divide the variance by δ^2. This is due to the fact that one

term, in the equation of the variance computation, does not depend on the effect size.

$_{391}$ and sometimes much larger than the bias and variance of Shieh's g_s and Cohen's d_s^*. [10]

$_{392}$ *When variances are unequal across groups.*   Figures 4 to 9 represent configurations

$_{393}$ where the equality of variances assumption is not met. According to our expectations, one

$_{394}$ observes that the bias of all estimators is approximately zero as long as the normality

$_{395}$ assumption is met (first column in the three Figures), and the further from the normality

$_{396}$ assumption (i.e. when moving from left to right in Figures), the larger the bias[11]. It might

$_{397}$ be considered surprising that the bias of Hedges' g_s remains very small troughout these

$_{398}$ conditions. As discussed in the section "Different measures of effect size", Hedges' g_s should

$_{399}$ be avoided when population variances and sample sizes are unequal across groups, because of

$_{400}$ the pooled error term. When pooling the estimates of two unequal population variances, the

$_{401}$ resulting estimator will be lower (in case of positive pairing) or larger (in case of negative

$_{402}$ pairing) as it should be. At the same time, when pooling two unequal population variances,

$_{403}$ the population effect size will also be lower (in case of positive pairing) or larger (in case of

$_{404}$ negative pairing) as it should be. As a consequence, the distorsion cannot be seen through

$_{405}$ the difference between the expected estimator and the population effect size measure. For

$_{406}$ this reason, the bias and variance of Hedges' g_s will not be taken into account in the

$_{407}$ following comparisons.

$_{408}$ We will observe that when variances are unequal across populations, Glass's g_s is

---

[10] We learn from Supplemental Material 3 that when the $\mu_1 - \mu_2 > 0$ (like in our simulations), all other

parameters being equal, an estimator is always less biased and variable when choosing a standardizer that is

positively correlated with $\bar{X}_1 - \bar{X}_2$. We also learn from Supplemental Material 3 that the smaller n_c, the

larger the magnitude of correlation between S_c and $\bar{X}_1 - \bar{X}_2$. When $cor(S_c, \bar{X}_1 - \bar{X}_2)$ is positive, the

positive effect of increasing the magnitude of the correlation is counterbalanced by the negative effect of

reducing n_c. On the other hand, when $cor(S_c, \bar{X}_1 - \bar{X}_2)$ is negative, the negative effect of increasing

magnitude of the correlation is amplified by the negative effect of decreasing n_c. This explain why the

difference between Glass's g_s and other estimators is larger when Glass's g_s is the least efficient estimator.

[11] When looking at the relative bias for all estimators, the maximum departure from zero is 0.0173 when

sample sizes are equal across groups, and 0.0274 when both sample sizes and variances differ across groups.

$_{409}$ sometimes better but also sometimes much worst than respectively Shieh's g_s and Hedges'

$_{410}$ g_s^*, both in terms of bias and variance. The performance of Glass's g_s highly depends on

$_{411}$ parameters that we cannot control (i.e. a triple interaction between the n-ratio, the SD-ratio

$_{412}$ and the correlation between the standardizer and the mean difference) and for this reason,

$_{413}$ we do not recommend using it. When the sample sizes ratio is close to 1, Shieh's g_s and

$_{414}$ Hedges' g_s^* are both appropriate but the further the sample sizes ratio is from 1, the larger

$_{415}$ the bias of Shieh's g_s in order that in the end, our favourite measure is Hedges' g_s^*.

$_{416}$ Figures 4 and 5 are dedicated to scenarios where population variances are unequal

$_{417}$ between groups and sample sizes are equal (condition c). In Figure 4, scenarios are

$_{418}$ subdivided as a function of the sample sizes and one can notice that all estimators are

$_{419}$ consistent, as their bias and variance decrease when the total sample size increases. In

$_{420}$ Figure 5, scenarios are subdivided as a function of the SD-ratio. Because the comparison

$_{421}$ pattern remains very similar for all sample sizes, we present only scenarios when sample sizes

$_{422}$ equal 20. One should first notice that for all estimators in Figure 5, the relative variance

$_{423}$ seems to be much larger when $S_2 > S_1$. [12] This information should not be taken into

$_{424}$ account because it is only an artefact of our simulation conditions combined with the way we

$_{425}$ computed the relative variance. [13]

$_{426}$ When samples are extracted from skewed distributions, the bias and variance of Glass's

$_{427}$ g_s are sometimes smaller and sometimes larger than the bias of Shieh's g_s and Hedges' g_s.

$_{428}$ This is mainly due to the fact that when two samples of same sizes are extracted from two

$_{429}$ skewed distributions with unequal variances (the two last columns in Figure 5), there will be

---

[12] The difference between the variance of estimators when the second group is 10 times larger than the first group was so large that we decided to not present it, for the sake of readability of the Figures.

[13] We previously mentioned that when dividing the variance by δ^2, we do not entirely remove the effect size effect. Actually, we introduce δ^2 in the denominator of the first term, in the equation of the variance computation. Because we performed our simulations in order that σ_1 always equals 1, the smaller S_2, the larger the population effect size and therefore, the lower the relative variance.

430 non-null correlations of opposite sign between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and the

431 standardizer of *all* estimators, depending on which population variance is the largest one [14].

432   Figures 6 to 9 are dedicated to scenarios where both sample sizes and population

433 variances differ across groups. Due to a high number of combinations between the sample

434 sizes-ratio and the SD-ratio in our simulations, we decided to present only some conditions.

435 Because equations in Table 3 revealed an interaction effect between the sample sizes ratio

436 and the SD-ratio on the bias and variance of Hedges' g_s^* and Shieh's g_s (see Supplemental

437 Material 1), we chose to present all configurations where the larger SD is 10 times larger

438 than the smaller SD (Figures 6 and 7), and configurations where the larger SD is twice

439 larger than the smaller SD (Figures 8 and 9), in order to compare the effect of the sample

440 sizes ratio on the bias and variance of all estimators when the SD-ratio is large

441 $(\frac{\sigma_1}{\sigma_2} = 10 \ or \ .1 \ )$ or medium $(\frac{\sigma_1}{\sigma_2} = 2 \ or \ .5)$.

442   When distributions are symmetric, the bias of Glass's g_s only depends on the size of

443 the control group and is therefore not impacted by neither the sample sizes ratio nor the

444 total sample size. When comparing Figures 6 to 9, one can also notice that the bias of

445 Glass's g_s does not depend on the SD-ratio either. Unlike the bias of Glass's g_s, its variance

446 depends on both sample sizes, but not in an equivalent way: most of the time, it is more

447 efficient, in order to reduce the variance of Glass's g_s, to add subjects in the control group.

448 Regarding Hedges' g_s^* and Shieh's g_s, their respective biases and variances depend on an

449 interaction effect between the sample sizes ratio and the SD-ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$: the sample

450 sizes ratio associated with the smallest bias and variance is not the same when the more

451 variable group is 10 times more variable than the other group (Figures 6 and 7) than when it

---

[14] When population variances are unequal, a non-null correlation occurs between standardizer estimates and $\bar{X}_1 - \bar{X}_2$. For standardizers computed based on both S_1 and S_2, the sign of the correlation between the standardizer and the mean 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 Supplemental Material 3.

⁴⁵² is only twice more variable (Figures 8 and 9). However, the respective biases and variances

⁴⁵³ of Hedges' g_s^* and Shieh's g_s are always smaller when there is a positive pairing between

⁴⁵⁴ sample sizes and variances. When samples are extracted from skewed distributions, the bias

⁴⁵⁵ and variance of Glass's g_s are sometimes smaller and sometimes larger than the bias of

⁴⁵⁶ Shieh's g_s and Hedges' g_s^*, due to a combination of three factors: (1) which group is the

⁴⁵⁷ largest one, (2) which group has the smallest standard deviation and (3) what is the

⁴⁵⁸ correlation between the standardizer and the mean difference.

Recommendations

⁴⁶⁰ We recommend using Hedges' g_s^* in order to assess the magnitude of the effect when

⁴⁶¹ comparing two independent means, because a) it does not rely on the equality of population

⁴⁶² variances asumption (unlike Hedges' g_s), b) it is always consistent (unlike Shieh's g_s), c) it is

⁴⁶³ easily interpretable (Hedges' g_s^* can be interpreted the same way as Hedges' g_s) and d) it

⁴⁶⁴ remains constant for any sample sizes ratio, even when population variances are unequal

⁴⁶⁵ across groups, as shown in our previously introduced Shiny App.

⁴⁶⁶ Hedges' g_s^* should always be associated with confidence interval. In order to help

⁴⁶⁷ researchers, we created *deffectsize*, a R package (see

⁴⁶⁸ https://github.com/mdelacre/deffectsize). The *datacohen_CI* function was built in order to

⁴⁶⁹ compute point estimators and confidence intervals based on raw data and the *cohen_CI*

⁴⁷⁰ function was built in order to compute point estimators and confidence intervals based on

⁴⁷¹ descriptive statistics (sample means, sample variances and sample sizes). By default,

⁴⁷² unbiased Cohen's g_s^* is computed but it is also possible to compute biased estimators

⁴⁷³ (e.g. Cohen's d_s^*) and/or to use a pooled error term as standardizer by assuming that the

⁴⁷⁴ equality of population variances is met (e.g. Hedges' g_s or Cohen's d_s, depending on whether

⁴⁷⁵ we choose to compute unbiased or biased estimator). Other functions (*datashieh_CI*,

⁴⁷⁶ *shieh_CI*, *dataglass_CI* and *glass_CI*) are available in order to compute Shieh's g_s (or

⁴⁷⁷ Shieh's d_s) and Glass's g_s (or Glass's d_s) as well as their respective confidence intervals, even

if we don't recommend to use them, as previously mentioned. For those who are not familiar

with R-gui, we also created a Shiny app in order to compute point estimators et confidence

estimators based on descriptive statistics:

https://mdelacre.shinyapps.io/deffectsize-Shinyapp/ [15].

In addition with standardized mean difference, we recommend to always report all

necessary descriptive statistics (i.e. sample means, sample SD and sample sizes).

References

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.

https://doi.org/10.22237/jmasm/1146456060

Altman, G. D. (2005). Why we need confidence intervals. *World Journal of Surgery,

29*, 554–556. https://doi.org/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.

https://doi.org/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.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness

---

[15] I would like to thank Aaron Caldwell for his help when creating the Figures of this Shiny App.

and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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.

https://doi.org/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. https://doi.org/10.3102/00346543042003237

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their
confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology,
14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting
and Clinical Psychology, 68*(1), 155–165. https://doi.org/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.
https://doi.org/10.1037/1082-989X.6.2.135

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

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

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

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals
around the standardized mean difference: Bootstrap and parametric confidence intervals.

Educational and Psychological Measurement, *65*(1), 51–69.

https://doi.org/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. https://doi.org/10.1037/1082-989X.13.2.110

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. https://doi.org/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. https://doi.org/10.3389/fpsyg.2013.00863

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

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. https://doi.org/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. https://doi.org/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/

Shieh, G. (2013). Confidence intervals and sample size calculations for the standardized mean difference effect size between two normal populations under heteroscedasticity. *Behavior Research Methods*, *45*, 955–967. https://doi.org/10.3758/s13428-012-0228-7

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. https://doi.org/10.4300/JGME-D-12-00156.1

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. https://doi.org/10.1007/bf02295644
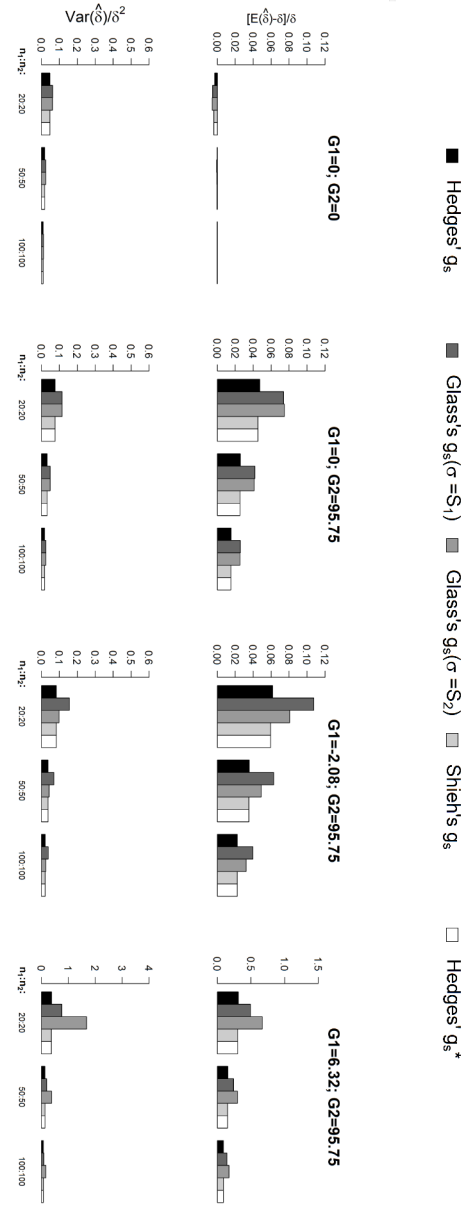
Figure 2. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)
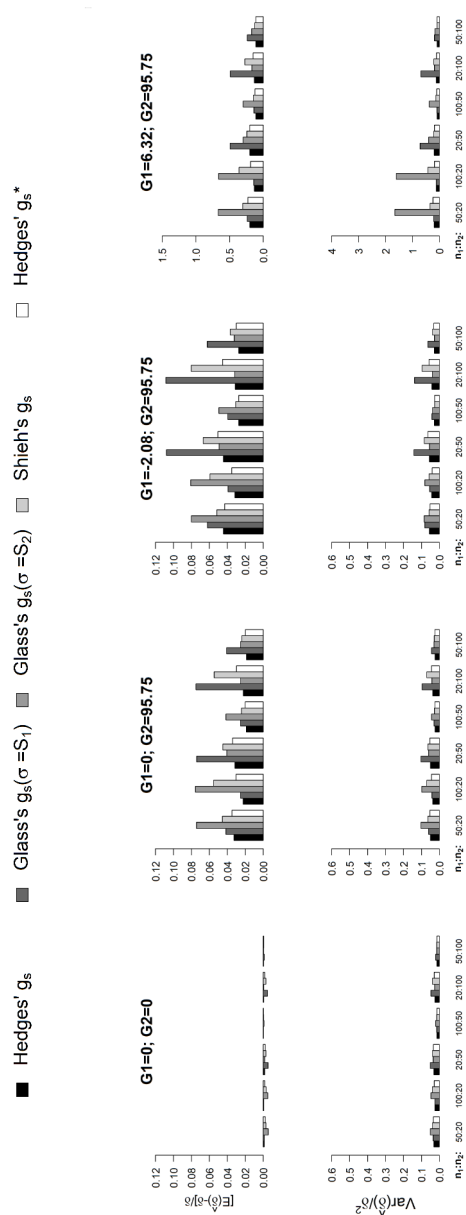
Figure 3. Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)
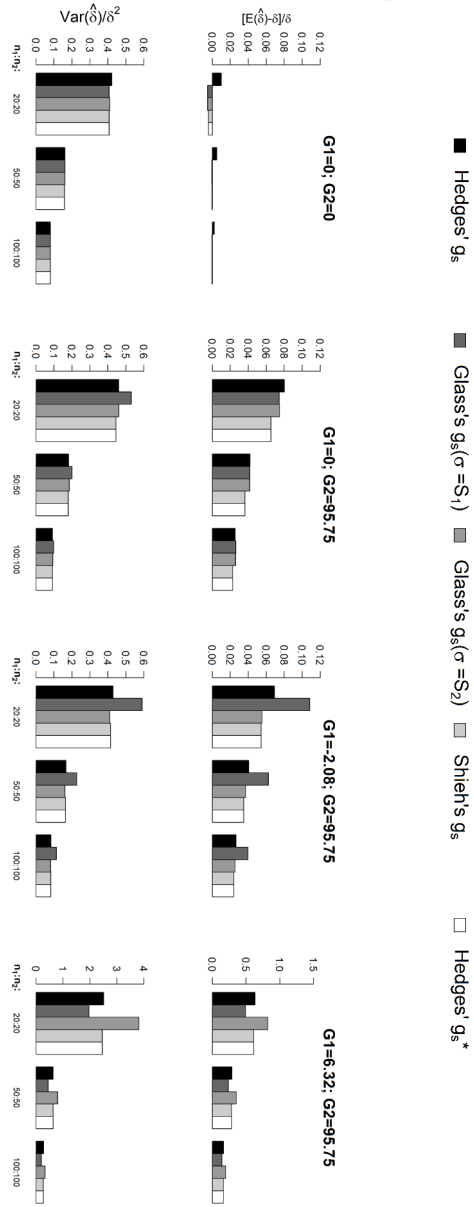
Figure 4. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c), as a function of n-ratio
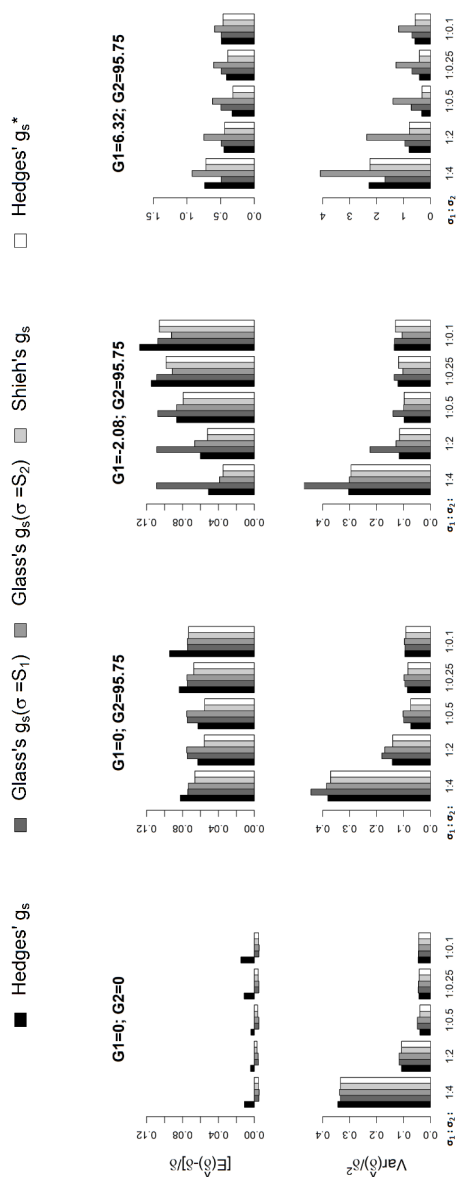
Figure 5. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the SD-ratio (when $n_1 = n_2 = 20$)
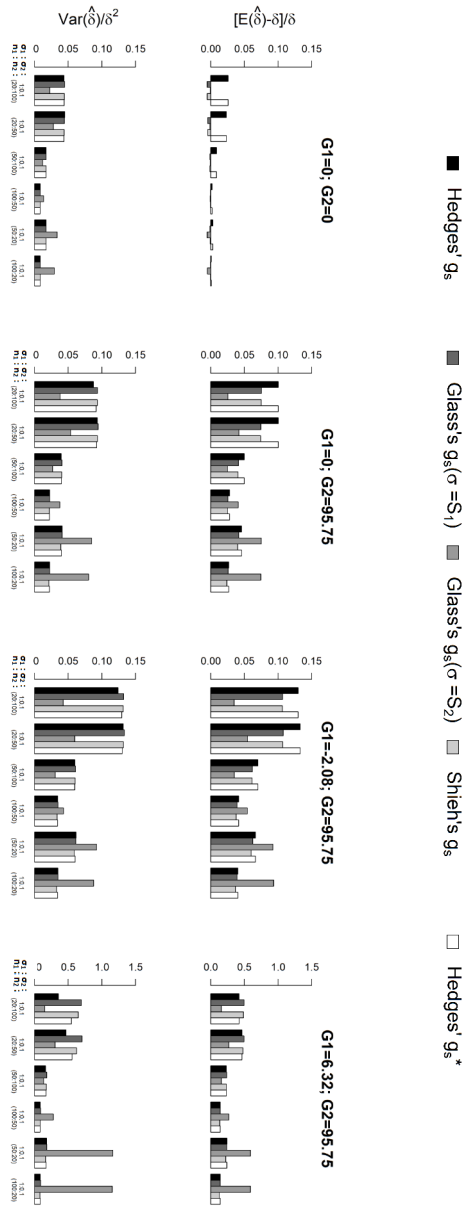
Figure 6. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 10$ and $\sigma_1 > \sigma_2$

Figure 7. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 10$ and $\sigma_1 < \sigma_2$
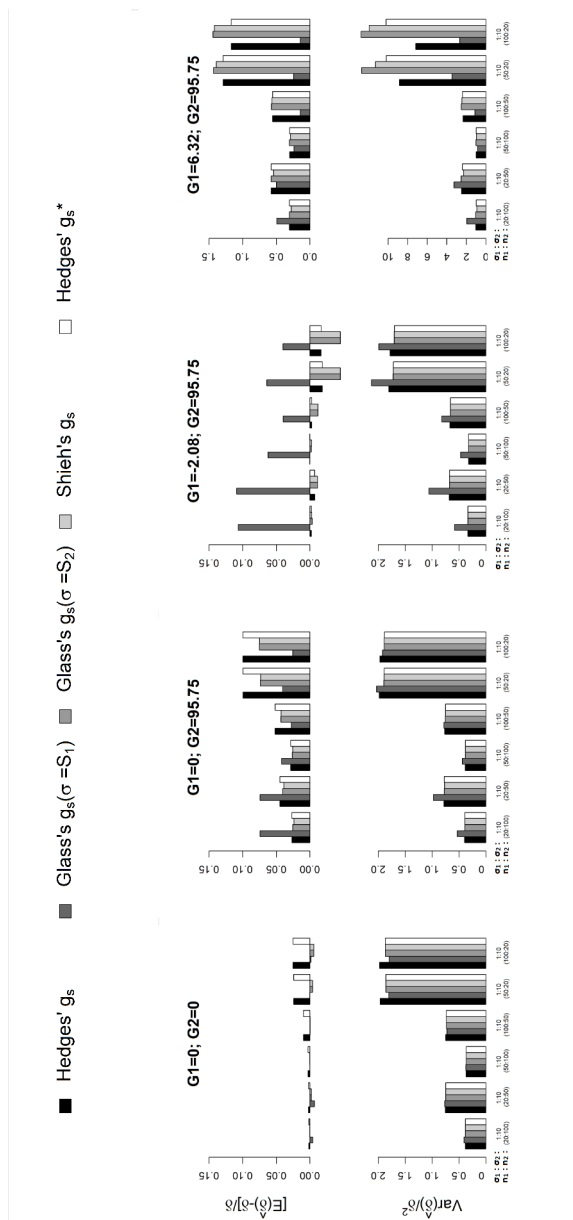
Figure 8. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 2$ and $\sigma_1 > \sigma_2$
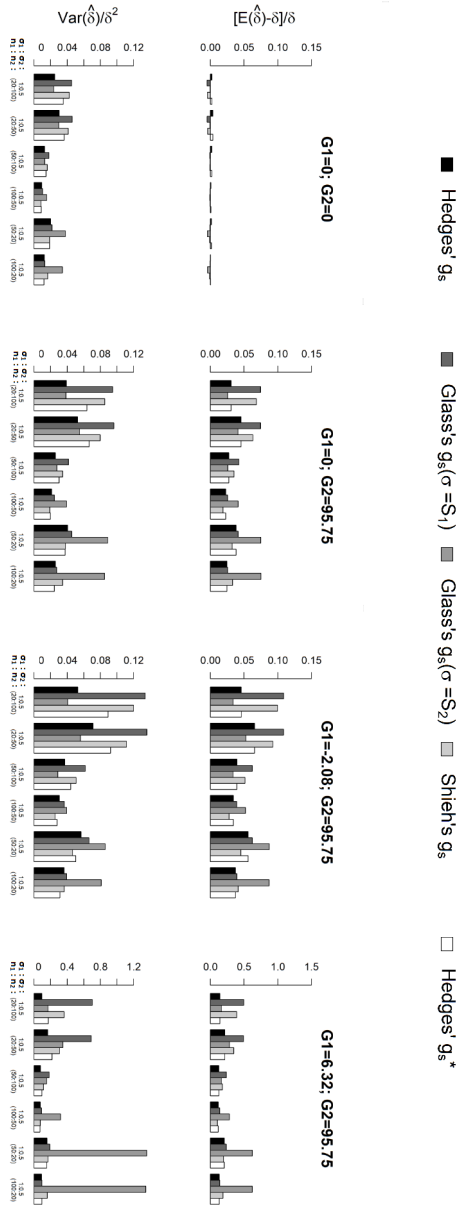
Figure 9. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), $\frac{\sigma_1}{\sigma_2} = 2$ and $\sigma_1 < \sigma_2$
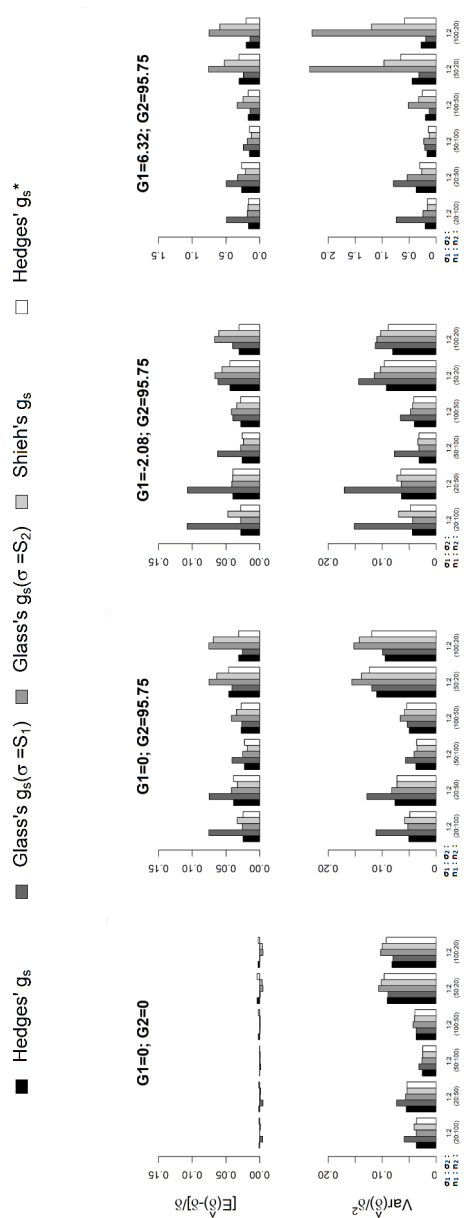
Appendix

587 **The bias of Cohen's d$_s$ is twice as large as the bias of Shieh's d$_s$ when**

588 **population variances and sample sizes are equal across groups: mathematical**

589 **demonstration.**

590   As mentioned in Table 1, the bias of Cohen's d_s is defined as

$$Bias_{Cohen's\ d_s} = \delta_{Cohen} \times \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} - 1 \right) \tag{16}$$

with

$$\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(n_1-1)\times\sigma_1^2+(n_2-1)\times\sigma_2^2}{n_1+n_2-2}}}$$

and

$$df_{Student} = n_1 + n_2 - 2$$

591   As mentioned in Table 2, the bias of Shieh's d_s is defined as

$$Bias_{Shieh's\ d_s} = \delta_{Shieh} \times \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} - 1 \right) \tag{17}$$

with

$$\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}} \quad (N = n_1 + n_2)$$

and

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

592   When $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$, δ_{Cohen} is twice larger than δ_{Shieh}, as shown below

593 in equations 18 and 19:

$$\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{2(n-1)\sigma^2}{2(n-1)}}} = \frac{\boldsymbol{\mu_1 - \mu_2}}{\boldsymbol{\sigma}} \tag{18}$$

594

$$\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{2\left(\frac{\sigma^2}{n/(2n)}\right)}} = \frac{\boldsymbol{\mu_1 - \mu_2}}{\boldsymbol{2\sigma}} \tag{19}$$

595 Moreover, degrees of freedom associated with Student's t-test and Welch's t-test are

596 identical, as shown below in equations 20 and 21:

$$df_{Student} = \mathbf{2(n-1)} \tag{20}$$

597

$$df_{Welch} = \frac{\left[2(\sigma^2/n)\right]^2}{\frac{2(\sigma^2/n)^2}{n-1}} = \mathbf{2(n-1)} \tag{21}$$

598 Equations 16 and 17 can therefore be redefined as follows:

$$Bias_{Cohen's\ d_s} = \frac{\mu_1 - \mu_2}{\sigma} \times \left(\frac{\sqrt{n-1} \times \Gamma\left(\frac{2n-3}{2}\right)}{\Gamma(n-1)} - 1\right) \tag{22}$$

599

$$Bias_{Shieh's\ d_s} = \frac{\mu_1 - \mu_2}{\boldsymbol{2\sigma}} \times \left(\frac{\sqrt{n-1} \times \Gamma\left(\frac{2n-3}{2}\right)}{\Gamma(n-1)} - 1\right) \tag{23}$$

600 We can therefore conclude that the bias of Cohen's d_s is twice larger than the bias of

601 Shieh's d_s.

602 **The variance of Cohen's d_s is four times larger than the bias of Shieh's d_s when**

603 **population variances and sample sizes are equal across groups: mathematical**

604 **demonstration.**

605   The variance of Cohen's d_s is defined in Table 1 as

$$Var_{Cohen's\ d_s} = \frac{N \times df_{Student}}{n_1 n_2 \times (df_{Student} - 2)} + \delta_{Cohen}^2 \left[\frac{df_{Student}}{df_{Student} - 2} - \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} \right)^2 \right] \tag{24}$$

606 and the variance of Shieh's d_s is defined in Table 2 as

$$Var_{Shieh's\ d_s} = \frac{df_{Welch}}{(df_{Welch} - 2)N} + \delta_{Shieh}^2 \left[\frac{df_{Welch}}{df_{Welch} - 2} - \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} \right)^2 \right] \tag{25}$$

607   We have previously shown in equations 20 and 21 that degrees of freedom associated

608 with Student's t-test and Welch's t-test equal $2(n-1)$, when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$.

609 As a consequence, the first term of the addition in equation 24 is 4 times larger than the first

610 term of the addition in equation 25:

$$\frac{N \times df_{Student}}{n_1 n_2 \times (df_{Student} - 2)} = \frac{2n \times 2(n-1)}{n^2 \times (2n-4)} = \frac{\mathbf{4(n-1)}}{\mathbf{n(2n-4)}}$$

611

$$\frac{df_{Welch}}{(df_{Welch} - 2)N} = \frac{2(n-1)}{2n(2n-4)} = \frac{\mathbf{n-1}}{\mathbf{n(2n-4)}}$$

612 We have also previously shown in equations 18 and 19 that δ_{Cohen} is twice larger than δ_{Shieh}

613 when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$ and, therefore, δ_{Cohen}^2 is four times larger than δ_{Shieh}^2. As

614 a consequence, the second term of the addition in equation 24 is also 4 times larger than the

615 second term of the addition in equation 25. Because both terms of the addition in equation

616 24 are four times larger than those in equation 25, we can conclude that the variance of

617 Cohen's d_s is four times larger than the variance of Shieh's d_s.