REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 295–313

WHICH EFFECT SIZE MEASURE IS APPROPRIATE FOR ONE-WAY AND TWO-WAY ANOVA MODELS? A MONTE CARLO SIMULATION STUDY

Authors: Soner Yigit
– Faculty of Agriculture, Animal Science Department,
Biometry and Genetics Unit, Canakkale Onsekiz Mart University,
Canakkale, Turkey
soneryigit@comu.edu.tr

Mehmet Mendes
– Faculty of Agriculture, Animal Science Department,
Biometry and Genetics Unit, Canakkale Onsekiz Mart University,
Canakkale, Turkey
mmendes@comu.edu.tr

Received: April 2016 Revised: September 2016 Accepted: September 2016

Abstract:

- It is very important to report some effect size measures that will show if the observed differences among the groups are also of practical significance along with statistical significance while reporting statistical analysis results. Performances of four commonly used effect size measures (Eta-Squared, Partial Eta Squared, Omega Squared and Epsilon Squared) were compared for one and two-way ANOVA models under 3000 different conditions. Results of simulation runs showed that the Epsilon and Omega-Squared estimates were quite unbiased when compared to Eta and Partial Eta-Squared which are directly reported by commonly used statistical packages while reporting ANOVA results. Thus, it could be concluded that reporting Epsilon or Omega-Squared is more appropriate to evaluate the practical significance of observed differences along with P-values.

Key-Words:

- *effect size measure; omega-squared; epsilon-squared; biased, simulation.*

AMS Subject Classification:

- 62-07.

1. INTRODUCTION

The Analysis of Variance Technique (ANOVA-F) is used in comparing the differences between two or more group means [1, 2]. However, it does not show how different the compared group means are from each other or how much of the difference occurred in the dependent variable results from the groups. In other words, while testing the statistical significance of the differences between the levels of independent variable, ANOVA-F test does not give any information about its practical significance [3]. On the other hand, in practice, there is a widespread belief that the smaller the P-value, which is used as the criterion of statistical significance, the more effective or the stronger the levels of the factor the effect of which is researched [4]. Nevertheless, statistical significance is affected by the size of the studied sample. Even very small differences could be found to be statistically significant with very large size samples, large effect sizes may not be found statistically significant with small size samples [5, 6]. Hays [7] reported that the effect size measures are as important as hypothesis testing. Recently, a significant portion of the scientific journals request reporting some effect size measures along with the P-value when reporting statistical analysis results [8] because calculating or estimating the effect size, along with helping in understanding how big the differences between the compared means are, could help in obtaining information about the practical significance of the observed difference and in determining what % of the variation of the analyzed property is described by the considered factor(s). Thus, while reporting analysis of variance results, reporting some effect size measures along with the P-values, which show statistical significance, provides significant benefits [9, 10]. For this purpose, different effect size measures are proposed [7, 11, 12, 13, 14, 15, 16]. The most popular effect size measures for analysis of variance models are found to be $\hat{\eta}^2$ (Eta-Squared), $\hat{\eta}_p^2$ (Partial Eta Squared), $\hat{\omega}^2$ (Omega Squared) and $\hat{\epsilon}^2$ (Epsilon Squared) [3, 8, 9, 17, 18]. However, it is remarkable that the performances of these effect size measures are shown for only one-way analysis of variance and that this is done under quite limited experimental conditions [8, 9, 17, 18]. Moreover, it is a reality that a significant number of the experiments conducted in practice involve in factorial designs. Thus, showing the performances of the aforementioned effect size measures in terms of in factorial design models, as well as the one-way analysis of variance model, would be beneficial. At the same time, contradicting results of the some of the limited studies comparing effect size measures (e.g. [17, 9]) could cause errors. Therefore, performances of the aforementioned effect size measures should be shown in detail under many conditions confronted in practice. Through this, it will be both possible to show the performances of the aforementioned effect size measures under many experimental conditions and to increase the opportunity of generalization of the obtained results. In the study conducted with this point of view, it is aimed to compare the performances of Eta-Squared ($\hat{\eta}^2$), Partial Eta Squared ($\hat{\eta}_p^2$), Omega Squared ($\hat{\omega}^2$) and Epsilon Squared ($\hat{\epsilon}^2$),

which are found as the most popular effect size measures in practice, for one and two-way analysis of variance models. By this means, it will be possible to determine the most convenient effect size measure or measures according to the considered experimental conditions.

2. MATERIAL AND METHOD

Materials for this study consists of random numbers generated by a Monte Carlo simulation technique. In the generation of the random numbers, the RNNOA, RNBET and RNCHI functions of IMSL library of Microsoft Fortran Power Station Developer Studio are used. In this study $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are compared in terms of their performances (bias) under different conditions such as group number or sub-group number, distribution shape, sample size, variance ratio and population effect size. Performances of these effect sizes are determined after 1.000.000 simulation experiments for each of the considered experimental conditions. Experimental conditions considered in the study for One-Way and Two-Way Analysis of Variance models are given together on Table 1 and Table 2.

Table 1: Experimental Conditions for the One-Way Anova.

Statistical model	$Y_{ij}=\mu + \alpha_i + e_{ij}$
Number of Group (k)	3, 4, 5 and 10
Distribution	N(0,1), $\beta(10,10)$, $\beta(5,10)$, $\beta(10,5)$ and $\chi^2(3)$
μ_1:μ_2:...:μ_k	0:0:...:0.30, 0:0:...:0.60, 0:0:...:0.90 and 0:0:...:1.20
σ_1^2:σ_2^2:...:σ_k^2	1:1:...:1, 1:1:...:9, and 1:1:...:20
Number of replication (n)	5, 10, 20, 30 and 50
Number of simulation	1.000.000

Table 2: Experimental Conditions for the Two-Way Anova.

Statistical model	$Y_{ijk}=\mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$
Experimental design (rxc)	2×2, 2×3, 4×2, 3×3, 4×3 and 4×4
Distribution	N(0,1), $\beta(10,10)$, $\beta(5,10)$, $\beta(10,5)$ and $\chi^2(3)$
μ_{11}:μ_{12}:...:μ_{rc}	0:0:...:0.30, 0:0:...:0.60, 0:0:...:0.90 and 0:0:...:1.20
σ_{11}^2:σ_{12}^2:...:σ_{rc}^2	1:1:...:1, 1:1:...:9, and 1:1:...:20
Number of replication (n)	2, 3, 5, 10 and 30
Number of simulation	1.000.000

In order to compare effect size measures in terms of their performances, firstly n numbers are generated from the distributions considered in the study.

Then, generated numbers are subjected to a transformation as $(X_{ij}\text{-}\mu)/\sigma$. Afterwards, certain constant numbers (0.3, 0.6, 0.9 and 1.2) are added to the last group or sub-group in order to create differences between population means. Finally, for all considered experiment conditions and in terms of all effect sizes, population effect size is estimated 1.000.000 times, then means and standard errors are calculated.

2.1. Effect Size Measures

In order to estimate population effect size, many effect size measures are developed. In this study, Eta-squared, Partial Eta-Squared, Omega-Squared, and Epsilon-Squared, which are found as the most popular effect size measures, are taken into consideration [3, 7, 11].

$$\hat{\eta}^2 = \frac{SS_{Effect}}{SS_{Total}}, \tag{2.1}$$

$$\hat{\eta}_p^2 = \frac{SS_{Effect}}{SS_{Total} + SS_{Error}}. \tag{2.2}$$

In a One Way ANOVA-F test, $\hat{\eta}^2$ and $\hat{\eta}_p^2$ are equal [19].

$$\hat{\epsilon}^2 = \frac{SS_{Effect} - df_{Effect} MS_{Error}}{SS_{Total}} \quad [11], \tag{2.3}$$

$$\hat{\omega}^2 = \frac{SS_{Effect} - df_{Effect} MS_{Error}}{SS_{Total} + MS_{Error}} \quad [7], \tag{2.4}$$

where SS_{Total}: Total sum of squares, SS_{Effect}: Sum of squares of effect, SS_{Error}: Error sum of squares, MS_{Error}: Mean square error and df_{Effect}: Degree of freedom of effect.

2.2. Determining Population Effect Size

When determining population effect sizes, Cohen's f value is considered. The relationship between population effect size and Cohen's f value is as follows.

$$\eta^2 = \frac{f^2}{1 + f^2}, \tag{2.5}$$

$$f = \frac{\sigma_\mu}{\sigma}, \tag{2.6}$$

$$\sigma_\mu = \sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \mu)^2}{k}}, \tag{2.7}$$

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^3 + \sigma_3^2}{k}}, \tag{2.8}$$

where σ_μ: Standard deviation of population means, σ: Pooled standard deviation, μ_i: i. population mean, μ: Mean of population means, k: Compared population number [20].

2.2.1. How to get population effect size in the One-Way fixed effects ANOVA model

Table 3: η^2 for the One Way ANOVA-F test.

k	$\mu_1 : \mu_2 :...: \mu_k$	$\sigma_1^2 : \sigma_2^2 : ... : \sigma_k^2$		
		1:1:...:1	1:1:...:9	1:1:...:20
3	0:0:...:0.3	0.01961	0.00543	0.00272
	0:0:...:0.6	0.07407	0.02135	0.01079
	0:0:...:0.9	0.15254	0.04679	0.02396
	0:0:...:1.2	0.24242	0.08027	0.04181
4	0:0:...:0.3	0.01660	0.00559	0.00293
	0:0:...:0.6	0.06323	0.02200	0.01160
	0:0:...:0.9	0.13185	0.04819	0.02573
	0:0:...:1.2	0.21260	0.08257	0.04485
5	0:0:...:0.3	0.01420	0.00551	0.00299
	0:0:...:0.6	0.05446	0.02167	0.01186
	0:0:...:0.9	0.11473	0.04748	0.02629
	0:0:...:1.2	0.18726	0.08140	0.04580
10	0:0:...:0.3	0.00803	0.00448	0.00279
	0:0:...:0.6	0.03138	0.01768	0.01105
	0:0:...:0.9	0.06795	0.03892	0.02452
	0:0:...:1.2	0.11473	0.06716	0.04278

If we want to compare the differences between three population means, it is found as follows
$\mu_1 = 0$, $\mu_2 = 0$ and $\mu_3 = 1.2$,
$\sigma_1^2 = 1$, $\sigma_2^2 = 1$ and $\sigma_3^2 = 20$,

$$\mu = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}}(\mu_\mathrm{i})}{\mathrm{k}} = \frac{0+0+1.2}{3} = 0.4,$$

$$\sigma_\mu = \sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \mu)^2}{k}} = \sqrt{\frac{(0-0.4)^2 + (0-0.4)^2 + (1.2-0.4)^2}{3}} = 0.56568,$$

$$\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^3 + \sigma_3^2}{k}} = \sqrt{\frac{1+1+20}{3}} = 2.70801,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.56568}{2.70801} = 0.20889,$$

$$\eta^2 = \frac{f^2}{1+f^2} = \frac{0.20889^2}{1+0.20889^2} = 0.04181.$$

Population effect sizes calculated in this way for the One Way Analysis of Variance are given on Table 3.

2.2.2. How to get population effect sizes for the Two-Way Fixed Effects ANOVA model

In a 2^2 factorial design,

	c_1	c_2	μ_i σ_i^2
r_1	$\mu_{11} = 0$ $\sigma_{11}^2 = 1$	$\mu_{12} = 0$ $\sigma_{12}^2 = 1$	$\mu_{1.} = 0$ $\sigma_{1.}^2 = 1$
r_2	$\mu_{21} = 0$ $\sigma_{21}^2 = 1$	$\mu_{22} = 1.2$ $\sigma_{22}^2 = 20$	$\mu_{2.} = 0.6$ $\sigma_{2.}^2 =?$

If $X = \{x_1, x_2, ..., x_N\}$, $Y = \{y_1, y_2, ..., y_N\}$ and $Z = \{x_1, x_2, ..., x_N, y_1, y_2, ..., y_N\}$, As is known, $\mu_Z = (\mu_X + \mu_Y)/2$. If $\mu_X = \mu_Y$, $\sigma_Z^2 = (\sigma_X^2 + \sigma_Y^2)/2$. However, if $\mu_X \neq \mu_Y$, then $\sigma_Z^2 \neq (\sigma_X^2 + \sigma_Y^2)/2$.
If $\mu_X \neq \mu_Y$ then,

$$\sigma_Z^2 = \frac{\mu_X^2 + \sigma_X^2 + \mu_Y^2 + \sigma_Y^2 - 2\mu_Z^2}{2}. \tag{2.9}$$

If k population is considered as one population, then the variance of the obtained new population is calculated as follows.

$$\sigma_Z^2 = \frac{\sum_{i=1}^{k}(\mu_i^2 + \sigma_i^2) - k\mu_Z^2}{k}. \tag{2.10}$$

This formula is empirically verified. In that case, since $\mu_{21} \neq \mu_{22}$ and $\sigma_{2.}^2 \neq (\sigma_{21}^2 + \sigma_{22}^2)/2$, then

$$\sigma_{2.}^2 = \frac{\mu_{21}^2 + \sigma_{21}^2 + \mu_{22}^2 + \sigma_{22}^2 - 2\mu_{2.}^2}{2} = \frac{0^2 + 1 + 1.20^2 + 20 - 2(0.6^2)}{2} = 10.86.$$

Similarly, it is found that $\sigma^2_{.2} = 10.86$.
In factorial experiments, following equality is valid for the population effect size.

$$\eta^2_{Model} = \eta^2_r + \eta^2_c + \eta^2_{rxc}. \tag{2.11}$$

Thus, in order to find the effect size in terms of interaction (η^2_{rxc}), first η^2_{Model}, η^2_r and η^2_c should be calculated.
Calculation of η^2_{Model}

$$\mu = \frac{\sum_i^r \sum_j^c \mu_{ij}}{rc} = \frac{0+0+0+1.2}{4} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^r \sum_j^c (\mu_{ij}-\mu)^2}{rc}}$$

$$= \sqrt{\frac{(0-0.3)^2+(0-0.3)^2+(0-0.3)^2+(1.2-0.3)^2}{(2)(2)}} = 0.51961,$$

$$\sigma = \sqrt{\frac{\sigma^2_{11}+\sigma^2_{12}+\sigma^2_{21}+\sigma^2_{22}}{rc}} = \sqrt{\frac{1+1+1+20}{(2)(2)}} = 2.39791,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.51961}{2.39791} = 0.21669,$$

$$\eta^2_{Model} = \frac{f^2}{1+f^2} = \frac{0.21669^2}{1+0.21669^2} = 0.04485.$$

Calculation of η^2_r

$$\mu = \frac{\sum_i^r \mu_{i.}}{r} = \frac{0+0.6}{2} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^r (\mu_{i.}-\mu)^2}{r}} = \sqrt{\frac{(0-0.3)^2+(0.6-0.3)^2}{2}} = 0.3,$$

$$\sigma = \sqrt{\frac{\sigma^2_{1.}+\sigma^2_{2.}}{r}} = \sqrt{\frac{1+10.86}{2}} = 2.43516,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.3}{2.43516} = 0.12319,$$

$$\eta^2_r = \frac{f^2}{1+f^2} = \frac{0.12319^2}{1+0.12319^2} = 0.01495.$$

Calculation of η^2_c

$$\mu = \frac{\sum_j^c \mu_{.j}}{c} = \frac{0+0.6}{2} = 0.3,$$

$$\sigma_\mu = \sqrt{\frac{\sum_i^c (\mu_{.j}-\mu)^2}{c}} = \sqrt{\frac{(0-0.3)^2+(0.6-0.3)^2}{2}} = 0.3,$$

$$\sigma = \sqrt{\frac{\sigma^2_{.1}+\sigma^2_{.2}}{c}} = \sqrt{\frac{1+10.86}{2}} = 2.43516,$$

$$f = \frac{\sigma_\mu}{\sigma} = \frac{0.3}{2.43516} = 0.12319,$$
$$\eta_c^2 = \frac{f^2}{1+f^2} = \frac{0.12319^2}{1+0.12319^2} = 0.01495.$$

Calculation of η_{rxc}^2

$$\eta_{rxc}^2 = \eta_{Model}^2 - \eta_r^2 - \eta_c^2 = 0.04485 - 0.01495 - 0.01495 = 0.01495.$$

Population effect sizes calculated in this way for Interaction Effect are given on Table 4.

Table 4: η_{rxc}^2 for the Two-Way ANOVA-F test.

k	$\mu_1 : \mu_2 :...: \mu_k$	$\sigma_1^2 : \sigma_2^2 : ... : \sigma_k^2$		
		1:1:...:1	1:1:...:9	1:1:...:20
2×2	0:0:...:0.3	0.00553	0.00186	0.00098
	0:0:...:0.6	0.02108	0.00733	0.00387
	0:0:...:0.9	0.04395	0.01606	0.00858
	0:0:...:1.2	0.07087	0.02752	0.01495
2×3	0:0:...:0.3	0.00494	0.00213	0.00120
	0:0:...:0.6	0.01905	0.00839	0.00474
	0:0:...:0.9	0.04045	0.01840	0.01052
	0:0:...:1.2	0.06667	0.03158	0.01832
3×3	0:0:...:0.3	0.00441	0.00234	0.00142
	0:0:...:0.6	0.01717	0.00924	0.00565
	0:0:...:0.9	0.03704	0.02032	0.01253
	0:0:...:1.2	0.06226	0.03501	0.02186
4×2	0:0:...:0.3	0.00418	0.00210	0.00125
	0:0:...:0.6	0.01624	0.00827	0.00494
	0:0:...:0.9	0.03488	0.01818	0.01096
	0:0:...:1.2	0.05832	0.03129	0.01911
4×3	0:0:...:0.3	0.00372	0.00224	0.00145
	0:0:...:0.6	0.01460	0.00885	0.00575
	0:0:...:0.9	0.03178	0.01953	0.01276
	0:0:...:1.2	0.05405	0.03377	0.02228
4×4	0:0:...:0.3	0.00315	0.00210	0.00144
	0:0:...:0.6	0.01239	0.00832	0.00573
	0:0:...:0.9	0.02719	0.01840	0.01274
	0:0:...:1.2	0.04669	0.03195	0.02228

3. RESULTS

In this study, five different distribution shapes, three different variance ratios, five different sample sizes, four different effect size magnitudes, four group combinations (k=3, 4, 5 and 10) in one-way analysis of variance analysis, six subgroup combinations (2×2, 2×3, 3×3, 4×2, 4×3 and 4×4) in two-way analysis of variance, totally 3000 different experimental conditions are considered. Thus, all of the results could not be presented in the essay. Obtained results are given on Figure 5-24 for One-way analysis of variance, on Figure 25-54 for Two-way analysis of variance, and on Supplementary Appendix together. Furthermore, some experiment results that reflect the results significantly are summarized on Figure 1 and 2 for the One-way analysis of variance and on Figure 3 and 4 for the Two-way analysis of variance.

3.1. Results of The One-Way Analysis of Variance

Comparing independent group means that are taken from normal distribution and variances of which are homogeneous, while $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give quite unbiased results, $\hat{\eta}^2$ gives quite biased results. Besides, as long as the variances are homogeneous, slight [$\beta(10, 10)$] or moderate [$\beta(5, 10)$ and $\beta(10, 5)$] deviations from normality does not affect the realized estimations in terms of the three effect sizes. Under these conditions, although there is a negligible difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$, $\hat{\epsilon}^2$ gives the most unbiased estimations. When variances are homogeneous, excessive skewness and kurtosis [$\chi^2(3)$] affect estimations of the three effect sizes negatively. However, both $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give more unbiased estimations compared to $\hat{\eta}^2$. Although $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give results quite close to each other, $\hat{\omega}^2$ gives more unbiased results compared to $\hat{\epsilon}^2$ under these experimental conditions. When variances are homogeneous, regardless of the distribution shape and sample size, as the number of groups increase, estimations of $\hat{\eta}^2$ diverge from η^2, whereas estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ approach to η^2. Additionally, depending on the increase in group number, differences between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ decrease gradually. For example; when n=10 and k=3, 4, 5 and 10, bias of $\hat{\eta}^2$ ranges between 4.80-6.45%, 5.76-6.95%, 6.39-7.30% and 7.95-8.30%, bias of $\hat{\epsilon}^2$ ranges between 0.40-1.3%, 0.29-0.97%, 0.24-0.72% and 0.08-0.28% and bias of $\hat{\omega}^2$ ranges between 0.87-1.0%, 0.62-0.71%, 0.43-0.54% and 0.18-0.21% the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ when variances are heterogeneous is smaller than when variances are homogeneous. In case variances get heterogeneous too, $\hat{\eta}^2$ gives quite biased and irregular results. Choosing compared groups from symmetric distributions [N(0,1) and $\beta(10, 10)$] and heterogeneous variances do not affect estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ negatively. In addition to this, the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ while variances are

heterogeneous is smaller than that while variances are homogeneous. However, bias of the distribution from symmetry increased the bias of the estimations made by $\hat{\epsilon}^2$ and $\hat{\omega}^2$ a little. This situation becomes much more significant especially when variances are excessively heterogeneous (20 times). However again, they give quite unbiased results compared to $\hat{\eta}^2$. When variances are heterogeneous, while an increase in group number negatively affects $\hat{\eta}^2$, but does not affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$ significantly. Regardless of the compared group number, distribution shape, variance ratios and population means, depending on the increase in sample size, it is seen that estimations gradually approach to η^2 in terms of the three effect size ($\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$). Furthermore, it is seen that the most biased results are given by $\hat{\eta}^2$ under all considered experimental conditions. Additionally, as the difference between means decreases, in other words as the population effect size decreases (η^2), while bias of the estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually decrease, bias of the estimations of $\hat{\eta}^2$ gradually increase. Regardless of the experimental conditions, as the sample size decreases (especially when n=5), estimations show severe bias in terms of $\hat{\eta}^2$ (Figure 1). $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are affected by the sample size less than $\hat{\eta}^2$ (Figure 1 and Figure 2).

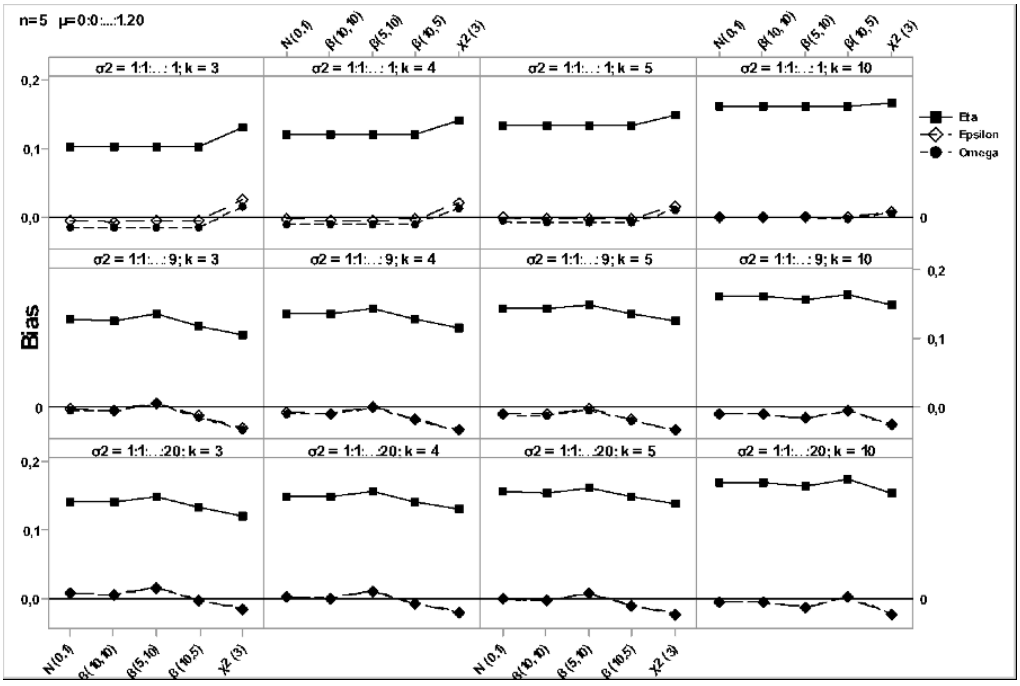

Figure 1: Bias for the One-Way ANOVA models when n=5 and μ=0:0:...:1.20.

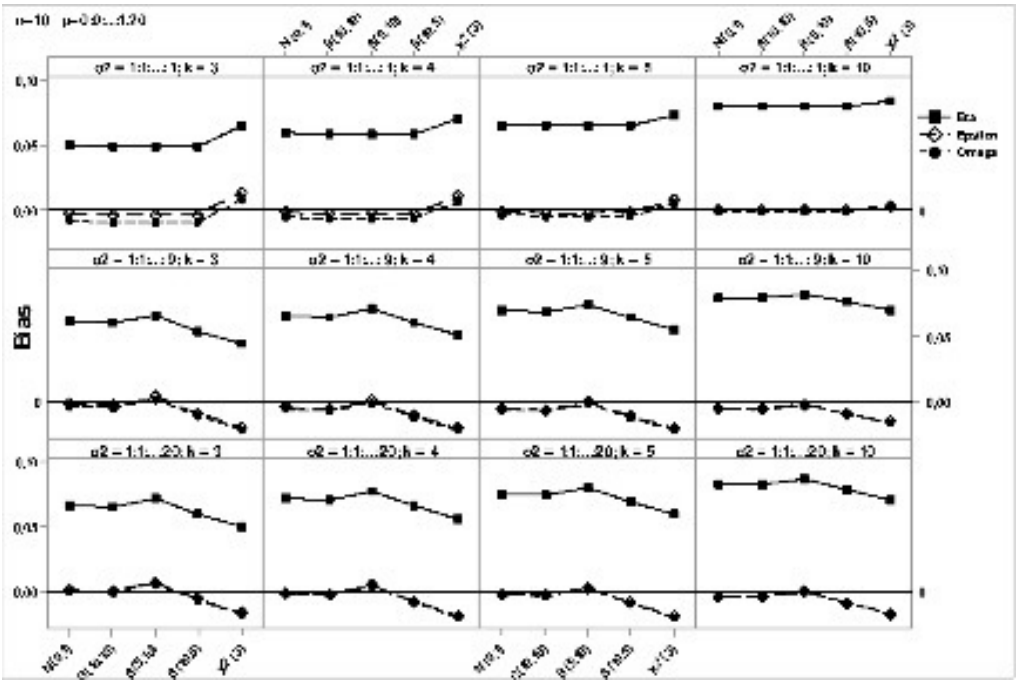

Figure 2: Bias for the One-Way ANOVA models when n=10 and μ=0:0:...:1.20.

3.2. Results of The Two-Way Analysis of Variance

With small size (n≤10) sub-groups that are taken from normal distribution and with homogeneous variances, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ show excessive bias (Figure 4). However, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give significantly unbiased results. When variances are homogeneous, having slight and moderate deviance from normality does not affect the estimations of the four effect size measures. Besides, $\hat{\epsilon}^2$ gives the most unbiased results. If there is excessive skewness and kurtosis [$\chi^2(3)$], $\hat{\omega}^2$ gives the most unbiased results. However, in both cases, the difference between them is negligible. When variances are heterogeneous, regardless of the distribution of the populations they are taken from, $\hat{\eta}_p^2$ gives the most biased results under all of the considered experimental conditions, and $\hat{\eta}^2$ follows it. Additionally, as the variances get heterogeneous, $\hat{\eta}_p^2$ and $\hat{\eta}^2$ approach each other. When variances are heterogeneous, in cases where the population sub-groups are taken from are N(0,1), $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give the most unbiased results. When variances are heterogeneous, slight deviance from normality [$\beta(10, 10)$] does not affect the performances of $\hat{\epsilon}^2$ and $\hat{\omega}^2$. However, increase of the deviance of the distribution from normality increased the bias of these two effect size measures as well. This situation is seen significantly when variances are excessively heterogeneous. When variances are generally homogeneous, regardless of the experimental conditions, in cases where

the number of studied sub-groups increase, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ gradually diverge from η^2, whereas estimations of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually approach to η^2. Furthermore, as the number of sub-group increase, the difference between $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gradually decreased. Considering the sub-groups with heterogeneous variances, regardless of the distribution shape, as the number of sub-groups increase, estimations of $\hat{\eta}_p^2$ and $\hat{\eta}^2$ gradually diverge from the population effect size. On the other hand, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not affected significantly from the increase in the sub-group number when variances are heterogeneous. Regardless of the experiment design (rxc), variance ratios, distribution shapes and sample size, the most biased estimations are made by $\hat{\eta}_p^2$, and $\hat{\eta}^2$ follows as a similar pattern (Figure 3 and 4).

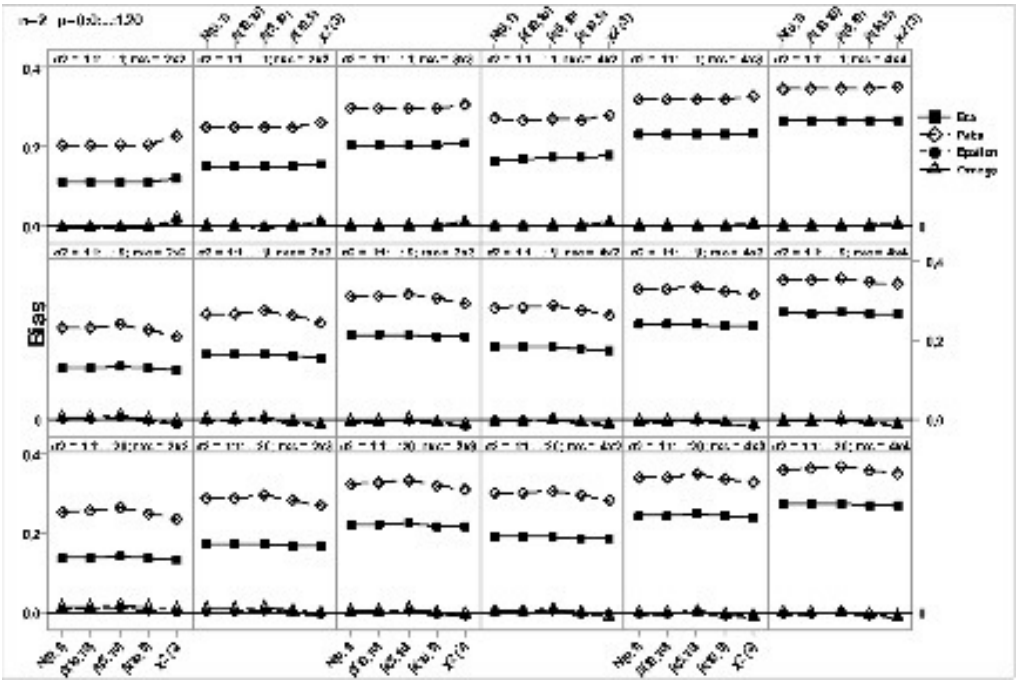

Figure 3: Bias for the Two-Way ANOVA models when n=2 and μ=0:0:...:1.20.

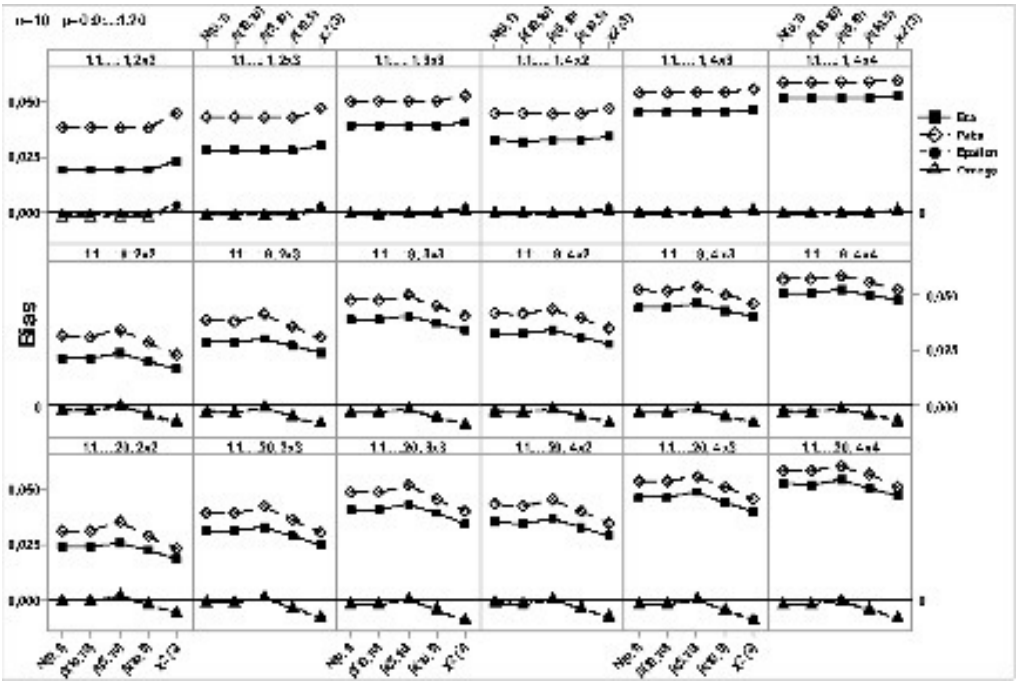

Figure 4: Bias for the Two-Way ANOVA models when n=10 and μ=0:0:...:1.20.

4. DISCUSSION

The Analysis of variance technique used most commonly in practice gives information about statistical significance only. It does not give information about practical significance of the factors and explained variance. Thus, when reporting analysis of variance results, reporting only P-values showing statistical significance will not be sufficient. Along with the P-values, some effect size measures such as $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ that show the practical significance and the share of the difference observed in the dependent variable explained by the considered factors should be reported. Thereby, understanding and interpreting the reported results in detail will be possible. There are many effect size measures developed for the purpose. However, it is an important shortcoming that performances of these effect size measures under many experimental conditions have not been shown in detail. Nevertheless, having detailed information about the performances of the effect size measures will provide insights to the researchers about which effect size measure they should report as a result of their studies. In the study conducted with this point of view, performances of $\hat{\eta}^2$, $\hat{\eta}_p^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$, which are found as the most popular effect size measures, are compared. Baguley [21] reported that

simple or unstandardized effect size measures are easier to compute and more robust than standardized effect size measures. Therefore, he has proposed to report simple effect size measures. However, in practice, standardized effect size measures have been commonly reported. For example, commonly used statistical package programs such as IBM SPSS, Minitab, Statistica and SAS report standardized effect size estimates along with P-values. We think that it is very easy to understand and interpret the effect size values for many authors and readers. That is why, in the simulation study conducted to compare performances of $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ effect size measures for one-way analysis of variance. Keselman [17] reported that $\hat{\eta}^2$ gives similar results as $\hat{\epsilon}^2$and $\hat{\omega}^2$ in case of small population effect size. On the other hand, considering the standard deviations of estimations, he reported that $\hat{\eta}^2$ is a better estimator compared to $\hat{\epsilon}^2$ and $\hat{\omega}^2$. Nonetheless, in all other studies, it is reported that using $\hat{\eta}^2$ in estimating population effect size gives quite biased results [18, 8, 9]. In the results of our study too, it is seen that $\hat{\eta}^2$ gives quite biased results in all considered experimental conditions. In his simulation study, Keselman [17] stated that as long as the assumption of homogeneity of variances is met, selecting samples from populations with high skewness ($\gamma_1 = 2$) and kurtosis ($\gamma_2 = 6$) does not affect the performances of $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ significantly. In their simulation study, Skidmore and Thompson [8] reported that even if the variances are heterogeneous, slight ($\gamma_1 = 0.5$ and $\gamma_2 = 0.5$ and moderate ($\gamma_1 = 1$ and $\gamma_2 = 3.75$)) level deviation from normality does not affect the performances of effect size measures ($\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$) significantly. As a result of our study too, it is seen that when variances are homogeneous as long as there is not excessive (χ^2) deviation from normality, shape of the distribution does not affect performances of the effect size measures significantly. However, in case variances are heterogeneous, it is seen that moderate [$\beta(5, 10)$ and $\beta(10, 5)$] and excessive (χ^2) deviations from normality affect effect size measures. Keselman [17] reported that while $\hat{\omega}^2$ could decisively provide estimations quite close to population effect size, $\hat{\epsilon}^2$ always produces estimations a little higher than that. However, Keselman did not report the number of replication in his study. In his simulation study aimed to compare some effect size measures, Okada [9] repeated Keselman's study with larger simulation number (1 million), at normal distribution and with different observation number combinations and stated that in all considered experimental conditions $\hat{\epsilon}^2$ gives more unbiased results than $\hat{\omega}^2$. In one of their studies, Glass and Hakstian [3] theoretically discussed whether $\hat{\epsilon}^2$ or $\hat{\omega}^2$ is unbiased and expressed that no matter how different their formulas are, both of them give similar results in practice. In the results of our study, in one-way variance analysis, as long as there is not excessive deviance from normality, it is seen that generally $\hat{\epsilon}^2$ gives the most unbiased results, and $\hat{\omega}^2$ follows it. However, in case of excessive deviances (χ^2) from normality, the most unbiased results are obtained by $\hat{\omega}^2$. On the other hand, in both situations, the difference between $\hat{\epsilon}^2$and $\hat{\omega}^2$ is negligible and confirms Glass and Hakstian [3]. As a result of the conducted simulation study, if the observation numbers in groups are equal and distributions are not excessively skewness and kurtosis, it is seen that heteroge-

neous variances do not affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$ almost at all. Carrol and Nordholm [18] reported similar results. However, both Carrol and Nordholm [18] and Skidmore and Thompson [8] reported that heterogeneous variances are especially effective at unequal sample sizes (direct and inverse pairing). As long as the variances are homogeneous, regardless of the considered experimental conditions, it is seen that as the compared group numbers increase, deviances of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ in estimations approach to zero. Thus, in these experimental conditions, an increase in group numbers positively affect $\hat{\epsilon}^2$ and $\hat{\omega}^2$. However, in case of heterogeneous variances, making assessments on whether the increase in group numbers have positive or negative effects on estimations could be misleading. On the other hand, regardless of if the assumption of homogeneity of variances is met or not, is significantly affected by an increase in group numbers. Results obtained under these conditions overlap with the findings of Skidmore and Thompson [8]. It is reported that as the population effect size decreases, biasness decreases too [17]. However, as the population effect size decreases, while $\hat{\eta}^2$ gives more biased results, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ gives more unbiased results. When estimating effect sizes related to interaction effects in factorial experiments, performance of $\hat{\eta}_p^2$ is investigated in addition to $\hat{\eta}^2$, $\hat{\epsilon}^2$ and $\hat{\omega}^2$. Whatever the experimental conditions considered in the study, as the sample size increases, estimations of the four effect sizes approach gradually to the population effect size. However, in all of the considered experimental conditions, it is seen that while $\hat{\eta}_p^2$ gives the most biased results, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ give the most unbiased results. Furthermore, while $\hat{\eta}_p^2$ and $\hat{\eta}^2$ are negatively affected by the increase in sub-group number, $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not negatively affected. In the meantime, it is remarkable that effects of both shape of distribution and variance rates on the considered effect size measures are generally similar to the ones in one-way variance analyses. On the other hand, our study has revealed the performances of the considered effect size measures in factorial ANOVA models. Thus, the study has fulfilled an important need in this field because factorial ANOVA design is commonly used in practice.

5. CONCLUSION AND RECOMMENDATIONS

Reporting statistical analysis results in an understandable and informative way is very important. Therefore, when reporting statistical analysis results, along with the P-value that shows statistical significance some effect size measures should be reported. While a statistically significant difference is not necessarily practically significant, a statistically non-significant difference is not necessarily practically non-significant. Notwithstanding, majority of researchers believe that the smaller the P-value is that shows the statistical significance, the larger and the more important the difference between the groups that are compared. However, the P-value does not provide any information about practical significance. Thus, in the results of the studies, along with the statistical significance (P-value), effect

size measures that provide information about the practical significance should necessarily be reported. However, it is remarkable that majority of the researchers who report effect size report $\hat{\eta}^2$ ($\mathbb{R}^2$) and $\hat{\eta}_p^2$ [22]. This is because commonly used statistics package programs such as Minitab, IBM SPSS, NCSS, Statistica etc. directly report $\hat{\eta}^2$ ($\mathbb{R}^2$) or $\hat{\eta}_p^2$ while reporting analysis of variance results. However, the most noteworthy thing is that reported effect size measures should represent population effect size as accurately as possible (unbiased). From this point forth, performances of the most commonly known effect size measures in practice are compared under many experimental conditions in one-way and two-way analysis of variance models. In the light of the acquired findings, concluding with following results is possible:

1. In both one factor and two factor experimental conditions, $\hat{\eta}^2$ gives quite biased results. Thus, since using $\hat{\eta}^2$ to estimate population effect size at the end of analysis of variance is quite misleading, reporting $\hat{\eta}^2$ should not be recommended.

2. Although $\hat{\eta}_p^2$ is used in experimental conditions considering more than one factor as an alternative to $\hat{\eta}^2$, it is seen that $\hat{\eta}_p^2$ gives more biased results than $\hat{\eta}^2$ in two factor experiments after 1.000.000 simulation experiments. Additionally, since $\hat{\eta}_p^2$ takes every effect separately in consideration ($SS_{Effect} + SS_{Error}$), total variation explained by the model could surpass 1 (100%) [23, 24]. This is a common situation in practice [25]. Since $\hat{\eta}_p^2$ estimates of effect size are biased, reporting it should not be recommended.

3. Although Okada [9] reported that relationships among Eta, Omega and Epsilon-squared is $\hat{\omega}^2 \leq \hat{\epsilon}^2 \leq \hat{\eta}^2$, this relation is not valid for every experimental condition. For example, the relationship between Epsilon and Omega squared is $\hat{\epsilon}^2 \leq \hat{\omega}^2$ when negative estimations are obtained regardless of experimental conditions.

4. Although it is seen that in some of the experimental conditions $\hat{\epsilon}^2$ and in some of the others $\hat{\omega}^2$ gives more unbiased results, the difference between these two measures is at a negligible level. It is seen that both $\hat{\epsilon}^2$ and $\hat{\omega}^2$ estimates population effect size in a quite unbiased fashion in all experimental conditions. Thus, it could be concluded that when estimating effect size in analysis of variance models and accordingly analyzing practical significance of the observed difference, using $\hat{\epsilon}^2$ or $\hat{\omega}^2$ is much truer and one of these measures should be reported.

5. Obtaining negative estimates in some experimental conditions (i.e. small effect size magnitude) may be considered a disadvantage of $\hat{\epsilon}^2$ and $\hat{\omega}^2$ estimates, although both measures give unbiased estimates almost all experimental conditions.

6. It is determined that it is a very important deficiency that $\hat{\epsilon}^2$ and $\hat{\omega}^2$ are not included in almost none of the commonly used statistics package programs although $\hat{\eta}^2$ and $\hat{\eta}_p^2$ has been reported to be quite biased in studies for 50 years. Thus, at least one of these two measures should be included in the libraries of commonly used package programs such as Minitab, SPSS etc.

REFERENCES

[1] Zar, H.J. (2010). *Biostatistical Analysis*, Pearson, New Jersey.

[2] Mendes, M. and Yigit, S. (2013). Comparison of ANOVA-F and ANOM tests with regard to type I error rate and test power, *Journal of Statistical Computation and Simulation*, **83**(11), 2093–2104.

[3] Glass, G.V. and Hakstian, A.R. (1969). Measures of Association in Comparative Experiments: Their Development and Interpretation, *American Educational Research Journal*, **6**(3), 403–414.

[4] Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, **5**(2), 241–301.

[5] Fan, X. (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin, *The Journal of Educational Research*, **94**(5), 275–282.

[6] Mendes, M. (2012). *Uygulamali Bilimler Icin Istatistik Ve Arastirma Yontemleri*, Kriter, Canakkale.

[7] Hays, W. (1963). *Statistics for Psychologists*, Rinehart Winston, New York.

[8] Skidmore, S.T. and Thompson, B. (2012). Bias and precision of some classical ANOVA effect sizes when assumptions are violated, *Behavior Research Methods*, **45**(2), 536–546.

[9] Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA, *Behaviormetrika*, **40**(2), 129–147.

[10] Cumming, G. and 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**(4), 532–574.

[11] Kelley, T.L. (1935). An Unbiased Correlation Ratio Measure, *Proceedings of the National Academy of Sciences of the United States of America*, **21**(9), 554–559.

[12] Maxwell, S.E.; Camp, C.J. and Arvey, R.D. (1981). Measures of strength of association: A comparative examination, *Journal of Applied Psychology*, **6**(5), 525–534.

[13] Keppel, G. (1982). *Design and Analysis: A Researcher's Handbook*, Prentice Hall, New Jersey.

[14] Olejnik, S. and Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations, *Contemporary Educational Psychology*, **25**(3), 241–286.

[15] Kirk, R.E. (2003). *The importance of effect magnitude*. In "Handbook of Research Methods in Experimental Psychology" (S.F. Davis, Ed.), Oxford, Blackwell, 83–105.

[16] Grissom, R. and Kim, J. (2005). *Effect Sizes for Research: A Broad Practical Approach*, Lawrence Erlbaum, New Jersey.

[17] Keselman, H.J. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared, *Canadian Psychological Review/Psychologie Canadienne*, **16**(1), 44–48.

[18] Carroll, R.M. and Nordholm, L.A. (1975). Sampling Characteristics of Kelley's epsilon and Hays' omega, *Educational and Psychological Measurement*, **35**(3), 541–554.

[19] Levine, T.R. and Hullett, C.R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research, *Human Communication Research*, **28**(4), 612–615.

[20] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, New Jersey.

[21] Baguley, T. (2009). Standardized or simple effect size: What should be reported?, *British Journal of Psychology*, **100**(3), 603–617.

[22] Fritz, C.O.; Morris, P.E. and Richler, J.J. (2012). Effect size estimates: current use, calculations, and interpretation, *Journal of Experimental Psychology: General*, **141**(1), 2–18.

[23] Cohen, J. (1973). Eta-Squared and Partial Eta-Squared in Fixed Factor Anova Designs, *Educational and Psychological Measurement*, **33**(1), 107–112.

[24] Haase, R.F. (1983). Classical and Partial Eta Square in Multifactor ANOVA Designs, *Educational and Psychological Measurement*, **43**(1), 35–39.

[25] Pierce, C.A.; Block, R.A. and Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs, *Educational and Psychological Measurement*, **64**(6), 916–924.