1 Argument to use the Hedges' g^* with the Welch's t-test : an investigation of Cohen's d and

2   related effect size estimators in terms of interpretability, bias, precision and robustness

3   Marie Delacre[1], Daniel Lakens[2], Christophe Ley[3], Limin Liu[4], & Christophe Leys[1]

4   [1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

5   [2] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven,

6   the Netherlands

7   [3] Université du Luxembourg, Department of Mathematics, Esch-sur-Alzette, Luxembourg

8   [4] Universiteit Gent, Department of Applied Mathematics, Computer Science and Statistics,

9   Gent, Belgium

10   Author Note

16   Correspondence concerning this article should be addressed to Marie Delacre, CP191,

17 avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

18 Argument to use the Hedges' g^* with the Welch's t-test : an investigation of Cohen's d and

19   related effect size estimators in terms of interpretability, bias, precision and robustness

20   Effect sizes are an important outcome of empirical research. Moving beyond decisions

21 about statistical significance, there is a strong call for researchers to report and interpret

22 effect sizes and associated confidence intervals. This practice is highly endorsed by the

23 American Psychological Association (APA) and the American Educational Research

24 Association (American Psychological Association, 2010; Duran et al., 2006).

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

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

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

28 pooled sample standard deviation (Peng et al., 2013; Shieh, 2013). This estimator is

29 available in many statistical software packages, such as SPSS and Stata. However,

30 computing the pooled sample standard deviation assumes that both sample variances are

31 estimates of a common population variance, which is known as the homogeneity of variance

32 assumption. It has been widely argued that there are many fields in psychology where this

33 assumption is ecologically unlikely (Delacre et al., 2017; Erceg-Hurn & Mirosevich, 2008;

34 Grissom, 2000). The question how to deal with the assumption of equal variances has been

35 widely explored in the context of hypothesis testing, and it is becoming increasingly

36 common to by default report a t-test that does not assume equal variances, such as Welch's

37 t-test.

38   However, the question which effect size to report when equal variances are not

39 assumed has received less attention. One possible reason is that researchers have not found

40 consensus on which of the available options should be used (Shieh, 2013). Even within the

41 very specific context of an estimate for the standardized sample mean difference there is

42 little agreement about which estimator is the best choice. In this article, we will review the

43 main candidates that have been proposed in the literature in the d family of effect sizes,

44 without (Cohen's d, Glass' d, Shieh's d and Cohen's d^*) and with correction for bias

45 (Hedges' g, Glass' g, Shieh's g and Hedges' g^*). We provide an R package and Shiny app

46 to compute relevant effect size measures and their confidence intervals.

47   Before reviewing the most important effect size measures in the d-family, we will first

48 list the different purposes effect size measures serve, and discuss the relationship between

49 effect sizes, statistical, and practical significance. Based on a detailed description of the

50 good properties an effect size measure should possess, we will evaluate these properties in

51 the Monte Carlo simulations we performed to compare the different effect size estimators

52 with correction for bias.

Three purposes of effect size estimators

54   The effect size is a measure of the magnitude of an effect. In the context of the

55 comparison of two groups based on their means, when the null hypothesis is the absence of

56 effect, d-family effect size estimators estimate the magnitude of the differences between

57 parameters of two populations groups are extracted from (e.g. the mean; Peng & Chen,

58 2014). Such a measure can be used for three different purposes.

59   First, effect size measures can be used for *interpretative* purposes. They allow

60 researchers to assess the practical significance of a result (i.e. statements about the

61 relevance of an effect in real life). In order to assess the meaningfulness of an effect, we

62 should be able to relate this effect size estimate with behaviors/meaningful consequences in

63 the real world (Andersen et al., 2007). This typically involves an analysis of the costs

64 (determined by a specific context) and the benefits (in part determined by the size of the

65 effect). It is important to remember an effect size is just a mathematical indicator of the

66 magnitude of a difference, which depends on the way a variable is converted into numerical

67 indicator. An effect size in itself is not a measure of the importance or the relevance of an

68 effect for real life (even if benchmarks for small, medium, or large effect sizes might have

contributed to such a misinterpretation; Stout & Ruble, 1995).

Second, effect size measures can be used for *comparative* purposes. They allow researchers to assess the stability of results across designs, analyses, and sample sizes. This includes statistically comparing and combining the results from two or more studies in a meta-analysis.

Third, effect size measures can be used for *inferential* purposes. Hypothesis tests and confidence intervals based on the same statistical quantity are directly related : 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. At the same time, the interval provides extra information about the precision of the sample estimate for inferential purposes (Altman, 2005; Ellis, 2010), and which effect sizes are excluded. The narrower the interval, the higher the precision, and the wider the confidence interval, the more the data lack precision. Effect size measures are also indirectly related to the hypothesis tests as effect sizes from previous studies can be used in an a-priori power analysis when planning a new study (Lakens, 2013; Prentice & Miller, 1992; Stout & Ruble, 1995; Sullivan & Feinn, 2012; Wilkinson, 1999).

Inferential properties of a good effect size estimator

The empirical value of an estimator (called the *estimate*) depends on the sample value. Different samples extracted from the same population will lead to different sample estimates of the population value. 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 useful, as it allows us to assess the qualities of an estimator. More specifically, three desirable properties a good estimator should possess for inferential purposes are : *unbiasedness*, *consistency* and *efficiency* (Wackerly et al., 2008).

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

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

96   if the distribution is centered around a value that is higher (or lower) than the true

97   population parameter (see Figure 1). In other words, examining the bias of an estimator

98   tells us if estimates are on average accurate. The *bias* of a point estimator $\hat{\delta}$ can be

99   computed as

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

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

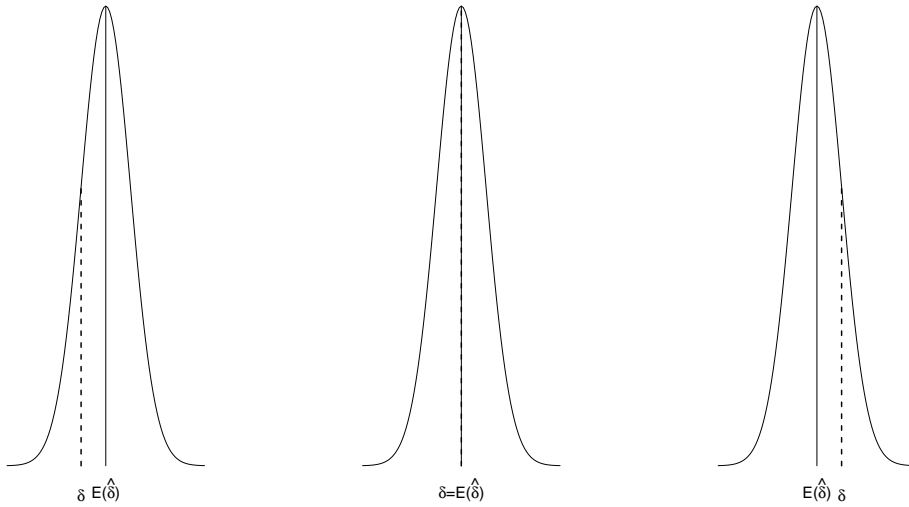101 true (population) parameter.

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

102   As we can see in Tables 1 and 2 the bias is directly related to the population effect

103 size. The larger the population effect size, the larger the bias. It is therefore also

104 interesting to examine the *relative bias*, defined as the ratio between the bias and the

105 population effect size :

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

106 While the bias informs us about the quality of estimates on average, in particular their

107 capacity of lying close to the true value, it says nothing about individual estimates.

108 Imagine a situation where the distribution of estimates is centered around the real

109 parameter but with such a large variance that some point estimates are very far from the

110 center. This would be problematic, since any single estimate might be very far from the

111 true population value. Therefore it is not only essential for an estimator to be unbiased,

112 but it is also desirable that the variability of its sampling distribution is small. Ideally,

113 sample estimates are close to the true population parameter. Among two unbiased

114 estimators $\hat{\delta}_1$ and $\hat{\delta}_2$, we therefore say that $\hat{\delta}_1$ is *more efficient* than $\hat{\delta}_2$ if

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

115 where $Var(\hat{\delta})$ is the variance of the sampling distribution of the estimator $\hat{\delta}$. Among all

116 unbiased estimators, the more efficient estimator will be the one with the smallest

117 variance.[1] The variance of an estimator $\hat{\delta}$ is a function of its size (the larger the estimator,

118 the larger the variance) and, therefore, we might be interested in evaluating the *relative*

119 *variance* as the ratio between the variance and the square of the population estimator :

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

120 Note that both unbiasedness and efficiency are very important when choosing an estimator.

121 In some situations, it might be better to have a slightly biased estimator with low variance,

122 (so that each estimate remains relatively close to the true parameter and one might be able

123 to apply bias correction techniques) rather than an unbiased estimator with a large

124 variance (Raviv, 2014).

125   Finally, the last property of a good point estimator is *consistency*. Consistency means

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

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

---

[1] The Cramér-Rao inequality provides a theoretical lower bound for the variance of unbiased estimators.
An estimator reaching this bound is therefore optimally efficient.

128 **Different measures of effect sizes**

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

130 conditions. The population effect size is defined as

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

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

132 $(j = 1, 2)$ and standard deviation σ. There exist different estimators of this effect size

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

134 sample means. When the equality of variances assumption is assumed, σ is estimated by

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

136 assumption cannot be assumed, alternatives to the pooled standard deviation are available.

137 In the next section we will present effect sizes that assume equal variances between groups

138 (Cohen's d and Hedges' g), and effect sizes that do not assume equal variances (Glass' d,

139 Shieh's d, Cohen's d^*, Glass' g, Shieh's g, and Hedges' g^*). For each effect size we will

140 provide information about their theoretical bias, variance, and consistency.

141 **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, where the sample mean difference is divided by

a pooled error term (Cohen, 1965) :

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

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

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

144 population variance (Keselman et al., 2008), which means a more accurate estimation of

145 the population variance can be achieved by pooling both estimates of this parameter

146   (i.e. S_1 and S_2). Since the larger the sample size, the more accurate the estimate, we give

147   more weight to the estimate based on the larger sample size. Cohen's d is directly related

148   to Student's t-statistic :

$$t_{Student} = \frac{Cohen's\ d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leftrightarrow Cohen's\ d = t_{Student} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6}$$

149   Under the assumption of normality and equal variances between groups, Student's

150   t-statistic follows a t-distribution with known degrees of freedom

$$df_{Student} = n_1 + n_2 - 2 \tag{7}$$

151   and noncentrality parameter[2]

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

152   where $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$, $\sigma_{pooled} = \sqrt{\frac{(n_1-1) \times \sigma_1^2 + (n_2-1) \times \sigma_2^2}{n_1 + n_2 - 2}}$ and σ_j is the standard deviation of the

153   j^{th} population ($j = 1, 2$). The relationship described in equation 6 and the theoretical

154   distribution of Student's t-statistic allow us to determine the sampling distribution of

155   Cohen's d, and therefore, its expectation and variance when the assumptions of normality

156   and equal variances are met. All these equations are provided in Table 1. For interested

157   readers, Supplemental Material 1 provides a detailed examination of the theoretical bias

158   and variance of Cohen's d based on Table 1, as well as the bias and variance of all other

159   estimators described later, based on Tables 2 and 3, with the goal to determine which

160   parameters influence the bias and variance of different estimators. The main results will be

161   discussed in the section "Monte Carlo Simulations : assessing the bias, efficiency and

162   consistency of 5 estimators" below.

    While Cohen's d is a consistent estimator, its bias and variance are substantial with

small sample sizes, even under the assumptions of normality and equal variances (Lakens,

---

[2] 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.

2013). In order to compensate for Cohen's d bias with small sample sizes, Hedges and
Olkin (1985) defined a bias-corrected version :

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

where $df_{Student}$ has been defined in equation 7 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 :

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

163 where N is the total sample size. Hedges' g is theoretically unbiased when the assumptions

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

165 both Cohen's d and Hedges' g depend on the same parameters (i.e. the total sample size

166 (N) and the sample sizes ratio $\left(\frac{n_2}{n_1}\right)$), Hedges' g is less variable, especially with small

167 sample sizes.[3]

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

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

170 this assumption (Cumming, 2013; Grissom & Kim, 2001; Grissom & Kim, 2005; Kelley,

171 2005; Shieh, 2013). When variances are unequal between groups, the expression in

172 equation 7 is no longer valid because both groups do not share a common population

173 variance. If we pool the estimates of two unequal population variances, the estimator of

174 effect size will be smaller as it should be in case of positive pairing (i.e. the group with the

175 larger sample size is extracted from the population with the larger variance) and larger as

176 it should be in case of negative pairing (i.e. the group with the larger sample size is

---

[3] In Table 1, one can see that the variance of Hedges' g equals the variance of Cohen's d, multiplied by
$\left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}\right]^2$. This term is always less than 1 and tends to 1 when the sample sizes tend to infinity
$(.52 \leq \left[\frac{\Gamma(\frac{df}{2})}{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}\right]^2 < 1$ for $3 \leq df < \infty)$. As a consequence, the larger the total sample size, the
smaller the difference between the variance of Cohen's d and Hedges' g.

177 extracted from the population with the smaller variance). Because the assumption of equal

178 variances across populations is rarely realistic in practice (Cain et al., 2017; Delacre et al.,

179 2017; Delacre et al., 2019; Erceg-Hurn & Mirosevich, 2008; Glass et al., 1972; Grissom,

180 2000; Micceri, 1989; Yuan et al., 2004), both Cohen's d and Hedges' g should be abandoned

181 in favor of an alternative robust to unequal population variances.

Table 1

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

	df	Expectation	Variance
Cohen's d	$N-2$	$\delta_{Cohen} \times c_f$	$\frac{N \times df}{n_1 n_2 \times (df-2)} + \delta_{Cohen}^2 \left[\frac{df}{df-2} - c_f^2 \right]$
		$\approx \frac{\delta_{Cohen}}{\left(1 - \frac{3}{4N-9}\right)}$	$\approx \frac{N \times df}{n_1 n_2 \times (df-2)} + \delta_{Cohen}^2 \left[\frac{df}{df-2} - \left(\frac{1}{1 - \frac{3}{4N-9}} \right)^2 \right]$
Hedges' g	$N-2$	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. $\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ and $c_f = \frac{\sqrt{\frac{df}{2}} \times \Gamma(\frac{df-1}{2})}{\Gamma(\frac{df}{2})}$; Cohen's d is a biased estimator, because its expectation differs from the population effect size. Moreover, the larger the population estimator (δ_{Cohen}), the larger the bias. Indeed, the bias is the difference between the expectation and δ_{Cohen}: $\delta_{bias} = \delta_{Cohen} \times (c_f - 1)$. On the other hand, Hedges' g is an unbiased estimator, because its expectation equals δ_{Cohen}; equations in this table 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 : (A) the Glass' d, (B) the Shieh's d and (C) the Cohen's d^*.

Glass' d. When comparing one control group with one experimental group, Glass et al. (1981) recommend using the standard deviation of the control group as standardizer. This yields

$$Glass' \; d = \frac{\bar{X}_e - \bar{X}_c}{S_c}$$

where $\bar{X}_e$ and $\bar{X}_c$ are 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 (Cumming, 2013). 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). As we shall see later through our simulations, this estimator also does not present good inferential properties. The distribution of Glass' d is defined as in Algina et al. (2006) :

$$Glass' \; d \sim \sqrt{\frac{1}{n_c} + \frac{\sigma_e^2}{n_e \times \sigma_c^2}} \times t_{df,ncp} \tag{8}$$

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

$$df = n_c - 1 \tag{9}$$

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

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 8, we can compute its

204 theoretical expectation and variance when the assumption of normality is met (see Table

205 2), and therefore determine which factors influence bias and variance, and how they do so

206 (see Supplemental Material 1).

Table 2

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

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

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 in this table require $df \geq 3$ and at least 2 subjects per group.

Shieh's *d*. Kulinskaya and Staudte (2007) were the first to recommend 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, proposed a modification of the exact *SD* of the sample mean difference :

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

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

$$Shieh's\ d = \frac{t_{Welch}}{\sqrt{N}} \leftrightarrow t_{welch} = Shieh's\ d \times \sqrt{N}. \tag{10}$$

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

$$df_{Welch} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2/n_2)^2}{n_2 - 1}} \tag{11}$$

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

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 10 and the theoretical distribution of Welch's *t*-statistic allow us to approximate the sampling distribution of Shieh's *d*. Based on the sampling distribution of Shieh's *d*, we can estimate its theoretical expectation and variance under the assumption of normality (see Table 2), and thereby determine which factors influence bias and variance, and how they do so (see Supplemental Material 1).

As demonstrated in Appendices 1 and 2, when variances and sample sizes are equal across groups, the biases and variances of Cohen's *d* and Shieh's *d* are identical except for multiplication by a constant. The same is true for the estimators δ_{Cohen} and δ_{Shieh} :

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

224

$$Bias_{Cohen's\ d} = 2 \times Bias_{Shieh's\ d} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \quad (13)$$

225

$$Var_{Cohen's\ d} = 4 \times Var_{Shieh's\ d} \quad \text{(considering } \sigma_1 = \sigma_2 \text{ and } n_1 = n_2) \quad (14)$$

226 We can deduce from equations 12, 13 and 14 that relative to their respective population

227 effect size, Cohen's d and Shieh's d are equally accurate. In other words, their relative bias

228 and variance are identical. This is a good illustration of our motivation to favor relative

229 bias and variance (previously defined in equations 2 and 4) over the most commonly used

230 raw bias and variance (previously defined in equations 1 and 3).

231   When sample sizes are not equal, according to the statistical properties of Welch's

232 statistic under heteroscedasticity, Shieh's d accounts for the allocation ratio of sample sizes

233 to each condition. The lack of generality caused by taking this specificity of the design into

234 account has led Cumming (2013) to question its usefulness in terms of interpretability :

235 when the mean difference $(\bar{X}_1 - \bar{X}_2)$, S_1, and S_2 remain constant, Shieh's d will vary as a

236 function of the sample sizes allocation ratio (unlike Cohen's d^* that we will define below).

237 At the population level, δ_{Shieh} also depends on the sample sizes allocation ratio, as

238 illustrated in the following shiny application :

239 https://effectsize.shinyapps.io/ShiehvsCohen/.

Cohen's d^*.   An effect size estimator based on the sample mean difference divided

by the square root of the non pooled average of both variance estimates was suggested by

Welch (1938). Here, we indicate the difference between Cohen's d (based on the pooled

standard deviations) and Cohen's d^* with an asterisk. This yields :

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

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

241 know the distribution of Cohen's d^* (Huynh, 1989) :

$$Cohen's\ d^* \sim \sqrt{\frac{2(n_2 \times \sigma_1^2 + n_1 \times \sigma_2^2)}{n_1 n_2 (\sigma_1^2 + \sigma_2^2)}} \times t_{df^*, ncp^*} \quad (15)$$

242 where df^* and ncp^* are defined as follows :

$$df^* = \frac{(n_1 - 1)(n_2 - 1)(\sigma_1^2 + \sigma_2^2)^2}{(n_2 - 1)\sigma_1^4 + (n_1 - 1)\sigma_2^4} \tag{16}$$

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

243 where $\delta_{Cohen}^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$. Using equation 15 we can compute its theoretical expectation and

244 variance when the assumption of normality is met (see Table 2), and therefore determine

245 which factors influence bias and variance, and how they do so (see Supplemental Material

246 1). This estimator has been widely criticized, because it results in a variance term of an

247 artificial population (i.e. since the variance term does not estimate the variance of one or

248 the other group, the composite variance is an estimation of the variance of an artificial

249 population; Grissom & Kim, 2001), and unless both sample sizes are equal, the variance

250 term does not correspond to the variance of the mean difference (Shieh, 2013). However,

251 we will show throughout the simulation section that this estimator exhibits very good

252 inferential properties. Moreover, it has a constant value across sample sizes ratios, as

253 shown in the Shiny App at https://effectsize.shinyapps.io/ShiehvsCohen/.

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

$$g = d \times \frac{\Gamma(\frac{\nu}{2})}{\sqrt{\frac{\nu}{2}} \times \Gamma(\frac{\nu - 1}{2})}$$

254 where the distinct values of ν are provided in equation 9 for Glass' g, in equation 11 for

255 Shieh's g and in equation 16 for Hedges' g^*. The three corrected estimators are

256 theoretically unbiased when the assumption of normality is met. Their variance is a

257 function of the same parameters as their biased equivalent. However, due to the correction

258 they have a smaller variance, especially with small sample sizes, as shown in Table 3. In

259 summary :

260   • The variances of Hedges' g^* and Shieh's g depend on the total sample size (N), their

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

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

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

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

265   How these parameters influence the variance of the estimators will be summarized

266   and illustrated in Monte Carlo simulations below.

Table 3

Expectation, bias and variance of Glass' g, Hedges' g*, and Shieh's g under the assumption that independent residuals are normally distributed.

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

Note. $c_f = \dfrac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \dfrac{\mu_c - \mu_e}{\sigma_c}$, $\delta_{Shieh} = \dfrac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N}+\frac{\sigma_2^2}{n_2/N}}}$ 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 in this table 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 Hedges' g, Glass' g (using respectively the sample SD of the first or second group as a standardizer), Hedges' g^* and Shieh's g.

A set of 100,000 datasets was generated for 1,008 scenarios as a function of different criteria. 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 G_2 = -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. G_2=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 $G_1 = -2.08$ and $G_1 = 6.32$. We also simulated samples extracted from population where $G_1 = 0$, in order to assess the

293 main effect of high kurtosis on the quality of estimators. All possible combinations of

294 population skewness and kurtosis and the number of scenarios for each combination are

295 summarized in Table 4. To provide the reader with an intuition about the four

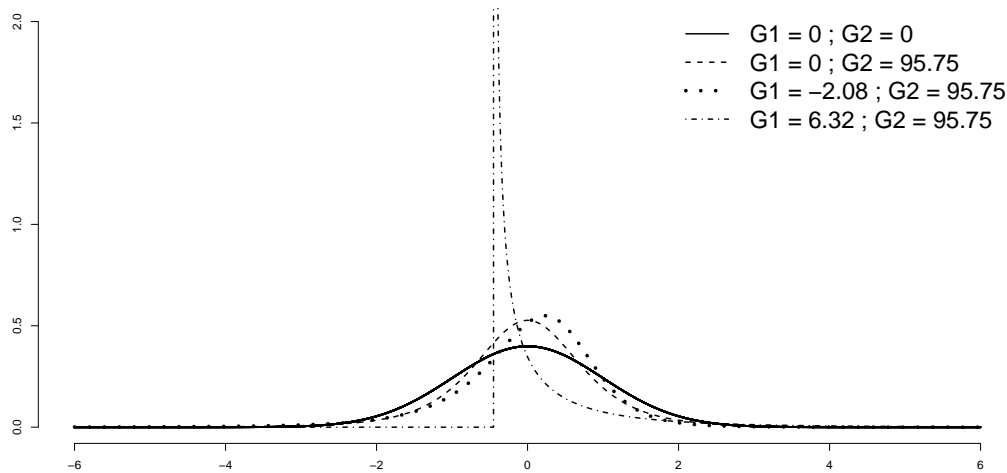296 distributions we consider, we have plotted them in Figure 2.

Figure 2. Plots of the four distributions considered in our simulations, with $\mu = 0$ and $\sigma = 1$. The undotted curve is the normal distribution.

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**

	Kurtosis		
TOTAL	**252**	**756**	**1008**

Note. Fisher's skewness (G1) and kurtosis (G2) are presented in Table 4. 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 4), 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_1 and n_2), the sample size ratio (n-ratio $= \frac{n_2}{n_1}$), the population SD-ratio (i.e. $\frac{\sigma_2}{\sigma_1}$), and the sample size and population variance pairing $\left(\frac{n_2}{n_1} \times \frac{\sigma_2}{\sigma_1}\right)$. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in steps of 1 (so does $\mu_1 - \mu_2$). Thus, we only consider a positive population effect size, which will result in positive bias of our estimators for reasons explained in Supplemental Material 1. [4] Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10, and therefore, the SD-ratio were 10, 4, 2, 1, .5, .25 or .1. The simulations for which both σ_1 and σ_2 equal 1 are the particular case of homoscedasticity, or equal population variances across groups. The sample sizes 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 (this is known as a balanced design). All possible combinations of n-ratio and population SD-ratio were simulated in order to distinguish

───────

[4] 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 their denominator. However, these extra simulations were included in the simulation checks, in Supplemental Material 2.

scenarios where both sample sizes and population variances are unequal across groups
(with positive pairing when the group with the largest sample size is extracted from the
population with the largest SD, and negative pairing when the group with the smallest
sample size is extracted from the population with the smallest SD) and scenarios with no
pairing between sample sizes and variances (sample sizes and/or population SD are equal
across all groups). In sum, the simulations grouped over different sample sizes yield 4
conditions (a, b, c and d) based on the n-ratio, population SD-ratio, and sample size and
population variance pairing, as summarized in Table 5. We chose to divide scenarios into
these 4 conditions because analyses in Supplemental Material 1 revealed main and
interaction effects of sample sizes ratio and SD-ratio on the bias and variance of some
estimators.

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 presenting the comparison of the estimators for each condition, it is useful to make some general comments.

1) We previously discussed 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. The same issue due to a change in scales occurs 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

339   of the raw bias and variance are available on Github at

340   https://github.com/mdelacre/Effect-sizes/.

341 2) For the sake of readability, the vertical axis differs across plots.

342 3) Throughout this section, we will *compare* the relative bias and variance of different

343   estimators, but we do not present bias and variance in absolute terms. We chose very

344   extreme (although realistic) conditions, and we know that none of the parametric

345   measures of effect size will be robust against such extreme conditions. Our goal is

346   therefore to study the robustness of the estimators against normality violations only

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

348   After these general remarks, we will analyze each condition separately. In all Figures

349 presented below, for different sub-conditions, the averaged relative bias and relative

350 variance of five estimators are presented. When describing the Glass' g estimators, we will

351 systematically refer to the "control group" as the condition the standardizer is based on

352 (i.e. the first group when using S_1 as standardizer, the second group when using S_2 as

353 standardizer). The other condition will be referred to as the "experimental group."

354   *When variances are equal across groups*

355   Figures 3 and 4 represent configurations where the equality of variances assumption is

356 met. According to our expectations, one observes that the bias of all estimators is

357 approximately zero as long as the normality assumption is met (first column in both

358 Figures).[5] However, the more the data generation process deviations from the normality

359 assumption (i.e. when moving from left to right in the Figures), the larger the bias in the

360 estimators.

361   We will observe that Glass' g should always be avoided when the equality of variance

---

[5] 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.

362 assumption is met. Hedges' g, Hedges' g^* and Shieh's g perform equally well as long as the

363 sample size ratio is close to 1 (condition a; see Figure 3). However, when designs are highly

364 unbalanced (condition b; see Figure 4), Shieh's g is not consistent anymore and while

365 Hedges' g^* remains consistent, Hedges's g is a better estimator. For interested readers,

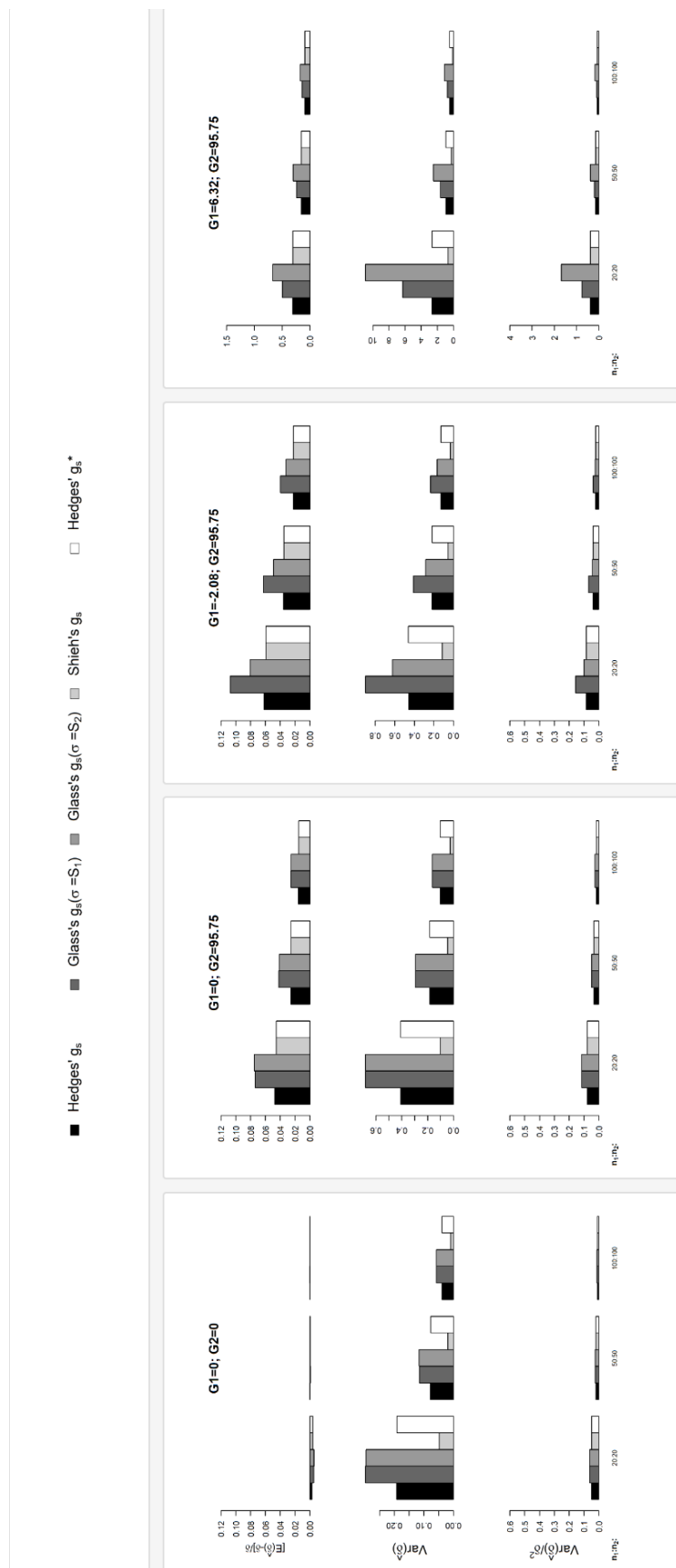366 these findings are detailed in the three paragraphs below.

Figure 3. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)
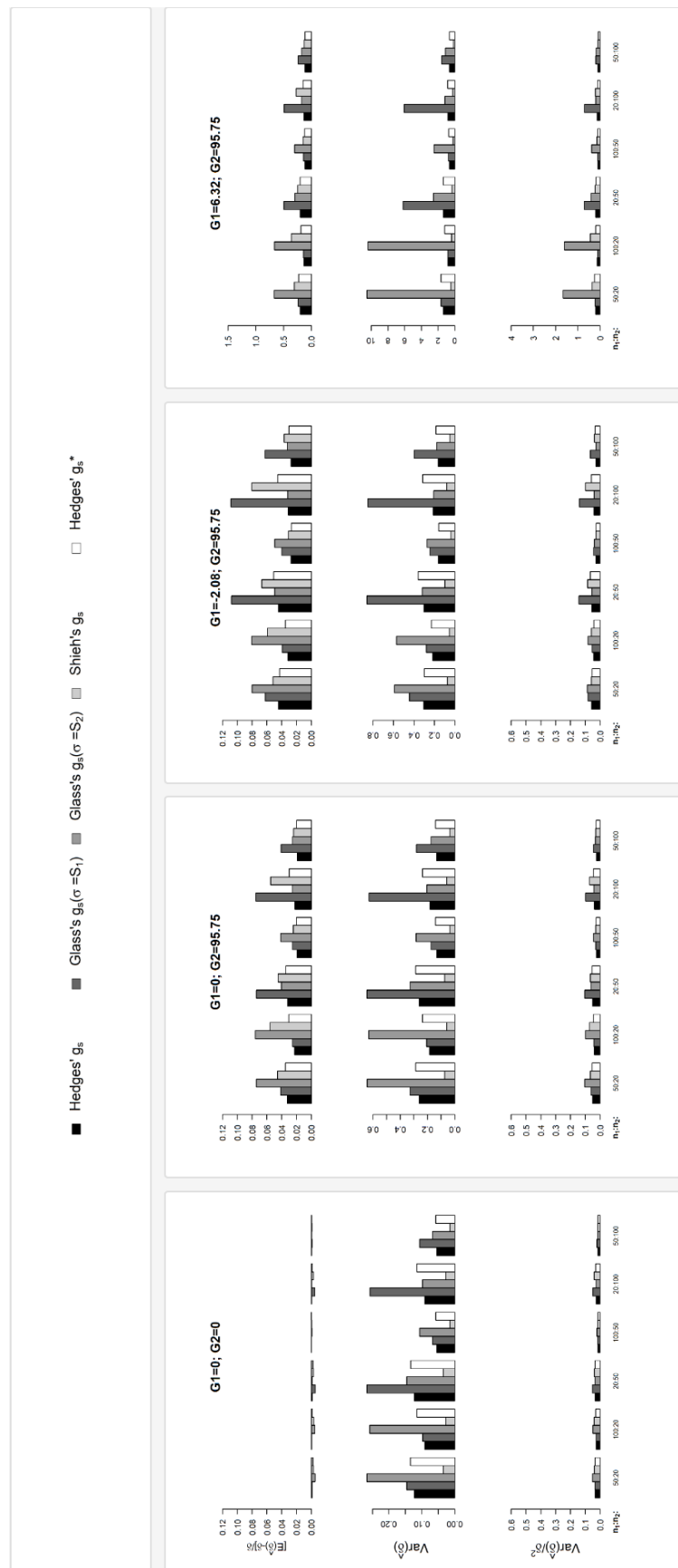
Figure 4. Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)

367   Figure 3 illustrates scenarios where both population variances and sample sizes are

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

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

370 from the normality assumption, both bias and variance of Hedges' g, Shieh's g and Hedges'

371 g^* are similar[6] and smaller than the bias and variance of Glass' g estimates using either S_1

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

373 Glass' g will show different bias and variance as a function of the chosen standardizer (S_1

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

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

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

377 in which situation a non-null correlation occurs between the sample mean difference

378 ($\bar{X}_1 - \bar{X}_2$) and the standardizer of compared estimators as well as the way this correlation

379 impacts the bias and variance of estimators.

380   Figure 4 illustrates scenarios where population variances are equal across groups, but

381 sample sizes are unequal (condition b). For any departures from the normality

382 assumptions, Hedges' g shows the smallest bias and variance. Hedges' g and Hedges' g^* are

383 consistent estimators (i.e. the larger the sample sizes, the lower the bias and the variance),

384 unlike Shieh's g and Glass' g. The bias of Glass' g does not depend either on the size of the

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

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

387 Glass' g depends on both sample sizes, but not in an equivalent way : in order to reduce

———

[6] While the bias and variance of Cohen's d, Cohen's d^* and Shieh's d are identical, the bias and variance of
Hedges' g are marginally different from the bias and variance of Hedges' g^* and Shieh's g (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 does not contain the sample variances (unlike the correction applied on both other
estimators), the bias and variance of Hedges' g are slighly different from the bias and variance of Hedges'
g^* and Shieh's g.

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

389 relative size of the experimental group decreases so does the variance, even when the total

390 sample size is decreased. Regarding Shieh's g, for a given sample size ratio, the bias and

391 variance will decrease when sample sizes increase. However, there is a large effect of the

392 sample sizes ratio such that when the sample sizes ratio moves away from 1 by adding

393 subjects, bias and variance might increase.[7] On the other hand, when the sample sizes

394 ratio moves closer to 1 by adding subjects, the bias will decrease.

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

396 two last columns in Figure 4), for a constant total sample size, Glass' g, Shieh's g and

397 Hedges' g^* will show different bias and variance depending on which group is the largest

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

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

400 n_1 and n_2 are respectively 20 and 50). This is due to a non-null correlation of opposite sign

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

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

403 under these configurations, the bias and variance of Glass' g are sometimes a bit smaller

404 and sometimes much larger than the bias and variance of Shieh's g and Cohen's d^*.[8]

----

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

zero, the larger the total sample size, the lower the variance, whether 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. In our

simulations the effect size is never zero. The effect size effect is partially visible in Figure 4 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.

[8] Supplemental Material 3 shows that when $\mu_1 - \mu_2 > 0$ (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 $\bar{X}_1 - \bar{X}_2$. Supplemental Material 3 also shows 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

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 all Figures),

and the further from the normality assumption (i.e. when moving from left to right in

Figures), the larger the bias.[9] It might be considered surprising that the bias of Hedges' g

remains very small throughout these conditions. As discussed in the section "Different

measures of effect size," Hedges' g 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 smaller

(in case of positive pairing) or larger (in case of negative pairing) than it should be. At the

same time, when pooling two unequal population variances, the population effect size will

also be smaller (in case of positive pairing) or larger (in case of negative pairing) as it

should be. As a consequence, the distortion 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 will not be taken into account in the following comparisons.

We will observe that when variances are unequal across populations, Glass' g

sometimes performs better, but also sometimes performs much worse than Shieh's g and

Hedges' g^*, both in terms of bias and variance. The performance of Glass' g highly

depends on parameters that we cannot control (i.e. a triple interaction between the n-ratio,

the SD-ratio and the correlation between the standardizer and the mean difference) and

for this reason, we do not recommend using it. When the sample sizes ratio is close to 1,

––––––

hand, when $cor(S_c, \bar{X}_1 - \bar{X}_2)$ is negative, the negative effect of increasing the magnitude of the correlation
is amplified by the negative effect of decreasing n_c. This explains why the difference between Glass' g and
other estimators is larger when Glass' g is the least efficient estimator.

[9] 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.

427 Shieh's g and Hedges' g^* are both appropriate but the further the sample sizes ratio is from

428 1, the larger the bias of Shieh's g in order that in the end, the measure that we believe

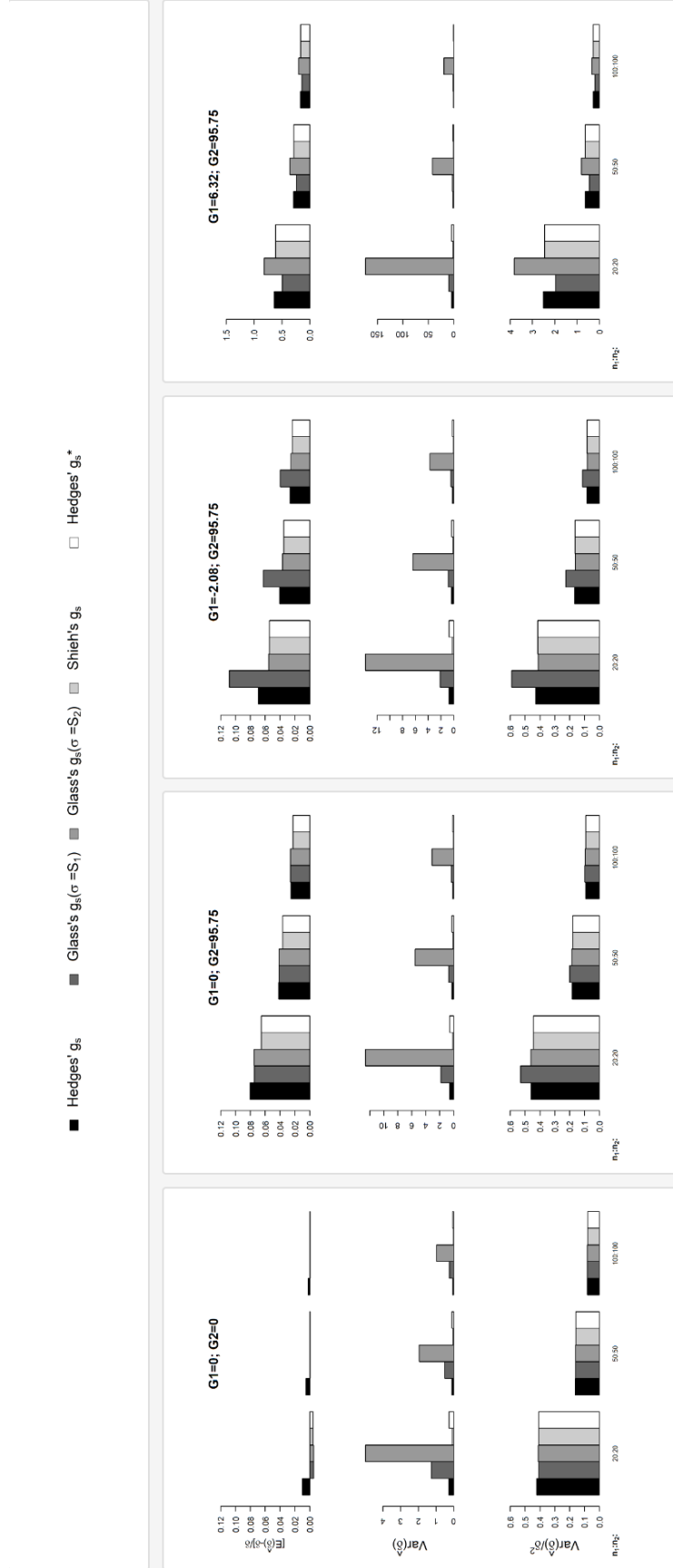429 performs best across scenarios is Hedges' g^*.

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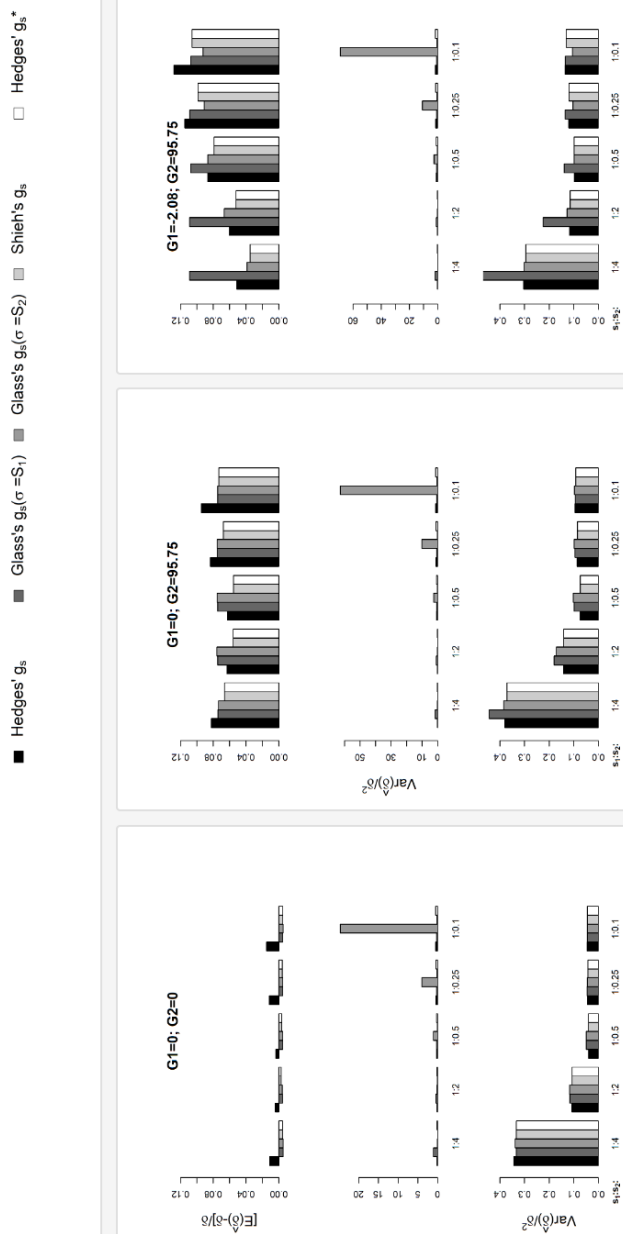
Figure 6. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the SD-ratio (when $n_1 = n_2 = 20$)

430   Figures 5 and 6 are dedicated to scenarios where population variances are unequal

431 between groups and sample sizes are equal (condition c). In Figure 5, scenarios are

432 subdivided as a function of the sample sizes and one can notice that all estimators are

433 consistent, as their bias and variance decrease when the total sample size increases. In

434 Figure 6, scenarios are subdivided as a function of the SD-ratio. Because the comparison

435 pattern remains very similar for all sample sizes, we present only scenarios when sample

436 sizes equal 20. One should first notice that for all estimators in Figure 6, the relative

437 variance seems to be much larger when $S_2 > S_1$.[10] This information should not be taken

438 into account because it is only an artefact of our simulation conditions combined with the

439 way we computed the relative variance.[11]

440   When samples are extracted from skewed distributions, the bias and variance of

441 Glass' g are sometimes smaller and sometimes larger than the bias of Shieh's g and Hedges'

442 g^*. This is mainly due to the fact that when two samples of same sizes are extracted from

443 two skewed distributions with unequal variances (the two last columns in Figure 6), there

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

445 the standardizer of all estimators, depending on which population variance is larger.[12]

446   Figures 7 to 10 are dedicated to scenarios where both sample sizes and population

———

[10] 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.

[11] 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 smaller the relative variance.

[12] 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.

⁴⁴⁷ variances differ across groups. Due to a high number of combinations between the sample

⁴⁴⁸ sizes ratio and the SD-ratio in our simulations, we decided to present only some

⁴⁴⁹ conditions. Because equations in Table 3 revealed an interaction effect between the sample

⁴⁵⁰ sizes ratio and the SD-ratio on the bias and variance of Hedges' g^* and Shieh's g (see

⁴⁵¹ Supplemental Material 1), we chose to present all configurations where the larger SD is 10

⁴⁵² times larger than the smaller SD (Figures 7 and 8), and configurations where the larger

⁴⁵³ SD is twice larger than the smaller SD (Figures 9 and 10), in order to compare the effect

⁴⁵⁴ of the sample sizes ratio on the bias and variance of all estimators when the SD-ratio is

⁴⁵⁵ large ($\frac{\sigma_2}{\sigma_1} = 10$ or .1) or medium ($\frac{\sigma_2}{\sigma_1} = 2$ or .5).
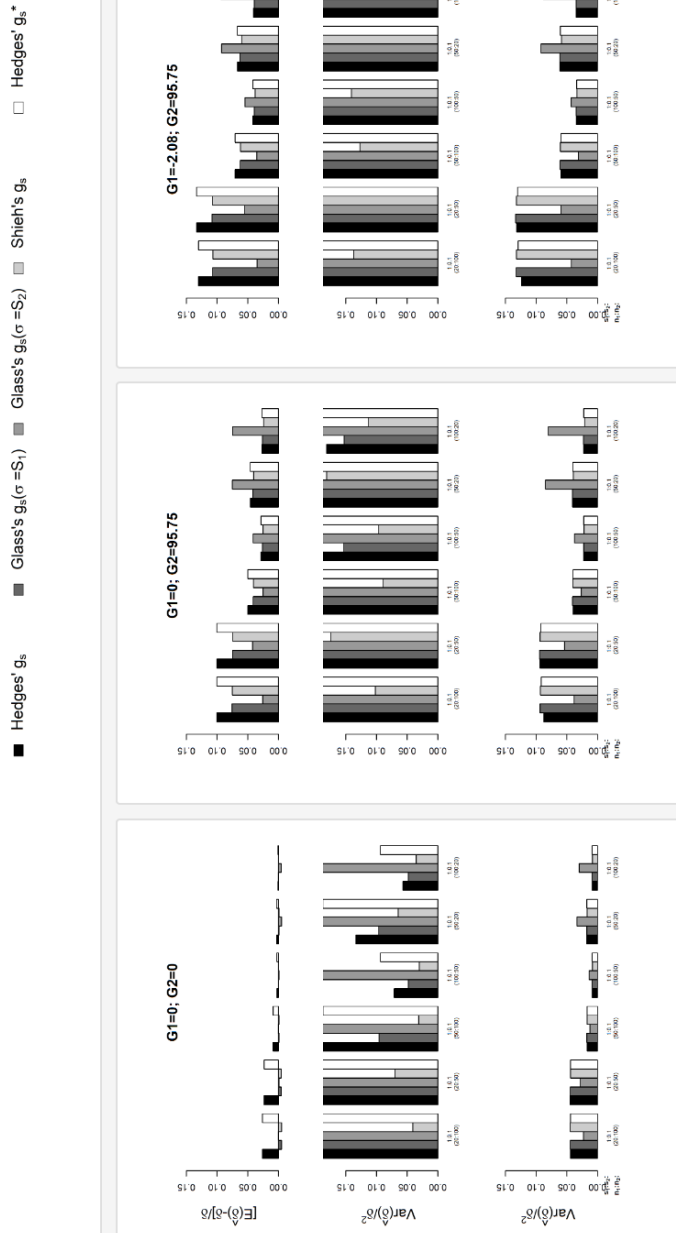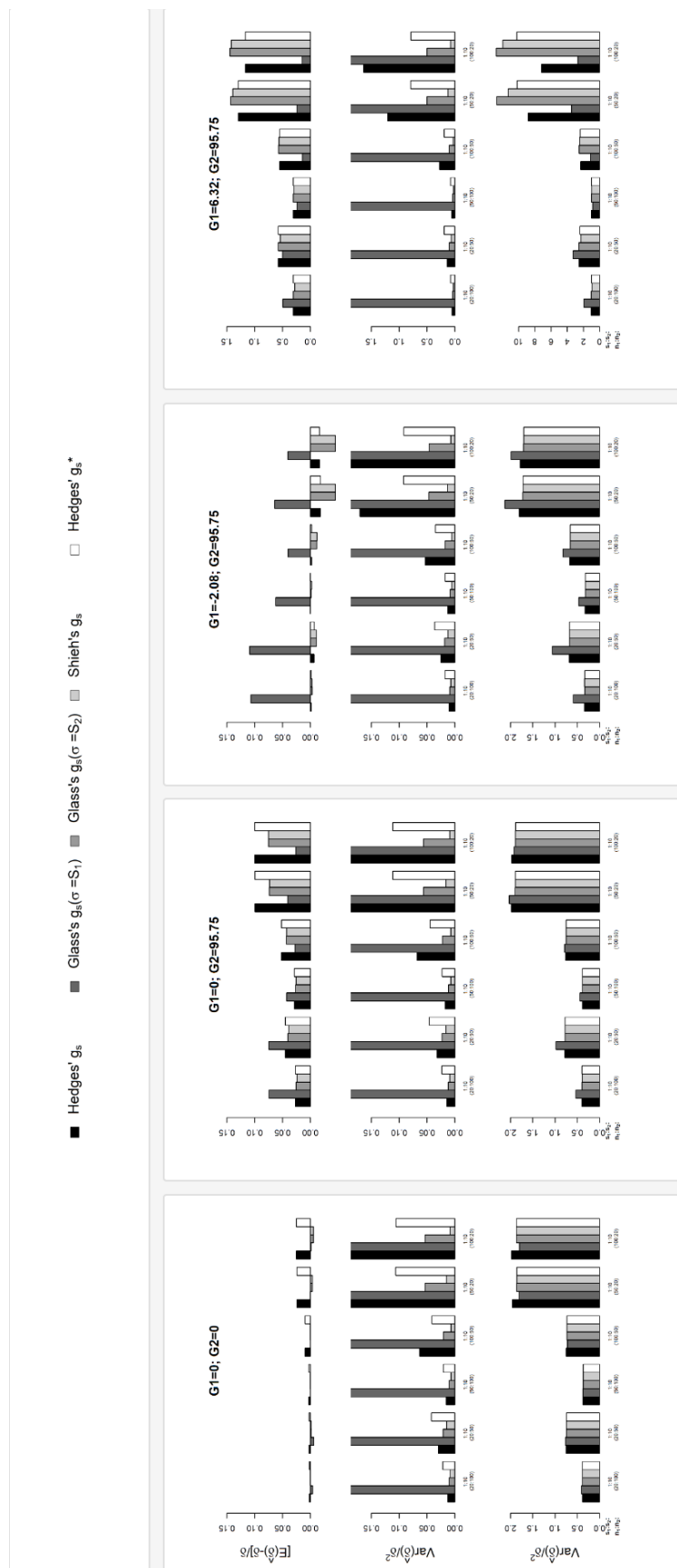
Figure 7. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is 10 times larger than σ_2

Figure 8. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_2 is 10 times larger than σ_1
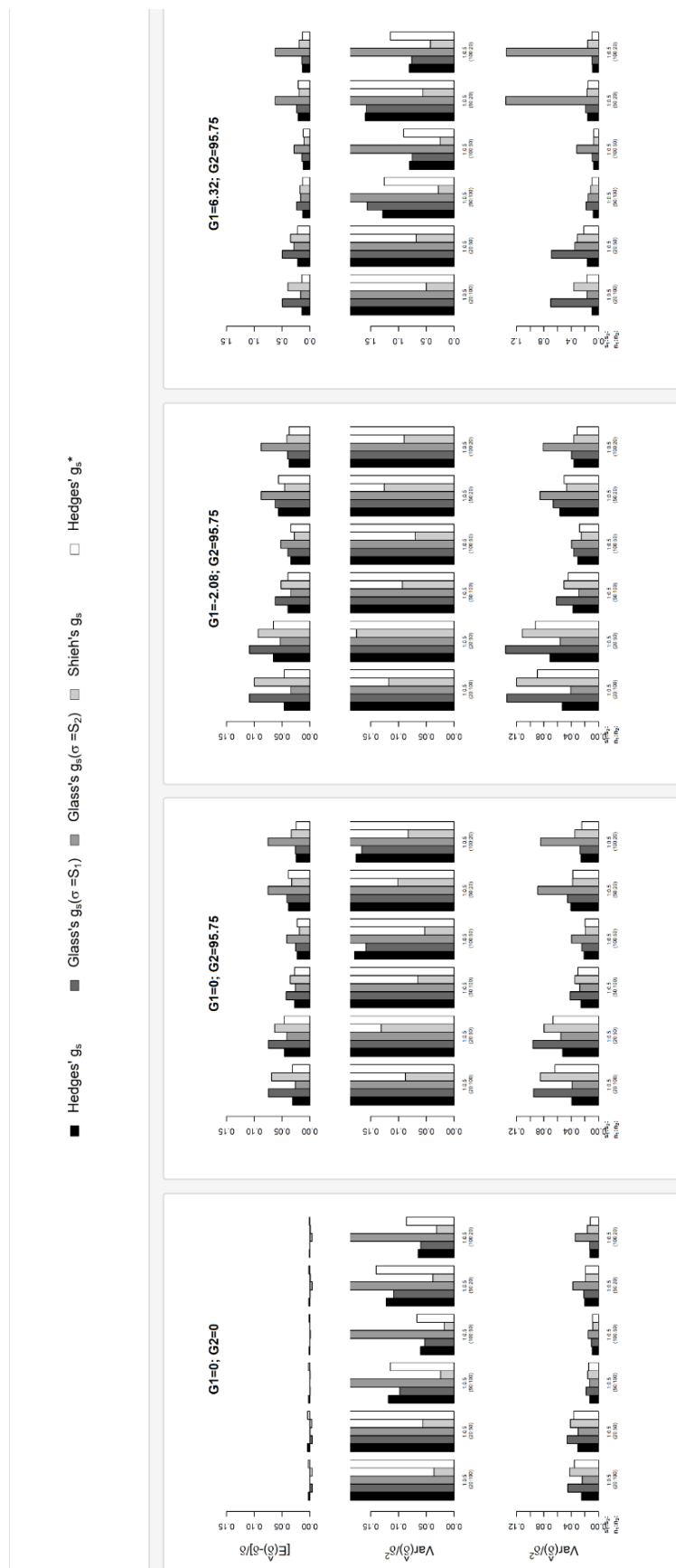
Figure 9. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is twice larger than σ_2
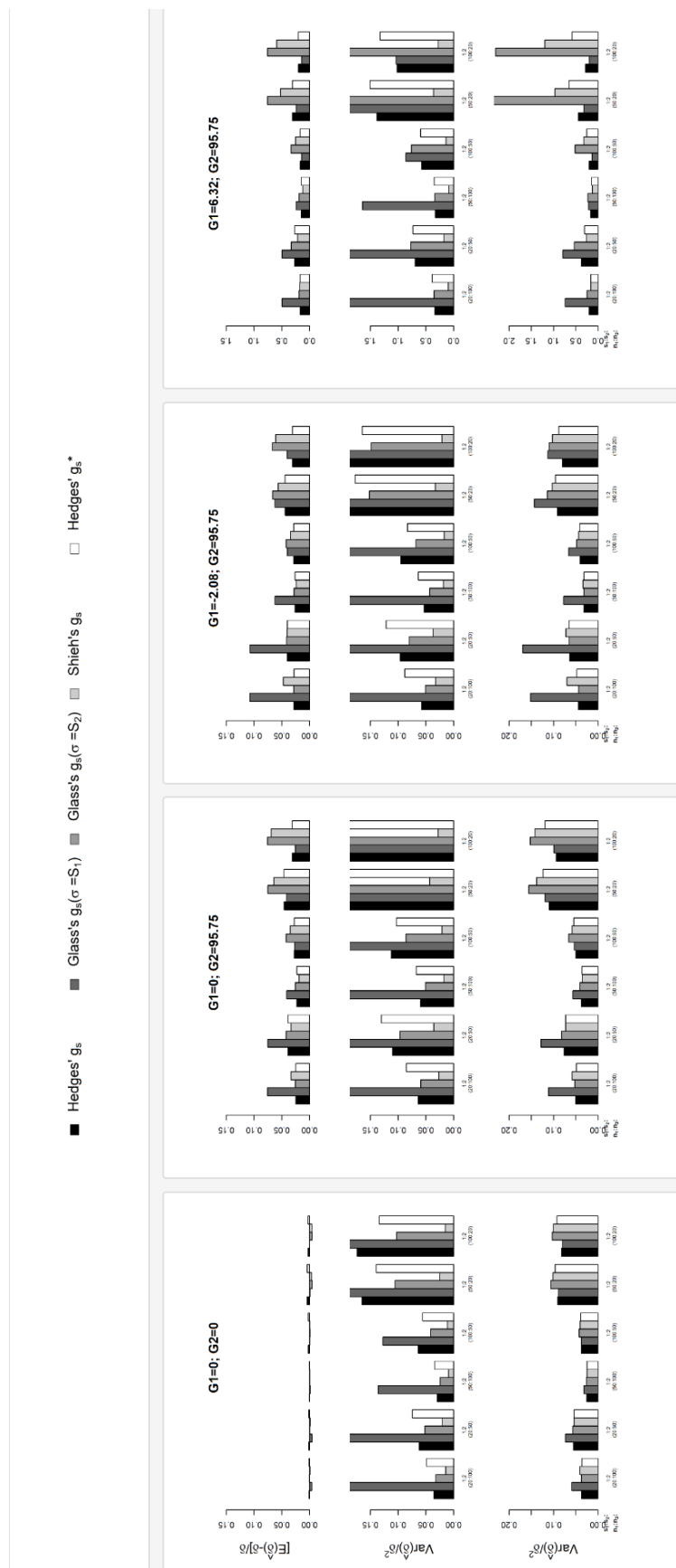
Figure 10. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_2 is twice larger than σ_1

456   When distributions are symmetric, the bias of Glass' g only depends on the size of

457 the control group and is therefore not impacted by either the sample sizes ratio or the total

458 sample size. When comparing Figures 7 to 10, one can also notice that the bias of Glass' g

459 does not depend on the SD-ratio either. Unlike the bias of Glass' g, its variance depends

460 on both sample sizes, but not in an equivalent way. In most scenarios it is more efficient, in

461 order to reduce the variance of Glass' g, to add subjects in the control group. Regarding

462 Hedges' g^* and Shieh's g, their respective biases and variances depend on an interaction

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

464 associated with the smallest bias and variance is not the same when the more variable

465 group is 10 times more variable than the other group (Figures 7 and 8) than when it is

466 only twice more variable (Figures 9 and 10). However, the respective biases and variances

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

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

469 bias and variance of Glass' g are sometimes smaller and sometimes larger than the bias of

470 Shieh's g and Hedges' g^*, due to a combination of three factors : (1) which group is larger,

471 (2) which group has the smallest standard deviation and (3) what is the correlation

472 between the standardizer and the mean difference.

473 **Recommendations**

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

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

476 variances assumption (unlike Hedges' g), b) it is always consistent (unlike Shieh's g), c) it

477 is easy to interpret (Hedges' g^* can be interpreted in the same way as Hedges' g) and d) it

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

479 across groups, as shown in the Shiny App at https://effectsize.shinyapps.io/ShiehvsCohen/.

480   Effect sizes estimates such as Hedges' g^* should always be reported with a confidence

481 interval. To help researchers compute Hedges' g^* and its confidence interval we created the

R package *deffectsize* (see https://github.com/mdelacre/deffectsize). The *datacohen_CI*

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

raw data and the *cohen_CI* function was built in order to compute point estimators and

confidence intervals based on descriptive statistics (sample means, sample variances and

sample sizes). By default, unbiased Hedges' g^* is computed but it is also possible to

compute biased estimators (e.g. Cohen's d^*) and/or to use a pooled error term as

standardizer by assuming that the equality of population variances is met (e.g. Hedges' g

or Cohen's d, depending on whether we choose to compute unbiased or biased estimator).

Other functions (*datashieh_CI*, *shieh_CI*, *dataglass_CI* and *glass_CI*) are available in

order to compute Shieh's g (or Shieh's d) and Glass' g (or Glass' d) as well as their

respective confidence intervals, even though we do not recommend to use these effect sizes

by default. Researchers who do not use R can use a Shiny app to compute point estimators

and confidence intervals based on descriptive statistics :

https://effectsize.shinyapps.io/deffsize/.

References

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery, 29*(5),
554–556. https://doi.org/10.1007/s00268-005-7911-0

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

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.
Paper presented at Annual Conference of the British Educational Research Association,
University of Exeter, Exeter, United-Kingdom.

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

Cumming, G. (2013). Cohen's *d* needs to be readily interpretable: Comment on shieh

(2013). *Behavior Research Methods*, *45*(4), 968–971.

https://doi.org/10.3758/s13428-013-0392-4

Duran, R. P., Eisenhart, M. A., Erickson, F. D., Grant, C. A., Green, J. L., Hedges, L. V.,

& Schneider, B. L. (2006). Standards for reporting on empirical social science research

in AERA publications: American Educational Research Association. *Educational*

Researcher, *35*(6), 33–40.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and*

the interpretation of research results. Cambridge, United-Kingdom: Cambridge

University Press.

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

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, CA.

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

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological*

Bulletin, *112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

Raviv, E. (2014, June 2). Bias vs. consistency. Retrieved July 12, 2021, 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*(4), 955–967.

https://doi.org/10.3758/s13428-013-0320-7

534   Stout, D. E., & Ruble, T. L. (1995). Assessing the practical significance of empirical results

535   in accounting education research: The use of effect size information. *Journal of*

536   *Accounting Education*, *13*(3), 281–298. https://doi.org/10.1016/0748-5751(95)00010-J

537   Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough.

538   *Journal of Graduate Medical Education*, *4*(3), 279–282.

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

540   Welch, B. L. (1938). The significance of the difference between two means when the

541   population variances are unequal. *Biometrika*, *29*(3), 350–362.

542   https://doi.org/10.2307/2332010

543   Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and

544   explanations. *American Psychologist*, *54*(8), 594–604.

545   https://doi.org/10.1037/0003-066X.54.8.594

Appendix

546 The bias of Cohen's _d_ is twice as large as the bias of Shieh's _d_ when population

547 variances and sample sizes are equal across groups: mathematical

548 demonstration.

549   As mentioned in Table 1, the bias of Cohen's d is defined as

$$Bias_{Cohen's\ d} = \delta_{Cohen} \times \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} - 1 \right) \tag{17}$$

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

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

550   As mentioned in Table 2, the bias of Shieh's d is defined as

$$Bias_{Shieh's\ d} = \delta_{Shieh} \times \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} - 1 \right) \tag{18}$$

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

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

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

552 below in equations 19 and 20:

$$\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{2(n-1)\sigma^2}{2(n-1)}}} = \boldsymbol{\frac{\mu_1 - \mu_2}{\sigma}} \tag{19}$$

553

$$\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{2\left(\frac{\sigma^2}{n/(2n)}\right)}} = \frac{\boldsymbol{\mu_1 - \mu_2}}{\boldsymbol{2\sigma}} \tag{20}$$

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

555 identical, as shown below in equations 21 and 22:

$$df_{Student} = \boldsymbol{2(n-1)} \tag{21}$$

556

$$df_{Welch} = \frac{[2(\sigma^2/n)]^2}{\frac{2(\sigma^2/n)^2}{n-1}} = \boldsymbol{2(n-1)} \tag{22}$$

557   Equations 17 and 18 can therefore be redefined as follows:

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

558

$$Bias_{Shieh's\,d} = \frac{\mu_1 - \mu_2}{\boldsymbol{2\sigma}} \times \left(\frac{\sqrt{n-1} \times \Gamma\left(\frac{2n-3}{2}\right)}{\Gamma(n-1)} - 1\right) \tag{24}$$

559   We can therefore conclude that the bias of Cohen's d is twice larger than the bias of

560 Shieh's d.

The variance of Cohen's d is four times larger than the bias of Shieh's d when population variances and sample sizes are equal across groups: mathematical demonstration.

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

$$Var_{Cohen's\ d} = \frac{N \times df_{Student}}{n_1 n_2 \times (df_{Student} - 2)} + \delta^2_{Cohen} \left[\frac{df_{Student}}{df_{Student} - 2} - \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} \right)^2 \right]$$
(25)

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

$$Var_{Shieh's\ d} = \frac{df_{Welch}}{(df_{Welch} - 2)N} + \delta^2_{Shieh} \left[\frac{df_{Welch}}{df_{Welch} - 2} - \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} \right)^2 \right]$$
(26)

We have previously shown in equations 21 and 22 that degrees of freedom associated with Student's t-test and Welch's t-test equal $2(n-1)$, when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$. As a consequence, the first term of the addition in equation 25 is 4 times larger than the first term of the addition in equation 26:

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

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

We have also previously shown in equations 19 and 20 that δ_{Cohen} is twice larger than δ_{Shieh} when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$ and, therefore, δ^2_{Cohen} is four times larger than δ^2_{Shieh}. As a consequence, the second term of the addition in equation 25 is also 4 times larger than the second term of the addition in equation 26. Because both terms of the addition in equation 25 are four times larger than those in equation 26, we can conclude that the variance of Cohen's d is four times larger than the variance of Shieh's d.