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 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

Argument to use the Hedges' g^* with the Welch's t-test : an investigation of Cohen's d and related effect size estimators in terms of bias, precision and robustness

Effect sizes are an important outcome of empirical research. Moving beyond decisions about statistical significance, there is a strong call for researchers to report and interpret effect sizes and associated confidence intervals. This practice is highly endorsed by the American Psychological Association (APA) and the American Educational Research Association (American Psychological Association, 2010; Duran et al., 2006).

In "between-subject" designs where individuals are randomly assigned into one of two independent groups and group scores are compared based on their means, the dominant estimator of effect size is Cohen's d, where the sample mean difference is divided by the pooled sample standard deviation (Peng et al., 2013; Shieh, 2013). This estimator is available in many statistical software packages, such as SPSS and Stata. However, computing the pooled sample standard deviation is tailor-made for the situation when both sample variances are estimates of a common population variance, which is known as the homogeneity of variance assumption. It has been widely argued that there are many fields in psychology where this assumption is ecologically unlikely (Delacre et al., 2017; Erceg-Hurn & Mirosevich, 2008; Grissom, 2000). The question how to deal with the assumption of equal variances has been widely explored in the context of hypothesis testing, and it is becoming increasingly common to by default report a t-test that does not assume equal variances, such as Welch's t-test.

However, the question which effect size to report when equal variances are not assumed has received less attention. One possible reason is that researchers have not found consensus on which of the available options should be used (Shieh, 2013). Even within the very specific context of an estimate for the standardized sample mean difference there is little agreement about which estimator is the best choice. In this article, we will review the

⁴⁴ main candidates that have been proposed in the literature in the *d* family of effect sizes,

⁴⁵ without (Cohen's *d*, Glass' *d*, Shieh's *d* and Cohen's d^*) and with correction for bias

⁴⁶ (Hedges' *g*, Glass' *g*, Shieh's *g* and Hedges' g^*). We provide an R package and Shiny app

⁴⁷ to compute relevant effect size measures and their confidence intervals.

⁴⁸   Before reviewing the most important effect size measures in the *d*-family, we will first

⁴⁹ list the different purposes effect size measures serve, and discuss the relationship between

⁵⁰ effect sizes, statistical, and practical significance. Based on a detailed description of the

⁵¹ good properties an effect size measure should possess, we will evaluate these properties in

⁵² the Monte Carlo simulations we performed to compare the different effect size estimators

⁵³ with correction for bias.

Three purposes of effect size estimators

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

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

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

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

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

⁶⁰   First, effect size measures can be used for *interpretative* purposes. They allow

⁶¹ researchers to assess the practical significance of a result (i.e. statements about the

⁶² relevance of an effect in real life). In order to assess the meaningfulness of an effect, we

⁶³ should be able to relate this effect size estimate with behaviors/meaningful consequences in

⁶⁴ the real world (Andersen et al., 2007). This typically involves an analysis of the costs

⁶⁵ (determined by a specific context) and the benefits (in part determined by the size of the

⁶⁶ effect). It is important to remember an effect size is just a mathematical indicator of the

⁶⁷ magnitude of a difference, which depends on the way a variable is converted into numerical

⁶⁸ indicator. An effect size in itself is not a measure of the importance or the relevance of an

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

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

71   Second, effect size measures can be used for *comparative* purposes. They allow

72 researchers to assess the stability of results across designs, analyses, and sample sizes. This

73 includes statistically comparing and combining the results from two or more studies in a

74 meta-analysis.

75   Third, effect size measures can be used for *inferential* purposes. Hypothesis tests and

76 confidence intervals based on the same statistical quantity are directly related : if the area

77 of the null hypothesis is out of the $(1 - \alpha)$-confidence interval, then the hypothesis test

78 would also result in a *p*-value below the nominal alpha level. At the same time, the interval

79 provides extra information about the precision of the sample estimate for inferential

80 purposes (Altman, 2005; Ellis, 2010), and which effect sizes are excluded. The narrower the

81 interval, the higher the precision, and the wider the confidence interval, the more the data

82 lack precision. Effect size measures are also indirectly related to the hypothesis tests as

83 effect sizes from previous studies can be used in an a-priori power analysis when planning a

84 new study (Lakens, 2013; Prentice & Miller, 1992; Stout & Ruble, 1995; Sullivan & Feinn,

85 2012; Wilkinson, 1999).

86   **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).

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

96   population parameter. An estimator is positively (or negatively) biased if the distribution

97   is centered around a value that is higher (or lower) than the true population parameter

98   (see Figure 1). In other words, examining the bias of an estimator tells us if estimates are

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

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

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

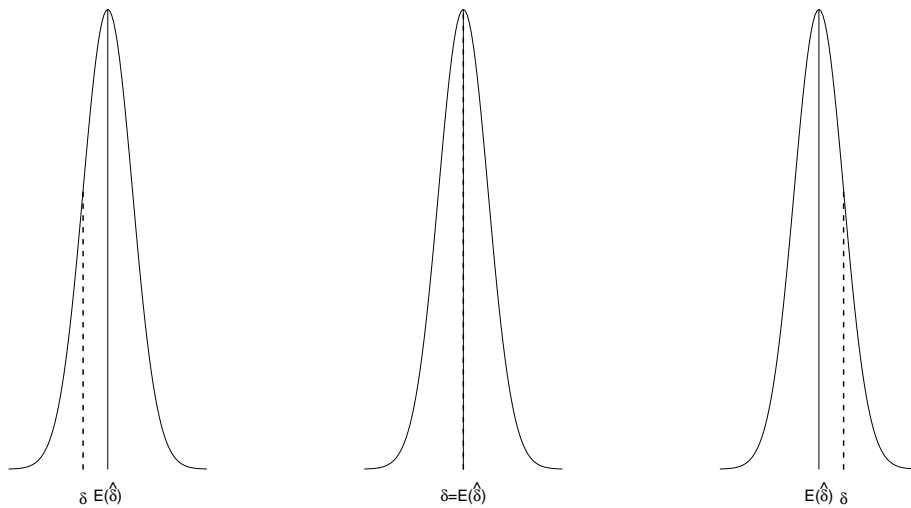101   true (population) parameter.

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. As we shall see in the sequel,

104   each effect size estimator actually estimates a different expression of the population effect

105   size (for instance, the expression of δ_{Cohen} is different from δ_{Glass}). It is therefore necessary

106   to examine the *relative bias*, defined as the ratio between the bias and the population effect

107 size :

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

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

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

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

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

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

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

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

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

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

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

117 where $Var(\hat{\delta})$ is the variance of the sampling distribution of the estimator $\hat{\delta}$. This variance

118 inequality of course only makes sense if two estimators $\hat{\delta}_1$ and $\hat{\delta}_2$ estimate the same

119 population quantity, which is not the case here as explained before (each effect size

120 estimator corresponds to a distinct population effect size expression). Therefore, we rather

121 have to consider the *relative variance*, defined as the ratio between the variance of an

122 estimator and the square of the corresponding population effect size δ :

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

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

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

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

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

127 variance (Raviv, 2014).

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

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

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

131   **Different measures of effect sizes**

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

133 conditions. In all generality, the population effect size is defined as

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

134 where μ_1 and μ_2 respectively stand for the population means of population 1 and 2, and σ

135 represents the "variability" of the combined observations from both populations. Typically

136 one assumes that both populations are independent and normally distributed (the

137 normality assumption however is often a too strong limitation; Delacre et al., 2017). In

138 some situations it is moreover assumed that the variances of both populations are equal, in

139 which case σ stands for the common standard deviation. There exist different estimators of

140 the effect size measure in expression 5. For all, the mean difference is estimated by the

141 difference $\bar{X}_1 - \bar{X}_2$ of both sample means. When the equality of variances assumption is

142 made, σ is estimated by pooling both sample standard deviations (S_1 and S_2). When the

143 equality of variances assumption cannot be made, the pooled standard deviation is a

144 possible but not recommendable choice as we explain later on, and alternatives are

145 available. In what follows, we shall present various effect size measures. For each effect

146 size, we will provide information about their theoretical bias, variance and consistency.

147 Simulation-based results will be discussed in the section "Monte Carlo Simulations :

148 assessing the bias, efficiency and consistency of 5 estimators" below.

149 **Cohen's *d* and Hedges' *g***

     When we have good reasons to assume equality of variances between groups then the

     most common effect size is Cohen's *d*, where the sample mean difference is divided by a

pooled error term (Cohen, 1965) :

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

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

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

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

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

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

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

156   to Student's t-statistic :

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

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

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

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

159   and noncentrality parameter[1]

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

160   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

161   j^{th} population ($j = 1, 2$). We attract the reader's attention to the act that δ_{Cohen} is the

162   expression of the effect size that the Cohen's d is estimating. The relationship described in

163   equation 6 and the theoretical distribution of Student's t-statistic allow us to determine the

164   sampling distribution of Cohen's d, and therefore, its expectation and variance when the

165   assumptions of normality and equal variances are met. All these equations are provided in

---

[1] 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.

166 Table 1. For interested readers, Supplemental Material 1 provides a detailed examination

167 of the theoretical bias and variance of Cohen's d based on Table 1, as well as the bias and

168 variance of all other estimators described later, based on Tables 2 and 3, with the goal to

169 determine which parameters influence the bias and variance of different estimators.

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,

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 :

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

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

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

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

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

174 sample sizes.[2]

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

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

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

———

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

178 2005; Shieh, 2013). If we pool the estimates of two unequal population variances, the

179 estimator of effect size will be smaller as it should be in case of positive pairing (i.e. the

180 group with the larger sample size is extracted from the population with the larger variance)

181 and larger as it should be in case of negative pairing (i.e. the group with the larger sample

182 size is extracted from the population with the smaller variance).

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. $N \geq 5$).

Three popular alternatives from the literature

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, σ_c^2 and σ_e^2 are the population variance of the control and experimental groups and and df and ncp are defined as follows :

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

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

204 where $\delta_{Glass} = \frac{\mu_c - \mu_e}{\sigma_c}$ and μ_c and μ_e are respectively the mean of the populations control

205 and experimental groups are extracted from. Note that here the population effect size

206 δ_{Glass} is based on the standard deviation σ_c of the control group. Thanks to equation 8, we

207 can compute the theoretical expectation and variance of Glass's d when the assumption of

208 normality is met (see Table 2), and therefore determine which factors influence bias and

209 variance, and how they do so (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 :

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

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

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

215   The exact distribution of Welch's t-statistic is more complicated than the exact distribution

216   of Student's t-statistic, but it can be approximated, under the assumption of normality, by

217   a t-distribution with degrees of freedom and noncentrality parameters (Welch, 1938)

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

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

218   where $\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}}$ is the population effect size measure estimated by Shieh's d.

219   The relationship described in equation 10 and the theoretical distribution of Welch's

220   t-statistic allow us to approximate the sampling distribution of Shieh's d. Based on the

221   sampling distribution of Shieh's d, we can estimate its theoretical expectation and variance

222   under the assumption of normality (see Table 2), and thereby determine which factors

223   influence bias and variance, and how they do so (see Supplemental Material 1).

224   As demonstrated in Appendices 1 and 2, when population variances and sample sizes

225   are equal across groups, the biases and variances of Cohen's d and Shieh's d are identical

226   except for multiplication by a constant. The same is true for the population effect size

227   expressions δ_{Cohen} and δ_{Shieh} :

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

228

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

229

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

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

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

232 and variance are identical.

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

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

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

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

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

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

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

240 illustrated in the following shiny application :

241 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 :

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

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

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

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

²⁴⁴ where df^* and ncp^* are defined as follows :

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

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

²⁴⁵ where $\delta_{Cohen}^* = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$ is the corresponding population effect size. Using equation 15 we

²⁴⁶ can compute its theoretical expectation and variance when the assumption of normality is

²⁴⁷ met (see Table 2), and therefore determine which factors influence bias and variance, and

²⁴⁸ how they do so (see Supplemental Material 1). This estimator has been widely criticized,

²⁴⁹ because it results in a variance term of an artificial population (i.e. since the variance term

²⁵⁰ does not estimate the variance of one or the other group, the composite variance is an

²⁵¹ estimation of the variance of an artificial population; Grissom & Kim, 2001). However, the

²⁵² same criticism would then apply to the celebrated Cohen's d because the pooled variance

²⁵³ also corresponds to the variance of an artificial population, unless when variances are equal.

²⁵⁴ But for equal variances the variance term of Cohen's d^* also yields σ^2, which is perfectly

²⁵⁵ interpretable. Moreover, we will show throughout the simulation section that this estimator

²⁵⁶ exhibits very good inferential properties. Moreover, it has a constant value across sample

²⁵⁷ sizes ratios, as 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 :

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

²⁵⁸ where the distinct values of ν are provided in equation 9 for Glass' g, in equation 11 for

²⁵⁹ Shieh's g and in equation 16 for Hedges' g^*. The three corrected estimators are

²⁶⁰ theoretically unbiased when the assumption of normality is met. Their variance is a

²⁶¹ function of the same parameters as their biased equivalent. However, due to the correction

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

263 summary :

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

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

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

267 • 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

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

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

270 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. $c_f = \dfrac{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}$; $\delta_{Glass} = \frac{\mu_c-\mu_e}{\sigma_c}$, $\delta_{Shieh} = \dfrac{\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N}+\frac{\sigma_2^2}{n_2/N}}}$ 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 in this table require $df \geq 3$ and at least 2 subjects per group.

274 Monte Carlo Simulations

275 **Assessing the bias, efficiency and consistency of 5 estimators.**

276 *Method.*

277 We performed Monte Carlo simulations using R (version 3.5.0) to assess the bias,

278 efficiency and consistency of Hedges' g, Glass' g (using respectively the sample SD of the

279 first or second group as a standardizer), Hedges' g^* and Shieh's g.

280 A set of 100,000 datasets was generated for 1,008 scenarios as a function of different

281 criteria. In 252 scenarios, samples were extracted from a normally distributed population

282 (in order to ensure the reliability of our calculation method) and in 756 scenarios, samples

283 were extracted from non normal population distributions. In order to assess the quality of

284 estimators under realistic deviations from the normality assumption, we referred to the

285 review of Cain et al. (2017). Cain et al. (2017) investigated 1,567 univariate distributions

286 from 194 studies published by authors in Psychological Science (from January 2013 to June

287 2014) and the American Education Research Journal (from January 2010 to June 2014).

288 For each distribution, they computed Fisher's skewness

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

289 and kurtosis

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

290 where n is the sample size and m_2, m_3 and m_4 are respectively the second, third and fourth

291 centered moments. They found values of kurtosis from $G_2 = $ -2.20 to 1,093.48. According

292 to their suggestions, throughout our simulations, we kept constant the population kurtosis

293 value at the 99th percentile of their distribution of kurtosis, i.e. $G_2 = 95.75$. Regarding

294 skewness, we simulated population parameter values which correspond to the 1st and 99th

295 percentile of their distribution of skewness, i.e. respectively $G_1 = $ -2.08 and $G_1 = 6.32$. We

296 also simulated samples extracted from population where $G_1 = 0$, in order to assess the

²⁹⁷ main effect of high kurtosis on the quality of estimators. All possible combinations of

²⁹⁸ population skewness and kurtosis and the number of scenarios for each combination are

²⁹⁹ summarized in Table 4. To provide the reader with an intuition about the four

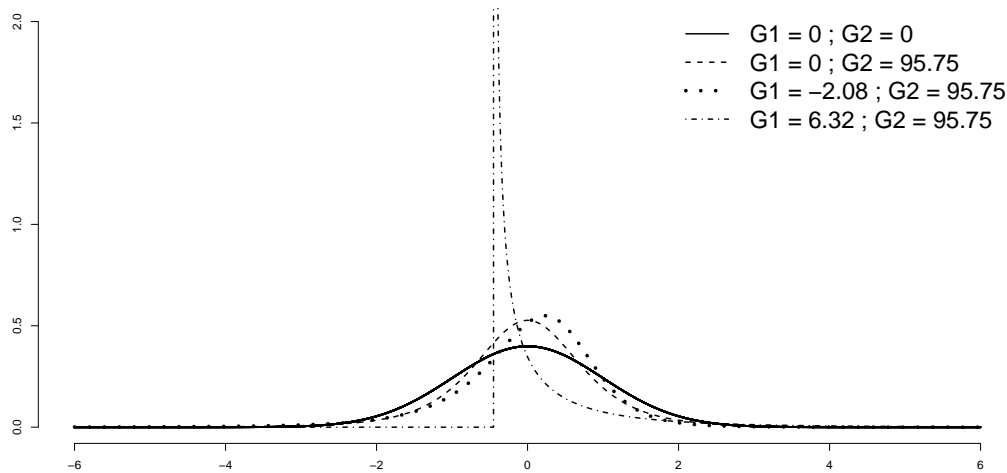³⁰⁰ distributions we consider, we have plotted them in Figure 2.

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**

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

302   combinations where both G1 and G2 equal 0 correspond to the normal case.

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

304   parameter values were chosen in order to illustrate the consequences of factors identified as

305   playing a key role on the variance of unbiased estimators. We manipulated the population

306   mean difference ($\mu_1 - \mu_2$), the sample sizes (n_1 and n_2), the sample size ratio (n-ratio $=$

307    $\frac{n_2}{n_1}$), the population SD-ratio (i.e. $\frac{\sigma_2}{\sigma_1}$), and the sample size and population variance pairing

308    $\left(\frac{n_2}{n_1} \times \frac{\sigma_2}{\sigma_1} \right)$. In our scenarios, μ_2 was always 0 and μ_1 varied from 1 to 4, in steps of 1 (so

309   does $\mu_1 - \mu_2$). Thus, we only consider a positive population effect size, which will result in

310   positive bias of our estimators for reasons explained in Supplemental Material 1.[3]

311   Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10, and therefore, the

312   SD-ratio were 10, 4, 2, 1, .5, .25 or .1. The simulations for which both σ_1 and σ_2 equal 1

313   are the particular case of homoscedasticity, or equal population variances across groups.

314   The sample sizes of both groups (n_1 and n_2) were 20, 50 or 100. When sample sizes of both

315   groups are equal, the n-ratio equals 1 (this is known as a balanced design). All possible

316   combinations of n-ratio and population SD-ratio were simulated in order to distinguish

---

[3] 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**
SD-ratio	**1**	a	b	b
	>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 explained why we need to consider here relative bias and relative variance instead of raw bias and variance. Thus, anytime we will mention the biases and variances in the results section, we will be referring to relative bias and variance.

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

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.

341   4) Because Cohen's δ is not an appropriate expression of the effect size when population

342   variances and sample sizes are unequal across groups, we will not plot the bias and

343   variance of Hedges' g in Figures 7 to 10.

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

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

346 variance of different estimators are presented. When describing the Glass' g estimators, we

347 will systematically refer to the "control group" as the condition the standardizer is based

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

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

350   *When variances are equal across groups*

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

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

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

354 Figures).[4] However, the more the data generation process deviations from the normality

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

356 estimators.

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

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

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

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

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

362 these findings are detailed in the three paragraphs below.

---

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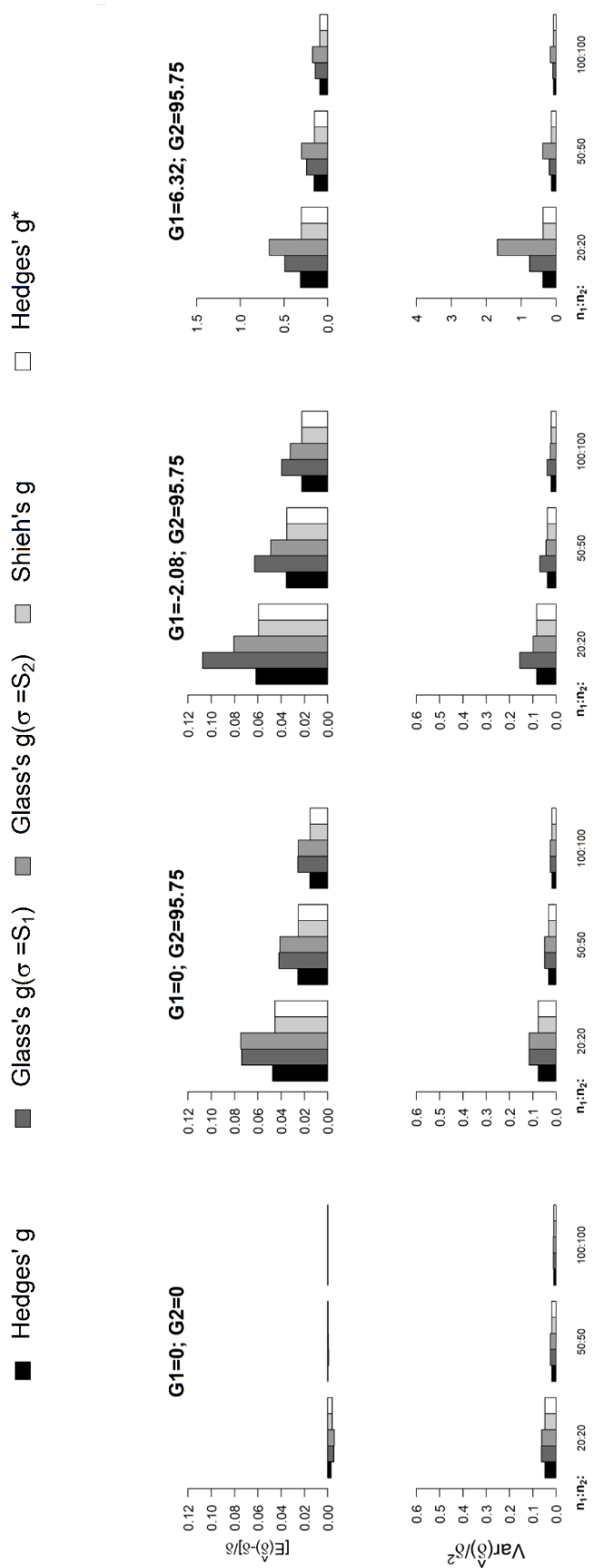
Figure 3. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are equal across groups (condition a)
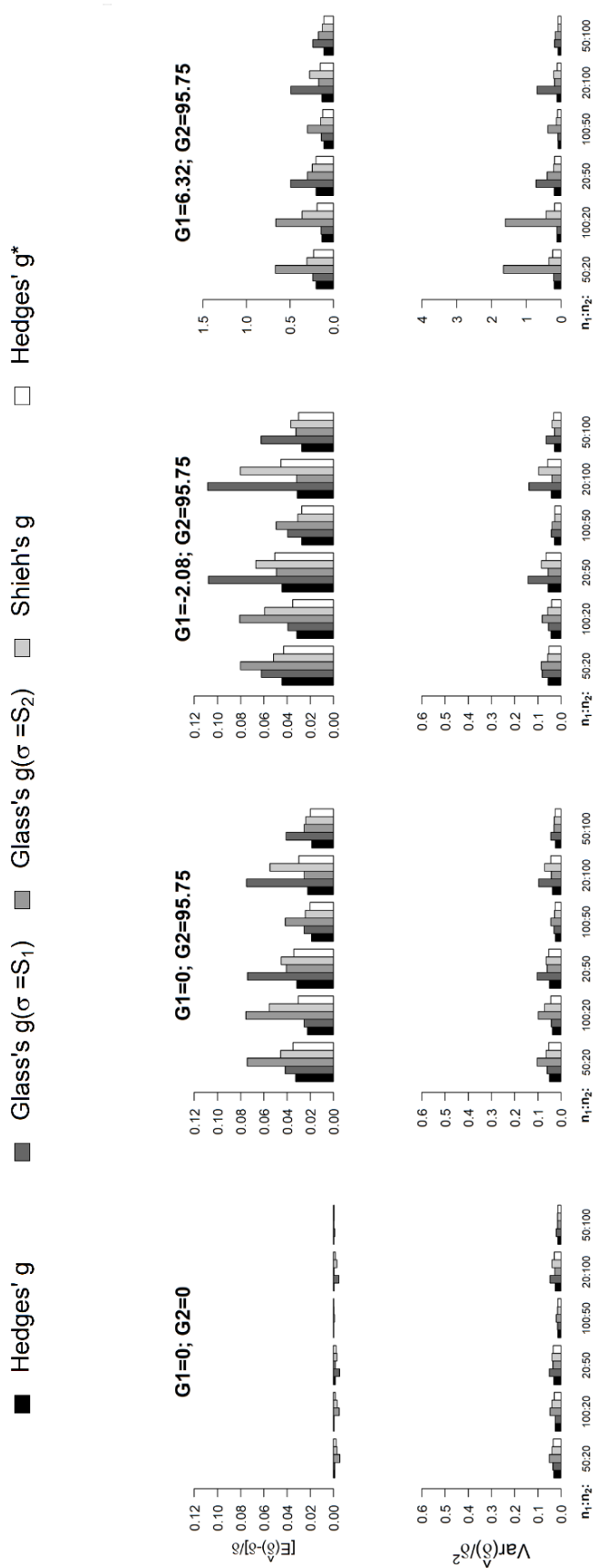
Figure 4. Bias and efficiency of estimators of standardized mean difference, when variances are equal across groups and sample sizes are unequal (condition b)

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

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

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

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

367 g^* are similar[5] and smaller than the bias and variance of Glass' g estimates using either S_1

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

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

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

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

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

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

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

375 impacts the bias and variance of estimators.

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

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

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

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

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

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

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

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

_____

[5] 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.

³⁸⁴ the variance, it is much more efficient to add subjects in the control group and when the

³⁸⁵ relative size of the experimental group decreases so does the variance, even when the total

³⁸⁶ sample size is decreased. Regarding Shieh's g, for a given sample size ratio, the bias and

³⁸⁷ variance will decrease when sample sizes increase. However, there is a large effect of the

³⁸⁸ sample sizes ratio such that when the sample sizes ratio moves away from 1 by adding

³⁸⁹ subjects, bias and variance might increase.[6] On the other hand, when the sample sizes

³⁹⁰ ratio moves closer to 1 by adding subjects, the bias will decrease.

³⁹¹   When samples are extracted from skewed distributions and have unequal sizes (the

³⁹² two last columns in Figure 4), for a constant total sample size, Glass' g, Shieh's g and

³⁹³ Hedges' g^* will show different bias and variance depending on which group is the largest

³⁹⁴ one (e.g. when distributions are right-skewed, the bias and variance of all these estimators

³⁹⁵ when n_1 and n_2 are respectively 50 and 20 are not the same as their bias and variance when

³⁹⁶ n_1 and n_2 are respectively 20 and 50). This is due to a non-null correlation of opposite sign

³⁹⁷ between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and their respective standardizers depending on

³⁹⁸ which group is the largest one, as detailed in Supplemental Material 3. One observes that

³⁹⁹ under these configurations, the bias and variance of Glass' g are sometimes a bit smaller

⁴⁰⁰ and sometimes much larger than the bias and variance of Shieh's g and Cohen's d^*.[7]

––––––

[6] 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.

[7] 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.[8]

We 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,

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

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

performs best across scenarios is Hedges' g^*.

---

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.

[8] 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.
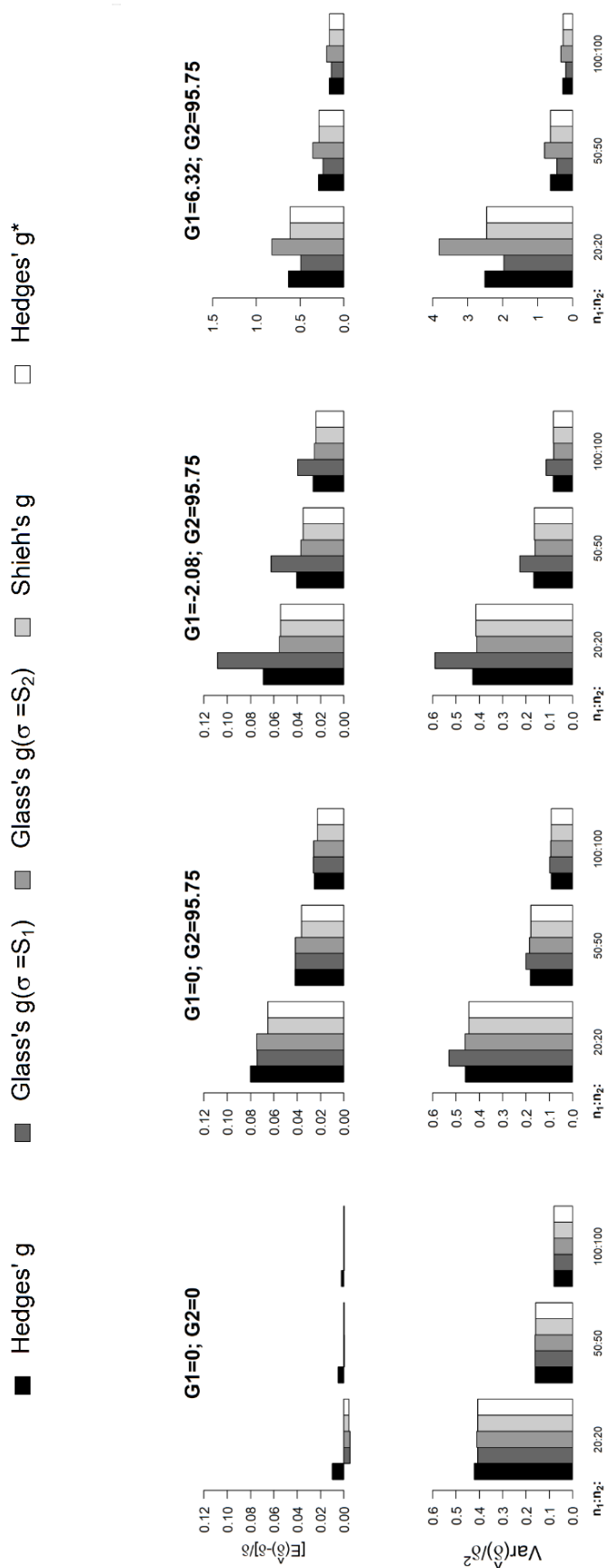
Figure 5. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c), as a function of sample sizes
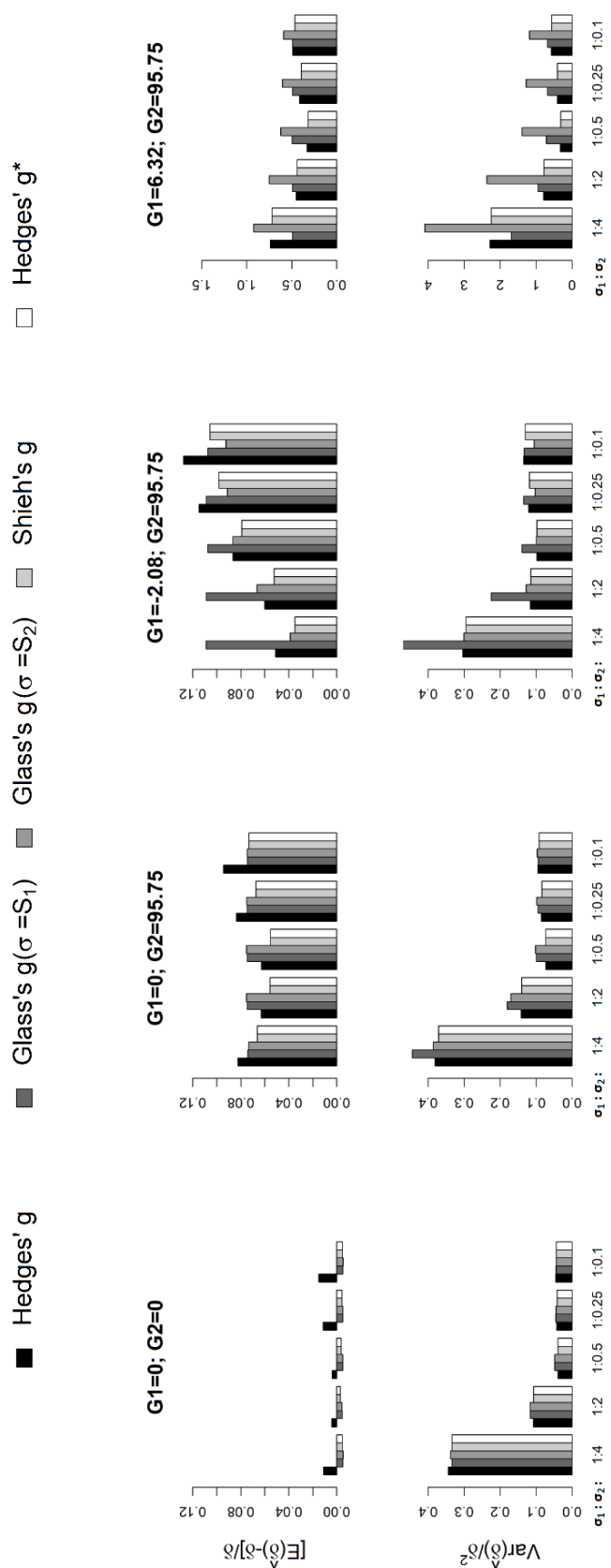
Figure 6. Bias and efficiency of estimators of standardized mean difference, when variances are unequal across groups and sample sizes are equal (condition c) as a function of the *SD*-ratio (when $n_1 = n_2 = 20$)

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

417 between groups and sample sizes are equal (condition c). In Figure 5, 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 6, 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

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

423 variance seems to be much larger when $S_2 > S_1$.[9] This information should not be taken

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

425 way we computed the relative variance.[10]

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

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

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

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

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

431 the standardizer of *all* estimators, depending on which population variance is larger.[11]

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

---

[9] 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.

[10] 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.

[11] 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.

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

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

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

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

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

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

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

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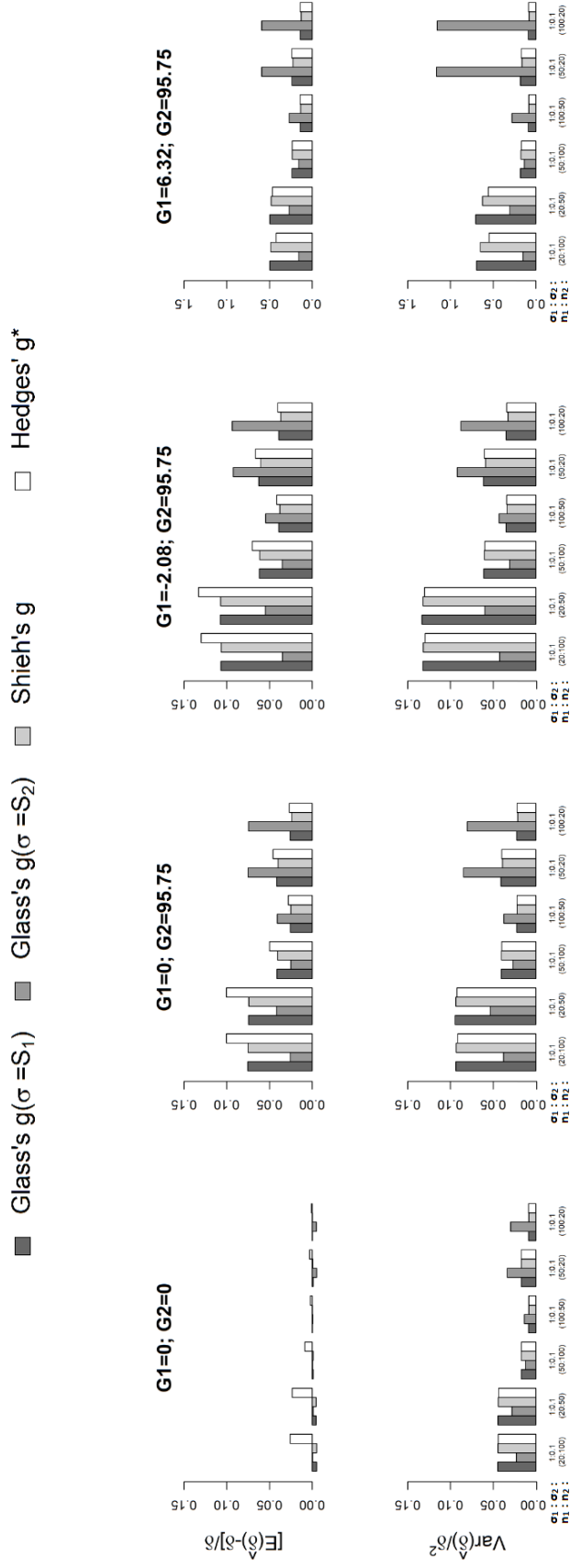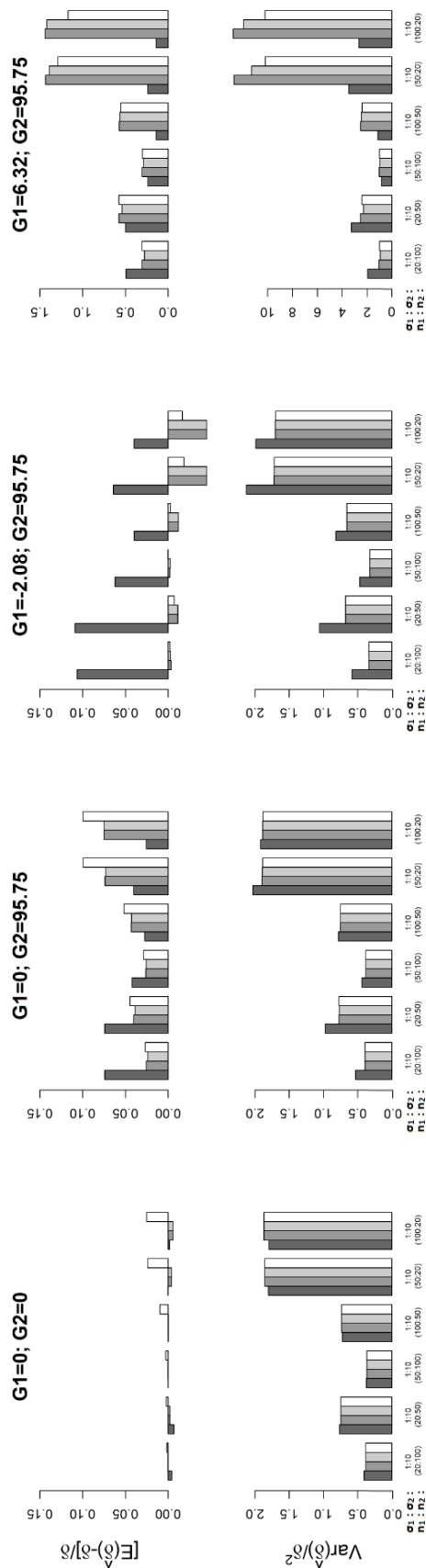
Figure 7. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is 10 times larger than σ_2

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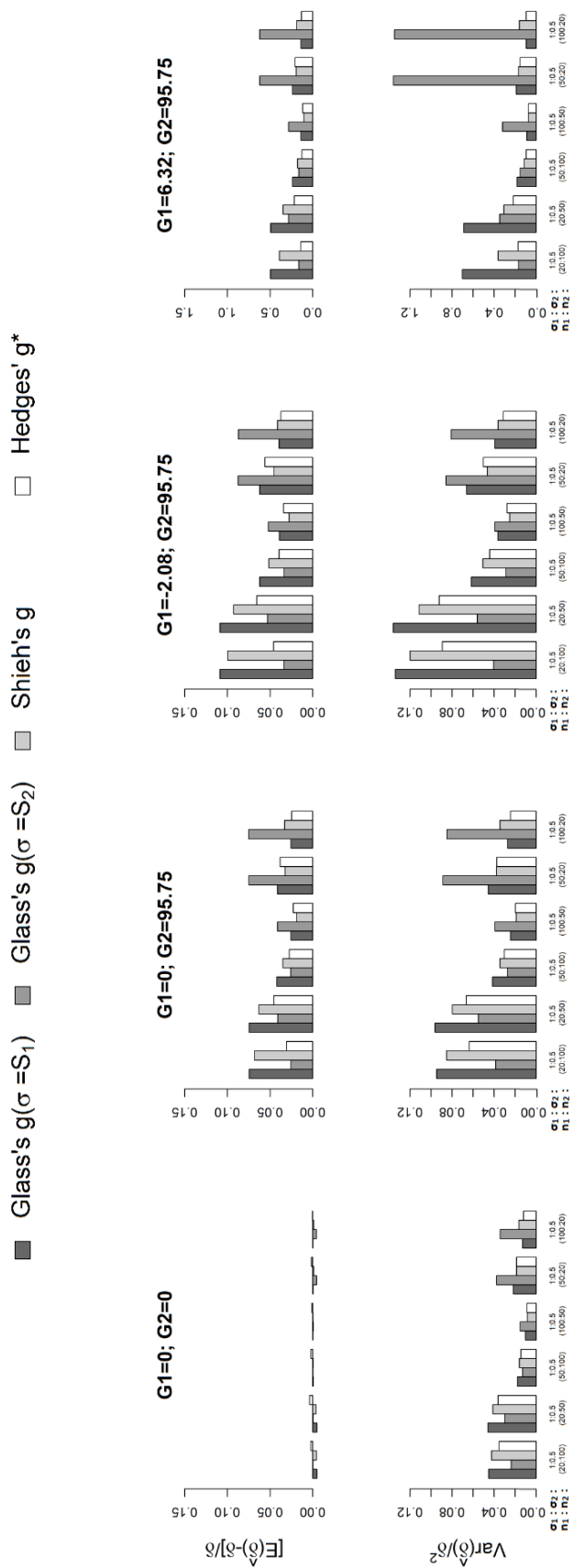
Figure 9. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_1 is twice larger than σ_2
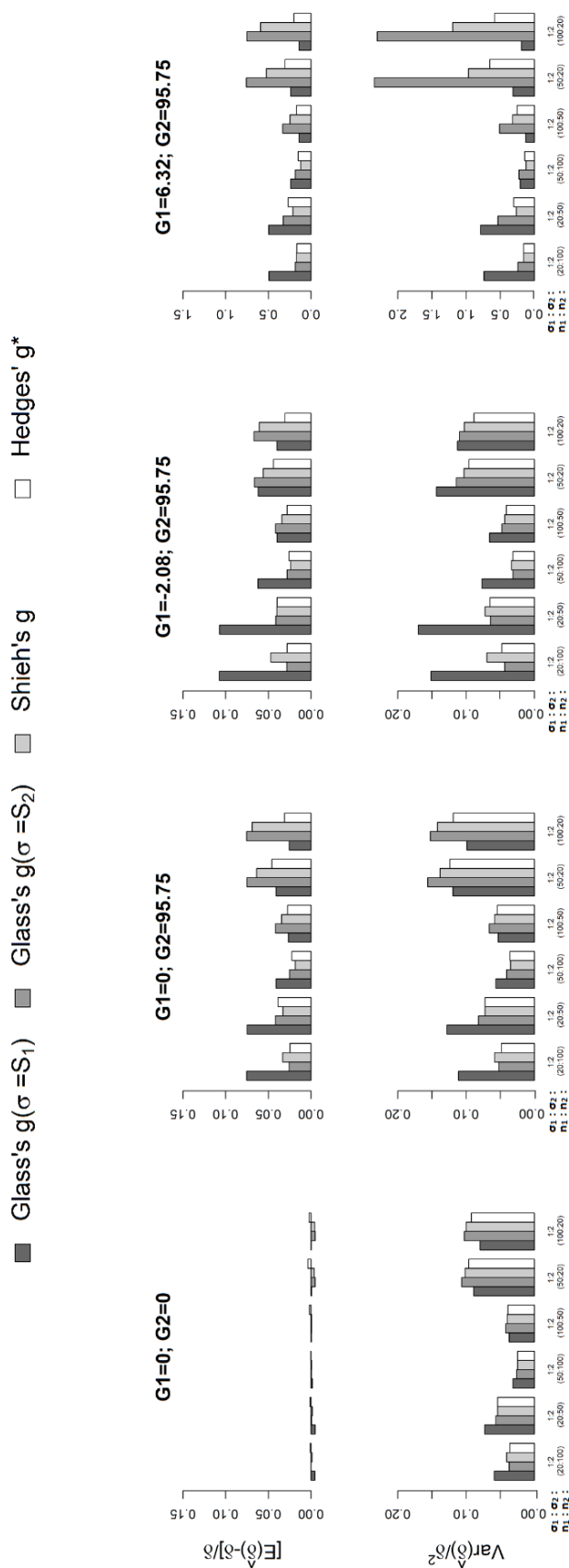
Figure 10. Bias and efficiency of estimators of standardized mean difference, when variances and sample sizes are unequal across groups (condition d), and σ_2 is twice larger than σ_1

When distributions are symmetric, the bias of Glass' g only depends on the size of the control group and is therefore not impacted by either the sample sizes ratio or the total sample size. When comparing Figures 7 to 10, one can also notice that the bias of Glass' g does not depend on the SD-ratio either. Unlike the bias of Glass' g, its variance depends on both sample sizes, but not in an equivalent way. In most scenarios it is more efficient, in order to reduce the variance of Glass' g, to add subjects in the control group. Regarding Hedges' g^* and Shieh's g, their respective biases and variances depend on an interaction 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 associated with the smallest bias and variance is not the same when the more variable group is 10 times more variable than the other group (Figures 7 and 8) than when it is only twice more variable (Figures 9 and 10). However, the respective biases and variances of Hedges' g^* and Shieh's g are always smaller when there is a positive pairing between sample sizes and variances. When samples are extracted from skewed distributions, the bias and variance of Glass' g are sometimes smaller and sometimes larger than the bias of Shieh's g and Hedges' g^*, due to a combination of three factors : (1) which group is larger, (2) which group has the smallest standard deviation and (3) what is the correlation between the standardizer and the mean difference.

Recommendations

We recommend using Hedges' g^* in order to assess the magnitude of the effect when comparing two independent means, because a) it does not rely on the equality of population variances assumption [12] (unlike Shieh's g), c) it is easy to interpret (Hedges' g^* can be interpreted in the same way as Hedges' g) and d) it remains constant for any sample sizes ratio, even when population variances are unequal across groups, as shown in

---

[12] The assumption of equal variances across populations is rarely realistic in practice (Cain et al., 2017; Delacre et al., 2017; Delacre et al. , 2019; Erceg-Hurn & Mirosevich, 2008; Glass et al., 1972; Grissom, 2000; Micceri, 1989; Yuan et al., 2004) (unlike Hedges' g), b) it is always consistent

465 the Shiny App at https://effectsize.shinyapps.io/ShiehvsCohen/.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

480 and confidence intervals based on descriptive statistics :

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

482   **References**

483 Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence intervals for an effect size

484   when variances are not equal. *Journal of Modern Applied Statistical Methods*, *5*(1),

485   2–13. https://doi.org/10.22237/jmasm/1146456060

486 Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery*, *29*(5),

487   554–556. https://doi.org/10.1007/s00268-005-7911-0

488 American Psychological Association. (2010). *Publication manual of the american

489   psychological association [APA]* (6th ed.). Washington, DC: American Psychological

490   Association.

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 and Exercise Psychology, *29*(5), 664–672.

https://doi.org/10.1123/jsep.29.5.664

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

kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior

Research Methods*, *49*(5), 1716–1735. 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.

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

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

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.

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

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

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 sizes for research: A broad practical approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

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.

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., Wilcox, R. R., & Deering, K. N. (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

543 Kulinskaya, E., & Staudte, R. G. (2007). Confidence intervals for the standardized effect

544   arising in the comparison of two normal populations. *Statistics in Medicine*, *26*(14),

545   2853–2871. https://doi.org/10.1002/sim.2751

546 Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A

547   practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12.

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

549 Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond cohen's *d*: Alternative effect size measures

550   for between-subject designs. *The Journal of Experimental Education*, *82*(1), 22–50.

551   https://doi.org/10.1080/00220973.2012.745471

552 Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA

553   and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*(2),

554   157–209. https://doi.org/10.1007/s10648-013-9218-2

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

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

557 Raviv, E. (2014, June 2). Bias vs. consistency. Retrieved July 12, 2021, from

558   https://eranraviv.com/bias-vs-consistency/

559 Shieh, G. (2013). Confidence intervals and sample size calculations for the standardized

560   mean difference effect size between two normal populations under heteroscedasticity.

561   *Behavior Research Methods*, *45*(4), 955–967.

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

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

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

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

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

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

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

569 Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with*

570   *applications* (7th ed.). Belmont, USA: Brooks/Cole.

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

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

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

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

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

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

Appendix

577 **The bias of Cohen's *d* is twice as large as the bias of Shieh's *d* when population**

578 **variances and sample sizes are equal across groups: mathematical**

579 **demonstration.**

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

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

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

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

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

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

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

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

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

583 below in equations 19 and 20:

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

584

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

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

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

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

587

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

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

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

589

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

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

591   Shieh's d.

[592] **The variance of Cohen's d is four times larger than the bias of Shieh's d when**

[593] **population variances and sample sizes are equal across groups: mathematical**

[594] **demonstration.**

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

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

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

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

[597]   We have previously shown in equations 21 and 22 that degrees of freedom associated

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

[599] As a consequence, the first term of the addition in equation 25 is 4 times larger than the

[600] first term of the addition in equation 26:

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

[601]

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

[602] We have also previously shown in equations 19 and 20 that δ_{Cohen} is twice larger than

[603] δ_{Shieh} when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$ and, therefore, δ^2_{Cohen} is four times larger than

[604] δ^2_{Shieh}. As a consequence, the second term of the addition in equation 25 is also 4 times

[605] larger than the second term of the addition in equation 26. Because both terms of the

[606] addition in equation 25 are four times larger than those in equation 26, we can conclude

[607] that the variance of Cohen's d is four times larger than the variance of Shieh's d.