Journal of Modern Applied Statistical Methods

Volume 22 | Issue 1 Article 3

Which Effect Size Calculation is the Best to Estimate the Population Effect Size in the Welch T Test?

Yi Zhou
Guangxi University, Nanning, China,
1045180219@qq.com

Xinyue Ren
Researcher, Ohio University, United States,
xr868414@ohio.edu

Gordon Brooks
Faculty at Ohio University, United States,
brooksg@ohio.edu

Recommended Citation

Yi Zhou, Xinyue Ren, Gordon Brooks (2023). Which Effect Size Calculation is the Best to Estimate the Population Effect Size in the Welch T Test?. Journal of Modern Applied Statistical Methods, 22(1), https://doi.org/10.56801/Jmasm.V22.i1.3

Journal of Modern Applied Statistical Methods
2023, Vol. 22, No. 1,
Doi: 10.56801/Jmasm.V22.i1.3

ISSN 1538 – 9472

Which Effect Size Calculation is the Best to Estimate the Population Effect Size in the Welch T Test?

Yi Zhou
Guangxi University, Nanning, China

Xinyue Ren
Researcher, Ohio University, United States

Gordon Brooks
Faculty at Ohio University, United States

The purpose of this study is to use Monte Carlo method to detect the most precise and least biased effect sizes calculations in a variety of conditions. The results show that there is no big difference to obtain effect sizes of using mean difference or trimmed mean difference as denominator. Cohen's d_A proves to be the less unbiased but more precise across all the conditions in Welch *t* test. It is worthwhile to notice that Hedges' *g* remains the same as Cohen's d_P across all the conditions of Welch *t* test. When group sample sizes are equal, no matter which population effect size formula are applied, Cohen's d_A, Cohen's d_P, and Hedges' *g* are the same estimates given the bias statistics.

Keywords: effect size, Monte Carlo, *R* software, Welch *t* test.

1. Introduction

In the field of social and behavioral sciences, effect size has been increasingly addressed as one of the important indicators when reporting and understanding the statistical results (APA, 2016; Campell, 1982; Cohen, 1994; Howell, 2010; Lakens, 2013; Maher, Markey, & Ebert-May, 2013). What is an effect size? Cohen (1988) defined effect size in his book, *Statistical Power Analysis for the Behavioral Sciences*, as follows:

Without intending any necessary implication of causality, it is convenient to use the phrase 'effect size' to mean '*the degree* to which the phenomenon is present in the population', or 'the degree to which the null hypothesis is false.' By the above route it can now readily be clear that when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific non-zero value in the population.* The larger this value, the greater degree to which the phenomenon under study is manifested (pp. 9-10, emphasis in original).

According to Ellis (2010), "an effect size refers to the magnitude of the result as it occurs, or would be found, in the population" (p. 4). The estimation of an effect size is pivotal to the interpretation of a study's results. Many researchers also state that

statistical significance testing, the foundation of quantitative research, may not be comprehensive or accurate enough to indicate the degree of the research results, the strength of the relationships among variables, and the implementation of the research findings in real-world situations (Howell, 2010; Maher et al., 2013; Vacha-Haase & Thompson, 2004). Because of the functionalities of the effect size in quantifying the magnitude of the intervention effect and the relationships among variables in a standardized metric, many experts agree the necessity of calculating and reporting effect sizes to supplement the limitations of significance testing and imply the practical meaningfulness of the research findings (Lakens, 2013; Maher et al., 2013; Rosnow & Rosenthal, 2003; Skidmore & Thompson, 2013; Vacha-Haase & Thompson, 2004; Walker, 2015).

Even though it has been a long time that the effect size has been viewed as an important part of research analyses, there are controversies about when reporting effect sizes among scholars (Leach & Henson, 2014; Roberts & Henson, 2002). The inclusion of effect size indices is still limited in many research journals, and the robust measures are still unsettled (Maher et al., 2013; Peng & Chen, 2014; Skidmore & Thompson, 2013). Due to the necessity of including effect sizes to better understand the research findings, there may be a priority to discuss which calculation should be appropriately selected to measure effect sizes in specific situations before reporting them.

Within the existing literature body pertaining to effect sizes, a variety of studies have been conducted to contribute to the detailed research of effect sizes (Cohen, 1988; Glass, 1976; Hedges, 1981). Cohen (1988) first proposed the effect size index: *d* to estimate the population effect size in the independent *t* test (see Equation 1). In 1976, Glass suggested the standard deviation (SD) in the control or reference group should be used to calculate the effect sizes for unbiased result. Based on what Glass had studied, Hedges (1981) proposed a corrected adjustment to calculate SD for the effect sizes. These studies have laid a solid foundation for the later researchers who seek to find out the detailed studies on effect sizes.

In the existing literature pertaining to meta-analysis, there have been two indices: *d*-index or *g*-index referring to the standardized mean difference measures. The *d*-index is typically related to *t*-tests or *F*-tests based on a comparison of two groups or experimental conditions (Cooper, 2017). The d-index is calculated based on small samples that might overestimate the magnitude of an effect in the population. Hedges (1980) suggested that *g*-index should be used when samples are smaller than 20. In line with the different techniques used in meta-analysis, there is no consensus that which estimation is the best precise to obtain effect size in *t*-tests.

Germane to the “forefathers” hard work, more current studies have addressed the significance of calculating and reporting effect sizes (Cohen, 1992; Cumming, 2012; Ellis, 2010; Kirk, 2013; Lakens, 2013; Maher et al., 2013; Nakagawa & Cuthill, 2007; Pek & Flora, 2018; Gorard, 2015). For example, some studies focused on the estimations of confidence intervals or computing confidence limits on effect sizes (Cumming & Finch, 2001; Fan & Thompson, 2001; Goulet-Pelletier & Cousineau, 2018; Howell, 2010; Wilcox, 2019). Several studies have listed some methods of

calculating effect sizes by using different SD as divisor (Gibbons et al., 1993; Lalongo, 2016). Several other studies were conducted to address the alternatives of Cohen's *d* (Cahan & Gamliel, 2011; Peng & Chen, 2014). Some researchers also discussed how to accurately interpret effect sizes when applying various measures (Lakens, 2013; Maher et al., 2013; Vacha-Haase & Thompson, 2004).

Independent *t* tests have been widely used to perform mean difference tests for various conditions, such as Student *t* test, Welch *t* test, and Yuen *t* test. However, when the assumptions of normality and homogeneity of equal variances are violated, many scholars believed that Welch *t* test is better to control Type I error (Delacre et al., 2017; Hayes & Cai, 2007; Zimmerman, 2004). Many researchers have discussed how to measure effect sizes in various situations, but few literatures could be found that clearly specified the best mean difference (MD) and SD to estimate the effect size when performing the Welch *t* test. The question about which SD and MD should be used to estimate effect size in the Welch *t* test has remained unsettled explicitly. Moreover, the fact that which effect size calculation is the best appropriate remains unsettled in the literature. The purpose of this study is to explore the more precise and least biased SD and MD combination to calculate effect sizes in Welch *t* test by using Monte Carlo method.

Guided by the research purpose, the research question is as follows:

Which effect size calculation is the best estimate of the population effect size in Welch *t* test?

2. Literature Review

In line with the effect sizes family, Cohen's *d*, η^2, and R^2 are the most frequently adopted in the social and behavioral sciences (Cooper, 2017; Ellis, 2010; Lakens, 2013; Peng & Chen, 2014). Cohen's *d* family effect sizes have been addressed for decades. Serval early studies have listed top three effect sizes calculations in d family (Kirk, 1996; Schmidt, 1996). Three methods to obtain effect sizes are the most discussed in the literature: Cohen's *d* (see Equation 1), Glass' Δ (see Equation 2), and Hedges' *g* (see Equation 3). However, few studies have been done to investigate which SD should be used as the most precise denominator to estimate the population effect size.

$$Glass'\Delta = \frac{M_1 - M_2}{S_{control}} \quad (1)$$

$$Cohen's\ d = \frac{M_1 - M_2}{S_{pooled}} \quad (2)$$

$$Hedges'g = \frac{M_1 - M_2}{S_{corrected}} \quad (3)$$

The current research methods, for instance, Monte Carlo simulation methods have been utilized in a number of studies to elucidate the best SD to estimate population effect sizes. In Goulet-Pelletier and Cousineau (2018), Monte Carlo simulations have

been used to detect the population effect sizes by using weighted pooled SD, unweighted regular (unweighted) SD, Hedges' g, and Glass's Δ based on the t distribution and the standard error (SE) estimations. Their results were revealed that in the scenario of two independent groups, Hedges' g is the best estimator overall, and the pooled SD is the best divider.

However, Cohen (1988) suggested that in the independent t test, for unequal variance and unequal sample size, the denominator to estimate the population effect size "requires the root mean square of σ_A and σ_B, that is, the square root of the mean of the two variances" (p. 44, see Equation 4). At the same time, Cohen also suggested another way of calculating pooled SD (see Equation 5) for the condition that "where M_A and M_B are the two sample means, and the usual pooled within sample estimate of the population standard deviation" (p. 67). Building on the prior works, Hedges and Olkin (1985) postulated an unbiased version of d calculation by using the Equation 6 to estimate the population effect size (p. 81).

In addition, for more recent studies, a variety of methods of calculating effect sizes by using different MDs and SDs have been elucidated (Vacha-Haase & Thompson, 2004; Gorard, 2015). More studies used Monte Carlo method to estimate population effect sizes in different analytical tests (Skidmore & Thompson, 2013; Rosnow & Rosenthal, 2003; Thompson, 2002).

$$S_A = \sqrt{\frac{(S_1^2 + S_2^2)}{2}} \tag{4}$$

$$S_P = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} \tag{5}$$

$$g \cong d\left(1 - \frac{3}{4(n_1 + n_2 - 9)}\right) \tag{6}$$

Researchers also summarized numerous common measures of effect size under two categories, including comparing differences among groups (such as d family, indicating standardized mean differences) and identifying the strength of the relationships (such as r family, describing the proportion of variance) (Lakens, 2013; Maher et al., 2013). For example, while comparing the differences between two groups by using a χ^2 test of homogeneity, the odds ratio can be used as a measure of effect size. Cohen's f can be utilized to report effect sizes when performing an analysis of variance test. While identifying the magnitude of the relationships, R^2 can be used to indicate the effect size in a multiple regression analysis. Lakens (2013) further discussed the usefulness of effect sizes in performing meta-analyses to compare results across studies and power analyses to determine reasonable sample sizes for the future studies. He mainly focused on how to use Cohen's d and eta squared (η^2) to calculate effect sizes for t-tests and ANOVAs under two situations: the differences between within-subjects and between-subjects designs. Rosnow and Rosenthal (2003) also examined how three effect size estimators, correlation (r related indices), difference (such as, Hedges' g or Cohen's d), and the odds ratio, were utilized in the field of experimental psychology. They concluded different

situations where researchers can intentionally select appropriate indices, including between and within subjects designs and effect sizes comparison.

Moreover, because of the weaknesses of three most commonly used indices, R^2, Cohen's *d*, and η^2 in generating robust and stable statistical analyses (Skidmore &Thompson, 2013), Peng and Chen (2014) mainly introduced multiple alternatives to Cohen's *d* within six categories. For example, while assuming normality and equal variance, Cohen's *d*, Glass' *g*, and Hedges' g_u can be used as standardized estimators; while normality only, Cohen's *d** and Keselman and colleagues' d_j may be used as standardized estimators. In addition to understand how to estimate effect sizes, it is important to appropriately report and interpret the results to avoid ambiguity. Vacha-Haase and Thompson (2004) further explained three rules of reporting effect sizes, including explicitly stating how effect sizes are measured; interpreting effect sizes while contextualize the research designs and limitations; and reporting confidence intervals and other related statistical results. They also provided suggestions about how to effectively interpret effect sizes. For instance, instead of solely relying on the fixed benchmarks proposed by Cohen, it is crucial to contextualize the study and compare the effect sizes across studies.

Skidmore and Thompson (2013) also discussed how to use three ANOVA effect sizes, η^2, ε^2, and ω^2, when assumptions are violated. In the Monte Carlo simulation study, through referring to previous studies, the researchers set the condition where all distributions having equal means (M = 100), Cohen's *d* for four nonnull conditions being 0.2, 0.5, 0.8, and 1.0, two-level, three-level, and four-level one-way being included, *k* being equal to 2, 3, and 4, and sample sizes being 24 and 48. During the replications, 5,000 samples were used to increase statistical accuracy. Three indices, Type 1 error rates, and power were calculated for total 4,050,000 samples. As a result, the study showed that η^2 is not an appropriate estimator for ANOVA because of large sampling error bias.

Many studies have shown the necessity of calculating and reporting effect sizes to increase the trustworthiness and understandability of the research results and indicate practical meaningfulness (Howell, 2010; Lakens, 2013; Maher, et al., 2013). However, the effect size is not a panacea and has its limitations as well. For instance, effect sizes are contextualized (Lakens, 2013). When they are affected by sampling strategies, the judgment of the practical significance may be biased. Therefore, while interpreting them, it is recommended to compare them across studies or by the common language effect size (Lakens, 2013; Vacha-Haase & Thompson, 2004). Moreover, the effect size alone is not sufficient to generate a comprehensive picture. For instance, observed effect sizes may overestimate the potential effect sizes in the population. In addition to the significance test, an appropriate confidence interval should also be measured when reporting the effect size (Howell, 2010; Lakens, 2013; Maher et al., 2013; Vacha-Haase & Thompson, 2004).

3. Methods

By application of the Monte Carlo method, the researchers set up a variety of situations to calculate effect sizes in Welch *t* test by using *R* statistical software. *R*, an open-source software, is one of the popular programming languages in statistical analysis. It contains favorable functions and has been widely used in various industries to perform statistical computing and generate data visualizations. In this study, there are 5 different group means for group 1 and group 2, in which group 1 is the control group while group 2 is the experimental group. Under this condition, the mean (M) and SD for group 1 are kept constant as 0 and 1, but the Ms for group 2 varies from 0, 0.2, 0.5, to 0.8, and SDs are from 1, 2, 3, to 4. In response to group sample sizes, group 2 is listed as 20, 40, 60, 80, 100, while group 1 sample sizes will be calculated by using maximum sample size of 120 for each condition subtracting group 2 sample sizes.

Moreover, apart from the MDs obtained from group 2 and group 1, 95% trimmed mean difference (TMD) between group 2 and group 1 are also included in this study. For the SDs, Hedges' *g*, Glass's Δ, pooled SD (see Equation 5), the square root of the mean of the variances between two groups (see Equation 4), and another three methods of calculating the estimated effect sizes by using the *R* package entitled Efficient Effect Size Computation, which is also named as "effsize" in short. There are three main functions in this package: cliff.delta, cohen.d., and VD.A.. In this Monte Carlo study, the function of "cohen.d" was applied to obtain the following estimates of the population effect sizes: (1) Cohen's d_P is calculated by using the MD as numerator and the Equation 5 as denominator; (2) Cohen's d_A is obtained by using the MD as numerator and the Equation 4 as denominator; (3) Glass' Δ: pooled SD = False (Using controlled group SD), hedges correction = False (Not using Hedges' *g* correction, the Equation 6); (4) Hedges' *g*: pooled SD = True (Not using controlled group SD), hedges correction = True (Using Hedges' *g* correction). (5) ES1: pooled SD = False (Using controlled group SD), hedges correction = True (Using Hedges' *g* correction, the Equation 6); (6) Cohen's d_{TA}: TMD is used as numerator and the Equation 4 as divider; (7) Cohen's d_{TP}: TMD is used as numerator and the Equation 5 as divider; (8) ES2: pooled SD = True (Using pooled SD), hedges correction = False (Not using Hedges' g correction).

For the names of effect sizes used within this study, Cohen's d_P is denoted as using the MD of two groups over Equation 5 as the denominator; Cohen's d_A is to be calculated by using the MD of two groups over Equation 4 as the divider; Cohen's d_{TA} is calculated by using TMD of two groups over Equation 4 as denominator; Cohen's d_{TP} is to use TMD of two groups over Equation 5 as divider; Hedges' *g*, Glass' Δ, ES1, and ES2 are also calculated by using R "effsize" package.

4. Results

In the field of Monte Carlo study, it is important for researchers to decide a reasonable number of replications needed for specific purpose. Based on the

previous Monte Carlo studies, Brooks et al. (2017) recommended to run 65,686 samples for 99% of confidence interval. They also mentioned that "greater precision" requires "larger number of replications" (p. 44). Therefore, in order to promote the accuracy of study results, 100,000 simulations were performed. After running 100,000 simulations, Table 1 displays the juxtaposition of the means for two population effect sizes statistics across all the conditions. It is self-evident that the popAES are kept consistent within each MD of relevant conditions while the popPES statistics are variant under each condition. That indicates that using Equation 4 as denominator yields least bias to estimate population effect sizes. In the following calculations of the bias and the RMSE statistics, both popAES and popPES will be applied as the population effect sizes statistics.

Based on the results of running 100,000 simulations, the estimates by using regular MDs or TMDs as nominators to calculate the effect sizes proved to be minor. Since it is more convenient for the researchers to obtain regular MDs rather than TMDs, we decided to remove the calculations of Cohen's $d3$ and Cohen's $d4$. In the following comparisons, four calculations will be applied within this study: Cohen's $d1$, Cohen's $d2$, Hedges' g, and Glass' Δ.

Table 1. The Summary of the Means of Population Effect Sizes under Different Conditions

N1	N2	S1	S2	M1	M2	popAES	popPES
100	20	1	1	0	0.2	0.2	0.2
		1	2			0.1265	0.1605
		1	3			0.0894	0.1322
		1	4			0.0686	0.1082
80	40	1	1	0	0.2	0.2	0.2
		1	2			0.1265	0.1417
		1	3			0.0894	0.1048
		1	4			0.0686	0.0819
60	60	1	1	0	0.2	0.2	0.2
		1	2			0.1265	0.1265
		1	3			0.0894	0.0894
		1	4			0.0686	0.0686
40	80	1	1	0	0.2	0.2	0.2
		1	2			0.1265	0.1153
		1	3			0.0894	0.0793
		1	4			0.0686	0.0602
20	100	1	1	0	0.2	0.2	0.2
		1	2			0.1265	0.1066
		1	3			0.0894	0.0720
		1	4			0.0686	0.0543

100	20	1	1	0	0.5	0.5	0.5
		1	2			0.3162	0.4106
		1	3			0.2236	0.3306
		1	4			0.1715	0.2706
80	40	1	1	0	0.5	0.5	0.5
		1	2			0.3162	0.3543
		1	3			0.2236	0.2619
		1	4			0.1715	0.2048
60	60	1	1	0	0.5	0.5	0.5
		1	2			0.3126	0.3126
		1	3			0.2236	0.2236
		1	4			0.1715	0.1715
40	80	1	1	0	0.5	0.5	0.5
		1	2			0.3126	0.2883
		1	3			0.2236	0.1983
		1	4			0.1715	0.1505
20	100	1	1	0	0.5	0.5	0.5
		1	2			0.3126	0.2666
		1	3			0.2236	0.1800
		1	4			0.1715	0.1357
100	20	1	1	0	0.8	0.8	0.8
		1	2			0.5060	0.6569
		1	3			0.3578	0.5289
		1	4			0.2744	0.4329
80	40	1	1	0	0.8	0.8	0.8
		1	2			0.5060	0.5669
		1	3			0.3578	0.4191
		1	4			0.2744	0.3278
60	60	1	1	0	0.8	0.8	0.8
		1	2			0.5060	0.5060
		1	3			0.3578	0.3578
		1	4			0.2744	0.2744
40	80	1	1	0	0.8	0.8	0.8
		1	2			0.5060	0.4612
		1	3			0.3578	0.3173
		1	4			0.2744	0.2407
20	100	1	1	0	0.8	0.8	0.8
		1	2			0.5060	0.4266
		1	3			0.3578	0.2881
		1	4			0.2744	0.2171

Note. popAES is calculated by using regular MD of two groups over Equation 4; while popPES is obtained by using regular MD of two groups over Equation 5.

Albeit eight effect size estimates included in this Monte Carlo study, we found that Cohen's d_P and ES2 remained exactly the same in all the conditions; Glass' Δ and ES1 were extremely close; the pairs of Cohens' d_P and Cohen's d_{TP}, Cohen's d_A and Cohen's d_{TA} shown no big difference respectively in this context. Regarding the difficulty of using TMDs in reality and the similar results shown as indicated previously, the commonly used effect size calculations will be postulated in the following discussion: Cohen's d_P, Cohen's d_A, Glass' Δ, and Hedges' g. Germane to the complex and different conditions in this Monte Carlo study, the following four scenarios will be discussed pertaining to their corresponding results: Scenario 1: When the MD is 0; Scenario 2: When the MD is 0.2; Scenario 3: When the MD is 0.5; Scenario 4: When the MD is 0.8.

4.1 Scenario 1: When the MD is 0

When the MD is 0, it implies that there is no statistically significant difference when the null hypothesis is true. Without statistical significance, some researchers argued that there was no need to report effect sizes (Sawilowsky, 2003; Sawilowsky & Yoon, 2002). However, a plethora of scholars have recommended that effect sizes should be reported and interpreted in the absence of statistical significance according to specific contexts (Cahan, 2000; Carver, 1993; Cumming & Finch, 2001; Harlow et al., 1997; Henson & Smith, 2000; Roberts & Henson, 2003). According to other researchers, Schmidt (1996) even stated that effect size estimates and confidence intervals are in preference to significant value (e.g. p value). In such situation, we suggest that researchers need to know which effect size calculation is the most unbiased and precise.

Table 2. The Summary of Bias and RMSE Statistics in Scenario 1 When MD is Zero (MD= 0

Statistics (popAES)	N1	N2	S1	S2	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ	Statistics (popPES)	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ
Bias	100	20	1	1	0.0008	0.0009	0.0008	0.0011	Bias	0.0008	0.0009	0.0008	0.0011
			1	2	-0.0013	-0.0010	-0.0013	-0.0007		-0.0013	-0.0010	-0.0013	-0.0007
			1	3	-0.0025	-0.0018	-0.0025	-0.0013		-0.0025	-0.0018	-0.0025	-0.0013
			1	4	-0.0004	-0.0002	-0.0004	-0.0002		-0.0004	-0.0002	-0.0004	-0.0002
	80	40	1	1	-0.0005	-0.0005	-0.0005	-0.0006		-0.0005	-0.0005	-0.0005	-0.0006
			1	2	-0.0001	0.00001	-0.0001	0.00001		-0.0001	0.00001	-0.0001	0.00001
			1	3	-0.0003	-0.0003	-0.0003	-0.0002		-0.0003	-0.0003	-0.0003	-0.0002
			1	4	-0.0006	-0.0012	-0.0006	-0.0004		-0.0006	-0.0005	-0.0006	-0.0004
	60	60	1	1	-0.0002	-0.0002	-0.0002	-0.0001		-0.0002	-0.0002	-0.0002	-0.0001
			1	2	0.0004	0.0004	0.0004	0.0003		0.0004	0.0004	0.0004	0.0003
			1	3	-0.0005	-0.0005	-0.0005	-0.0004		-0.0005	-0.0005	-0.0005	-0.0004
			1	4	-0.0002	-0.0002	-0.0002	-0.0001		-0.0002	-0.0002	-0.0002	-0.0001
	40	80	1	1	-0.0004	-0.0004	-0.0004	-0.0004		-0.0004	-0.0004	-0.0004	-0.0004
			1	2	-0.0003	-0.0003	-0.0003	-0.0002		-0.0003	-0.0003	-0.0003	-0.0002

			1	3	-0.0006	-0.0007	-0.0006	-0.0005		-0.0006	-0.0007	-0.0006	-0.0005
			1	4	0.0001	0.0001	0.0001	0.0001		0.0001	0.0001	0.0001	0.0001
	20	100	1	1	-0.0002	-0.0002	-0.0002	-0.0002		-0.0002	-0.0002	-0.0002	-0.0002
			1	2	-0.0001	-0.0002	-0.0001	-0.0001		-0.0001	-0.0002	-0.0001	-0.0001
			1	3	0.00001	-0.0001	0.00001	0.00001		0.00001	-0.0001	0.00001	0.00001
			1	4	-0.0001	-0.0002	-0.0001	-0.0001		-0.0001	-0.0002	-0.0001	-0.0001
RMSE	100	20	1	1	0.2476	0.2483	0.2451	0.2583	RMSE	0.2467	0.2483	0.2451	0.2584
			1	2	0.3816	0.3003	0.3792	0.2426		0.3816	0.3003	0.3792	0.2426
			1	3	0.4565	0.3158	0.4536	0.2383		0.4565	0.3158	0.4536	0.2383
			1	4	0.5019	0.3239	0.4987	0.2378		0.5019	0.3239	0.4987	0.2378
	80	40	1	1	0.196	0.1962	0.1947	0.1994		0.1960	0.1962	0.1947	0.1994
			1	2	0.2410	0.2160	0.2394	0.1724		0.2410	0.2160	0.2394	0.1724
			1	3	0.2593	0.2222	0.2577	0.1664		0.2593	0.2222	0.2577	0.1664
			1	4	0.2689	0.2257	0.2672	0.1651		0.2689	0.2257	0.2672	0.1651
	60	60	1	1	0.1842	0.1842	0.1831	0.1858		0.1842	0.1842	0.1831	0.1858
			1	2	0.1839	0.1839	0.1827	0.1462		0.1839	0.1839	0.1827	0.1462
			1	3	0.1849	0.1849	0.1837	0.1383		0.1849	0.1849	0.1837	0.1383
			1	4	0.1851	0.1851	0.1839	0.1352		0.1851	0.1851	0.1839	0.1352
	40	80	1	1	0.1955	0.1957	0.1943	0.1963		0.1955	0.1957	0.1943	0.1963
			1	2	0.1589	0.1740	0.1578	0.1381		0.1589	0.1740	0.1578	0.1381
			1	3	0.1491	0.1679	0.1481	0.1254		0.1491	0.1679	0.1481	0.1254
			1	4	0.1446	0.1648	0.1437	0.1203		0.1446	0.1648	0.1437	0.1203
	20	100	1	1	0.2468	0.2486	0.2452	0.2472		0.2468	0.2486	0.2452	0.2472
			1	2	0.1611	0.1909	0.1600	0.1511		0.1611	0.1909	0.16	0.1511
			1	3	0.1360	0.1688	0.1352	0.126		0.1360	0.1688	0.1352	0.1260
			1	4	0.1258	0.1589	0.1250	0.1160		0.1258	0.1589	0.1250	0.1160

As shown in Table 2 and Figure 1, the results of using PopAES or PopPES are postulated in an agreement across all the conditions when MD is 0. Given that no MD is detected, Glass' Δ is the best estimation in terms of the smallest values of the RMSE statistics. However, in line with the bias statistics, there is no big difference no matter which calculation is used. In the following scenarios that MDs vary from 0.2 to 0.8, the same patterns have been showcased pertinent to the bias and RMSE statistics. Therefore, the patterns were only displayed when MD is 0.5 respectively in A, B, C, and D patterns in Figure 2.

4.2 Scenario 2: When the MD is 0.2

In this context, Table 3 displays that some differences have been shown pertinent to which population effect size calculation to be used. When used popAES as population effect size, Cohen's d_A proves to be the most unbiased calculation to estimate the population effect size according to the bias statistics; however, taking into consideration of the RMSE statistics, Glass' Δ is marked as the most precise estimate compared with the other three calculations. For the impact of group sizes,

there is no difference to use Cohen's d_A, Cohen's d_P or Hedges' g when the group sizes are equal. The diagnostic statistics tend to be liberal each other when group sizes are dramatically unequal (largest /smallest > 2) or large variances are associated with the small group sizes. Across all the conditions, it is noteworthy that the results of Cohen's d_P and Hedges' g maintain constant in this scenario.

Table 3. The Summary of the Bias and the RMSE Statistics in Scenario 2 When the MD is 0.2

Statistics (popAES)	N1	N2	S1	S2	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ	Statistics (popPES)	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ
Bias	100	20	1	1	0.0013	0.0025	0.0001	0.0087	Bias	0.0013	0.0025	0.0001	0.0087
			1	2	0.0403	0.0041	0.0392	-0.0216		0.0025	-0.0337	0.0015	-0.0594
			1	3	0.0431	0.0017	0.0423	-0.0210		0.0004	-0.0411	-0.0005	-0.0637
			1	4	0.0429	0.0030	0.0422	-0.0161		0.0033	-0.0366	0.0025	-0.0557
	80	40	1	1	0.0005	0.0006	-0.0008	0.0032		0.0005	0.0006	-0.0008	0.0032
			1	2	0.0167	0.0017	0.0158	-0.0244		0.0015	-0.0135	0.0006	-0.0397
			1	3	0.0178	0.0024	0.0171	-0.0208		0.0025	-0.013	0.0018	-0.0361
			1	4	0.0135	0.0002	0.0129	-0.0183		0.0001	-0.0131	-0.0004	-0.0317
	60	60	1	1	-0.0005	-0.0005	-0.0018	0.0007		-0.0005	-0.0005	-0.0018	0.0007
			1	2	0.0006	0.0006	-0.0002	-0.0256		0.0006	0.0006	-0.0002	-0.0256
			1	3	0.001	0.001	0.0005	-0.0218		0.001	0.001	0.0005	-0.0218
			1	4	0.0007	0.0007	0.0003	-0.0180		0.0007	0.0007	0.0003	-0.018
	40	80	1	1	0.0021	0.0023	0.0008	0.0028		0.0021	0.0023	0.0008	0.0028
			1	2	-0.0103	0.0008	-0.0111	-0.0256		0.0008	0.0120	0.0001	-0.0144
			1	3	-0.0093	0.0008	-0.0098	-0.0220		0.0008	0.0109	0.0003	-0.0119
			1	4	-0.0075	0.0011	-0.0079	-0.0178		0.0010	0.0095	0.0006	-0.0094
	20	100	1	1	0.0004	0.0015	-0.0008	0.0007		0.0004	0.0015	-0.0008	0.0007
			1	2	-0.0185	0.0016	-0.0191	-0.0251		0.0014	0.0214	0.0007	-0.0053
			1	3	-0.0163	0.0013	-0.0168	-0.0218		0.0011	0.0187	0.0006	-0.0043
			1	4	-0.0142	0.0002	-0.0145	-0.0184		0.0002	0.0145	-0.0002	-0.0041
RMSE	100	20	1	1	0.247	0.249	0.2454	0.2612	RMSE	0.2470	0.2490	0.2454	0.2612
			1	2	0.3828	0.3	0.3803	0.2436		0.3807	0.3018	0.3783	0.2498
			1	3	0.4589	0.3161	0.4559	0.2395		0.4569	0.3188	0.4539	0.2470
			1	4	0.5065	0.3259	0.5033	0.2399		0.5047	0.3279	0.5015	0.2457
	80	40	1	1	0.1955	0.1958	0.1942	0.2000		0.1955	0.1958	0.1942	0.2000
			1	2	0.2415	0.2160	0.2399	0.1742		0.2409	0.2164	0.2394	0.1770
			1	3	0.2613	0.2234	0.2596	0.1687		0.2607	0.2238	0.259	0.1713
			1	4	0.2690	0.2256	0.2673	0.1660		0.2687	0.226	0.267	0.168
	60	60	1	1	0.1849	0.1849	0.1837	0.1871		0.1849	0.1849	0.1837	0.1871
			1	2	0.1854	0.1854	0.1842	0.1497		0.1854	0.1854	0.1842	0.1497
			1	3	0.1850	0.1850	0.1838	0.1401		0.1850	0.1850	0.1838	0.1401
			1	4	0.1859	0.1859	0.1847	0.1370		0.1859	0.1859	0.1847	0.1370

	40	80	1	1	0.1959	0.1962	0.1946	0.1969		0.1959	0.1962	0.1946	0.1969
			1	2	0.1603	0.1753	0.1594	0.1415		0.1600	0.1757	0.1590	0.1399
			1	3	0.1487	0.1672	0.1478	0.1268		0.1484	0.1503	0.1475	0.1255
			1	4	0.1451	0.1650	0.1442	0.1218		0.1449	0.1653	0.1440	0.1208
	20	100	1	1	0.2479	0.2498	0.2463	0.2484		0.2479	0.2498	0.2463	0.2484
			1	2	0.1630	0.1920	0.1620	0.1540		0.1619	0.1931	0.1609	0.1521
			1	3	0.1370	0.1687	0.1362	0.1278		0.1360	0.1697	0.1351	0.1260
			1	4	0.1262	0.1584	0.1255	0.1170		0.1254	0.1591	0.1246	0.1157

4.3 Scenario 3: When the MD is 0.5

Regarding the previous condition, this condition is set when the MDs are 0.5 across all the conditions. As displayed in Table 4 and Figure 2, the Glass's Δ still remains as the most precise calculation given to the RMSE statistics since they are the smallest values observed no matter which population effect size is used. However, there is no agreement in the bias statistics when using different population effect sizes. For the popAES results, Cohen's d_P and Hedges' g agree each other across all the conditions in this scenario. However, Cohen's d_A proves to be the least biased calculation. For the popPES results, there are no differences between Cohen's d_P and Hedges' g calculations, which are also tested as the best estimates.

Table 4. The Summary of the Bias and the RMSE Statistics in Scenario 3 When the MD is 0.5

Statistics (popAES)	N1	N2	S1	S2	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ	Statistics (popPES)	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ
Bias	100	20	1	1	0.0044	0.0071	0.0012	0.0222	Bias	0.0044	0.0071	0.0012	0.0222
			1	2	0.0962	0.0065	0.0935	-0.0571		0.0018	-0.0879	-0.0008	-0.1514
			1	3	0.1141	0.0086	0.1119	-0.0490		0.0071	-0.0984	0.005	-0.1560
			1	4	0.1040	0.0054	0.1022	-0.0419		0.0049	-0.0937	0.0032	-0.1410
	80	40	1	1	0.0036	0.0039	0.0004	0.0101		0.0036	0.0039	0.0004	0.0101
			1	2	0.0407	0.0034	0.0409	0.0385		0.0027	-0.0347	0.0004	-0.0999
			1	3	0.0422	0.0039	0.0405	-0.0533		0.0039	-0.0344	0.0022	-0.0917
			1	4	0.0375	0.0038	0.0362	-0.0434		0.0042	-0.0295	0.0028	-0.0768
	60	60	1	1	0.0035	0.0035	0.0003	0.0069		0.0035	0.0035	0.0003	0.0069
			1	2	0.002	0.002	0.00001	-0.0635		0.0020	0.0020	0.00001	-0.0635
			1	3	0.0025	0.0025	0.0011	-0.0547		0.0025	0.0025	0.0011	-0.0547
			1	4	0.0021	0.0021	0.001	-0.0448		0.0021	0.0021	0.001	-0.0448
	40	80	1	1	0.0034	0.0037	0.0002	0.0050		0.0034	0.0032	0.0002	0.005
			1	2	-0.0256	0.0023	-0.0274	-0.0637		0.0024	0.0303	0.0005	-0.0357
			1	3	-0.0231	0.0023	-0.0244	-0.0550		0.0022	0.0275	0.0009	-0.0297
			1	4	-0.0190	0.0023	-0.0199	-0.0447		0.0021	0.0233	0.0011	-0.0237
	20	100	1	1	0.0035	0.0061	0.0003	0.0041		0.0035	0.0061	0.0003	0.0041
			1	2	-0.0478	0.0019	-0.0495	-0.0644		0.0018	0.0515	0.0001	-0.0148

			1	3	-0.0427	0.0009	-0.0439	-0.0561		0.0008	0.0444	-0.0003	-0.0126
			1	4	-0.0351	0.0008	-0.0360	-0.0458		0.0007	0.0366	-0.0002	-0.0100
RMSE	100	20	1	1	0.2491	0.2525	0.2475	0.2746	RMSE	0.2491	0.2525	0.2475	0.2746
			1	2	0.3954	0.3036	0.3924	0.2533		0.3835	0.3160	0.3811	0.2896
			1	3	0.4731	0.3185	0.4697	0.2456		0.4592	0.3332	0.4562	0.2868
			1	4	0.5129	0.3245	0.5094	0.242		0.5023	0.3377	0.4991	0.2769
	80	40	1	1	0.1984	0.1989	0.1971	0.2078		0.1984	0.1989	0.1971	0.2078
			1	2	0.2459	0.2179	0.2440	0.1854		0.2425	0.2206	0.2410	0.2013
			1	3	0.2653	0.2247	0.2634	0.1768		0.2620	0.2273	0.2603	0.1919
			1	4	0.2716	0.2259	0.2689	0.1709		0.2691	0.2278	0.2673	0.1822
	60	60	1	1	0.1869	0.1869	0.1857	0.1917		0.1869	0.1869	0.1857	0.1917
			1	2	0.1863	0.1863	0.1851	0.1618		0.1863	0.1863	0.1851	0.1618
			1	3	0.1856	0.1856	0.1844	0.1493		0.1856	0.1856	0.1844	0.1493
			1	4	0.1863	0.1863	0.1851	0.1433		0.1863	0.1863	0.1851	0.1433
	40	80	1	1	0.1988	0.1993	0.1975	0.2011		0.1988	0.1993	0.1975	0.2011
			1	2	0.1628	0.1760	0.1621	0.1539		0.1608	0.1786	0.1598	0.1446
			1	3	0.1517	0.1688	0.1510	0.1377		0.1500	0.1710	0.1490	0.1297
			1	4	0.1463	0.1652	0.1455	0.1287		0.1451	0.1669	0.1442	0.1229
	20	100	1	1	0.2493	0.2528	0.2477	0.2501		0.2493	0.2528	0.2477	0.2501
			1	2	0.1694	0.1925	0.1689	0.1656		0.1625	0.1992	0.1614	0.1533
			1	3	0.1433	0.1696	0.1428	0.1386		0.1368	0.1753	0.1359	0.1273
			1	4	0.1303	0.1585	0.1298	0.1244		0.1255	0.1627	0.1247	0.1161

The group sizes are still an impact to the estimation of the population effect sizes. When big variance relates to small group sizes (largest/smallest > 2), the four calculations based on two diagnostic statistics are showcased to be variant each other while they tend to be less biased and better precise when group sizes are equal or large variances are connected with large group sizes. Particularly, when group sizes are equal, three calculations are convergent each other: Cohen's d_P, Cohen's d_A, and Hedges' g.

4.4 Scenario 4: When the MD is 0.8

As shown in Table 5, note that the four calculations based on two diagnostics statistics prove to be more largely liberal compared with the two previous scenarios, the bias and the RMSE statistics are the larger compared with the previous two scenarios when large group sizes are to the large variances and the ratio between the largest group sizes over smallest group sizes are over 2. Three calculations are constant each other: Cohen's d_P, Cohen's d_A, and Hedges' g when the group sizes are equal.

Table 5. The Summary of the Bias and the RMSE Statistics in Scenario 3 When the MD is 0.8

Statistics (popAES)	N1	N2	S1	S2	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ	Statistics (popPES)	Cohen's *d*1	Cohen's *d*2	Hedges' *g*	Glass' Δ
Bias	100	20	1	1	0.0062	0.0104	0.0011	0.0346	Bias	0.0062	0.0104	0.0011	0.0346
			1	2	0.1575	0.0133	0.1533	-0.0889		0.0065	-0.1376	0.0023	-0.2399
			1	3	0.1806	0.0122	0.1772	-0.0797		0.0095	-0.1589	0.0061	-0.2508
			1	4	0.1674	0.0095	0.1646	-0.0664		0.0089	-0.149	0.0061	-0.2249
	80	40	1	1	0.0048	0.0054	-0.0004	0.0153		0.0048	0.0054	-0.0004	0.0153
			1	2	0.0661	0.0063	0.0625	-0.0982		0.0052	-0.0547	0.0016	-0.1591
			1	3	0.0652	0.0043	0.0626	-0.0869		0.0039	-0.057	0.0012	-0.1482
			1	4	0.0585	0.0048	0.0564	-0.0704		0.0051	-0.0486	0.003	-0.1238
	60	60	1	1	0.0056	0.0056	0.0004	0.0107		0.0056	0.0056	0.0004	0.0107
			1	2	0.0036	0.0036	0.0003	-0.1015		0.0036	0.0036	0.0003	-0.1015
			1	3	0.0042	0.0042	0.0019	-0.0873		0.0042	0.0042	0.0019	-0.0873
			1	4	0.0041	0.0041	0.0023	-0.0711		0.0041	0.0041	0.0023	-0.0711
	40	80	1	1	0.0052	0.0059	0.0001	0.0077		0.0052	0.0059	0.0001	0.0077
			1	2	-0.0419	0.0026	-0.0448	-0.1028		0.0028	0.0474	-0.0001	-0.0580
			1	3	-0.0388	0.0016	-0.0408	-0.0895		0.0016	0.0421	-0.0004	-0.0490
			1	4	-0.0309	0.0030	-0.0324	-0.0720		0.0028	0.0367	0.0012	-0.0383
	20	100	1	1	0.0048	0.0091	-0.0003	0.0058		0.0048	0.0091	-0.0003	0.0058
			1	2	-0.0766	0.0030	-0.0793	-0.1031		0.0028	0.0824	0.00001	-0.0237
			1	3	-0.0678	0.0021	-0.0696	-0.0893		0.0019	0.0718	0.0001	-0.0196
			1	4	-0.0554	0.0023	-0.0568	-0.0726		0.0019	0.0596	0.0005	-0.0152
RMSE	100	20	1	1	0.2526	0.2588	0.2509	0.2983	RMSE	0.2526	0.2588	0.2509	0.2983
			1	2	0.4146	0.3068	0.4107	0.2670		0.3835	0.336	0.3811	0.3477
			1	3	0.4955	0.3215	0.4915	0.2561		0.4615	0.3584	0.4585	0.3495
			1	4	0.5310	0.3266	0.5271	0.2491		0.5040	0.3589	0.5008	0.3290
	80	40	1	1	0.2030	0.2041	0.2017	0.2213		0.2030	0.2041	0.2017	0.2213
			1	2	0.2541	0.2213	0.2517	0.2040		0.2454	0.2279	0.2438	0.2394
			1	3	0.2711	0.2261	0.2688	0.1909		0.2632	0.2331	0.2615	0.2255
			1	4	0.2767	0.2273	0.2746	0.1806		0.2705	0.2323	0.2687	0.2074
	60	60	1	1	0.1917	0.1917	0.1904	0.2007		0.1917	0.1917	0.1904	0.2007
			1	2	0.1893	0.1893	0.1880	0.1828		0.1893	0.1893	0.1880	0.1828
			1	3	0.1871	0.1871	0.1859	0.1653		0.1871	0.1871	0.1859	0.1653
			1	4	0.1876	0.1876	0.1864	0.1545		0.1876	0.1876	0.1864	0.1545
	40	80	1	1	0.2025	0.2036	0.2011	0.2069		0.2025	0.2036	0.2011	0.2069
			1	2	0.1683	0.1781	0.1681	0.1756		0.163	0.1843	0.162	0.1538
			1	3	0.1562	0.1702	0.1558	0.1558		0.1513	0.1754	0.1503	0.1366
			1	4	0.1486	0.1655	0.1480	0.1407		0.1454	0.1695	0.1444	0.1268
	20	100	1	1	0.2520	0.2582	0.2504	0.2535		0.2520	0.2582	0.2504	0.2535
			1	2	0.1809	0.194	0.1812	0.1854		0.1639	0.2108	0.1629	0.1558
			1	3	0.1535	0.1708	0.1536	0.1557		0.1378	0.1853	0.1369	0.1291

			1	4	0.1382	0.1599	0.1381	0.1375		0.1267	0.1707	0.1259	0.1177

Given the RMSE statistics, the Glass' Δ is kept as the most precise calculation across all the conditions when the MD is 0.8. However, there is no agreement for the use of popAES or popPES pertaining to the four estimates. When used the popAES, Cohen's d_A is examined as the least biased calculation while Cohen's d_P and Hedges' g are regarded as the most unbiased ones when the popPES is applied.

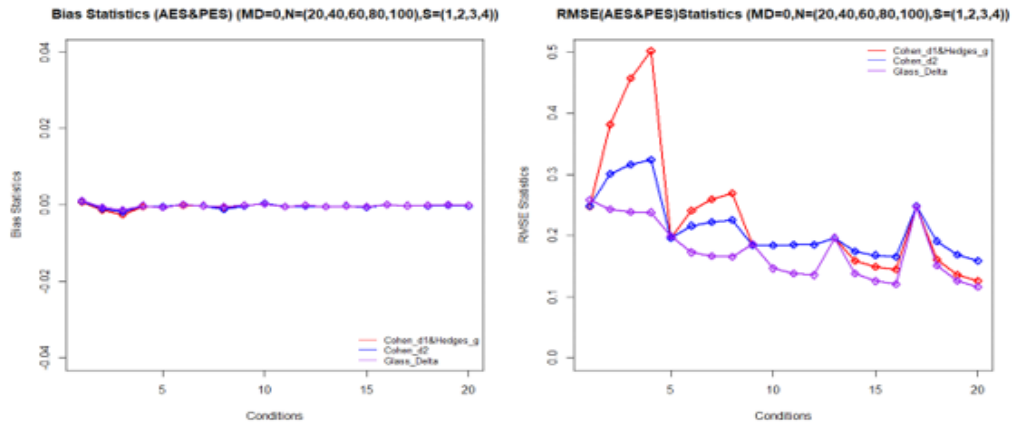
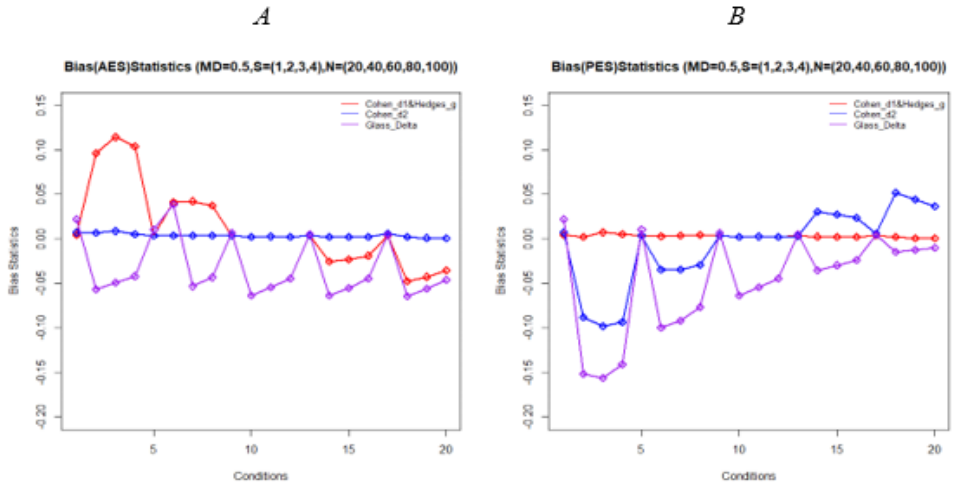


Figure 1. The Bias and the RMSE Statistics under Scenario 1 When the MD = 0



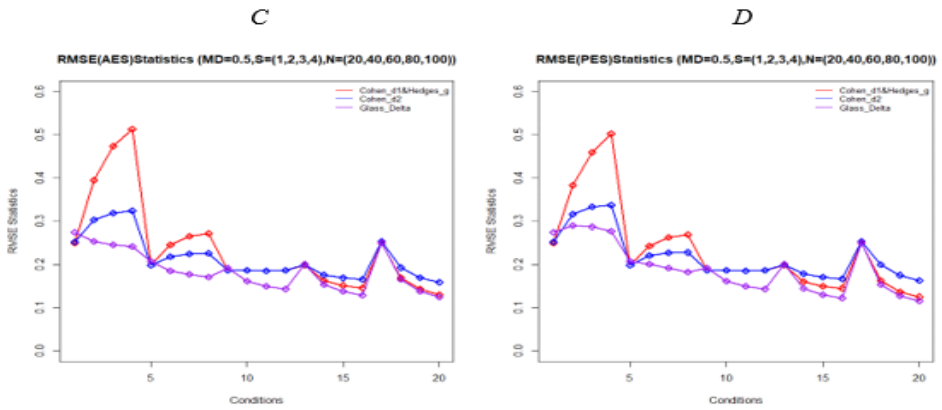


Figure 2 The Bias and RMSE Statistics under Scenario 3 When the MD is 0.5

Note. The following attached table provides the corresponding conditions in Figure 1 and Figure 2.

Condition	N1	N2	S1	S2
1	100	20	1	1
2			1	2
3			1	3
4			1	4
5	80	40	1	1
6			1	2
7			1	3
8			1	4
9	60	60	1	1
10			1	2
11			1	3
12			1	4
13	40	80	1	1
14			1	2
15			1	3
16			1	4
17	20	100	1	1
18			1	2
19			1	3
20			1	4

5. Discussion

5.1 The Divider to Calculate Population Effect Size

Based on the results from this Monte Carlo study, the myth of which SD is the better denominator to estimate the population effect size is solved. The SD calculated by Equation 4 (see Table 1) is proved to be the most consistent and least variant across all the combinations of relevant conditions in this study. This result is one of the contributions to the effect size literature.

5.2 The Best Estimate When MD = 0

Under the condition that there is no significant MD detected in the results, Glass' Δ is shown to be the most precise calculation across all the conditions to detect the standardized MD (see Figure 1). For the least unbiased calculation, these four methods have been proved no big difference under the examination of the bias statistics. The results are useful for the researchers who obtain a non-significant *p* values to detect the most precise or the least unbiased estimate of the population effect sizes.

5.3 The Best Calculation in the Welch *t* Test

Welch *t* test has been regarded as "unequal variances *t* test". When heterogeneity exists in all possible conditions, there has been proved to be disagreements among these four calculations when using different formula to calculate the population effect sizes, then obtaining the corresponding bias and the RMSE statistics across all the possible combinations.

5.3.1 When PopAES is applied:

Cohen's d_A is regarded as the best estimate when judged from the bias statistics (see Figure 2-A). Cohen's d_A is calculated by using regular MDs between two groups as nominator and Equation 4 as the divider. Given to the RMSE statistics (See Figure 1-C), Glass' Δ is the most precise calculation when using the controlled group SD as the denominator and the regular MD as the nominator. However, when small group sizes are associated with large variances or the ratio between two group sizes is over 2, the four calculations are marked dramatically different from each other; the variation tends to be less different each other when small group sizes are related to small variances. Across all the possible conditions in Welch *t* test, Hedges' *g* do not impact the estimate in that Hedges' *g* yields the same results as Cohen's d_P calculation which regular MD is divided by Equation 4. For the better precision estimate, Cohen's d_A generally performed a better job than Cohen's d_P regarding the RMSE statistics.

5.3.2 When PopPES is applied:

In this context of using popPES as the population effect size (see Figure 2-B), Cohen's d_P and Hedges' *g* are proved to be the best calculations across all the possible conditions of Welch *t* test judged from the bias statistics. For Cohen's d_P, it is obtained by the use of regular MD as nominator and the Equation 5 as

denominator. Given the RMSE statistics (see Figure 2-D), Glass' Δ is tested as the most precise estimate consistently. When the large group sizes are connected to small variances or the group sizes are substantively unequal (largest/smallest > 2), the bias and the RMSE statistics are inflated dramatically for these four calculations. However, when the small group sizes meet small variances, the differences among the four calculations become less liberal each other. When compared the better precision of Cohen's d_A and Cohen's d_P, Cohen's d_A proved to be more stable than Cohen's d_P across all the unconditional situations.

5.3.3 When Group Sizes are Equal:

Sample size is the most "sensitive" component in the field of social and behavioral sciences, which has been extensively discussed in the literature (Stevens, 1999). That is because even a slightest MD would be detected with large sample size, while a substantial MD would not be investigated with a relatively small sample size (Cortina & Nouri, 2000). Therefore, the magnitude of the effect under the equal sample sizes of both groups is one of the interesting conditions that the researchers want to detect. As stated previously, the combination of group sizes and the variances is influential to the two diagnostic statistics in this study. When the group sizes are equal, Cohen's d_A, Cohen's d_P, and Hedges' g are constant no matter which population effect size calculation is used, and no matter which diagnostic statistics is tested. However, in this condition, Cohen's d_A, Cohen's d_P, and Hedges' g are the better estimate compared with Glass' Δ based on the bias statistics; while Glass' Δ is the better precise calculation according to the RMSE statistics.

6. Conclusion

In this Monte Carlo study, the primary purpose is to detect which estimate is the least biased and the most precise one to predict population effect size in Welch t test. As stated earlier, Cohen's two equations to estimate population effect sizes proved to be a relatively variation between each other. Therefore, two different methods of calculating population effect sizes are taken into consideration: popAES (using Equation 4 as divider) and popPES (using Equation 5 as divider). In line with the diagnostic statistics, the bias and the RMSE statistics are applied to ascertain the least unbiased and the most precise estimate.

Based on the 100,000 simulations, the results postulated as follows:

6.1 When There was No MD

When there was no significant MDs detected (MD = 0), no matter which popAES or popPES was applied and which diagnostic statistics was used, Glass' Δ proved to be the most consistent in most of the conditions unless Cohen's d_A, Cohen's d_P, and Hedges' g are the better estimates than Glass' Δ when sample sizes are equal. In addition, pertinent to the bias statistics, four calculations turned out to no substantial difference across all the conditions, however, Glass' Δ was shown the most precise estimate given the RMSE statistics. What's more, in both diagnostic results, when the larger group sample is associated with smaller variances or the ratio of two group

sample is bigger than 2, Cohen's d_P and Hedges' g were seen as the most deviate estimates compared with the other two, that is, Cohen's d_A and Glass' Δ. While when the smaller group sample size is related to smaller variances, the four estimates proved to be more stable, flat, and not deviate from each other.

6.2 When the MDs are Not Zero

Across all the unconditional situations of Welch *t* test, the way of calculating the population effect size and which diagnostic statistics used do matter in this study. The following parts will be summarized based on the results of 100,000 simulations.

6.2.1 When Using PopAES

Cohen's d_A proved to be the most unbiased estimate in these conditions which was displayed as the most stable one across all the conditions. In the conditions that the larger group sample size match with smaller variances and the ratio between two group sizes is above 2, Cohen's d_P, Hedges' g, and Glass' Δ were detected substantially deviate from each other, typically Cohen's d_P and Hedges' g, which were showcased to be worse and worse with the larger MDs. However, when group sample sizes are equal, Cohen's d_A, Cohen's d_P, and Hedges' g were the better estimates than Glass' Δ. However, when smaller group sample sizes are connected with smaller variances, Cohen's d_P and Hedges' g became less deviate but Glass' Δ proved to be the most biased one. When group sample sizes are equal, except Glass' Δ, the other three calculations proved to be the least unbiased: Cohen's d_P, Cohen's d_A, and Hedges' g.

There was seen an inconsistency between the bias and the RMSE statistics. Regarding the RMSE statistics, across all the conditions, Glass' Δ proved to be the most precise calculation albeit the different group sample sizes and variances respectively. The same situation was detected when the smaller group sample size went with larger variances and the ratio between two group sample sizes is bigger than 2, four estimates dramatically deviated from each other, particularly, the least precise estimates were Cohen's d_P and Hedges' g calculations. The situations became better when larger group sample sizes went with larger variances. When the group sample sizes were equal, Cohen's d_P, Cohen's d_A, and Hedges' g were kept consistent.

6.2.2 When Using the PopPES

The previous two situations were based on the popAES. In this context, the Equation 5 was used as divider to calculate population effect size, which is denoted as popPES in this study.

Cohen's d_P and Hedges' g proved to be always the best estimate across all the conditions. When group sample sizes are equal, Cohen's d_A, Cohen's d_P, and Hedges' g proved the same estimate except Glass' Δ. In addition, Cohen's d_A and Glass' Δ were worse and worse, especially Glass' Δ, when the larger group sample sizes were connected with smaller variances, the ratio between two group sample

sizes is bigger than 2, and the MDs increased. The impact started to be less when smaller group sample sizes were related to smaller variances.

When the group sample sizes are equal, there was no difference no matter which estimates were used: Cohen's d_P, Cohen's d_A, and Hedges' g. However, Glass' Δ in this particular condition, was not a good estimate.

Given the RMSE statistics, Glass' Δ seemed to be the most precise estimate across all the conditions in this context. Cohen's d_P and Hedges' g turned out to be the least precise ones when the larger group sample sizes are associated with larger variances and the ratio between two group sample sizes is bigger than 2. The situation has been better when the smaller group sample sizes go with smaller variances. To notice that Cohen's d_P, Cohen's d_A, and Hedges' g proved to be the same estimates when the group sample sizes are equal.

When all the possible conditions relevant to examine the most precise calculation to estimate population effect size have been explored in this study, the long-standing myth is now tackled. As noted in the previous discussion, no matter which population effect size used, Cohen's d_A proved to be the less biased and more precise across all the unconditional Welch t test when compared with Cohen's d_P, Hedges' g, and Glass' Δ.

With this in mind, in real life context, the research phenomena prove to be far more complex than what has been observed in research. It is never clear for the researchers that the population effect sizes and variances are far less ascertained. Therefore, in Welch t test, Cohen's d_A (using Equation 4 as denominator) is proved to be a bit less biased but far more precise than the other three estimates: Cohen's d_P, Hedges' g, and Glass' Δ. What the researchers have found in this study will potentially be served as a strong reference when dealing with the estimate population effect size based on the samples. It is surprised to detect that Hedges' g did not impact at all in this study, which enjoyed the same function of using Cohen's d_P, that is, Equation 5 was applied. However, in the true experimental design, Cohen's d_P and Hedges' g may be better choices where both groups are sampled from the same population before conducing the random assignment, that is, the pooled SD shown as Equation 5 is a better denominator to estimate the population effect size in this context in which both samples come from the same population in the experimental conditions.

For the future study in this perspective, some special conditions which have been detected during this Monte Carlo study are two situations where MDs are equal, but the ratio of the sample sizes between two groups is 2 or 0.5 and the ratio of the variances between two groups is 1/3. In prementioned conditions, all the effect size calculations turned out to be the least biased and the most precise. However, due to the limited time and delimited boundary of the study, the researchers did not investigate in-depth to seek for the answers to these two special conditions. For getting easier access to effect size calculations, we provide you the following website (Ellis, 2009):

http://www.polyu.edu.hk./mm/effectsizefaqs/calculator/claculator.html

Pertaining to confidence intervals of effect sizes, Howell, D. C. (2010) suggested the following link: http://www.psy.latrobe.edu.au/esci.

In conclusion, effect sizes are important indicators to identify the magnitude of the intervention effect and describe the strength of the relationships among variables. The study aims to examine multiple choices that the future researchers may consider while reporting their research findings. Among these various options, there is no "one size fits all" measure of effect sizes, researchers may need to make their decisions of indices according to their specific contexts, such as research purposes, questions, and designs.

References

APA (2010). Publication Manual of the American Psychological Association, 6th Edition. Washington, DC: American Psychological Association.

Brooks, G. P., Diaz, E. A., & Johanson, G. A. (2017). A precision-based and adaptive approach to number of replications for Monte Carlo studies of robustness and power. General Linear Model Journal, 43(1), 31-49.

Cahan, S. (2000). Statistical significance is not a "Kosher Certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. Educational Researcher, 29(1), 31-36.

Cahan, S., & Gamliel, E. (2011). First among others? Cohen's d vs. alternative standardized mean group difference measures. Practical Assessment, Research & Evaluation, 16(10), 1-6.

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. Journal of Applied Psychology, 67(6), 691-700.

Carver, R. P. (1993). The case against statistical significance testing, revisited. The Journal of Experimental Education, 61: 287-292.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 45(12), 997-1003.

Cooper, H. (2017). Research synthesis and meta-analysis: A step-by-step approach (5th ed.). Thousand Oaks, CA: Sage.

Cortina, J. & Nouri, H. (2000). Effect size for ANOVA designs. Thousand Oaks: CA: Sage.

Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educational and Psychological Measurement, 60(4), 532-575.

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.

Ellis, P. D. (2009). Effect size calculator. Retrieved from http://www.polyu.edu.hk./mm/effectsizefaqs/calculator/claculator.html

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power. Meta-analysis, and the interpretation of research results. Cambridge, UK: Cambridge University Press.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. Educational and Psychological Measurement, 61(4), 517-531.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. Journal of Educational Statistics, 18(3), 271-279.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Gorard, S. (2015). Introducing the mean absolute deviation "effect" size. International Journal of Research & Method in Education, 38(2), 105-114.

Goulet-Pelletier, J. C., & Cousineau, Denis. (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.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Lawrence Erlbaum.

Hayes, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. British Journal of Mathematical and Statistical Psychology, 60(2), 217-244.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 106-128.

Hedges, L. V. & Olkin, I. (1985). Statistical Methods for Meta-Analysis. San Diego, CA: Academic Press.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force report and current trends. Journal of Research and Development in Education, 33, 285-296.

Howell, D. C. (2010). Confidence intervals on Effect Size. Retrieved from https://www.uvm.edu/~dhowell/methods/Supplements/Confidence%20Intervals%20on%20Effect%20Size.pdf

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 1-12.

Lalongo, C. (2016). Understanding the effect size and its measures. Biochemica Medica, 26(2), 150-163.

Leach, L. F., & Henson, R. K. (2014). Bias and precision of the squared canonical correlation coefficient under nonnormal data condition. Journal of Modern Applied Statistical Methods, 13(1), 110-139.

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: Effect size analysis in quantitative research. CBE-Life Sciences Education, 12, 345-351.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. Biological Reviews, 82(4), 591-605.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. Psychological Methods, 23(2), 208-225.

Peng, C. Y. J., & Chen, L. T. (2014). Beyond Cohen's d: Alternative effect size measures for between-subject designs. The Journal of Experimental Education, 82(1), 22-50.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. Educational and Psychological Measurement, 62(2), 241-253.

Roberts, J. K., & Henson, R. K. (2003). Not all effects are created equal: A rejoinder to Sawilowsky. Journal of Modern Applied Statistical Methods, 2(1), 226-230.

Rosnow, R. L. & Rosenthal, R. (2003). Effect sizes for experimenting psychology. Canadian Journal of Experimental Psychology, 57(3), 221-237.

Sawilowsky, S. S. (2003). Deconstructing arguments from the case against hypothesis testing. Journal of Modern Applied Statistical Methods, 2(2), 467-474.

Sawilowsky, S. S., & Yoon, J. S. (2002). The trouble with trivials ($p > .05$). Journal of Modern Applied Statistical Methods, 1(1), 143-144.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 2, 115-129.

Skidmore, S. T., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. Behavioral Research, 45, 536-546.

Stevens, J. P. (1999). Intermediate Statistics: A Modern Approach (2nd ed.). Madwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2002). “Statistical”, “practical”, and “clinical”: How many kinds of significance do counselors need to consider? Journal of Counseling & Development, 80, 64-71.

Vacha-Haase, T. & Thompson, B. (2004). How to estimate and interpret various effect sizes. Journal of Counseling Psychology, 51(4), 473-481.

Walker, D. A. (2015). JMASM34: Two group program for Cohen 's d, Hedges’ g, η2, radj2, ω2, ε2, confidence intervals, and power. Journal of Modern Applied Statistical Methods, 14(2), 282-292.

Wilcox, R. (2019). A robust nonparametric measure of effect size based on an analog of Cohen’s d, plus inferences about the median of the typical difference. Journal of Modern Applied Statistical Methods, 17(2), 1-18.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. British Journal of Mathematical and Statistical Psychology, 57(1), 173-181.