Behaviormetrika
Vol.40, No.2, 2013, 129–147

IS OMEGA SQUARED LESS BIASED? A COMPARISON OF THREE MAJOR EFFECT SIZE INDICES IN ONE-WAY ANOVA

Kensuke Okada*

The purpose of this study is to find less biased effect size index in one-way analysis of variance (ANOVA) by performing a thorough Monte Carlo study with 1,000,000 replications per condition. Our results show that contrary to common belief, epsilon squared is the least biased among the threemajorindices, while omega squared produces the least root mean squared errors, for all conditions. Although eta squared results in the least standard deviation, this does not necessarily make it a good estimator because a considerable amount of bias still occurs when the sample size is small.

1. Introduction

The importance of investigating and reporting not only the results of null hypothesis statistical testing (NHST) but also the magnitude of the effect involved is widely accepted in psychological and behavioral research. The measure of the magnitude of effect is called *effect size* (ES). Complete reporting of the estimates of appropriate effect sizes is one of "the minimum expectations for all APA journals" (American Psychological Association, 2009, p.33) and a requirement for 24 scholarly journals (Natesan & Thompson, 2007), including five educational research journals (Alhija & Levy, 2009). The rationale for requiring reporting effect size includes allowing readers to quantitatively evaluate the practical importance of study results and to incorporate these results in future meta-analyses (Cumming & Finch, 2001).

Although the importance of reporting and interpreting effect sizes is widely recognized, the detailed characteristics of effect size indices have received relatively little attention. For instance, there are more than seventy varieties of effect size indices (Kirk, 2003). The fact that effect size indices are not unique requires further research.

In this paper, we consider one-way fixed effect analysis of variance (ANOVA) with independent samples. The total population variance σ_{t}^2 is divided into the sum of variance between groups, σ_{b}^2, and the variance within groups, σ_{w}^2:

$$\sigma_{\mathrm{t}}^2 = \sigma_{\mathrm{b}}^2 + \sigma_{\mathrm{w}}^2, \tag{1}$$

where σ_{b}^2 represents the variance due to different levels of the independent variable,

Key Words and Phrases: effect size, bias, ANOVA, eta squared, epsilon squared, omega squared

∗ Department of Psychology, Senshu University.
Mail Address: ken@psy.senshu-u.ac.jp

This work was supported in part by grants from the Japan Society for the Promotion of Science (24730544, 090100000119, 23300310) and a grant of Strategic Research Foundation Grant-aided Project for Private Universities from MEXT Japan (2011–2015 S1101013).

Correspondence concerning this article should be addressed to Kensuke Okada, Department of Psychology, Senshu University, 2–1–1, Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214–8580, Japan.

and σ_{w}^2 represents the variance that cannot be attributed to the independent variable (i.e., the error term). The population effect size, η^2, is defined as

$$\eta^2 = \frac{\sigma_{\mathrm{b}}^2}{\sigma_{\mathrm{t}}^2} = 1 - \frac{\sigma_{\mathrm{w}}^2}{\sigma_{\mathrm{t}}^2}. \tag{2}$$

η^2 represents the amount of variance in the dependent variable explained by the independent variable, ranging from 0 (no effect) to 1 (maximum effect). This is typically the quantity of interest in ANOVA procedures (Graham, 2008). The population effect size, η^2, is an unknown parameter, and as such, must be estimated from samples.

There are three major sample effect size indices in ANOVA (Grissom & Kim, 2004; D. Matsumoto, Kim, & Grissom, 2011; Keppel, 1982; Olejnik & Algina, 2000): eta squared ($\hat{\eta}^2$); epsilon squared ($\hat{\varepsilon}^2$; Kelley, 1935); and omega squared ($\hat{\omega}^2$; Hays, 1963)[1)]. These sample indices correspond to the same population parameter, η^2, in Equation 2. They are defined as follows (Maxwell, Camp, & Arvey, 1981, p.527):

$$\hat{\eta}^2 = \frac{\mathrm{SS_b}}{\mathrm{SS_t}}, \tag{3}$$

$$\hat{\varepsilon}^2 = \frac{\mathrm{SS_b} - \mathrm{df_b MS_w}}{\mathrm{SS_t}}, \tag{4}$$

$$\hat{\omega}^2 = \frac{\mathrm{SS_b} - \mathrm{df_b MS_w}}{\mathrm{SS_t} + \mathrm{SS_w}}, \tag{5}$$

where $\mathrm{SS_t}$ is the total sum of squares; $\mathrm{SS_b}$ is the sum of squares between groups; $\mathrm{SS_w}$ is the sum of squares within groups; $\mathrm{df_b}$ is the degree of freedom between groups; $\mathrm{MS_w}$ is the mean sum of squares within groups; and $\mathrm{SS_t} = \mathrm{SS_b} + \mathrm{SS_w}$.

Among these three effect size indices, $\hat{\eta}^2$ was developed as a descriptive index while $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ were intended for inferential purposes (Maxwell et al., 1981). The nature of $\hat{\eta}^2$ is easily understood as it is given by replacing the population variance parameters in Equation 2 (σ_{t}^2 and σ_{b}^2) with the corresponding sample sum of squares ($\mathrm{SS_t}$ and $\mathrm{SS_b}$). Meanwhile, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are inferential indices which correct the bias in estimating the population effect size, η^2. Both $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are constructed by substituting the variance component parameters of η^2 with its bias-corrected sample estimators.

The idea of Kelley's (1935) $\hat{\varepsilon}^2$ is simply to substitute the population parameters σ_{t}^2 and σ_{w}^2 in Equation 2 with their corresponding unbiased estimators. To be specific, he estimated σ_{t}^2 with $\mathrm{SS_t}/(\mathrm{n}-1)$ and σ_{w}^2 with $\mathrm{MS_w}$, where n is the sample size. On the other hand, Hays (1963) used the relationship

$$\sigma_{\mathrm{t}}^2 = \sigma_{\mathrm{w}}^2 + \frac{\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{n_j} \alpha_{\mathrm{j}}^2}{\mathrm{n}}, \tag{6}$$

where $\mathrm{n_j}$ represents the number of observations in the j-th of J levels (groups) and

[1)] Note that $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ (Equations 3 to 5) are sample counterparts of the same parameter, η^2 (Equation 2). This popular notation of effect size may seem unusual because unlike the typical meaning of the hat symbol, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are estimators of η^2.

α_j represents the deviation of each group-specific mean from the grand mean, to reformulate the Equation 2 as

$$\eta^2 = \frac{\sum_{j=1}^{J} n_j \alpha_j^2}{n\sigma_w^2 + \sum_{j=1}^{J} n_j \alpha_j^2}. \quad (7)$$

Then, he estimated the numerator and denominator of Equation 7 with their respective unbiased estimators, $SS_b - df_b MS_w$ and $SS_t + MS_w$, to derive the formula for $\hat{\omega}^2$. Thus, because their decompositions and sample estimators are different, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are also different.

Although their original derivations were different, the resultant forms of the three indices are similar to each other. Equations 3 to 5 show that $\hat{\varepsilon}^2$ is given by subtracting $df_b MS_w$ from the numerator of $\hat{\eta}^2$, while $\hat{\omega}^2$ is given by adding SS_w to the denominator of $\hat{\varepsilon}^2$. Because every term in Equations 3 to 5 are greater than or equal to zero, the following inequality holds true for all three indices:

$$\hat{\omega}^2 \leq \hat{\varepsilon}^2 \leq \hat{\eta}^2. \quad (8)$$

It is known that $\hat{\eta}^2$ overestimates the population effect size, η^2, because its numerator, SS_b, is inflated by some error variability (Grissom & Kim, 2004). Previous studies such as those by Snyder and Lawson (1993) and Maxwell and Delaney (2004) give a detailed discussion of this bias. Note that $\hat{\eta}^2$ is also known as the coefficient of determination or R^2, which is also known for its upward bias. Also, $\hat{\varepsilon}^2$ is equivalent to the adjusted R^2 (Ezekiel, 1930).

Both $\hat{\varepsilon}^2$ and $\hat{\eta}^2$ replaces the parameters of the variance ratio $\eta^2$with their corresponding sample estimators. However, a ratio of unbiased estimators is generally not an unbiased estimator of the ratio (Olkin & Pratt, 1958). In fact, it is shown that none of the three sample effect size measures ($\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$; Equations 3 to 5) is an unbiased estimator of η^2 (Darlington, 1968). It is also shown that as sample size tends to infinity, $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ all converge toward the same value (Maxwell et al, 1981). However, little is known about their comparative behaviors in realistic finite (small) sample settings. In order to investigate the finite sample properties of these effect size measures, Monte Carlo experiments are required.

For several decades, researchers believed that $\hat{\omega}^2$ is the least biased index among the three, followed closely by $\hat{\varepsilon}^2$, and then $\hat{\eta}^2$. Keselman (1975) studied the performance of $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ in a Monte Carlo study using Cohen's (1969) definitions of small, medium, and large population effects under a correct model, that is, one in which the assumptions of ANOVA and effect sizes are correct. Keselman (1975) states, "the mean values for omega squared are consistently closer to the population treatment magnitudes while the mean values for epsilon squared are always slightly larger than omega squared but smaller than the mean values for eta squared" (p.47). In their textbook on effect size, Grissom and Kim (2004) agree with Keselman's findings, citing that "a somewhat less biased alternative estimator of η^2 is $\hat{\varepsilon}^2$, and a more nearly unbiased estimator is $\hat{\omega}^2$" (p.121). A similar description is provided by D. Matsumoto

et al. (2011). Olejnik and Algina (2000) explain the order of bias among the three indices:

> Epsilon squared corrects the numerator of eta squared by subtracting the error mean square from the contrast sum of squares. Omega squared further adjusts epsilon squared by adding the error mean square sum to the total sum of squares in the denominator of epsilon squared (p.262)[2)].

These commonly-held beliefs, however, raise several questions. First, $\hat{\omega}^2$ was not originally intended as the index that "further adjusts" $\hat{\varepsilon}^2$. Although it is easy to derive such an interpretation from the indices' formula as shown in Equations 4 and 5, $\hat{\omega}^2$ and $\hat{\varepsilon}^2$ are actually derived independently. Hays (1963), for instance, did not use $\hat{\varepsilon}^2$ at all in his derivation, nor tried to "further correct" $\hat{\varepsilon}^2$. Thus, although the ordinal relationship shown in Equation 8 holds, it is possible that $\hat{\omega}^2$ underestimates the population effect size more than $\hat{\varepsilon}^2$ does.

Second, another Monte Carlo study by Carroll and Nordholm (1975) produced results that somewhat contradicted those of Keselman (1975). Putting aside the fact that their study used a smaller number of replications per condition than Keselman's which may have resulted in more sampling errors, Carroll and Nordholm's (1975) results implied that while "$\hat{\omega}^2$ is slightly negatively biased" (p.548), "any bias in $\hat{\varepsilon}^2$ is not evident" (p.549). However, their overall conclusion was that "it was found that $\hat{\omega}^2$ and $\hat{\varepsilon}^2$ were very similar" (p.553) and they did not investigate further.

Since these two studies were performed more than thirty years ago, there were no other comparable studies conducted to evaluate the performance of these ANOVA effect size indices. Although some researchers such as Maxwell et al. (1981) and Snyder and Lawson (1993) discussed the values of $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ for a single dataset extensively, they did not attempt to measure the extent of bias among these indices using a Monte Carlo study. Owing to the lack of research on the bias of these indices, a consensus on which index is most appropriate for certain cases has not yet been reached (Kline, 2004, p.100).

One of the reasons for the unreliable conclusions of studies such as Keselman (1975) and Carroll and Nordholm (1975) is the limited capabilities of computers during that period, which only allowed a relatively small number of replications per condition. The 5,000 and 1,000 replications per condition used by Keselman (1975) and Carroll and Nordholm (1975), respectively, may not be sufficient for today's standards.

Another limitation of these older studies is that they evaluated the effect size indices in terms only of their means and standard deviations. Since standard deviation is the mean squared deviation from the sample mean and not from the true value, η^2, it may be insufficient for evaluating the sample effect size indices. Because standard deviation does not take η^2 into account, there may be cases in which the standard deviation is small but samples are substantially biased from the true value.

This study aims to contribute to the current body of literature by overcoming the

[2)] Although Olejnik and Algina (2000) also showed the equations for $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ within this quotation, we omitted them to avoid confusion because their notation is different from that in our study.

two main limitations of these older studies. We compare the performance of these three effect size indices by evaluating the bias and errors under the correct model using modern computers. We use the same condition as Keselman's (1975) numerical experiment, increasing the number of replications per condition to 1,000,000 to ensure an exhaustive comparison of the effect size indices. We also add a sample size condition to the numerical experiment to assess the effect of sample size to the bias of the effect size indices.

In addition, we also calculated the root mean squared errors (RMSEs). While standard deviation measures the square root of the average squared discrepancy from the sample mean, RMSE measures the square root of the average squared discrepancy from the true value. Note that if the sample mean of the estimator exactly corresponds to the true value in the population (i.e., there is zero bias), the standard deviation is the same as the RMSE. In reality, however, this is not usually the case. By using the RMSE, we account for errors in estimating the population effect size.

2. Method

As we have mentioned earlier, our numerical experiment is based on Keselman's (1975) study. We use a one-way, four-level ANOVA model with independent samples where the population effect size magnitude is controlled according to Cohen's (1962) criteria of small, medium, and large population effects. After reviewing a volume of the *Journal of Abnormal and Social Psychology*, Cohen (1962) argues that the large experimental treatment effects were those in which the treatment means differed by .40 standard deviation unit. Medium and small treatment effects are also operationally defined by Cohen as those in which the means differ by .25 and .10 standard deviation unit, respectively[3)].

The variability of population means within four levels is controlled as either maximum or intermediate. Table 1 (a) shows the specific population mean values under all three effect sizes multiplied by two mean variability conditions in population. The treatment means are dichotomized at the end points of their range difference to create maximum variability among them (Cohen, 1969, p.270). For instance, for four treatment means differing by .40 standard deviation unit (i.e., large treatment effect), the range of the differences is 1.20, and therefore $\mu_1 = 0.00$, $\mu_2 = 0.00$, $\mu_3 = 1.20$, and $\mu_4 = 1.20$. Intermediate variability occurs when the means are equally spaced over their range. For instance, for the same treatment means condition, $\mu_1 = 0.00$, $\mu_2 = 0.40$, $\mu_3 = 0.80$, and $\mu_4 = 1.20$. Note that the values in this table are the same as those in Table 1 of Keselman's (1975) study. The population standard deviation of all conditions is 1. The population (true) effect size for each condition is also shown in Table 1 (b).

Our study differs from Keselman's (1975) in several ways. First, owing to the capa-

[3)] Note that these values are different from what is implied by so-called Cohen's (1969) criteria of large, medium and small effects for effect sizes d or R^2. We used these values based on Cohen (1962) in order to obtain comparability with Keselman's (1975) study.

Table 1:
(a) Population Means for Each Effect Size Magnitude and Mean Variability Condition

ES Magnitude	**Mean Variability**							
	Maximum				Intermediate			
	μ_1	μ_2	μ_3	μ_4	μ_1	μ_2	μ_3	μ_4
Large	0.00	0.00	1.20	1.20	0.00	0.40	0.80	1.20
Medium	0.00	0.00	0.75	0.75	0.00	0.25	0.50	0.75
Small	0.00	0.00	0.30	0.30	0.00	0.10	0.20	0.30

(b) Population Effect Size for Each Conditiong

ES Magnitude	**Mean Variability**	
	Maximum	Intermediate
Large	.26471	.16667
Medium	.12329	.07246
Small	.02200	.01235

bilities of the modern computer, the number of random samples per one condition is substantially higher at 1,000,000. In Keselman's study, the relatively smaller number of replications has led to a number of potential errors. For instance, $\hat{\varepsilon}^2$ overestimates the population effect size in large treatments and maximum variability conditions and underestimates it in other conditions (Keselman, 1975, Table 2). Because the hypothesized model is correct in all the conditions, this result can be due to the sampling errors owing to the small number of replications. With 1,000,000 replications, we expect that sampling error in our study will be very minimal. Second, our study also considers sample size per analysis as a third condition of the numerical experiment. In contrast, Keselman (1975) does not show the sample size used in his numerical experiment; he only states that "n_j observations" are used (Keselman, 1975, p.47), not showing nor controlling the actual value of n_j. In our study, we controlled the number of observations per group from 10 to 100 in increments of 10, enabling us to consider the effect of sample size on the performance of the effect size indices. Each of the four groups includes an equal number of observations.

Third, we evaluated our results in terms of three measures: bias, standard deviation, and RMSE. While Keselman (1975) only reported the means and standard deviations of the effect size indices, we also included the indices' RMSEs as an important indicator of their performance. In addition, we explicitly calculated the biases of the effect size indices. Although it is possible that Keselman (1975) and Carroll and Nordholm (1975) calculated the biases from their results, they did not explicitly show it.

The sample effect size statistic, $\hat{\theta}^2$, denotes one of the indices $\hat{\eta}^2$, $\hat{\varepsilon}^2$, or $\hat{\omega}^2$; $\hat{\theta}_i^2$ denotes their sample values in the i-th replication; and $\hat{\theta}^2$ denotes their means over n_{rep} replications. In this study, as noted above, $n_{rep} = 1,000,000$. The bias of the estimator, $\hat{\theta}^2$, is given by

$$\text{bias}(\hat{\theta}^2) = (\hat{\theta}^2 - \eta^2). \tag{9}$$

Table 2: Resulting Biases for the Effect Size Indices

Mean Variability	ES Magnitude	ES Index	Sample size per group										Keselman (1975)
			10	20	30	40	50	60	70	80	90	100	
Maximum	Large	$\hat{\eta}^2$	.05439	.02671	.01779	.01338	.01058	.00885	.00757	.00663	.00588	.00529	.0512
		$\hat{\varepsilon}^2$	−.00235	−.00126	−.00077	−.00050	−.00052	−.00038	−.00034	−.00028	−.00027	−.00024	.0003
		$\hat{\omega}^2$	−.00686	−.00361	−.00235	−.00170	−.00148	−.00119	−.00102	−.00089	−.00081	−.00072	−.0039
	Medium	$\hat{\eta}^2$	.06570	.03227	.02146	.01599	.01286	.01059	.00914	.00800	.00711	.00640	.0184
		$\hat{\varepsilon}^2$	−.00189	−.00106	−.00066	−.00056	−.00036	−.00042	−.00029	−.00025	−.00021	−.00020	−.0006
		$\hat{\omega}^2$	−.00427	−.00234	−.00153	−.00122	−.00089	−.00086	−.00067	−.00059	−.00051	−.00046	−.0016
	Small	$\hat{\eta}^2$	.07482	.03690	.02453	.01829	.01464	.01219	.01043	.00914	.00814	.00731	.0043
		$\hat{\varepsilon}^2$	−.00045	−.00025	−.00013	−.00016	−.00010	−.00008	−.00009	−.00006	−.00004	−.00005	−.0003
		$\hat{\omega}^2$	−.00083	−.00049	−.00030	−.00029	−.00021	−.00017	−.00016	−.00012	−.00010	−.00010	−.0003
Inter-mediate	Large	$\hat{\eta}^2$	.06204	.03050	.02029	.01521	.01206	.01004	.00863	.00757	.00672	.00605	.0309
		$\hat{\varepsilon}^2$	−.00223	−.00119	−.00074	−.00052	−.00051	−.00043	−.00033	−.00027	−.00025	−.00022	−.0005
		$\hat{\omega}^2$	−.00536	−.00284	−.00186	−.00137	−.00119	−.00100	−.00082	−.00070	−.00063	−.00056	−.0022
	Medium	$\hat{\eta}^2$	.07015	.03453	.02296	.01710	.01372	.01136	.00978	.00855	.00758	.00683	.0143
		$\hat{\varepsilon}^2$	−.00130	−.00072	−.00044	−.00041	−.00027	−.00029	−.00019	−.00017	−.00017	−.00014	.0004
		$\hat{\omega}^2$	−.00273	−.00151	−.00097	−.00082	−.00060	−.00057	−.00043	−.00038	−.00035	−.00031	.0001
	Small	$\hat{\eta}^2$	.07574	.03737	.02484	.01854	.01484	.01235	.01058	.00924	.00823	.00740	.0020
		$\hat{\varepsilon}^2$	−.00026	−.00014	−.00006	−.00009	−.00005	−.00005	−.00004	−.00005	−.00002	−.00003	−.0001
		$\hat{\omega}^2$	−.00043	−.00027	−.00015	−.00017	−.00011	−.00010	−.00008	−.00009	−.00006	−.00006	−.0001

The standard deviation of the estimator is calculated by

$$\mathrm{SD}(\hat{\theta}^2) = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n_{rep}}} \left(\hat{\theta}^2_{\mathrm{i}}, -\hat{\theta}^2\right)^2}{\mathrm{n_{rep}}}}. \tag{10}$$

The root mean squared error (RMSE) is calculated by

$$\mathrm{RMSE}(\hat{\theta}^2) = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n_{rep}}} \left(\hat{\theta}^2_{\mathrm{i}} - \eta^2\right)^2}{\mathrm{n_{rep}}}}. \tag{11}$$

Apart from the three major instances cited above, our study differs from Keselman (1975) in other, albeit minor, ways. We generated a random number of the artificial datasets from normal distributions with predetermined means shown in Table 1 and standard deviation of 1 using the Mersenne twister algorithm (M. Matsumoto & Nishimura, 1998). This algorithm provides fast and high-quality pseudorandom numbers and is adopted as a default random number generator by many statistical computer programs including R and Matlab. We therefore expect the quality of the random numbers to be better than in previous studies, during which time the Mersenne twister algorithm has not yet been developed. Also, we did not include a condition when the distribution of the population is exponential rather than normal in order to intensively study the performance of the indices under the correct model. This is because it is often not reasonable to assume that psychological data is exponentially distributed. In fact, Keselman (1975) only briefly described the results for an exponential distribution. Moreover, he found that "the above results [for normal distribution] are also descriptive of the data when sampling observations from the non-normal exponential distribution" (p.47). Based on the above analysis, we concluded that we do not need to include the exponential distribution conditions.

We performed the entire analysis using the software, R 2.13.1 (R Foundation for Statistical Computing, 2011). We included the R code used in this study in the Appendix for other researchers hoping to conduct further studies in this area. Note that by definition, the estimates $\hat{\omega}^2$ and $\hat{\varepsilon}^2$ occasionally take negative values. Although negative sample effect size is sometimes reported as 0, Fidler and Thompson (2001) argued that any obtained negative effect size value should be reported as it is to facilitate interval estimation. Based on their argument, and also following Keselman (1975) and Carroll and Nordholm (1975), we used all the estimates in $\mathrm{n_{rep}} = 1,000,000$ replications.

3. Results and Discussion

3.1 Bias

The resultant biases in all the conditions are summarized in Table 2 and plotted in Figure 1. Our most important and impressive finding is that $\hat{\varepsilon}^2$ has the least absolute bias among the three indices, surpassing $\hat{\omega}^2$ *in all the conditions*, in contrast with

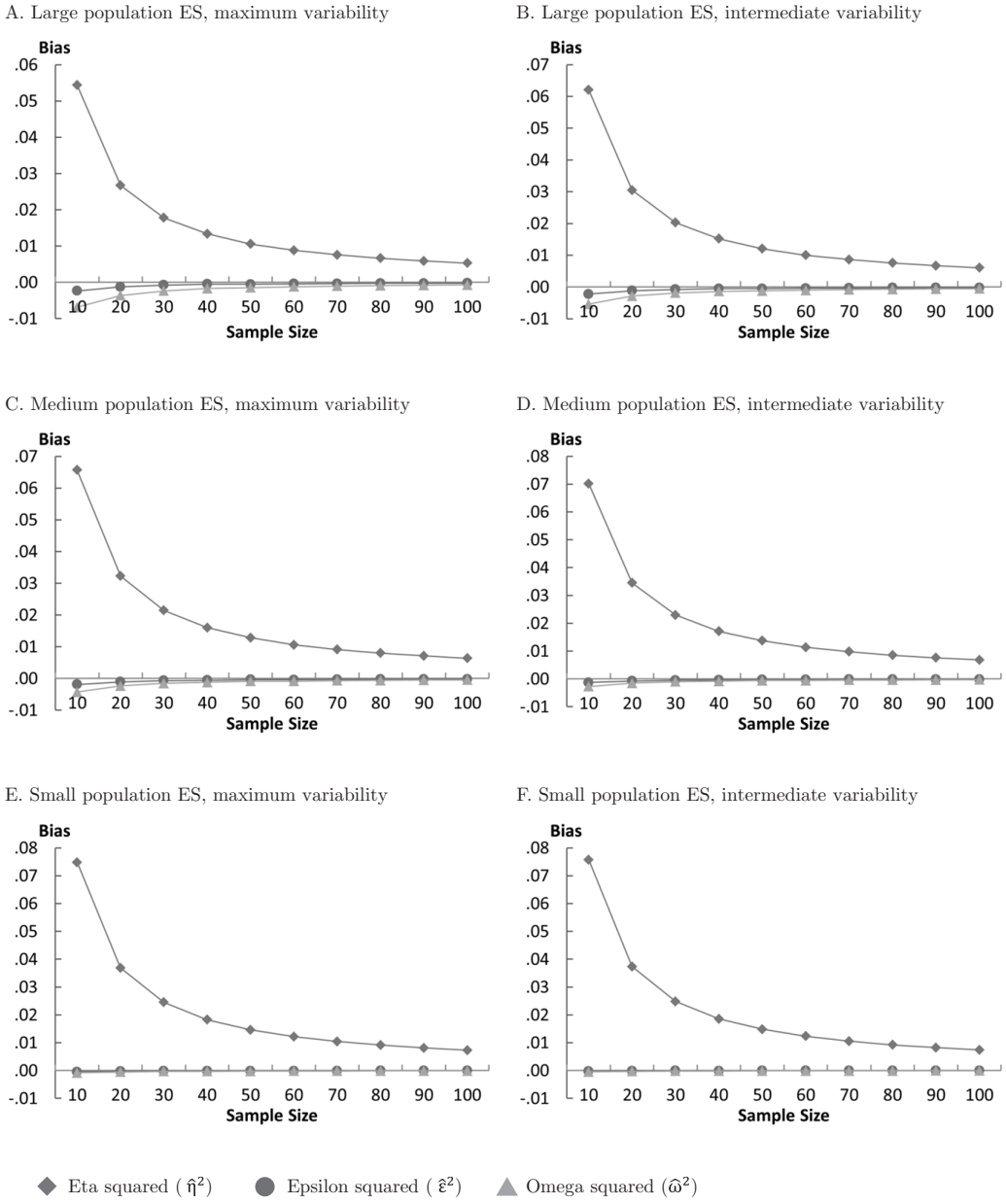

Figure 1: Plot of the results shown in Table 2. The x-axis is the sample size per group and the y-axis is the resulting biases for the effect size indices.

the commonly-held belief that $\hat{\omega}^2$ is less biased than $\hat{\varepsilon}^2$ (Grissom & Kim, 2004; D. Matsumoto, et al., 2011). As shown in Table 2 and illustrated in Figure 1, the bias of $\hat{\varepsilon}^2$ is the least among the three in every sample size conditions, and $\hat{\varepsilon}^2$ always slightly underestimated the true effect size (i.e., $\hat{\varepsilon}^2$ has a slight negative bias); $\hat{\omega}^2$ also has a negative bias whose magnitude is always greater than $\hat{\varepsilon}^2$'s; and $\hat{\eta}^2$ has a considerable

large positive bias, even when sample size is 100. Note that in general underestimation is preferable to overestimation which can erroneously identify an effect that does not in fact exist.

Since all of the above findings apply to all other combinations of population effect size and variability conditions, we can therefore conclude that in terms of bias, $\hat{\varepsilon}^2$ is the best index among three. Note that for all three effect size indices, the absolute bias becomes larger when the sample size is small.

Although Keselman (1975) did not expressly calculate biases, we can compute them from the results presented in his paper. Therefore, we also calculated the biases from the results of Keselman's study, as shown at the rightmost column of Table 2. We found that Keselman's results show the same ordinal pattern of the absolute bias as ours, $\left|\text{bias}(\hat{\varepsilon}^2)\right| < \left|\text{bias}(\hat{\omega}^2)\right| < \left|\text{bias}(\hat{\eta}^2)\right|$, although this relation is obscured when population effect size is small owing to the sampling errors brought about by the small number of replications. Also, unconformities, such as the seemingly positive bias of $\hat{\varepsilon}^2$ and $\hat{\omega}^2$, or the seemingly lesser bias of $\hat{\omega}^2$ than $\hat{\varepsilon}^2$, are sometimes found in biases calculated from Keselman's results. The results of our Monte Carlo study, meanwhile, do not show any of the unconformities.

3.2 Standard Deviation

The resulting standard deviations for all the conditions are summarized in Table 3 and illustrated in Figure 2. Contrary to our results for bias, our findings show that in all conditions, $\hat{\eta}^2$ has the least standard deviation, followed by $\hat{\omega}^2$, then $\hat{\varepsilon}^2$. Similar findings were also found by Keselman (1975), as shown in the rightmost column of Table 3, although this relation is obscured under small population effect conditions due to sampling errors. From this, Keselman (1975) stated that "eta squared could be considered the more efficient estimator (p.47)." However, it is important to remember that $\hat{\eta}^2$ has a considerably large bias when the sample size is small, and that an estimator would have a small standard deviation even if its estimates were consistently off deviate from the true value. It is therefore insufficient to determine the quality of an estimator based only on its standard deviation. Root mean squared errors (RMSEs) are needed as additional indicators.

3.3 RMSE

The resulting RMSEs for all the conditions are shown in Table 4 and illustrated in Figure 3. In all the conditions, $\hat{\omega}^2$ has the least RMSEs, followed by a narrow margin by $\hat{\varepsilon}^2$, then $\hat{\eta}^2$. Our results show that the difference between $\hat{\omega}^2$ and $\hat{\varepsilon}^2$ is almost always small, although this relationship is consistent across conditions. $\hat{\eta}^2$ tends to have the highest RMSEs among the indices. The difference between $\hat{\eta}^2$ and the other two indices also tends to be larger when the population effect size is small.

Although $\hat{\eta}^2$ has the least standard deviation among three, we found that it actually has the largest RMSE which means that although the variability of $\hat{\eta}^2$ is relatively

Table 3: Resulting Standard Deviations for the Effect Size Indices

Mean Variability	ES Magnitude	ES Index	Sample size per group										Keselman (1975)
			10	20	30	40	50	60	70	80	90	100	
Maximum	Large	$\hat{\eta}^2$	.11078	.07863	.06427	.05562	.04978	.04546	.04211	.03941	.03710	.03525	.1078
		$\hat{\varepsilon}^2$	.12001	.08173	.06593	.05669	.05054	.04603	.04257	.03978	.03741	.03552	.1148
		$\hat{\omega}^2$	.11861	.08125	.06568	.05652	.05042	.04594	.04250	.03972	.03736	.03547	.1135
	Medium	$\hat{\eta}^2$	.09931	.06854	.05543	.04781	.04267	.03884	.03590	.03357	.03162	.02997	.0590
		$\hat{\varepsilon}^2$	.10758	.07124	.05687	.04873	.04332	.03934	.03629	.03389	.03189	.03019	.0603
		$\hat{\omega}^2$	.10570	.07060	.05652	.04851	.04316	.03922	.03619	.03381	.03182	.03014	.0599
	Small	$\hat{\eta}^2$	.07031	.04224	.03212	.02667	.02319	.02079	.01894	.01756	.01641	.01545	.0121
		$\hat{\varepsilon}^2$	.07617	.04391	.03296	.02718	.02354	.02105	.01915	.01773	.01655	.01557	.0120
		$\hat{\omega}^2$	.07451	.04342	.03270	.02702	.02343	.02097	.01909	.01767	.01651	.01553	.0120
Inter-mediate	Large	$\hat{\eta}^2$	.10526	.07373	.05997	.05181	.04628	.04224	.03907	.03656	.03441	.03263	.0744
		$\hat{\varepsilon}^2$	.11403	.07664	.06153	.05280	.04698	.04278	.03949	.03690	.03470	.03288	.0768
		$\hat{\omega}^2$	.11223	.07602	.06119	.05259	.04683	.04266	.03940	.03683	.03464	.03283	.0762
	Medium	$\hat{\eta}^2$	.08818	.05883	.04698	.04023	.03572	.03246	.02991	.02793	.02626	.02488	.0360
		$\hat{\varepsilon}^2$	.09553	.06115	.04819	.04101	.03627	.03287	.03023	.02820	.02648	.02507	.0364
		$\hat{\omega}^2$	.09366	.06053	.04786	.04079	.03612	.03276	.03014	.02812	.02642	.02501	.0362
	Small	$\hat{\eta}^2$	.06568	.03755	.02768	.02250	.01928	.01710	.01545	.01420	.01321	.01237	.0062
		$\hat{\varepsilon}^2$	.07115	.03903	.02840	.02293	.01957	.01732	.01562	.01434	.01332	.01247	.0062
		$\hat{\omega}^2$	.06957	.03858	.02818	.02280	.01948	.01725	.01556	.01429	.01328	.01244	.0062

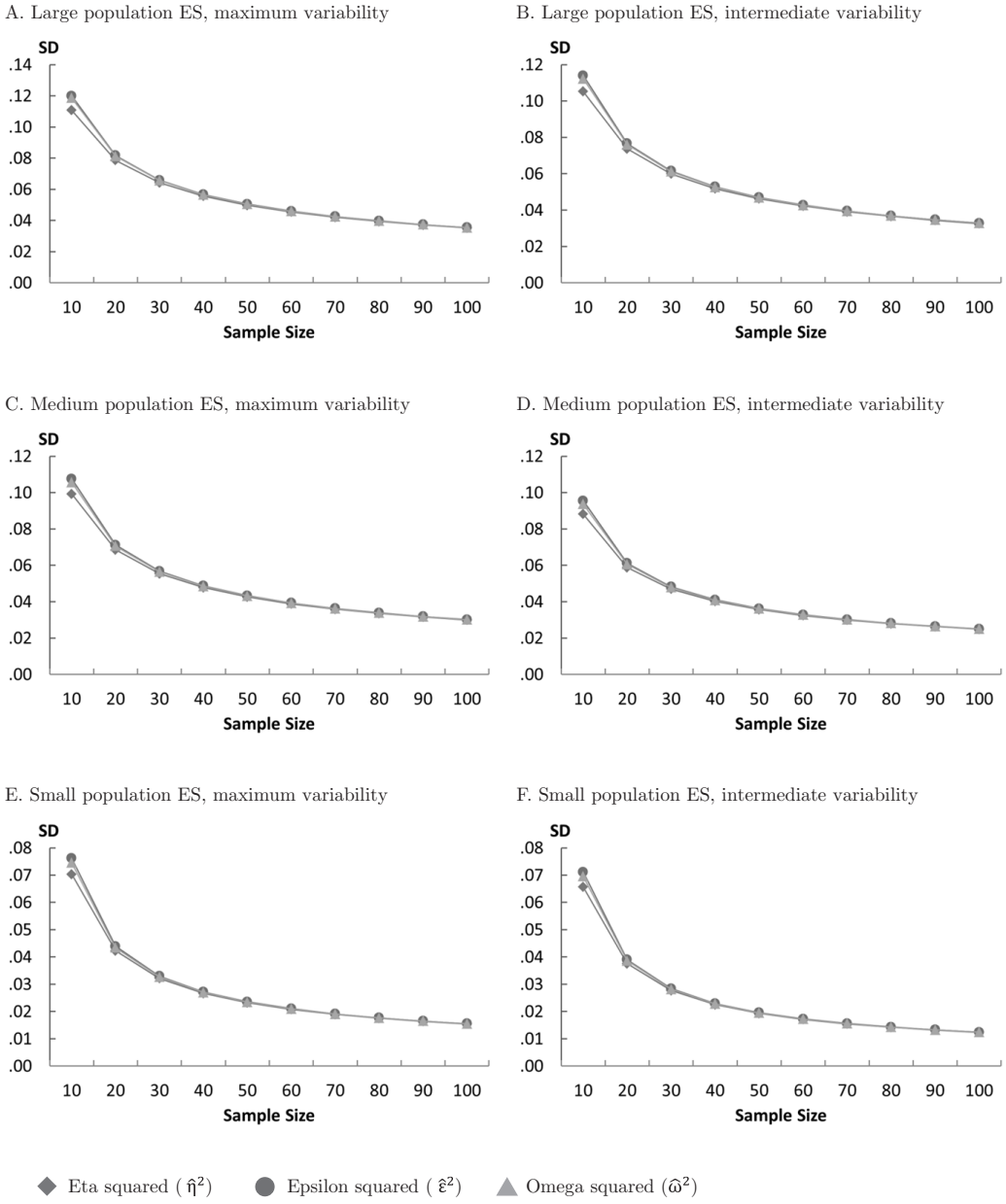

Figure 2: Plot of the results shown in Table 3. The x-axis is the sample size per group and the y-axis is the resulting standard deviations for the effect size indices.

small, it consistently deviates from the true value, reinforcing the poor quality of $\hat{\eta}^2$ as an estimate of population effect size. Note that unlike the means and standard deviations we have shown in Tables 2 and 3, we did not show the RMSE values from Keselman's (1975) findings in the rightmost column of Table 4 because there was not enough information to calculate them.

Table 4: ResultingRoot Mean Squared Errors for the Effect Size Indices

Mean Variability	ES Magnitude	ES Index	Sample size per group									
			10	20	30	40	50	60	70	80	90	100
Maximum	Large	$\hat{\eta}^2$	.12341	.08304	.06669	.05720	.05089	.04631	.04279	.03996	.03756	.03564
		$\hat{\varepsilon}^2$	.12003	.08174	.06594	.05669	.05054	.04603	.04257	.03978	.03741	.03552
		$\hat{\omega}^2$	.11881	.08133	.06572	.05655	.05044	.04596	.04251	.03973	.03737	.03548
	Medium	$\hat{\eta}^2$	.11907	.07576	.05944	.05042	.04456	.04026	.03704	.03451	.03241	.03064
		$\hat{\varepsilon}^2$	.10760	.07125	.05687	.04874	.04332	.03934	.03629	.03389	.03189	.03020
		$\hat{\omega}^2$	.10578	.07064	.05654	.04852	.04317	.03923	.03620	.03382	.03182	.03014
	Small	$\hat{\eta}^2$	.10267	.05609	.04042	.03234	.02742	.02410	.02163	.01980	.01832	.01709
		$\hat{\varepsilon}^2$	.07617	.04391	.03296	.02718	.02354	.02105	.01915	.01773	.01655	.01557
		$\hat{\omega}^2$	.07452	.04342	.03270	.02702	.02343	.02097	.01909	.01767	.01651	.01553
Inter-mediate	Large	$\hat{\eta}^2$	.12218	.07979	.06331	.05399	.04782	.04342	.04001	.03733	.03506	.03319
		$\hat{\varepsilon}^2$	.11405	.07665	.06153	.05281	.04699	.04278	.03950	.03690	.03470	.03288
		$\hat{\omega}^2$	.11236	.07608	.06122	.05261	.04685	.04267	.03941	.03683	.03464	.03283
	Medium	$\hat{\eta}^2$	.11268	.06821	.05229	.04372	.03827	.03439	.03147	.02921	.02733	.02580
		$\hat{\varepsilon}^2$	.09554	.06115	.04819	.04101	.03627	.03287	.03024	.02820	.02648	.02507
		$\hat{\omega}^2$	.09370	.06055	.04787	.04080	.03612	.03276	.03015	.02813	.02642	.02501
	Small	$\hat{\eta}^2$	.10025	.05298	.03720	.02916	.02433	.02109	.01872	.01694	.01556	.01442
		$\hat{\varepsilon}^2$	.07115	.03903	.02840	.02293	.01957	.01732	.01562	.01434	.01332	.01247
		$\hat{\omega}^2$	.06957	.03858	.02818	.02280	.01948	.01725	.01556	.01429	.01328	.01244

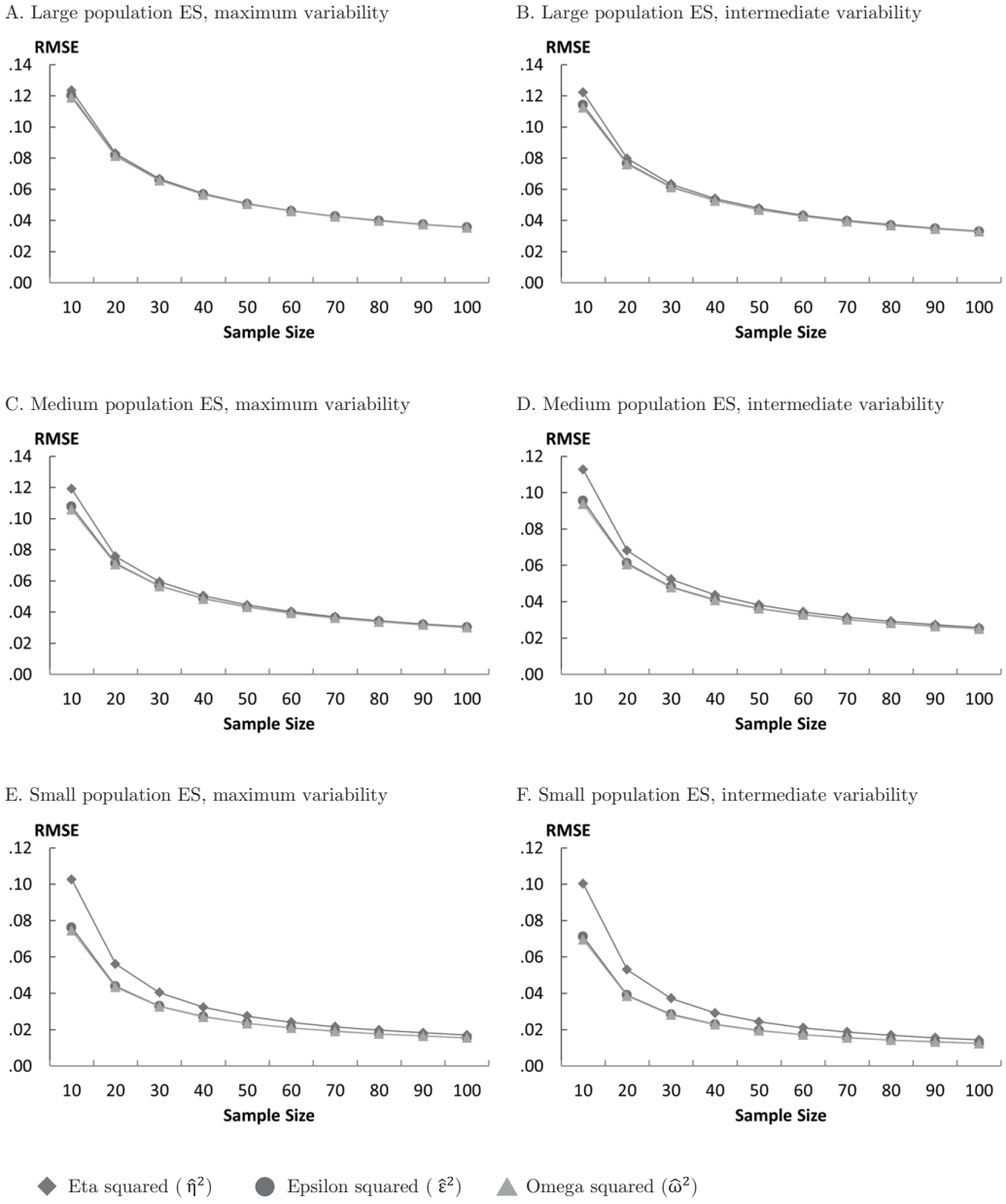

Figure 3: Plot of the results shown in Table 4. The x-axis is the sample size per group and the y-axis is the resulting root mean squared errors for the effect size indices.

4. Conclusion

In this study, we evaluated the performance of three major effect size indices in ANOVA using an intensive Monte Carlo study with 1,000,000 replications per condition, the results of which are summarized as follows.

First, contrary to the common belief that $\hat{\omega}^2$ is a lesser-biased version of $\hat{\varepsilon}^2$, we found that $\hat{\varepsilon}^2$ is the best effect size index in terms of bias. Second, we found $\hat{\omega}^2$ to be the best index among the three in terms of RMSE. Since previous studies did not calculate RMSEs, this is an important new finding. Although the difference in the performance of $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ is not very big, it is neither negligibly small especially when the sample size is small. Moreover, these two findings were consistent without any exception across all of the experimental conditions.

Third, although $\hat{\eta}^2$ is the best index in terms of standard deviation, we found that it is the worst in terms of both bias and RMSE. This means that the replicated values of $\hat{\eta}^2$ consistently deviate from the true population effect size value. As with previous studies, we found a positive bias of $\hat{\eta}^2$ which is robust and remains even when the sample size per condition is 100. Care is needed when interpreting $\hat{\eta}^2$ in psychological and behavioral science where ANOVA is often applied to data without a large sample size. Because of $\hat{\eta}^2$'s positive bias, analysts may incorrectly assume that the effect under study is large when it is not, especially when the sample size is small. Therefore, although it is the most frequently reported index, we do not recommend using $\hat{\eta}^2$ for inferential purposes, especially when the sample size is small.

Overall, our results indicate that $\hat{\varepsilon}^2$ is a promising effect size index because bias is often considered as a primary measure in evaluating statistics and the difference in RMSEs between the two indices is relatively small compared to bias.

Reducing the bias that may be present in an estimator is considered one of the most important tasks of statistical inference (Schucany, Gray, & Owen, 1971). Because there is no exception in the resulting order of bias, that is, $\left|\text{bias}(\hat{\varepsilon}^2)\right| < \left|\text{bias}(\hat{\omega}^2)\right| < \left|\text{bias}(\hat{\eta}^2)\right|$, in all our $2\times3\times10 = 60$ Monte Carlo experiment conditions (with 1,000,000 replications per condition), we are most certain about this result. Therefore, our results indicate that $\hat{\varepsilon}^2$ is a promising effect size, because it has the least bias, and the discrepancy between $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ is relatively subtle in RMSE compared to bias.

The inconsistent results reported in the previous study may be due to an old and weak random number generator. Keselman (1975) used the IBM 360 scientific subroutine package which generates random numbers by using the linear congenital algorithm. However, it is known that this algorithm can produce systematically incorrect results in Monte Carlo methods owing to subtle correlations in generated random numbers (Ferrenberg, Landau, & Wong, 1992). On the other hand, the current study uses Mersenne twister, which is known for creating uncorrelated high-quality sequences of random numbers (Gentle, 2003).

In reality, the most frequently reported index in psychological studies is $\hat{\eta}^2$, followed by $\hat{\omega}^2$. Although existing literature on effect size mentions $\hat{\varepsilon}^2$, it has relatively less attention than the other two indices. This could be partly due to the fact that the $\hat{\eta}^2$ value is often reported in commercial statistical software (Pierce, Block, & Aguinis, 2004). For instance, Fritz, Morris, and Richler's (2011) review of effect size describes both $\hat{\eta}^2$ and $\hat{\omega}^2$ in detail, but only mentions $\hat{\varepsilon}^2$ briefly, stating that "however, ε^2 is rarely reported, and we do not discuss it further here" (p.11). Kline (2004) discussed the three indices in a similar manner. Our study, however, clearly indicates that $\hat{\varepsilon}^2$,

as an effect size index, deserves more attention.

This study is not without limitations. In this study, we only considered cases when the assumptions of ANOVA in deriving effect size indices are correct. The number of conditions with violated assumptions is too large for a single undertaking. Thus, we chose to focus our attention to the case where the assumptions are correct, which has never been studied this extensively before. Note that previous studies did not find a substantial difference in the results for both cases. As we stated in our methodology, previous researchers considered cases under other assumptions. Keselman (1975) considered cases where the population distribution is exponential rather than normal and found little difference between the two. Meanwhile, Carroll and Nordholm (1975) considered cases where the population variance is heterogeneous and discovered that "heterogeneity of variances had negligible effects on the estimates under conditions of equal n" (p.553). It is important to note, however, that in these studies, both the number of replications and the variety of violations are limited. As we have mentioned earlier, there are many more conditions in which these assumptions are violated, which future studies should explore and consider. Another limitation of our study is that we only used a one-way, independent ANOVA, the most basic ANOVA model. Although we expect that similar results as ours can be derived from factorial designs (i.e., models with more than one factor) and repeated measurement designs, future studies need to further verify this expectation.

Appendix

The R code used in our study for large effect magnitudes and maximum variability is presented below. Note that this code can be easily adapted for other conditions by changing the value of `muvec`.

```
muvec <- c(0.00,0.00,1.20,1.20)
meanmu <- mean(muvec)
sigb <- sum((muvec-meanmu)^2)/4
eta2p <- sigb/(sigb+1)
k <- length(muvec)
nsim <- 1000000
njs <- c(10,20,30,40,50,60,70,80,90,100)
BIASmat <- matrix(NA,nrow=length(njs),ncol=3)
rownames(BIASmat) <- njs
colnames(BIASmat) <- c("eta2","epsilon2","omega2")
RMSEmat <- matrix(NA,nrow=length(njs),ncol=3)
rownames(RMSEmat) <- njs
colnames(RMSEmat) <- c("eta2","epsilon2","omega2")
SDmat <- matrix(NA,nrow=length(njs),ncol=3)
rownames(SDmat) <- njs
colnames(SDmat) <- c("eta2","epsilon2","omega2")
```

```
niter <- 1
for (nj in njs){
   x <- matrix(NA,nrow=nj,ncol=4)
   eta2 <- rep(NA,nsim)
   epsilon2 <- rep(NA,nsim)
   omega2 <- rep(NA,nsim)
   for (ii in 1: nsim){
      y <- c(rnorm(n=nj,mean=muvec[1],sd=1),
      rnorm(n=nj,mean=muvec[2],sd=1),
      rnorm(n=nj,mean=muvec[3],sd=1),
      rnorm(n=nj,mean=muvec[4],sd=1))
      x <- as.factor(c(rep("mu1",nj),rep("mu2",nj),
      rep("mu3",nj),rep("mu4",nj)))
      res <- anova(aov(y~x))
      res <- as.matrix(res)
      SSb <- res[1,2]
      SSt <- res[1,2] + res[2,2]
      MSw <- res[2,3]
      eta2[ii] <- SSb/SSt
      epsilon2[ii] <- (SSb - 3*MSw)/SSt
      omega2[ii] <- (SSb - 3*MSw)/(SSt+MSw)
   }
   BIASmat[niter,1] <- mean(eta2) - eta2p
   BIASmat[niter,2] <- mean(epsilon2) - eta2p
   BIASmat[niter,3] <- mean(omega2) - eta2p
   RMSEmat[niter,1] <- sqrt(sum((eta2-eta2p)^2)/nsim)
   RMSEmat[niter,2] <- sqrt(sum((epsilon2-eta2p)^2)/nsim)
   RMSEmat[niter,3] <- sqrt(sum((omega2-eta2p)^2)/nsim)
   SDmat[niter,1] <- sqrt(sum((eta2-mean(eta2))^2)/nsim)
   SDmat[niter,2] <- sqrt(sum((epsilon2-mean(epsilon2))^2)
   /nsim)
   SDmat[niter,3] <- sqrt(sum((omega2-mean(omega2))^2)
   /nsim)
   niter <- niter+1
}
```

References

Alhija, F. N., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, 69, 245–265. doi: 10.1177/0013164408315266

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington DC: American Psychological Association.

Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelly's and Hays'. *Educa-*

tional and Psychological Measurement, 35, 541–554. doi: 10.1177/001316447503500304

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153. doi: 10.1037/h0045186

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

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*, 61, 532–574. doi: 10.1177/0013164401614002

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161–182. doi:10.1037/h0025471

Ezekiel, M. J. B. (1930). *Methods of Correlational Analysis*. New York: Wiley.

Ferrenberg, A. M., Landau, D. P., & Wong, Y. J. (1992). Monte Carlo simulations: hidden errors from "good" random number generators. *Physical Review Letters*, 69, 3382–3384. doi:10.1103/PhysRevLett.69.3382

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-andrandom-effects effect sizes. *Educational and Psychological Measurement*, 61, 575–604. doi:10.1177/0013164401614003

Fritz, C. O., Morris, P. E., & Richler, J. J. (2011). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141, 2–18. doi: 10.1037/a0024338

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2 ed). New York: Springer.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics*, 33, 485–506. doi: 10.3102/1076998607306151

Grissom, R. J., & Kim, J. J. (2004). *Effect sizes for research: A broad practical approach*. New York: Psychology Press.

Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart, and Winston.

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21, 554–559.

Keppel, G. (1982). *Design and analysis*. Englewood Cliffs, NJ: Prentice Hall.

Keselman, H. J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 16, 44–48. doi: 10.1037/h0081789

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (83–105). Oxford, UK: Blackwell.

Kline, R. B. (2004). *Beyond significance testing*. Washington, DC: American Psychological Association.

Matsumoto, D., Kim, J. J., & Grissom, R. J. (2011). Effect sizes in cross-cultural research. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology*. New York: Cambridge University Press.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3–30. doi: 10.1145/272991.272995

Maxwell, S. E., Camp, J. C., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, 66, 525–534. doi: 10.1037/0021-9010.66.5.525

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective* (2 ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Natesan, P., & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. *Educational and Psychological Mea-*

surement, 67, 59–72. doi: 10.1177/0013164406292028

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241–286. doi: 10.1006/ceps.2000.1040

Pierce, C. A., Block, R. A. & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, 64, 916–924.

R Foundation for Statistical Computing. (2011). R: A Language and Environment for Statistical Computing. Available from http://www.R-project.org/

Schucany, W. R., Gray, H. L., & Owen, D. B. (1971). On bias reduction in estimation. *Journal of the American Statistical Association*, 66, 524–533.

wSnyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, 61, 334–349.

(Received May 4 2013, Revised July 22 2013)