**Okada (2013)**

**Is Omega Squared Less Biased?**

* The measure of magnitude of effect is called effect size. It is the minimum expectations for all APA journals
* The detailed characteristics of effect size indices have received relatively little attention. There are more than seventy varieties of effect size indices (Kirk, 2003)
* In this paper, we consider a one-way fixed effect ANOVA with independent samples.
* Eta squared 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 – cite someone different).
* There are three major sample effect size indices in ANOVA (Grissom & Kim, 2004; Matsumoto, Kim, & Grissom, 2011; Keppel, 1982; Olejnik & Algina, 2000): eta squared, epsilon squared, and omega squared.
* #gives formulas for each effect size#
* It is known that eta squared overestimates the population effect size, because its numerator, SSb is inflated by some error variability (Grissom & Kim, 2004).
  + Snyder and Lawson (1993) and Maxwell and Delaney (2004) give a detailed discussion of this bias.
* It is also shown that as sample size tends to infinity, eta, epsilon, and omega converge toward the same value (Maxwell et al., 1981). However, little is known about their comparative behaviors in realistic finite small sample settings.
* Keselman (1975) studied the performance of eta, epsilon, omega in a onte carlo study using 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 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 for eta squared”.
  + Grissom and Kim (2004) agree with Keselman’s findings “a somewhat less biased alternative estimator of population effect size is epsilon, and a more nearly unbiased estimator is omega.
  + Matsumoto et al (2011) agrees
* Olejnik & Algina (2000) explain the order of bias among the three idices
  + “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
* Carroll and Nordholm (1975) monte carlo study contradicted Keselman (1975). Putting aside the fact that their study used a smaller number of replications per condition than Keselman which may have resulted in more sampling errors, their results implied that while omega is slightly negatively biased, any bias in epsilon is not evident.
* **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)**
* The 5,000 and 1,000 replications per condition used by Keselman and Carrol/Nordholm, respectively, may not be sufficient for today’s standards.
  + **They also evaluated the effect size indices in terms only of their means and standard deviations. Since SD is the mean squared deviation from the sample mean and not from the true value population eta, it may be insufficient for evaluating the sample effect size indices. Because SD does not take pop eta into account, there may be cases in which the SD is small but samples are substantially biased from the true value.**
* **Current study compared the performance of ES by evaluating the bias and errors under the correct model using modern computers. Used the same condition as Keselman, but increased the number of replications per condition to 1 million. We also added a sample size condition to assess effect of sample size to the bias of ES indices.** 
  + **In addition, we also calculated the root mean squared errors (RMSEs). While SD 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.** 
    - **If the sample mea of estimator exactly corresponds to true value, there is zero bias and SD=RMSE.**
    - **By using RMSE, we account for errors in estimating the population ES.**
* **In Keselman’s study, the relatively smaller number of replications has led to a number of potential errors. For instance, epsilon overestimates the population effect size in large treatments and maximum variability conditions and underestimates it in other conditions. Because the hypothesized model is correct in all the conditions, this result can be due to the sampling errors to the small number of replications. With 1 million, we expect that sampling error in our study will be very minimal.**
* **Generated a random number of the artificial datasets from normal distributions with predetermined means shown in table 1 and SD of 1 using the Mersenne twister algorithm.**
* **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 and repeated measures designs, future studies need to further verify this expectation.**

**Kirk (2003)**

* NHST doesn’t tell researchers what they want to know. We want to know the probability that the null hypothesis is true given that we have obtained a set of data. What NHST tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true.
  + Second criticism is that All null hypotheses are false. Tukey (1991) wrote the effects of A and B are always different – in some decimal place – for any A and B. Thus asking ‘Are the effects different’ is foolish.
  + Third criticism is making a dichotomous decision from a continuum of uncertainty. By adopting a fixed level of significance, a researcher turns a continuum of uncertainty into a dichotomous reject-do not reject decision (Frick, 96; Grant, 62; Rossi, 97; Wickens, 98)
* Measures of effect magnitude
  + Cohen’s d, f, g, h, q, w
  + Glass’s g and Hedges g
* Measures of strength of association
  + r r2 R R2 epsilon, eta, comega, Cohen’s f2, cramer’s v, fisher’s Z
  + Provide a measure of the strength of the association between the independent and dependent variables. Hays (1963) introduced omega squared.
  + O’Grady (1982) states that omega squared may underestimate the true proportion of explained variance.

**Maxwell & Delaney (2004)**

* One can argue, as Hays (1994) does, that what is of most interest is the proportion of variance in the population that would be accounted for by the treatments. If this is granted, then characteristics of r2 as an estimator must be considered. In this regard, recall that the numerator of R2 depends on the variability among the group means….However, even if the population-group means were identical, the sample means would almost certainly differ from each other. Thus, although in the population the treatments may account for no variance, R2 would nonetheless be expected to be greater than zero because of this sampling variability in the observed means. This positive bias of R2 can be estimated and is a decreasing function of sample size.
  + The other measures of association like omega squared attempt to correct for this positive bias by shrinking the numerator.
  + Maxwell et al. (1981) show that the value of R2 is typically within .02 of omega squared.

**Keselman (1975)**

* Omega squared is a more accurate estimator while eta squared has the smallest sampling variability

**Olejnik & Algina (2000)**

* **Some ES are not comparable across different designs.**
* Omega squared meaures problem is they were derived from the variance components obtained from the expected mean squares for the sources of variation in the model. The expected mean squares assume a balanced design. In many applied research contexts sample sizes are not equal and are generally disproportionate. Vaughan and Corballis (1969) cautioned against the use of omega squared in these situations. Carroll and Nordholm (1975), however, found that in a single-factor design, unequal n had little effect on the estimation of epsilon or omega if variances are equal, and with equal n heterogeneous variances had little impact on the estimation of these effect sizes. But unequal n and heterogeneous variances lead to an overestimation of the effect size, and Carroll and Hordholm cautioned against their use in these situations.
* Fern and Monroe (1996)
  + 1. Low reliability increases the error variance and puts a limit on the amount of variance that can be explained by an explanatory variable. Two measures of effect from two studies of the same explanatory variable can be substantially different if the outcome measures used have substantially different reliabilities.
  + 2. Population heterogeneity can reduce the magnitude of the effect size measure. The effect sizes computed in two studies that differ with respect to the variability of the outcome measure may not be comparable.
  + 3. In fixed effects models the magnitude to the omnibus measures of effect size depends on the specific levels of the variables studied. If different levels of the explanatory variable are investigated, the measure of effect will not be comparable.
  + 4. The quantification of the strength of a treatment is generally unknown. Sechrest and Yeaton (1982) argue that without knowledge of the strength of a treatment the proportion of variance accounted for is meaningless.
  + 5. The range of treatments included in a study can increase or reduce the proportion of variation explained.

**Carrol and Nordholm (1975)**

* Statistics were found to have rather large standard errors when small samples were used. Large samples are recommended for accuracy of estimation. Epsilon and omega were negligibly biased. Heterogeneity of variances had negligible effects on estimates, but combination of heterogeneity of variance and unrepresentative sample sizes yielded poor estimates.

**Okada & Hoshino (2016)**

* **Effect sizes depend on the choice of the number of levels and their ranges in experiments.**
* **The variance-accounted-for effect sizes would be substantially affected by the basic research design such as the number of levels.**

**Okada (2016)**

* **A bias-corrected estimate can take a negative value, and of course, a negative variance ratio does not make sense.**
* **Negative estimates occur more tha half the time under some reasonable conditions. Treating the obtained negative estimates as zero causes substantial overestimation of even bias-corrected estimators when the sample size and population effect are not large.**
* **Suggestion is to report as is.**

**Lakens (2013)**

* Effect sizes are the most important outcome of empirical studies.
* Allow researchers to present the magnitude of the reported effects in a standardized metric which can be understood regardless of the scale that was used to measure the dependent variable….Allows to communicate the practical significance of their results instead of only reporting the statistical significance.
* ES allow researchers to draw meta-analytic ocnclusions by comparing standardized effect sizes across studies.
* ES can be used when planning a new study.
* The d family are based on the difference between observations divided by the standard deviation of these observations. The r family ES describe the proportion of variance that is explained by a group membership. R is calculated from the sum of squares divided by the sum of squares for other factors in the design.
* Population effect sizes are almost always estimated on the basis of samples, and all population effect size estimates based on sample averages overestimates the true population effect (Thompson, 2006).
* **Keppel (1991) recommended partial eta squared to improve the comparability of effect sizes between studies.**

**Thompson (2007)**

* Sample-based estimates of population effect sizes tend to be positively biased, because the ordinary least squares OLS analyses cannot discriminate between variability in the sample data that are real (represent true population variability) and sample data that are unique to a given sample (does not exist in the population, or if any other sample).
* Sample data have more idiosyncrasies (more sampling error variance) when a) sample size is small, b) the number of measured variables is large, and c) population effect sizse is small.
* Roger Kirk (2003)
  + It is evident that the current practice of focusing exclusively on a dichotomous reject-nonreject decision strategy of NHST can actually impede scientific progress… in fact, focusing on p values and rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypotheses and are practically significant. The focus of research should be on our scientific hypotheses, what data tell us about the magnitude of effects, the practical significance, and the steady accumulation of knowledge.