Journal of Methods and Measurement in the Social Sciences,
Vol. 8, No. 1, 3-19, 2017

A Reassessment of ANOVA Reporting Practices: A Review of Three APA Journals

Yuanyuan Zhou
Texas A&M University

Susan Troncoso Skidmore
Sam Houston State University

Historically, ANOVA has been the most prevalent statistical method used in educational and psychological research and today ANOVA continues to be widely used. A comprehensive review published in 1998 examined several APA journals and discovered persistent concerns in ANOVA reporting practices. The present authors examined all articles published in 2012 in three APA journals (*Journal of Applied Psychology*, *Journal of Counseling Psychology*, and *Journal of Personality and Social Psychology*) to review ANOVA reporting practices including *p* values and effect sizes. Results indicated that ANOVA continues to be prevalent in the reviewed journals as a test of the primary research question, as well as to test conditional assumptions prior to the primary analysis. Still, ANOVA reporting practices are essentially unchanged from what was previously reported. However, effect size reporting has improved.

Keywords: ANOVA; *F*-test; reporting practices; statistical techniques

The prevalence of analysis of variance (ANOVA) in the educational and psychological literature has been well-documented (Skidmore & Thompson, 2010). Researchers have pronounced ANOVA to be the "the foundation of entire curricula in research methods courses in the social and behavioral sciences" (Gamst, Meyers, & Guarino, 2008, p.5) and "probably the most used statistical technique in psychological research" (Howell, 2011, p. 407). A survey of quantitative doctoral curricula in psychology further supports the expansive nature of ANOVA usage as most programs reported providing comprehensive coverage of the technique (Aiken, West, Sechrest, & Reno, 2008). Similarly, in a recent review of syllabi of APA-accredited doctoral programs in psychology, the prevalence of the teaching of both univariate and multivariate analysis of variance was found to be indicative of the foundational nature of these techniques in doctoral-level courses in statistics (Ord, Ripley, Hook, & Erspamer, 2016). And in a national review of undergraduate psychology programs, ANOVA was also a central component, especially beyond the introductory level courses (Friedrich, Buday, & Kerr, 2000).

Most commonly, ANOVA is a statistical model for analyzing mean differences across *k* groups, but its utility goes beyond means testing. For example, Generalizability Theory is built upon the ANOVA framework, wherein the "facets" in "G" theory are analogous to the "factors" in the analysis of variance. Similarly, the modern definition of intraclass correlation coefficient (*ICC*) fits within the framework of random-effects ANOVA. Through the use of ANOVA, one is able to parcel out the between-group variance and within-group variance. Traces of ANOVA can also be observed in hierarchical linear modeling, where the random intercept model is equivalent to a one-way random-effects ANOVA. Other usages of ANOVA include experimental designs that adopt ANOVA theory, such as random-effects ANOVA designs, mixed-effects ANOVA designs. In the present study, the term ANOVA specifically refers to statistical models with a continuous dependent variable and at least one categorical independent variable.

As in all parametric analyses, the conclusions drawn from ANOVA results are dependent upon the extent to which statistical assumptions are met. Numerous works have reported the lack of robustness of the *F* ratio (and resulting $p_{\text{calculated}}$) in the presence of an unbalanced design and heterogeneous variance (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). Monte Carlo evidence has demonstrated the negative impact heterogeneity of variance can also have on estimates of practical significance in one-way ANOVA designs (Keselman, 1975; Skidmore & Thompson, 2013). Of course, the assumption of homogeneity of variance is never fully met in applied research. A practical question for researchers then is “whether the plausible violations of the assumptions have serious consequences on the validity of probability statements [and estimates of effect sizes] based on the standard assumptions” (Glass et al., 1972, p. 237).

Statistical reporting reform efforts have acknowledged the necessity of an estimate of practical significance (i.e., effect sizes) to supplement null hypothesis statistical significance testing (NHSST) (e.g., Nakagawa & Cuthill, 2007; Schmidt, 1996; Thompson, 1996). At present, editorial policies at more than 20 journals require authors to report estimates of effect sizes (Grissom & Kim, 2012). In the latest American Psychological Association (APA) publication manual, effect sizes have been identified as a necessary element to “convey the most complete meaning of results” (American Psychological Association., 2010, p. 33). More recently, the journal, *Basic and Applied Social Psychology* (*BASP*), has banned *p*-values and even confidence intervals and instead requires authors to provide “strong descriptive statistics, including effect sizes” (Trafimow & Marks, 2015, p. 1).

Methodological research reviews are frequently used to identify trends in quantitative research practice. Such reviews of practice are important in that “journals both create and mirror their fields” (Silverman, 1987, p.40). All too often methodological research reviews have found a substantial gap between recommended inferential methods and the methods adopted by applied researchers. For instance, Keselman et al. (1998) presented a review of researchers’ ANOVA practices in prominent journals from 1994 and 1995, including validity assumptions, sample sizes, and effect size indices. The review results indicated that researchers (1) rarely verified that ANOVA distributional assumptions were satisfied, (2) typically used regular ANOVA tests that were not robust to assumption violations, (3) rarely reported effect size statistics, and (4) rarely performed power analyses to determine the sample size requirements. Understanding researcher practices provide an opportunity to make recommendations about best practices, offer guidance on graduate training, and provide a basis for what statistical knowledge is needed to read, engage in, and contribute to a field. Today, almost 20 years after the Keselman et al. (1998) review, the present authors investigate the extent to which ANOVA practices have changed.

Objective of the Present Review

The objective of the present review is threefold. First, the current state of ANOVA reporting practices is provided. Second, a comparison is made between present and prior ANOVA reporting practices (i.e., Keselman et al., 1998). Finally, recommendations are offered regarding the remaining pernicious issues in ANOVA reporting practices.

Method

Three APA journals were chosen for this review: *Journal of Applied Psychology* (*JAP*)*, Journal of Counseling Psychology* (*JCP*)*, Journal of Personality and Social Psychology* (*JPSP*). These journals have been selected as target journals in previous methodological reviews (e.g., Edgington, 1964, 1974; Kieffer, Reese, & Thompson, 2001), and thus provide for a longitudinal comparison. Further, because these are APA journals, it is expected that APA reform efforts, including appropriate reporting of statistical and practical significance, are encouraged in these publications.

Inclusion/Exclusion Criteria

All articles published in 2012 from *JAP*, *JCP*, and *JPSP* were collected. A total of 87 entries were located for *JAP*, 61 for *JCP*, and 147 for *JPSP*. In the preliminary electronic review, 19 key terms were used for screening. The following key terms were selected as they are often associated with the ANOVA *F*-test and alternatives to the *F*-test thereby maximizing the probability of identifying potential articles for inclusion:

> ANOVA, OVA, Factorial, Non-Factorial, *F*-test, Omnibus test, One-Way, Two-Way, Multi-Way, Brown-Forsythe, Welch, Mann-Whitney *U*, Kruskal-Wallis, Friedman, *t*-test, ANCOVA, Means test, James, *Post hoc*.

Articles identified for potential inclusion then underwent a manual review. Articles that did not contain any of the key terms were excluded from further review. This procedure resulted in a total of 224 articles (*JAP*: 68, *JCP*: 39, and *JPSP*: 117). The manual review further excluded six articles: five of which were qualitative research articles and one article was retracted due to academic fraud. The remaining 218 articles were subjected to a manual coding process. A coding scheme was developed and transferred to an online questionnaire. Six variables were coded for the 218 articles that underwent a manual coding process: (1) types of means tests; (2) types of *F*-tests; (3) textbooks or article references used to justify the use of ANOVA; (4) statistical packages reported; (5) terms used to qualify effect sizes; and (6) number of other analysis of variance means tests conducted, aside from between-subjects fixed-effects ANOVA, (i.e., MANOVA, MANCOVA, mixed-effects ANOVA, and repeated measures). Each element of the coding scheme is discussed in greater detail in the results section. Eighty-two articles that reported the use of between-subjects fixed-effects ANOVA *F*-tests, which are the most popular data-analytic technique among all analysis of variance *F*-tests, underwent the full coding process.

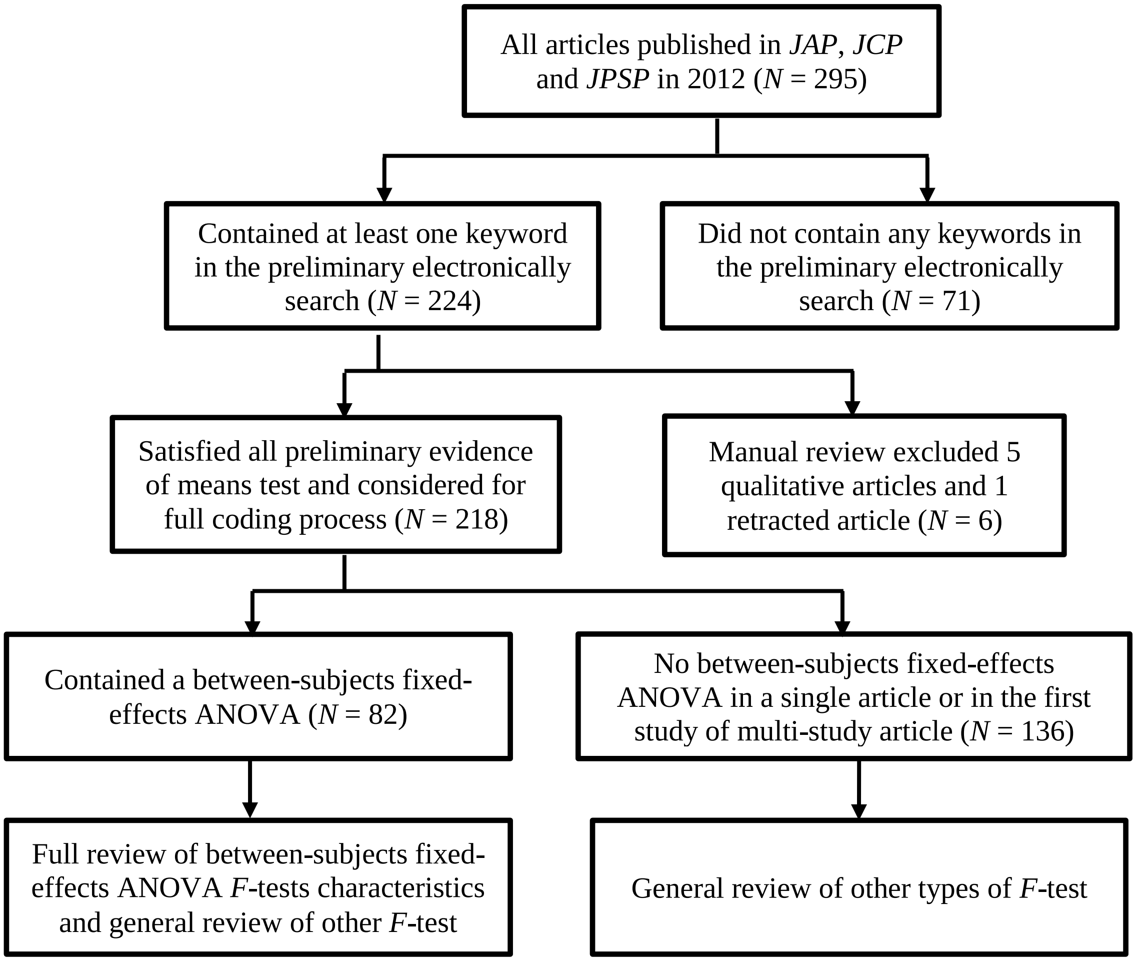
In the full coding process, an additional 13 characteristics were coded: (1) number of between-subjects ANOVA *F*-tests; (2) reported statistical violation assumptions; (3) reported method to test for violation assumptions; (4) reported methods to deal with assumption violations; (5) number of ways; (6) number of levels for each way; (7) design type (factorial or non-factorial); (8) reported *post hoc* tests; (9) ratio of the largest to smallest standard deviation; (10) ratio of the largest to smallest group size; (11) sample sizes; (12) ways *p* values were reported; and (13) reported effect sizes. For additional details, see Table 1.

Table 1
Coding Scheme for Online Questionnaire

Coding questions	Options
General Review	
Types of mean test	*t*-test, *F*-test, Brown and Forsythe, Welch, Mann-Whitney *U,* Kruskal-Wallis, Friedman, planned contrasts, others
Types of *F*-test	Between-subjects ANOVA(independent *t*-test), ANCOVA, MANOVA, Mixed ANOVA(repeated measures, paired *t*-test), others
Textbook or article references to justify the use of ANOVA	Not given, others
Report of statistical packages	none was given, SPSS, STATA, SAS, R, other
Use the terms of "small", "medium", or "large" to quantify effect sizes	None; Cohen (1988); Cohen (1990); Cohen (1992a); Cohen (1992b); Cohen (1994); others
Number of uncoded *F*-tests research techniques	MANOVA, mixed ANOVA, repeated measures, paired *t*-test, random-effects ANOVA
Full Review	
Number of between-subjects fixed-effects ANOVAs	1, 2, 3, 4, 5, more than 5
Report of statistical violation assumptions	None, independence of observations, variance homogeneity, distribution (normality)
Report of methods to deal with assumption violations	Nothing because assumptions were not mentioned; nothing because assumptions were not violated; transformation; use of nonparametric analyses; winsorizing and trimming; converting continuous variables to categorical variables.
Report of method to test for violation assumptions	None, Levene's, Bartlett's test, test was run, but no name was given
Indicate the number of ways, levels, and design type in ANOVA.	
Report of *post hoc* tests	None, LSD, Bonferroni, Sidak, Scheffe, Tukey, Duncan, Hochberg, Gabriel, Walter-Duncan, Dunnet, others
Ratio of the largest to smallest standard deviation	
Ratio of the largest to smallest group size	
Sample sizes	
The way *p* values were reported	$p < .05$, $p < .01$, $p < .001$, $p > .\#\#\#$, $p = .\#\#\#$, *n*s, others
Reported effect sizes	None, η^2, partial η^2, ξ^2, ω^2, *d*, *f*, *r*, other

The first step of the coding process was to exclude those articles that did not include any keywords. However, an article that contained a keyword during the electrical review process did not necessarily contain *F*-tests. For example, an article that contained the keyword "Friedman" did not necessarily have the Friedman test. "Friedman" might have been an author of this article or the author of referenced articles. Therefore, the number of articles containing any *F*-tests was smaller than the original 224 articles identified for review. Also, MANOVAs, between-subjects random-effects ANOVAs and repeated measures (including mixed-effects ANOVA and paired *t*-test) were excluded from the full coding process. Finally, to maintain consistency in the coding process and because some articles contained multiple studies per article, the decision was made to code the first five ANOVAs within the first study when multiple studies were reported per article. Therefore, if the study or the first study contained more than five ANOVAs, only the first five ANOVAs were coded. This decision underestimated the proportion of ANOVA *F*-tests used in the three journals because the authors did not code ANOVA *F*-tests reported other than the first study and did not code the remaining ANOVA *F*-tests when the first five ANOVA *F*-tests were coded. Consequently, the resulting total number of articles that contained at least one between-subjects fixed-effects ANOVA and thus were subjected to the full coding process was 82. Figure 1 presents the article inclusion criteria and the article review process.

Figure 1. Article inclusion criteria decision sequence.


All articles published in JAP, JCP and JPSP in 2012 (N = 295)
Contained at least one keyword in the preliminary electronically search (N = 224)
Did not contain any keywords in the preliminary electronically search (N = 71)
Satisfied all preliminary evidence of means test and considered for full coding process (N = 218)
Manual review excluded 5 qualitative articles and 1 retracted article (N = 6)
Contained a between-subjects fixed-effects ANOVA (N = 82)
No between-subjects fixed-effects ANOVA in a single article or in the first study of multi-study article (N = 136)
Full review of between-subjects fixed-effects ANOVA F-tests characteristics and general review of other F-test
General review of other types of F-test

Articles selected for inclusion were evaluated twice. During the first round of coding, the primary coder met with the secondary coder every other week to discuss issues that were encountered during the process of coding to ensure consistency. The secondary coder also randomly coded 10% of the articles. Any discrepancies in coding were discussed and resolved. The first coder then reviewed all of the articles a second time to identify and correct any possible errors that may have occurred during the initial round of coding. Finally, the two coders discussed any discrepancies between the first and second rounds of coding until a consensus was reached.

Results

General Review

Overview of the types of *F*-tests. Among the 218 articles that underwent manual review, 53% (n = 116) of the articles reported analysis of variance *F*/*t*-tests, of which 82 articles contained between-subjects fixed-effects ANOVA *F*-tests, nine articles contained between-subjects random-effects ANOVA *F*-tests, 40 articles contained within-subject ANOVA *F*-tests (including mixed-effects ANOVAs and repeated measures), and 15 articles contained MANOVA/MANCOVA *F*-tests. Because many articles reported more than one type of analysis of variance *F*-test (for instance, an article might have contained both between-subjects fixed-effects ANOVA *F*-tests and between-subjects random-effects ANOVA *F*-tests), the sum of all types of *F*-tests is greater than the total number of articles that reported analysis of variance tests (see Table 2 for details).

Table 2
Journal Source and Frequency of OVA Reported

Journal/ Statistic	Between-subjects fixed-effects ANOVA / independent *t*-test	Between-subjects random-effects ANOVA	Mixed-effects ANOVA / repeated measures / paired *t*-test	MANOVA / MANCOVA
JAP	20	9	6	6
JCP	16	0	3	5
JPSP	46	0	31	4
Total	82	9	40	15
Percentage	37.61%	4.13%	18.35%	6.88%

Note. Percentage reflects the N = 218 preliminarily reviewed quantitative articles. The sum of the percentages does not equal to 100% because: (a) not all of the 218 articles contain an analysis of variance *F*/*t*- test, and (b) some articles reported more than one type of *F*/*t*-test.

Between-subjects fixed-effects ANOVA *F*-tests were the most popular *F*-tests among all analysis of variance techniques. The between-subjects random-effects ANOVA *F*-tests were also fairly common, as 22.48% of the total reviewed articles (i.e., 4.13% for between-subjects random-effects ANOVA and 18.35% mixed-effects/repeated measures ANOVAs) contained analysis of variance *F*-tests that treated at least one way as random. Discernible patterns by analysis of variance technique were observed in the three journals. In *JAP*, to estimate homogeneity within groups, one-way random ANOVA *F*-tests were used to calculate the intraclass correlation coefficient; while in *JPSP*, because personal traits are often the major research focus, mixed ANOVA designs or repeated measures were frequently used. On the other hand, MANOVA/ MANCOVA was a rarity across the three journals.

Use of references. Among the 218 articles that underwent manual review, only six articles cited references to justify the use of ANOVA techniques in the method section. The citations were used in relation to possible assumption violations, to justify the use of alternative methods, and to assess the suitability of data aggregation. For example, three articles, one each in *JAP*, *JCP*, and *JPSP*, used references to justify the use of alternative methods when ANOVA assumptions were violated (e.g., heterogeneity of variance and non-normality) and when an unbalanced design was present. In addition, an article in *JPSP* cited a reference as the criterion to select a covariate. Two articles in *JAP* cited a reference focused on data aggregation (viz., Bliese, 2000) to justify aggregating data for analysis. Still, the overwhelming majority of the articles that used ANOVA did not justify their analytical choices with a reference in the method section.

Benchmarks for effect sizes. Even though Cohen (1988) cautioned against the inherent subjectivity of using terms such as "small", "medium", and "large" for effect sizes in the presence of existing literature that could more accurately describe the magnitude of an effect within a particular discipline, these nebulous terms still occasionally appear in published articles. Among the 218 manually reviewed articles, 23 articles used these terms, and 10 articles justified the use of these terms with references. Cohen (1988) recommended his benchmarks only be used "when no better basis for estimating the ES index is available" (p. 25), however, nine out of the 10 articles that provided references to support the use of the benchmarks cited Cohen. Specifically, two articles cited Cohen (1988) and seven articles referenced Cohen (1992).

Software packages. Reporting the statistical software used for data analysis was not common in the three journals examined in the present study. Among the 218 reviewed articles, 80% did not mention anything about the package used for data analysis. Among those articles that reported the name of the software, SPSS was noted most frequently (n = 22), followed by SAS (n = 12), HLM (n = 5), MPLUS (n = 4), LISREL (n = 3), R (n = 1), and STATA (n = 1).

Full Review

The proportion of different means tests. Among the 82 articles that contained between-subjects fixed-effects ANOVAs a total of 261 means tests were documented, of which, 138 (52.9%) were traditional ANOVA *F*-tests, 108 (41.4%) were independent *t*-tests, seven (2.7%) were ANCOVA *F*-tests, two were Welch alternative *F*-tests, five were planned contrast tests, and one was a nonparametric bootstrapping analysis test. See Table 3 for additional details. It appears that most researchers adopted the traditional ANOVA *F*- and *t*- tests (94.3% of all documented means tests). The use of ANCOVA was not as common as the other two types of means tests in the three reviewed journals.

Assessment of validity assumptions. When researchers use ANOVA as an analytic technique, a very important first step is to verify the distributional assumptions. If these assumptions are not reasonably met, results generated from the ANOVA test will "at best, [be] somewhat different from what they should be and, at worst, worthless" (Keselman et al., 1998, p. 351). For most (94.3%) of the means tests reported, researchers neglected to provide *any* information on statistical assumptions tested. Only 2.7% (n = 7) of the authors addressed the possible violation of homogeneity of variance, 4.6% (n = 12) addressed possible violations of normality, and no authors addressed the independence assumption. Additional details are provided in Table 4.

Table 3
The Proportion of Various Means Tests in JAP, JCP, and JPSP

Journal	*t*-test	ANOVA	ANCOVA	Welch test	Planned Contrasts	Other
JAP	38	34	3	0	0	0
JCP	26	16	1	0	0	0
JPSP	44	88	3	2	5	1
Total	108	138	7	2	5	1
Proportion (%)	41.4	52.9	2.7	0.8	1.9	0.4

Note: The proportion reflects N = 261 means tests. The sum of the proportions did not exactly equal 100% due to rounding errors.

When authors did report using tests that evaluated ANOVA assumptions, authors usually did not report the name of the tests that were used to assess the assumption. Researchers seemed more concerned with non-normality issues, even though heterogeneity of variance within an unbalanced design may result in more severe departures from the true values (Glass, Peckham, & Sanders, 1972; Skidmore & Thompson, 2013). In only two means tests did researchers directly mention that Levene's test was used to test the homogeneity of variance assumption, and these occurrences came from a single article. When assumptions were violated, transformation was the most frequently reported resolution; eight tests used this adjustment to obtain a more robust estimate. Winsorizing and trimming were the next most frequently used methods, which were reported four times in the 261 identified means tests. One means test used nonparametric analysis, and two failed to report what procedure was used to address the violation issue.

The ways and levels for ANOVA means tests. Displayed in Table 5 are the 261 documented means tests distributed based on the number of ways and levels. The most commonly used ANOVA was the one-way ANOVA, which comprised 80.8% (n = 211) of the documented means tests. In the one-way ANOVA, 82.9% (n = 175) were two-group mean difference tests, 11.4% (n = 24) were three-group mean difference tests, 3.3% (n = 7) were four-group mean difference tests, and 1.4% (n = 3) did not provide enough information to be able to determine a description. The use of two-factor ANOVA means tests in psychological research was used quite often too; 17.2% (n = 45) of documented means tests were two-way ANOVA means tests. Within the two-way ANOVAs, 86.7% were 2 × 2 ANOVA, 6.7% were 2 × 3 ANOVA, 2.2% were 2 × 4 ANOVA, and for 4.4%, it was impossible to determine. Three-way ANOVA and four-way ANOVA were only occasionally used in psychological research and had no more than two levels in each way. These results are especially useful to methodologists who often conduct simulation studies to estimate how violations of assumptions will affect the accuracy of statistical results. Incorporating various degrees of deviation from distributional assumptions (i.e., non-normality, heterogeneity of variance, and dependent residuals) for the most commonly used research designs in behavioral science research maximizes the value of Monte Carlo simulation results.

Table 4
How Assumption Violations Were Addressed

What statistical violation assumptions were mentioned?			What tests (if any) were performed to test for / violation assumptions?			How were assumptions violations addressed?		
	n	%		*n*	%		*n*	%
none	246	94.30%	None	246	94.30%	nothing was done/ assumptions were not reported	246	94.30%
independence of observations	0	0.00%	Levene's	2	0.80%	something was done but didn't mention what procedure was used	2	0.80%
homogeneity of variance	7	2.70%	Shapiro-Wilks	0	0.00%	transformation	8	3.10%
distribution (normality)	12	4.60%	Bartlett's test	0	0.00%	use of nonparametric analyses	1	0.40%
			Test was run but no name was given	13	5.00%	winsorizing and trimming	4	1.50%

Note. The percentage reflects the 261 means tests. The sum of percentage for the first column was not equal to 100% because one ANOVA test addressed two types of violations.

ANOVAs appearing in articles were used in different ways. Some ANOVAs were used to answer the main research question (e.g., an ANOVA test for an ANOVA research design), while other ANOVAs were used for testing the equality of groups prior to the main research question (e.g., if the results differed by gender). Thus, all 261 ANOVA means tests were coded as "main research question ANOVA test" for the former ANOVA tests, and the latter ANOVA tests were coded as "conditional assumptions ANOVA test".

Among one-way ANOVA means tests, 95 out of 211 were used for main research questions. In two-way ANOVA means tests, 36 out of 45 were used for main research questions. All three or more way ANOVA means tests were used for main research questions.

Group size. Researchers did not often disclose group sizes when reporting ANOVAs results. Indeed, more than 80% (n = 210) of the means tests reported no information about the group sizes. Among the 51 means tests where group sizes were discernible, only six means tests were balanced (equal number of participants across groups), the other 45 means tests all had unequal group sizes. Of those unbalanced designs, 23 had a group size ratio (from high to low) smaller than two, 18 had the ratios between two and 10, and four had the ratios larger than 10. The largest ratio in an unbalanced ANOVA observed in the reviewed articles was 70! These results are especially disheartening in light of previous findings where even group size ratios as small as two in the presence of heterogeneous variance resulted in biased effect size estimates (Skidmore & Thompson, 2013). The distribution of group size ratios is shown in Figure 2.

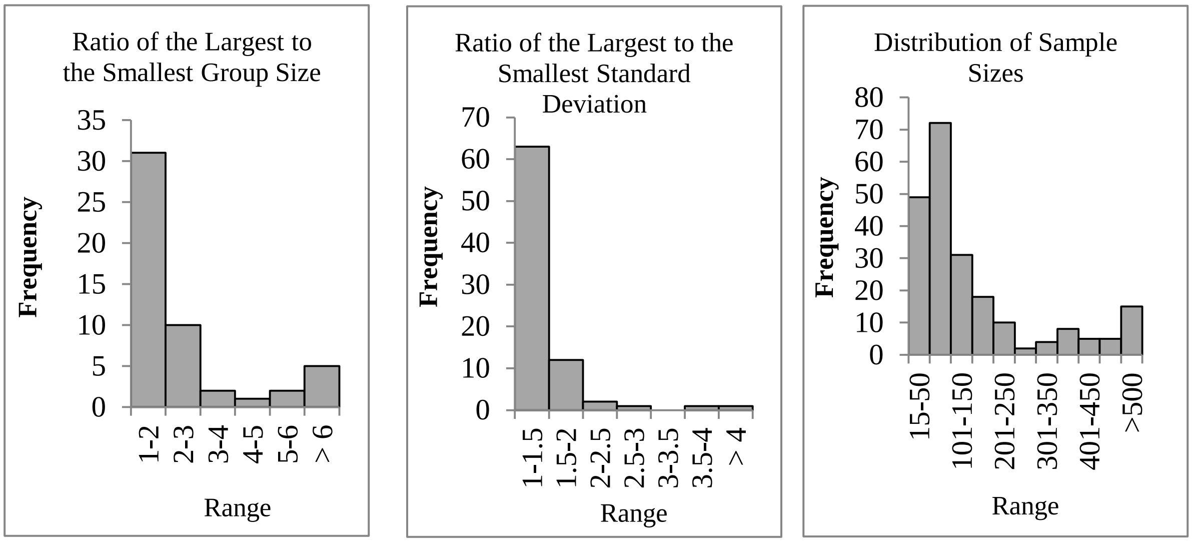
Table 5
Frequency of Ways and Levels for Reported ANOVAs

Number of Ways	Frequency	% (way)	Number of Levels	Frequency	% (level)
One-way	211	80.8%	Two-group	175	82.9%
			Three-group	24	11.4%
			Four-group	7	3.3%
			Five-group	2	0.9%
			Not mentioned	3	1.4%
Two-ways	45	17.2%	2 × 2	39	86.7%
			2 × 3	3	6.7%
			2 × 4	1	2.2%
			2 × ?	2	4.4%
Three-ways	4	1.5%	2 × 2 × 2	4	100.0%
Four-ways	1	0.4%	2 × 2 × 2 × 2	1	100.0%

Note. The percentage of the number of factors reflects the total number of documented means tests (N = 261). The percentage of the number of levels reflects the total number of means tests within the same number of ways (i.e., n = 211 one-way, n = 45 two-way, n = 4 three-way, and n = 1 four-way).

Variance. In comparison to the group size, more researchers disclosed the standard deviation or variance for each group. Still, the overall number of means tests where information about standard deviation or variance was reported was small. In only 31% (n = 80) of the means tests was the standard deviation for each group reported. Among the 80 means tests where the standard deviation was reported, 63 means tests had a ratio of the standard deviation (from high to low) smaller than 1.5, 12 means tests had a ratio between 1.5 to two, and five means tests had a ratio greater than two. The largest ratio of standard deviation in the coded articles was 8.2! The distribution of standard deviation ratios is shown in Figure 2. For the remaining 70% of the means tests reported researchers either did not conduct tests (e.g., Levene's or Bartlett's tests) to validate the homogeneity of variance assumption or did not report doing so. It is important to note that when the standard deviation of each group varies too much, especially in the presence of an unbalanced design, the p-values and effect sizes (e.g., $R^2$and η^2) generated are essentially meaningless.

Figure 2. The distributions of ratios for group size and standard deviation (from the largest to the smallest) and the distribution of sample sizes.



Sample sizes. The given sample size reported in a study is often different than the sample size reported in a given analysis as variables used in each analysis can have different levels of missingness. Therefore, as *F*-tests are the object of interest, the sample size tabulated in the present study were specifically for each *F*-test. Most coded ANOVA *F*-tests had moderately large sample sizes. Among the 261 means tests, 220 reported the total sample sizes that ranged from 15 to 27,565, of which, 50 means tests had sample sizes that ranged between 15 and 50, 72 means tests had sample sizes that ranged between 51 to 100, 31 means tests had sample sizes that ranged between 101 to 150, 18 means tests had sample sizes that ranged between 151 to 200, and the other 49 means tests had sample sizes greater than 200. Only eight means tests used sample sizes smaller than 30. The distribution of sample sizes is displayed in Figure 2.

Pairing. "Pairing" refers to the situation where heterogeneity of variance exists together with unequal group sizes. There are two types of "pairing" possible when using ANOVA to test mean differences across unequal group sizes: "positive pairing", defined as the larger group having the larger variance and the smaller group having the smaller variance, and "negative pairing", defined as larger group having smaller variance and smaller group having larger variance. Previous simulation studies have revealed that when negative pairings exist, estimates of effect sizes have positive sampling errors bias; when positive pairing exist, estimates of effect sizes have negative sampling errors bias (Skidmore & Thompson , 2013). Among the 261 tests, article authors provided enough information to discern nine positive pairings and nine negative pairings for 18 ANOVAs. For all other means tests, it was not possible to discern the type of pairing because only the variance for each group, only the sizes for each group, or neither group variance nor size was provided.

***p*-Value and Effect size.** Researchers' reliance on *p*-values continues to dominate statistical reporting practices. Article authors often neglected to report the group sizes, group variances, validity of ANOVA assumptions, effect sizes, and other pertinent information, but they always reported the p-value. Indeed, all of the documented 261 means tests reported p-values for ANOVA tests. Because multi-way ANOVAs may have more than one F-tests with more than one corresponding p-value, all 292 p values were documented, of which, 36 p values were reported as "$p < .05$", 52 p values were reported as "$p < .01$", 50 p values were reported as "$p < .001$", 13 p values were reported as "*ns*" without providing a value or range, 35 p values were reported as "$p > .\#\#\#$", and 11 p-values were reported as $p < .\#\#\#$ ("*.###*" denotes a value other than the commonly used benchmarks such as .05, .01, and .001). Only 95 reported the exact p-values. Perhaps these reporting practices are reflective of dichotomous thinking about p-values. Thompson (1989) has noted that such thinking can lead researchers to mistakenly believe that the import of their study results is determined by whether $p_{\text{calculated}}$ is greater than α_{critical} or not. In support of this premise, some researchers only reported "*ns*" (i.e., non-significant) or "significant" without providing the critical α as a criterion. Reporting p-value as "*ns*" or "significant" demonstrates a lack of transparency in reporting because the same value, for example, $p = .03$, can be considered "*ns*," if $\alpha_{\text{critical}} = .01$ but can be claimed as "significant," if $\alpha_{\text{critical}} = 0.05$.

Compared to the reporting of p-values, effect size reporting is woefully inadequate. Among the 261 means tests, 119 means tests reported effect sizes, which is 45.6% of the total documented means tests. Among those that reported effect size, 53 (44.5%) reported partial η^2, 39 (32.8%) reported Cohen's d, 25 (21.0%) reported η^2, and two (1.7%) reported ξ^2. Thus, partial η^2 was the most frequently reported effect size. Similarly, in a review of effect size reporting practices from articles published in 2002 from a sample of 10 educational research journals, partial η^2 was the most frequently reported effect size when analysis of variance procedures were used (Alhija & Levy, 2009). In comparison, Kirk (1996) reviewed four 1995 volumes of APA journals, including *JAP* and *JPSP*, and noted the type of effect size reported. In *JAP*, η^2 was reported six times, ω^2 was reported four times, and Cohen's d or Hedges' g was reported three times. In *JPSP*, η^2 and η were reported five times and Cohen's d or Hedges' g was reported four times. These numbers, however, are confounded by the fact that Kirk used the designation of 'variance-accounted-for' to record some effect sizes, which could have included R^2, η^2, and ω^2, because authors failed to identify the type of statistic that was used. Kirk noted this practice as "one of the many examples of sloppy reporting in the literature" (p. 753). Nonetheless, Kirk recorded variance-accounted-for effect sizes 19 times in *JAP* and 43 times in *JPSP*.

***Post hoc* test.** Excluding 219 means tests that did not need *post hoc* comparison (175 one-way two levels comparison, and 44 multi-way two levels comparison) and five means tests with an unclear number of levels, there was a total of 37 means tests for which a *post hoc* test could be reported. Thus, out of the 37 documented means tests that could have reported a *post hoc* test, 20 (54.1%) reported *post hoc* tests: two were LSD, one was Bonferroni, six were Tukey, and 17 were not reported. Thus, for 17 (46.0%) means tests with more than two levels *post hoc* tests were not reported.

Discussion

The original intent of this study was to determine to what extent test assumptions were met for these means tests, however, such determination was not possible because most article authors did not provide sufficient information such as means, standard deviations, and group sizes. It follows then that it is also not possible to understand the reason for researcher's lack of reporting. Possible reasons for researchers' failure to report assumptions tested could be simple oversight, journal space limitations, or, hopefully not commonly the case, because assumptions were not tested.

Among the coded different types of means tests, the majority were the regular ANOVA F-test, in the contrast, the use of ANCOVA were very uncommon (seven out of 261 documented means tests). The low frequency might due to that the very strict assumptions for ANCOVA (i.e., homogeneity of regression, extremely reliable measurement of covariates, and interpretable residualized dependent variables) are difficult to meet (Thompson, 2004), and further restricted its usage. Further, criticisms of ANCOVA misuse might also deter researchers' more frequent usage (Bartlett, 1949; Evans & Anastasio, 1968; Porter & Raudenbush, 1987; Ree & Carretta, 2006; Schneider, Avivi-Reich, & Mozuraitis, 2015).

Conclusions and Recommendations Concerning the ANOVA Practices

In educational and psychological research, the ANOVA *F*-test has historically been identified as the most popular data-analytic technique (Edgington, 1974; Goodwin & Goodwin, 1985; Howell, 2011). Indeed, because of its prevalence and versatility, ANOVA has been one of the most critical components of graduate statistical training in psychology (Aiken et al., 1990; Aiken, West, & Millsap, 2008; Ord et al., 2016). Although there is some evidence that ANOVA is not as prevalent as it once was (Skidmore & Thompson, 2010), the results of the present study demonstrate that ANOVA *F*-tests are nonetheless widely used in a variety of contexts. In fact, Skidmore & Thompson (2010), noted that between 1990 and 1997, across the 10 education and psychology journals reviewed, roughly 40% of the articles reported the use of ANOVA and *t*-tests. Similarly, in our findings, in 39% of the articles (i.e., 116 out of the 295 articles in *JAP*, *JCP*, and *JPSP* in 2012), *F*-tests were reported. Further, in a review focused on ANOVA techniques published almost two decades ago, ANOVA was "used most frequently within the context of one-way and factorial between-subjects univariate designs" (Keselman et al., 1998, p. 353). This trend can be observed in our study results as well, where 94.3% of the means tests identified were between-subjects fixed-effects ANOVAs. Whether ANOVA is used as the primary analytical method or to test conditional assumptions, such as whether or not a group lost due to attrition differs from the remaining participants, ANOVA continues to be regularly used in undergraduate (Friedrich, Buday, & Kerr, 2000) and graduate training (Ord, Ripley, Hook, & Erspamer, 2016) and in social science in research (Aiken et al., 2008; Howell, 2011; Meyers et al., 2008).

If for no other reason than the prevalence of the use of ANOVA across disciplines, ANOVA reporting practices have a great impact on the field. Although ANOVA has been around since the early 1920's (David, 1995) and cautions regarding ANOVA practices were voiced over 45 years ago (Glass, Peckham, & Sanders, 1972), reporting practices for ANOVA *F*-tests in the three reviewed journals remain problematic. Twenty years ago, Keselman et al. (1998) noted that most behavioral science researchers automatically conduct "standard" analyses, which rely on strict assumptions and may result in misleading or erroneous findings when assumptions are violated. In our results, it is clear that most researchers either did not feel the need to test for ANOVA assumption violations or did test, but failed to report their results. Thus, it is apparent that statistical assumptions are taken for granted to the extent that no more than 6% of the reported means test results even mentioned statistical assumptions. Of those researchers who attended to the importance of verifying ANOVA assumptions, non-normality was more likely to be of concern than heterogeneity of variance and unequal group sizes. Although a balanced design is not an assumption, unbalanced designs have implications in the presence of heterogeneity of variance. In the present study, an overwhelming majority of researchers (88.2%) who reported group sizes reported unbalanced designs. Yet simulation research results have provided evidence of the deleterious effects of variance heterogeneity on both Type I error rates (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996) and effect sizes (Skidmore & Thompson, 2013).

There has been some progress in ANOVA reporting practices. The reporting of effect sizes in conjunction with *p*-values has markedly increased. During the time Keselman et al. (1998) and Kirk (1996) conducted their reviews (i.e., articles published in the 1994 or 1995 issues) effect sizes were almost never reported. At the time the present review was conducted effect sizes were reported alongside *p*-values for nearly half of the ANOVA tests reported. This is consistent with a prior review of effect size reporting practices (Alhija & Levy, 2009) in journals where effect size reporting was explicitly required, where 47% of the *t*-test and 69% of the ANOVA results reporting included effect sizes. Similarly, in another review focused on education and psychology journals published between 2005 and 2007, 49% of articles provided effect sizes (Sun, Pan, & Wang, 2010). Nonetheless, no researchers reported confidence intervals, nor confidence intervals for effect sizes, even though methodologists have recommended using confidence intervals to replace statistical significance tests (e.g., Meehl, 1997; Thompson, 1999, 2002). Of those who reported effect sizes, partial η^2 was the most commonly reported effect sizes.

The prevalence of partial η^2 highlights another concern: reporting of the statistical package used for data analysis, which is seldom disclosed. The rapid growth of statistical software has provided more options for quantitative data analysis. However, there are always concerns about the accuracy of the software because researchers have identified errors, even for the most popular software, like SPSS (Levine & Hullett, 2002). If researchers do not report the package adopted for data analysis, tracking those mistakes is not possible. Still another concern with the prevalence of partial η^2 is the confusion that has been noted in textbooks and in the literature regarding η^2 and partial η^2 (viz., Richardson, 2010). Interestingly, Jacob Cohen (1973) had explicated the differences between η^2 and partial η^2 almost 40 years prior to Richardson's article and yet, historically, one of the most commonly used statistical software, SPSS, had mislabeled the partial η^2 as η^2 (Levine & Hullett, 2002). This error has been corrected in more recent SPSS versions. Still, when using the "General Linear Model" option, regardless of the number of ways in the ANOVA, the reported effect size is labeled as "partial η^2". Of course, in a one-way ANOVA partial η^2 and η^2 are the equivalent; nevertheless, in a one-way ANOVA, η^2 is the appropriate label (Cohen, 1973). In the present study, although 80.8% of the means tests were one-way ANOVAs, the most commonly reported effect size was partial η^2. So it seems as if statistical programs influence what researchers report. Therefore, it is important to acknowledge that researchers should not take for granted that statistical software, and the corresponding output, is correct. Speaking of the topic of reasonable expectations for statistics, Bruce Thompson (2006) noted, "Formulas for descriptive statistics were not transmitted on stone tablets given to Moses, nor otherwise divinely authored. Instead, different human people developed various formulas as ways to characterize quantitative data" (p. 32). Similarly, it is important to recognize the fallibility of the output provided by statistical software. After all, it is not the responsibility of the statistical software, but rather the researcher, to think.

The squeaky wheel of effect size reporting has indeed made a difference in reporting practices. On the other hand, what has not been emphasized is almost never observed—reporting of statistical assumption results. The question for all to reflect upon is, are assumptions merely a statistical nuisance that can be ignored or taken for granted? Or instead, is the testing of assumptions of statistical tests a necessary pre-requisite step to the veracity of interpreted results?

Author Notes: Yuanyuan Zhou is an Assessment Analyst in the Office of Medical Education at Texas A&M University College of Medicine. Her research interests center on statistical assumption violations; effect sizes; psychometric validation and refinement; and applications of advanced statistical methods in medical and higher education. Susan Troncoso Skidmore is an associate professor in educational leadership at Sam Houston State University. Her research interests center on statistical assumption violations; effect sizes; evidence-based practices; cultures of assessment; and the recruitment, development, and retention of underrepresented groups in secondary and post-secondary institutions.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32-50.

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

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger III, H. L., Scarr, S., . . . Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*, 721-734.

Alhija, F. N.-A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, 69, 245-265, doi:10.1177/00131644083152 66.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, 5, 207-212.

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

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician, 49*, 121-133. doi:10.2307/2684625

Edgington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist, 19*, 202-203.

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist, 29*, 25-26.

Evans, S. H., & Anastasio, E. J. (1968). Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, *69*, 225.

Friedrich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and commentary on undergraduate programs. *Teaching of Psychology, 27*, 248-257.

Gamst, G., Meyers, L. S., & Guarino, A. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. New York, NY: Cambridge University Press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.

Goodwin, L. D., & Goodwin, W. L. (1985). An analysis of statistical techniques used in the journal of educational psychology, 1979-1983. *Educational Psychologist, 20*, 13-21.

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 and Behavioral Statistics, 17*, 315-339.

Howell, D. C. (2011). *Fundamental statistics for the behavioral sciences* (7th ed.). Belmont, CA: Wadsworth, Cengage Learning.

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . 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.

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

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*, 612-625. doi:10.1111/j.1468-2958.2002.tb00828.x

Lix, L. M., Keselman, J. C., & Keselman, H. (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.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). *What if there were no Significance Tests* (pp. 393-425). Mahwah, NJ: Lawrence Erlbaum.

Ord, A. S., Ripley, J. S., Hook, J., & Erspamer, T. (2016). Teaching statistics in APA-accredited doctoral programs in clinical and counseling psychology: A syllabi review. *Teaching of Psychology*, 43, 221-226. doi:10.1177/00986283

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology*, 34, 383–392.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods*, 9, 99–112.

Schneider, B. A., Avivi-Reich, M., & Mozuraitis, M. (2015). A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Frontiers in Psychology*, 6, 1-12. doi:10.3389/fpsyg.2015.00474

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. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, 45, 536-546. doi: 10.3758/s13428-012-0257-2

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102, 989-1004, doi:10.1037/a0019507.Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22*, 2-6.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25-32.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 165-181. doi:10.1177/095935439992006

Willson, V. L. (1980). Research techniques in" AERJ" articles: 1969 to 1978. *Educational Researcher, 9*, 5-10.