

Bias and precision of some classical ANOVA effect sizes when assumptions are violated

Susan Troncoso Skidmore · Bruce Thompson

Published online: 6 October 2012
© Psychonomic Society, Inc. 2012

Abstract Previous simulation research has focused on evaluating the impact of analytic assumption violations on statistics related to the F test and associated $p_{\text{CALCULATED}}$ values. The present article evaluated the bias of classical estimates of practical significance (i.e., effect size sample estimators $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) in a one-way between-subjects univariate ANOVA when assumptions are violated. The simulation conditions modeled were selected on the basis of prior empirical research. Estimated (1) sampling error bias and (2) precision computed for each of the three effect size estimates for the 5,000 samples drawn for each of the 270 (5 parameter Cohen's d values \times 3 group size ratios \times 3 population distribution shapes \times 3 variance ratios \times 2 total ns) conditions were modeled for each of the $k = 2, 3$, and 4 group analyses. Our results corroborate the limited previous related research and suggest that $\hat{\eta}^2$ should *not* be used as an ANOVA effect size estimator, even though $\hat{\eta}^2$ is the only available choice in the menus in most commonly available software.

Keywords Effect size · Practical significance · Analysis of variance · Homogeneity of variance · Type I error · Power · Eta squared · Epsilon squared · Omega squared

Analysis of variance (ANOVA) was a term first used by Sir Ronald Fisher in 1918 (see David, 1995). Fisher conceived of the ANOVA as a way to analyze differences in crop yields across agricultural plots (Gamst, Meyers, & Guarino, 2008). The ANOVA is a parametric statistical technique that explores mean differences on a single response variable

across two or more groups on each of one or more ways or factors. Reviews of statistical techniques in the literature empirically demonstrate the long-standing popularity of ANOVA techniques within the social sciences (Edginton, 1964, 1974; Elmore & Woehlke, 1996; Kieffer, Reese, & Thompson, 2001; Skidmore & Thompson, 2010). The ANOVA is the most popular inferential analysis technique for between-subjects univariate designs and was used in 93.3 % of the between-subjects univariate articles reviewed by Keselman et al. (1998).

As with all statistical techniques, the integrity of ANOVA results is contingent upon the extent to which the assumptions of the ANOVA are met. When the outcome variable scores exhibit independence, normality, and homogeneity of variance across groups, the ANOVA assumptions are satisfied. Unfortunately, empirical studies suggest that "researchers rarely verify that validity assumptions are satisfied . . . and . . . typically use analyses that are nonrobust to assumption violations" (Keselman et al., 1998, p. 350).

In practice, the question is not whether ANOVA assumptions are perfectly met but, rather, whether assumptions are sufficiently well met that reasonable confidence can be vested in the ANOVA statistics. Of course, not all these statistical assumptions are equally vital. For example, it is well known that the F test is robust to "mild departures from normality" (Harwell, Rubinstein, Hayes, & Olds, 1992, p. 316) and, more generally, that the F test is relatively insensitive to normality assumption violations under conditions of equal group sizes (cf. Glass, Peckham, & Sanders, 1972; Lix, Keselman, & Keselman, 1996).

The behavior of Type I error rates under heterogeneity of variance conditions is also well documented in the literature (Glass et al., 1972; Harwell et al., 1992). As summarized by Glass et al. (1972) and corroborated by Harwell et al. (1992), when groups are equal in size (i.e., a balanced design) but given heterogeneity of variance, there is a slight increase in the Type I error rate. Differential and more

S. Troncoso Skidmore
Sam Houston State University,
Huntsville, TX, USA

B. Thompson (✉)
Texas A&M University,
College Station, TX, USA
e-mail: bruce-thompson@tamu.edu

pronounced effects are observed given both groups unequal in size (i.e., an unbalanced design) and heterogeneity of variance. In negative pairing, when smaller sample sizes are paired with larger variances on the outcome variable, Type I error rates are markedly inflated as against the nominal alpha level. In cases of positive pairing, when smaller sample sizes are paired with smaller variances, Type I error rates conversely are less than the nominal level. Some researchers have recommended under these circumstances the use of alternatives to the ANOVA, such as the James and the Welch tests (Lix et al., 1996).

Keselman et al. (1998) commented on the severity of violations to ANOVA assumptions:

Without the assumptions (or barring strong evidence that adequate compensation for them has been made), it can be—and has been—shown that the resulting significance probabilities (p values) are, at best, somewhat different from what they should be and, at worst, worthless. (p. 351)

Yet on average, in published research, the highest standard deviation tends to be roughly twice as large as the lowest standard deviation across groups (Keselman et al., 1998). And for published one-way designs, positive pairings (i.e., largest outcome variable variance occurs in the largest sized group) were present roughly a third of the time (31.3 %), and negative pairings (i.e., smallest outcome variable variance in the largest sized group) were present roughly a fifth (22.1 %) of the time (Keselman et al., 1998). Thus, assumption violations should be explored by all researchers, and the extent to which violations are present should be matched against empirical literature that details the extent to which Type I error rates and/or power are impacted before making a judgment as to whether a different analytical tool better suited to the characteristics of the data needs to be used.

Moving beyond statistical significance testing

While an ANOVA can be used to test the statistical significance of group mean differences, a second and at least equally important use of the ANOVA is to estimate the practical significance, or the magnitude of effect, of group mean differences. Previous researchers have focused primarily on understanding the impact of violation assumptions on both power and the p values for null hypothesis statistical significance testing.

Of course, beginning in the late 1980s, psychologists increasingly emphasized the importance of effect size reporting and interpretation. Fiona Fidler (2005) reviewed this evolution in her comprehensive 70,000+ word dissertation, *From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine, and Ecology*.

By 1994 the American Psychological Association (APA) Publication Manual first mentioned and "encouraged" (p. 18) effect size reporting, in APA (2001) it noted that failure to report effect sizes was "a defect" (p. 5), and in 2010 it noted that

historically, researchers in psychology have relied heavily on null hypothesis statistical significance testing (NHST) as a starting point for many (but not all) of its analytical approaches. APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals and extensive description are needed to convey the most complete meaning of results. (p. 33)

Indeed, in 2002, Fidler (2002) noted that, "of the major American associations, only all the journals of the American Educational Research Association [AERA] have remained silent on all these issues" (p. 754). But in 2006, the AERA spoke, and published its standards requiring effect size reporting in all AERA journals (AERA, 2006).

Purpose of the present study

Previous Monte Carlo ANOVA simulation research focused on evaluating the impact of assumption violations on statistics related to the F test and associated $p_{\text{CALCULATED}}$ values. For example, Wilcox has published extensively on the robustness (or lack thereof) of the F test under assumption violations (Wilcox, 1995; Wilcox, Charlin, & Thompson, 1986) and has suggested the use of more robust methods (Wilcox, 1993; Wilcox & Keselman, 2003).

A few researchers have considered the effects of assumption violations on ANOVA-related effect sizes. In one study, Wilcox (2006) examined the robustness of one measure of effect size, Cohen's d , which is relevant in the two-group one-way ANOVA. In cases where there is a contaminated normal distribution (see Tukey, 1960), "Cohen's d can mask a large effect size" (Wilcox, 2006, p. 355). Wilcox (2006) found that when the tails of the distribution are thicker, as in a contaminated normal distribution or in the presence of outliers, indices of effect size, such as Cohen's d , can be distorted.

One study on the robustness of estimates of practical significance was published over 30 years ago by Carroll and Nordholm (1975). Means across nonnull conditions were held constant, while the within-population variances were adjusted to achieve η^2 parameters of .05, .15, .40, and .75 within the context of a three-level one-way fixed effects ANOVA under both balanced and unbalanced conditions. Carroll and Nordholm found that just as heterogeneity of variance in unbalanced designs causes serious distortions to power and Type I error rates in ANOVA, the most serious

distortions to $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ (i.e., sample estimators) occurred when variances were unequal across unbalanced designs.

Keselman (1975) investigated the (1) bias and (2) precision of three ANOVA effect sizes ($\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) with respect only to robustness against distribution assumption violations. He found that "omega squared is a more accurate estimator [i.e., smallest bias] of the true population magnitude while eta squared has the smallest sampling variability [i.e., greatest precision]" (p. 47).

The purpose of the present article is to move beyond the robustness of estimates of statistical significance (Type I error rates and power) to evaluate the robustness of estimates of practical significance (i.e., effect sizes $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) in a one-way between-subjects univariate ANOVA. We sought to understand the utility of these effect sizes in the presence of assumption violations.

Method

To the extent possible, the conditions for the present Monte Carlo investigation were chosen on the basis of previous simulation research findings that demonstrated a need either to investigate a particular condition or to investigate particular researcher practices empirically shown to be common in the extant literature. Thus, the conditions modeled here are based on what previous research indicates should have an impact on result integrity while still maintaining an ecologically valid footing by grounding our simulations in the framework of typical researcher practices.

Computing the effect sizes

While our study also allowed for confirmation of previous findings regarding the behavior of estimates of statistical significance (i.e., Type I error rates and power) under ANOVA assumption violations, our focus was on the behavior of estimates of practical significance ($\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$). Eta squared ($\hat{\eta}^2$) is an effect size that is uncorrected for sampling error influences and quantifies the "proportion of the variance in the population that is accounted for by variation in the treatment" (Grissom & Kim, 2005, p. 121). Eta squared is given by $\hat{\eta}^2 = SS_{MODEL} / (SS_{TOTAL})$, or in the case of a one-way design, can also be computed as $[(k-1) * (F)] / \{[(k-1) * (F)] + n - k\}$, where k is the number of groups and n is the total sample size (Wilcox, 1987). It is well known that $\hat{\eta}^2$, like \hat{R}^2 (Yin & Fan, 2001) and \hat{r}^2 (Skidmore & Thompson, 2011; Wang & Thompson, 2007), is positively biased. To correct this bias, Kelley (1935) and Hays (1981) developed $\hat{\varepsilon}^2$ and $\hat{\omega}^2$, respectively. Epsilon squared is given by $\hat{\varepsilon}^2 = [SS_{MODEL} - (k-1) * (MS_{ERROR})] /$

(SS_{TOTAL}) , or equivalently by $(F-1) / \{F + [(n-k)/(k-1)]\}$ (Carroll & Nordholm, 1975). Omega squared is given by $\hat{\omega}^2 = [SS_{MODEL} - (k-1) * (MS_{ERROR})] / (SS_{TOTAL} + MS_{ERROR})$ or, equivalently, by $(F-1) / \{F + [(n-k+1)/(k-1)]\}$ (Carroll & Nordholm, 1975). The presence of the F test statistic in the formulas underscores the relationships between all parametric analyses within the general linear model.

Furthermore, given (1) that assumption violations impact $F_{CALCULATED}$, and (2) the potential use of $F_{CALCULATED}$ when computing ANOVA effects, assumption violations clearly must also impact effect size estimates. What is less clear is the degree to which ANOVA effect sizes are robust to assumption violations and whether certain ANOVA effect sizes may be more or less robust than others to assumption violations.

Population effect sizes used in the simulation

Cohen (1988) himself eschewed the thoughtless fixation on effect size benchmarks, noting that "these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible [*italics added*]" (p. 532) in result interpretation, because interpretation ought to be context specific (Thompson, 2002, 2006b). Cohen developed his benchmarks on the basis of his impressions of what he thought might be about the smallest ($d = 0.2$), the largest ($d = 0.8$), and the average ($d = 0.5$) effects across all published social science research. And Cohen felt that "typicality" was not a useful index of result import. Typicality is useful to simulation study design. Therefore, Cohen's benchmarks are considered in setting the population parameter values.

In the present study, all distributions had equal means set to 100.0 for the null condition (i.e., the null condition where population Cohen's $d = 0.0$). Cohen's d provides a standardized effect size for two population means. When there are more than $k = 2$ means, Cohen's f is an analogous effect size. Cohen's f is "the standard deviation of the standardized k population means" (1988, p. 276). In determining the necessary mean differences to obtain the given Cohen's f value, Cohen's (1988) pattern 1 was used, where one mean is at each end of the range and "the remaining $k - 2$ means are all at the midpoint" (p. 277). Cohen's f , under pattern 1, is given by formula 8.2.8 provided in Cohen's text: $f = d * \text{sqrt}(1/2k)$. For the $k = 2$ case, the formula reduces to $d * 0.5$. Thus, Cohen's f , in this case, is half of d . As the number of groups increase, the multiplier to obtain f decreases. Interested readers can refer to Cohen (1988) for a full explanation. For the present study, we maintained the five Cohen's d conditions constant and converted from Cohen's f value as appropriate when $k > 2$. The four nonnull conditions are Cohen's d equal to 0.20, 0.50, 0.80, and 1.00.

Numbers of groups (k) modeled in the simulation

Wilcox et al. (1986) demonstrated that when there were four groups in the one-way ANOVA, the F test was not as robust as when there were two groups. Thus, the condition of number of groups is important to consider when evaluating the robustness of the F test. In a review of Monte Carlo studies, the number of groups (k) examined by researchers in prior ANOVA simulation studies focusing on Type I error rates and power varied between 2 and 10 for equal group sizes and between 2 and 6 for unequal group sizes, with 3 groups being the most commonly examined (Harwell et al., 1992). Therefore, two-level, three-level, and four-level one-way situations were modeled in the present study.

Total ns modeled in the simulation

Kieffer et al. (2001) examined quantitative articles in 10 volumes of the *American Educational Research Journal* and *Journal of Counseling Psychology* and found that median sample sizes in some years were as low as 76 and 43, respectively. Of course, some individual articles had even lower sample size than these volume medians.

Previous Monte Carlo simulation studies on the robustness of ANOVA to assumption violations modeled relatively small total ns . For example, Hsu's (1938) largest total sample size was 20, and Box's (1954) largest total sample size was 25. A later study (Donaldson, 1968), which was highlighted as "exemplary" in Glass et al. (1972, p. 265), examined a minimum sample size of 8 for the two-group ANOVA and a maximum total sample size of 128 for the four-group ANOVA.

In a meta-analytic summary of 28 Monte Carlo studies of ANOVA dynamics, Harwell et al. (1992) found that Monte Carlo researchers used an average total sample size of 111 ($SD = 154$) across the simulation studies reviewed, with a minimum total sample size of 8 and a maximum of 750. In our simulation study, we modeled total sample ns of 24 and 48.

Group sizes modeled in the simulation

In a review of the analytical practices of educational researchers, Keselman and his colleagues (1998) explained that one-way designs made up 58.3 % of the 61 between-subjects univariate studies they located in their review. Furthermore, in the 23 one-way studies with an unbalanced design, the ratio of the largest to the smallest group size was greater than 3:1 in 43.5 % of the studies.

For $k = 2$, we modeled group sizes of 12:12, 8:16, and 6:18 for a total n of 24 and of 24:24, 16:32, and 12:36 for a total n of 48. For $k = 3$, we modeled group sizes of 8:8:8, 6:6:12, and 5:5:14 for a total n of 24 and of 16:16:16,

12:12:24, and 10:10:28 for a total n of 48. For $k = 4$, we modeled group sizes of 6:6:6:6, 5:5:5:9, and 4:4:4:12 for a total n of 24 and of 12:12:12:12, 10:10:10:18, and 8:8:8:24 for a total n of 48.

Within-group outcome variable variances modeled

Heterogeneity of variance is a serious assumption violation in the ANOVA (Harwell et al., 1992; Lix et al., 1996). Nevertheless, published one-way designs are known to have an average ratio of the largest to the smallest standard deviation of 2:1 (Keselman et al., 1998). Furthermore, it is well documented not only that homogeneity of variances is an important assumption, but also that the ways in which sample sizes are paired with heterogeneous variances produce different results. Smaller sample sizes paired with larger variances (negative pairing) produce larger Type I error rates; larger sample sizes paired with larger variances (positive pairing) produce a lower Type I error rate (e.g., Harwell et al., 1992).

We studied both homogeneous and heterogeneous variance conditions. And for heterogeneous variance situations, we studied both negative and positive pairings of variances with group sizes.

For the equal variance conditions, parameter σ^2 was set equal to 225.0 within each group for $k = 2, 3$, and 4. Outcome variable variances for unequal negative pairing conditions were set at $\sigma^2 = 360.0$ and 90.0; 385.7, 192.8, and 96.4; and 400.0, 200.0, 200.0, and 100.0, for $k = 2, 3$, and 4, respectively. Outcome variable variances for unequal positive pairing conditions were set at $\sigma^2 = 90.0$ and 360.0; 96.4, 192.8, and 385.7; and 100.0, 200.0, 200.0, and 400.0, for $k = 2, 3$, and 4, respectively. Thus, in all cases the average variance was equal to 225.0.

Shape conditions modeled in the simulation

As was previously noted, mild departures from normality have negligible effects on the F test (Harwell et al., 1992); thus, conditions in the present study were chosen to represent normal (i.e., coefficient of skewness = coefficient of kurtosis = 0.0), mildly deviant (i.e., coefficient of skewness = coefficient of kurtosis = 0.5), and moderately deviant (coefficient of skewness = 1.0, coefficient of kurtosis = 3.75) distribution shapes. The population data with the desired shape parameters were generated using Vale and Maurelli's (1983) multivariate extension of Fleishman's (1978) procedure.

To confirm that the program was working in the intended manner, populations of 100,000 scores were generated for each of the three distributional shape conditions we modeled. The population parameters closely matched the expected coefficients of skewness and kurtosis.

Replications

To minimize the standard error of the simulation in exploring Type I error rates and the robustness of the ANOVA effect sizes across assumption violations, 5,000 samples (see Robey & Barcikowski, 1992) were drawn for each of the 270 (5 parameter d values \times 3 group size ratios \times 3 population distribution shapes \times 3 variance ratios \times 2 total ns) conditions modeled for each of the $k = 2, 3$, and 4 group analyses. According to the *SAS for Monte Carlo Studies: A Guide for Quantitative Researchers*, 5,000 replications for the ANOVA analyses performed provide “reasonable accuracy” (Fan, Felsovalyi, Sivo, & Keenan, 2001, p. 130). Thus, we modeled a total of $270 \times 3 = 810$ conditions. Detailed SAS programming explanations, including related sample programs, can be found in *SAS for Monte Carlo Studies* (Fan et al., 2001).

Three indices of practical significance ($\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) and two indices related to statistical significance (Type I error rates and power) were computed for each of the 4,050,000 ($5,000 \times 270 \times 3$) samples. Estimated (1) sampling error bias and (2) precision were computed for each of the three effect size estimates.

Simulation baseline check

When ANOVA assumptions are met, the expectation is that the actual Type I error rates should closely match the nominal α level. Similarly, when ANOVA assumptions are met, the expectation is that theoretical power levels should agree with actual obtained power. When a Monte Carlo study is conducted, providing both Type I error rates and theoretical versus empirical power estimates for the null condition provides evidence that the simulation study was correctly conducted. Glass et al. (1972) recommended that such “baseline checks” of the entire simulation procedure should be performed and reported” (p. 282). We performed three such baseline checks for our simulation.

First, we computed actual versus expected Type I error rates for (1) parameter Cohen's $d = 0.0$ (2) with perfect homogeneity of variance and (3) normally distributed outcome variables for the 18 (3 group size ratios \times 2 total $ns \times k = 2, 3$, and 4 group analyses) simulation conditions relevant when ANOVA assumptions are perfectly met for the null case. Across the 5,000 samples in each of these 18 cases, the actual empirical Type I error rates ranged from 0.043 to 0.054 ($M = 0.050$, $SD = 0.004$). Thus, the empirical Type I error rates were, as expected, close to or equal to 0.05 when nominal $\alpha = .05$.

Second, for the 24 simulation conditions in which parameter Cohen's d did not equal 0.0, we compared our actual empirical power values when ANOVA assumptions were perfectly met (i.e., normality and homogeneity of variance) with balanced designs with theoretically expected power

estimates obtained using *G*Power* (Version 3.1.0; Faul et al. 2007). The deviations of actual minus theoretically expected power values were quite small ($M = -0.001$, $SD = 0.004$).

As a third and final confirmation that our simulation worked correctly, simulation results for the null case (i.e., parameter Cohen's $d = 0.0$) for the most severe cases of assumption violations are presented, when unequal samples sizes were paired with heterogeneous variances. Table 1 presents Type I error rates for both the smaller and larger total sample size conditions across $k = 2, 3$, and 4. Our results closely match the findings in previous simulation studies of effects on Type I error rates of ANOVA assumption violations (cf. Glass et al., 1972; Harwell et al., 1992).

Results

The estimated bias due to sampling error was computed for each of the 4,050,000 sets of three effect size estimates, by subtracting the parameter η^2 values from the individual sample $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ values. Thus, positive sampling error bias values indicate that sample estimates overestimated the parameter and negative parameter sampling error bias values indicate that sample estimates underestimated the parameter. Finally, precision was estimated by computing the standard deviations of each of the 5,000 estimates within the 270 conditions modeled each for $k = 2, 3$, and 4.

Table 1 Impact of heterogeneity of variance on Type I error rates

k	Group size proportion	Variance ratio	Type I error	
			Smaller n (24)	Larger n (48)
2	1:1	1:4	0.054	0.052
2	1:2	1:4	0.018	0.017
2	1:3	1:4	0.009	0.005
2	1:1	4:1	0.058	0.055
2	1:2	4:1	0.113	0.105
2	1:3	4:1	0.155	0.145
3	1:1:1	1:2:4	0.058	0.053
3	1:1:2	1:2:4	0.030	0.022
3	1:1:3	1:2:4	0.016	0.013
3	1:1:1	4:2:1	0.062	0.058
3	1:1:2	4:2:1	0.099	0.100
3	1:1:3	4:2:1	0.129	0.130
4	1:1:1:1	1:2:2:4	0.066	0.059
4	1:1:1:2	1:2:2:4	0.035	0.033
4	1:1:1:3	1:2:2:4	0.015	0.016
4	1:1:1:1	4:2:2:1	0.056	0.057
4	1:1:1:2	4:2:2:1	0.086	0.086
4	1:1:1:3	4:2:2:1	0.132	0.129

Note. Cohen's $d = 0.0$, shape = normal, $\alpha = .05$

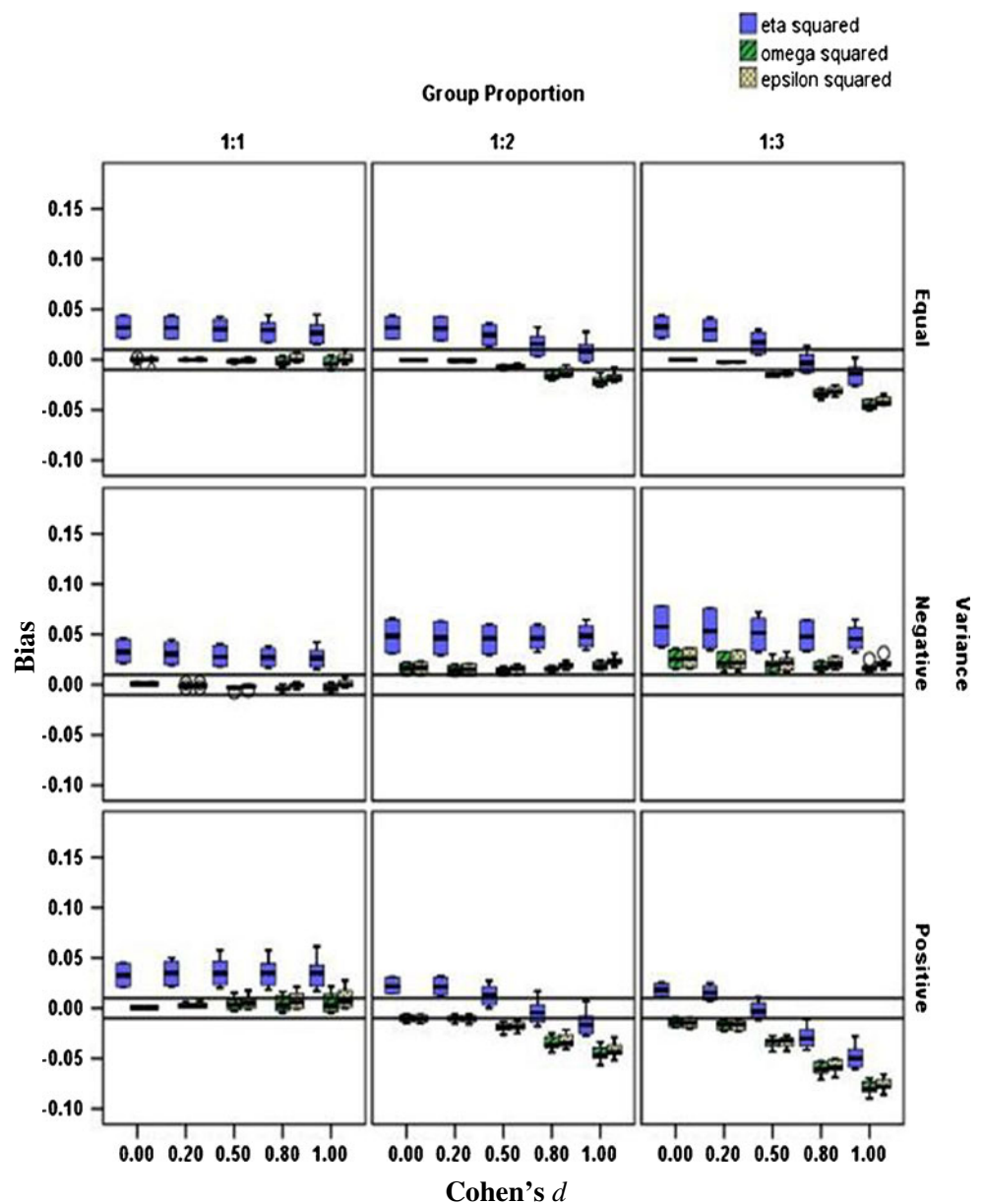
Effect size sampling error bias

The dispersions of the 1,350,000 (i.e., 270 simulation conditions \times 5,000 replications) sampling error biases were relatively homogeneous across the three numbers of groups (i.e., $k = 2, 3$, and 4) and the three ANOVA effect sizes (i.e., $\hat{\eta}^2$, $\hat{\epsilon}^2$, and $\hat{\omega}^2$). For $k = 2$, $SD_{\eta^2 \text{ SQUARED}} = 0.094$, $SD_{\epsilon^2 \text{ SQUARED}} = 0.095$, and $SD_{\omega^2 \text{ SQUARED}} = 0.097$. For $k = 3$, $SD_{\eta^2 \text{ SQUARED}} = 0.103$, $SD_{\epsilon^2 \text{ SQUARED}} = 0.106$, and $SD_{\omega^2 \text{ SQUARED}} = 0.109$. For $k = 4$, $SD_{\eta^2 \text{ SQUARED}} = 0.111$, $SD_{\epsilon^2 \text{ SQUARED}} = 0.118$, and $SD_{\omega^2 \text{ SQUARED}} = 0.121$. Considerably more descriptive statistics for the simulation results are available from the senior author.

The shape conditions modeled resulted in minimal impact to effect size estimates. As was expected, the smaller

total n condition resulted in greater variability than did the larger total n condition for each of the effect size estimates. The greatest amount of bias was present when heterogeneous variances were paired with unbalanced designs. Figures 1, 2 and 3 present box-and-whisker plots for sampling error biases of the three ANOVA effect size formulas across the $k = 2, 3$, and 4 number-of-groups cases for balanced and unbalanced design conditions (with positive and negative pairing) across the five Cohen's d conditions. For reference purposes, each figure includes horizontal lines drawn at 0.0 ± 0.01 to discriminate between relatively biased versus unbiased estimates (see Kromrey & Hines, 1996; Yin & Fan, 2001). Thus, estimates outside the two horizontal lines represent situations in which the use of a particular formula resulted in

Fig. 1 Box-and-whisker plots for sampling error biases across variance homogeneity/heterogeneity, group size proportions, and values of Cohen's d for the $k = 2$ group ANOVA case



biased effect size estimates for the conditions examined in the present study. For example, in the two-group case with the equal variance condition in a balanced design, both $\hat{\epsilon}^2$ and $\hat{\omega}^2$ were unbiased estimators across the five Cohen's d conditions examined. However, in the unbalanced design condition (1:3), with negative variance pairing, all the effect size formulas provided positively biased results.

Effect size precision

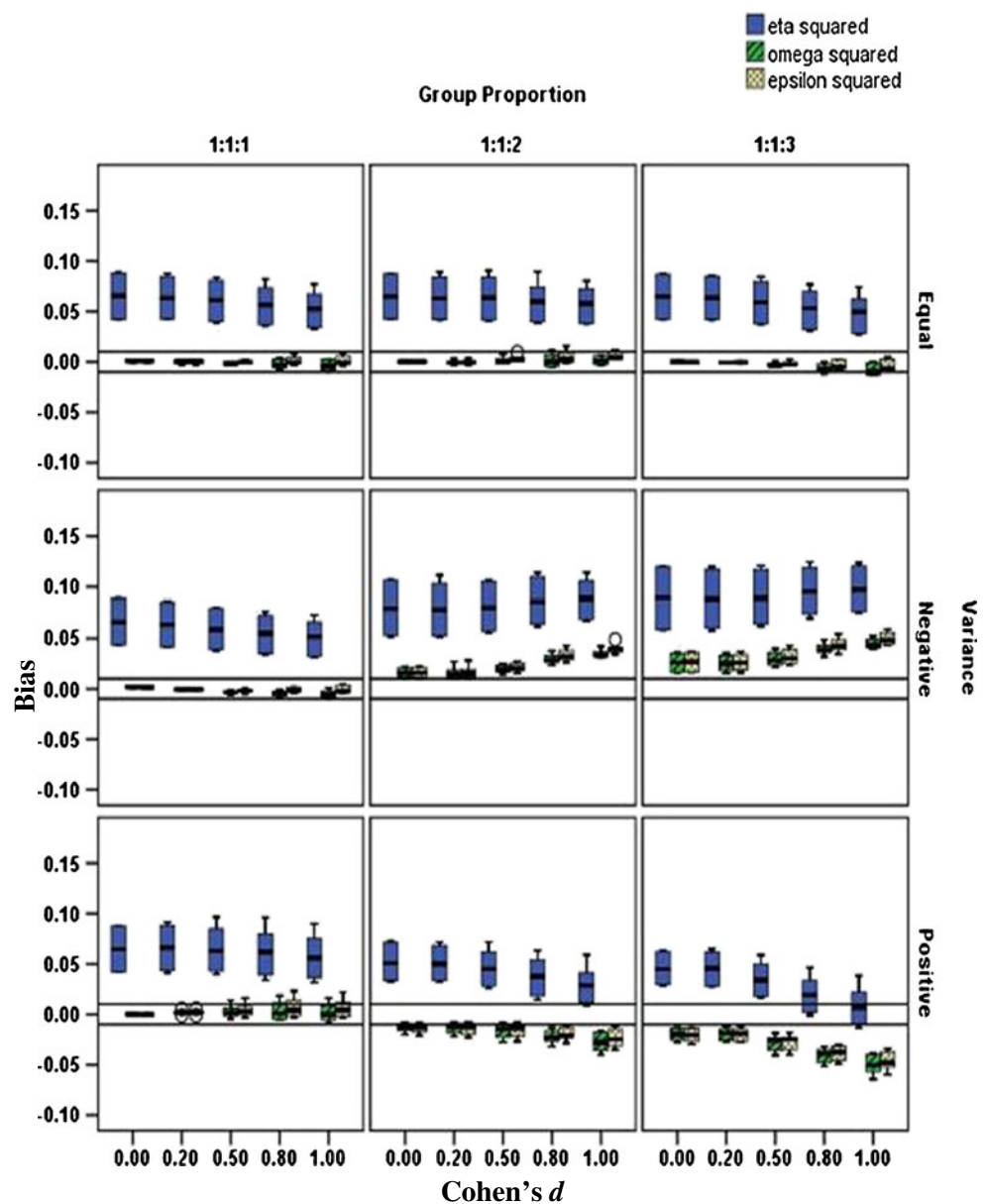
Researchers want sample estimates of population effect sizes to be unbiased, but we would also like the estimates to be precise (i.e., for a given design, we prefer estimates that are narrowly dispersed over repeated sampling). As indices of precision, we computed the SD s for the 5,000

effect size estimates within each of the 270 (5 parameter d values \times 3 group size ratios \times 3 population distribution shapes \times 3 variance ratios \times 2 total n s) simulation conditions for all three ANOVA effect sizes.

Across these 270 simulation conditions, for $k = 2$, $SD_{\eta^2 \text{ SQUARED}}$ for precision = 0.037, $SD_{\epsilon^2 \text{ SQUARED}} = 0.038$, and $SD_{\omega^2 \text{ SQUARED}} = 0.039$. For $k = 3$, $SD_{\eta^2 \text{ SQUARED}}$ for precision = 0.033, $SD_{\epsilon^2 \text{ SQUARED}} = 0.036$, and $SD_{\omega^2 \text{ SQUARED}} = 0.037$. For $k = 4$, $SD_{\eta^2 \text{ SQUARED}}$ for precision = 0.031, $SD_{\epsilon^2 \text{ SQUARED}} = 0.037$, and $SD_{\omega^2 \text{ SQUARED}} = 0.038$.

The variability in precisions across the design features were largely explained by Cohen's d values (i.e., 0.00, 0.20, 0.50, 0.80 or 1.00) and samples sizes (i.e., $n = 24$ or 48). For the $k = 2$ group situation, across all three ANOVA effect sizes, the Cohen's d main effect and the sample size main

Fig. 2 Box-and-whisker plots for sampling error biases across variance homogeneity/heterogeneity, group size proportions, and values of Cohen's d for the $k = 3$ group ANOVA case



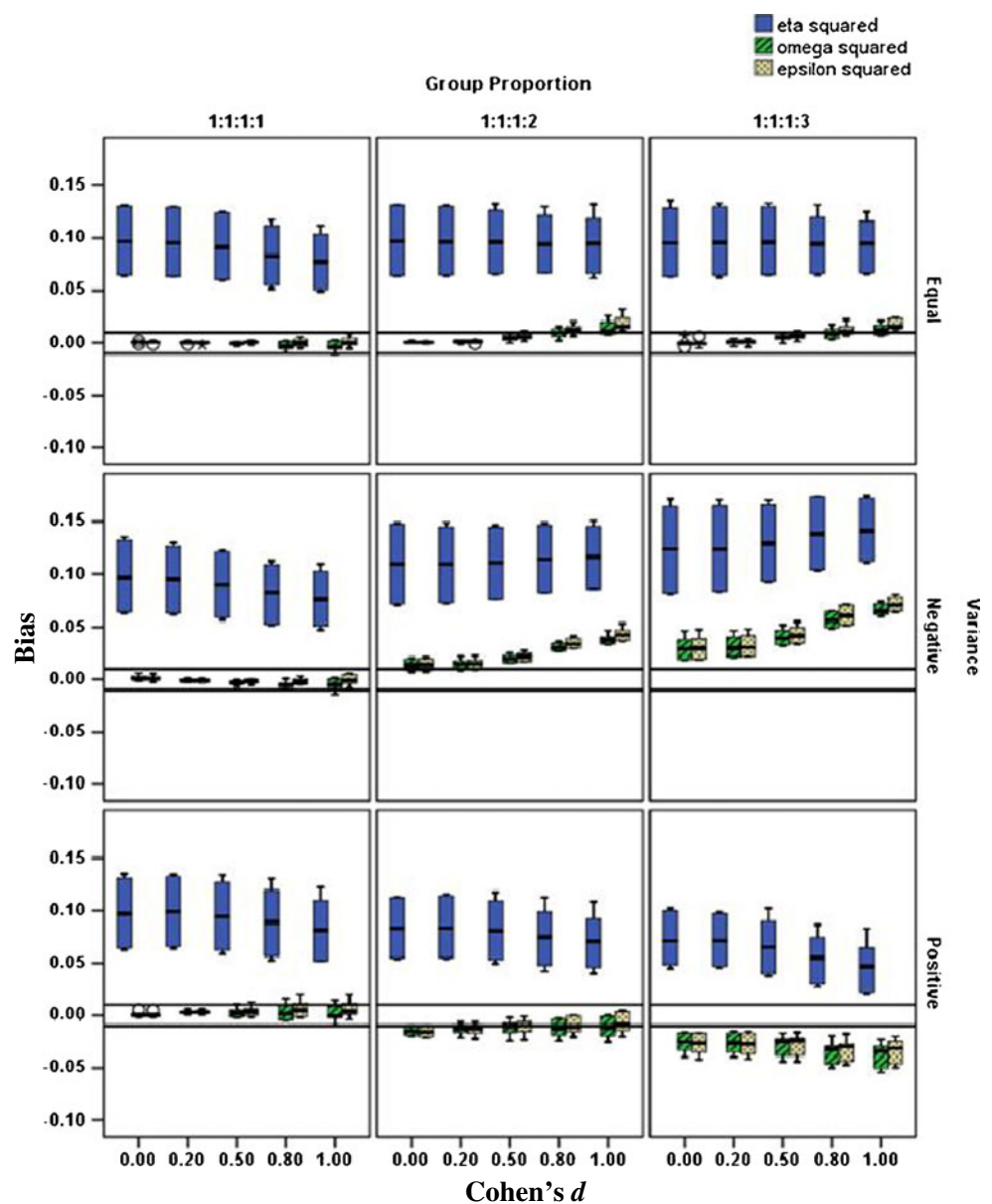
effect explained roughly 60 % and 20 % of the variability in precision. For the $k = 3$ group situation, across all three ANOVA effect sizes, the Cohen's d main effect and the sample size main effect explained roughly 50 % and 35 % of the variability in precision. For the $k = 3$ group situation, across all three ANOVA effect sizes, the Cohen's d main effect and the sample size main effect explained roughly 40 % and 50 % of the variability in precision. Sampling error variance accounts for the fact that Cohen's d and sample size impact effect size estimates. All other design conditions held constant, sampling error variance is greater in samples with smaller n s than in samples with larger n s. Similarly, the larger the effect size, the less sampling error variance. A clear and extended description of these phenomena can be found in (Thompson, 2006a).

Discussion

Our Monte Carlo simulation results suggest a number of important conclusions regarding the characteristics of three ANOVA effect sizes (i.e., $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) under conditions when analytic assumptions have been met or assumptions are violated to varying degrees. This information will be increasingly important to scholars as researchers come off what has historically been a low baseline of effect size reporting in the published literature (cf. Snyder & Thompson, 1998; Thompson, 1999; Thompson & Snyder, 1998) in response to the admonitions of various standards (e.g., AERA, 2006; APA, 2010; Wilkinson & APA Task Force, 1999).

Of course, methodologists have long recognized that the most commonly used effect size estimators (e.g., Cohen's d)

Fig. 3 Box-and-whisker plots for sampling error biases across variance homogeneity/heterogeneity, group size proportions, and values of Cohen's d for the $k = 4$ group ANOVA case



are not robust. Thus, Algina et al. (2005) recommended using a robust version of Cohen's d , with 20 % trimmed means and the square root of a 20 % Winsorized variance. Similarly, in the presence of heterogeneous variances and nonnormal data, Keselman et al. (2008a, 2008b) advocated using an approximate df test statistic based on trimmed means and Winsorized variances. Zhang and Schoeps (1997) proposed two nonparametric estimators of effect size. Among the endearing qualities of the proposed estimators are that they are easy to calculate, robust, and relatively efficient.

Unfortunately, robust statistical methods have only minimally penetrated the contemporary practices of applied researchers. Penetration has been slowed, first, by the limited space afforded methodology within doctoral curricula (Aiken, West, & Millsap, 2008; Capraro & Thompson, 2008; Henson & Williams, 2006). Second, the software commonly used by applied researchers affords few—and often nonoptimal—analytic choices. Thus, Pierce, Block, and Aguinis (2004) noted that "because common statistical software packages such as SPSS only report eta-squared values and not omega-squared or epsilon-squared values in their ANOVA output files, many researchers in education and psychology report eta-squared values" (p. 918). To assist researchers in reporting epsilon-squared and omega-squared values, we provide an Excel sheet to easily calculate $\hat{\epsilon}^2$ and $\hat{\omega}^2$ from information currently available in common statistical software such as SPSS (see <http://www.shsu.edu/~sts008/>).

Precision

The *precisions* of the three ANOVA effect sizes (i.e., $\hat{\eta}^2$, $\hat{\epsilon}^2$, and $\hat{\omega}^2$) across the 270 (5 parameter d values \times 3 group size ratios \times 3 population distribution shapes \times 3 variance ratios \times 2 total ns) simulation conditions were small and similar for all three effects. These results suggest that precision is not a relevant consideration with respect to differential preferences among the three ANOVA effect sizes we studied.

Sampling error bias

Our results support a number of conclusions. First, as Figs. 1, 2 and 3 indicate, across the five values of Cohen's d and the balanced and the two unbalanced designs we considered, when the homogeneity of variance assumption was met, $\hat{\eta}^2$ had considerable positive sampling error bias, especially for the $k = 3$ and 4 designs. Second, also when the homogeneity of variance assumption was met, both $\hat{\epsilon}^2$, and $\hat{\omega}^2$ tended to have little bias, especially for the $k = 3$ and 4 designs.

Third, across the $k = 2, 3$, and 4 group designs, when variances were heterogeneous and involved unbalanced designs with negative pairings, all three estimators tended

to have positive sampling error biases. Fourth, across the $k = 2, 3$, and 4 group designs, when variances were heterogeneous and involved unbalanced designs with positive pairings, even the $\hat{\epsilon}^2$ and $\hat{\omega}^2$ estimators tended to have negative sampling error bias and not to function as well as they did in the remaining simulation design conditions.

Limitations

Of course, as in any study, our foci were necessarily limited. We did not investigate effect sizes in multiway designs (see Kirk, 1995, pp. 397–399). Furthermore, we studied only classical effect sizes, rather than robust analogs of these estimates (e.g., Algina et al., 2005; Keselman et al., 2008a, 2008b). It is also worth noting that classical effect sizes enjoy widespread use in the contemporary applied social sciences, while robust estimates to date are almost never reported, however unfortunate this reality may be.

Summary

Overall, our results corroborate the limited previous research (Carroll & Nordholm, 1975; Keselman, 1975) and suggest that $\hat{\eta}^2$ should *not* be used as an ANOVA effect size estimator, because across the range of conditions we examined, $\hat{\eta}^2$ had considerable sampling error bias, as reported in Figs. 1, 2 and 3. Of course, this recommendation flies directly in the face of both common analytic practice and the commonly available software choices, as noted previously (Pierce et al., 2004). We look forward to a day when researchers will be less susceptible to the appeal of point-and-click menus within commonly used software and more willing to venture into the world of simple calculations or even software syntax and robust statistics.

References

- Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement and methodology in psychology. *American Psychologist*, 63, 32–50. doi:10.1037/0003-066X.63.1.32
- Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317–328. doi:10.1037/1082-989X.10.3.317
- American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, 35(6), 33–40. doi:10.3102/0013189X035006033
- American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 25, 484–498. doi:10.1214/aoms/1177728717
- Capraro, R. M., & Thompson, B. (2008). The educational researcher defined: What will future researchers be trained to do? *The Journal of Educational Research*, 101, 247–253.
- Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ϵ^2 and Hay's ω^2 . *Educational and Psychological Measurement*, 35, 541–554. doi:10.1177/001316447503500304
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, 49, 121–133. doi:10.2307/2684625
- Donaldson, T. S. (1968). Robustness of the *F*-test to errors of both kinds and the correlation between the numerator and denominator of the *F*-ratio. *Journal of the American Statistical Association*, 63, 660–676. doi:10.2307/2284037
- Edginton, E. S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, 19, 202–203. doi:10.1037/h0039177
- Edginton, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist*, 29, 25–26. doi:10.1037/h0035846
- Elmore, P. B., & Woehlke, P. L. (1996, April). *Research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978–1995*. Paper presented at the Annual meeting of the American Educational Research Association, New York.
- Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2001). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Fidler, F. (2002). The fifth edition of the APA *Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, 62, 749–770. doi:10.1177/001316402236876
- Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine, and ecology*. Doctoral dissertation, University of Melbourne. www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf
- Fisher, R. A. (1918). The causes of human variability. *The Eugenics Review*, 10, 213–220.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521–532. doi:10.1007/BF02293811
- Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. Cambridge: Cambridge University Press.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumption underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, 42, 237–288.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. New York: Psychology Press.
- Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17, 315–339. doi:10.2307/1165127
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart and Winston.
- Henson, R. K., & Williams, C. (2006, April). *Doctoral training in research methodology: A national survey of education-related degrees*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Hsu, P. L. (1938). Contribution to the theory of "Student's" *t*-test as applied to the problem of two samples. *Statistical Research Memoirs*, 2, 1–24.
- Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21, 554–559. doi:10.1073/pnas.21.9.554
- 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
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008a). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, 13, 110–129. doi:10.1037/1082-989X.13.2.110
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008b). Supplemental materials to 129a. A SAS program to implement a general approximate degrees of freedom solution for inference and estimation. <http://dx.doi.org/10.1037/1082-989X.13.2.110.supp>
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350–386.
- Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in "AERJ" and "JCP" articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education*, 69, 280–309. doi:10.1080/00220970109599489
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). New York: Brooks/Cole.
- Kromrey, J. D., & Hines, C. V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. *The Journal of Experimental Education*, 64, 240–266.
- Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "F" test. *Review of Educational Research*, 66, 579–619.
- 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. doi:10.1177/0013164404264848
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45, 283–288.
- Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*, 70, 777–795. doi:10.1177/0013164410379320
- Skidmore, S. T., & Thompson, B. (2011). Choosing the best correction formula for the Pearson r^2 effect size. *The Journal of Experimental Education*, 79, 257–278. doi:10.1080/00220973.2010.484437
- Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly*, 13, 335–348. doi:10.1037/h0088990 [Named the 1999 APA Division 16 Fellows' Article of the Year].
- Thompson, B. (1999). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children*, 65, 329–337.

- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 24–31. doi:[10.3102/0013189X031003025](https://doi.org/10.3102/0013189X031003025)
- Thompson, B. (2006a). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford.
- Thompson, B. (2006b). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583–603). Washington, DC: American Educational Research Association.
- Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development*, 76, 436–441.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In S. G. Olkin, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford, CA: Stanford University Press.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, 48, 465–471. doi:[10.1007/BF02293687](https://doi.org/10.1007/BF02293687)
- Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *The Journal of Experimental Education*, 75, 109–125. doi:[10.3200/JEXE.75.2.109-125](https://doi.org/10.3200/JEXE.75.2.109-125)
- Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, 38, 29–60. doi:[10.1146/annurev.ps.38.020187.000333](https://doi.org/10.1146/annurev.ps.38.020187.000333)
- Wilcox, R. R. (1993). Robustness in ANOVA. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 345–374). New York: Marcel Dekker.
- Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size. *Review of Educational Research*, 65, 51–77.
- Wilcox, R. R. (2006). Graphical methods for assessing effect size. *The Journal of Experimental Education*, 74, 353–367. doi:[10.3200/JEXE.74.4.351-367](https://doi.org/10.3200/JEXE.74.4.351-367)
- Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, F^* statistics. *Communications in Statistics: Simulation and Computation*, 15, 933–944. doi:[10.1080/03610918608812553](https://doi.org/10.1080/03610918608812553)
- Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8, 254–274. doi:[10.1037/1082-989X.8.3.254](https://doi.org/10.1037/1082-989X.8.3.254)
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. doi:[10.1037/0003-066X.54.8.594](https://doi.org/10.1037/0003-066X.54.8.594)
- Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, 69, 203–224. doi:[10.1080/00220970109600656](https://doi.org/10.1080/00220970109600656)
- Zhang, Z., & Schoeps, N. (1997). On robust estimation of effect size under semiparametric models. *Psychometrika*, 62, 201–214. doi:[10.1007/BF02295275](https://doi.org/10.1007/BF02295275)