

論文

The Role of Effect Size Reporting in Post-ANOVA Multiple Comparison: An Examination of Reporting Practices and a Simulation Study on Sample Size Impact¹

宮崎 啓² Kei Miyazaki

Abstract

The purpose of this study is to consider the role of effect size reporting in post-ANOVA multiple comparison analyses. The study has two folds. The first phase examines the number of effect size reports published in the JACET Journal from 2020 to 2024. It highlights the gradual increase in effect size reporting within ANOVA, while noting ongoing needs for discussions in post-ANOVA contexts. The second phase is a simulation study which further illustrates how sample size impacts significance estimates, revealing discrepancies between p-values and sample size. The findings emphasize the importance of reporting effect sizes alongside p-values to enhance the interpretability of multiple comparisons in TESOL and EFL research.

Keywords: Effect size reporting, ANOVA, Multiple Comparison, TESOL

1. Introduction

The promotion of effect size reporting in the field of Teaching English to Speakers of Other Languages (TESOL) and English as a Foreign Language (EFL) has gained over the past few decades. A primary concern driving this call is that the results derived from significance testing, such as t-tests and Analysis of Variance (ANOVA), can be heavily influenced by sample size, among other factors (Larson-Hall, 2016). By 2020, it had become increasingly standard practice to include effect size measures in the reporting of these statistical tests, as indicated by guidelines from the American Psychological Association (APA, 2020).

However, a notable gap persists in the TESOL literature, particularly in the reporting of effect sizes in post-ANOVA multiple comparison analysis, an area frequently utilized in educational research. Instead of effect sizes, researchers often draw conclusions solely based on whether or not a p-value exceeds the significant level. This paper aims to shed light on this critical issue by exploring two key areas: (a) an examination of effect size reporting in multiple comparison analyses published in the Japan

¹ I would like to thank the anonymous researcher of Tokai University for his cooperation of coding in this study.

² Professor, Department of English at Tokai University.

Association of College English Teachers (hereafter JACET) Journal over the past five years, and (b) a simulation study using the author's prior research data and the imaginary data to illustrate the impact of sample size on p-value.

By clarifying these points, this study seeks to enhance the understanding of effect size as an essential component of rigorous statistical reporting, thereby encouraging TESOL researchers and students, particularly those in graduate programs, to adopt more informative approaches in their research practices.

2. Background

2.1 Effect size

Statistical tests predominantly rely on a few critical factors: sample size, significance level, effect size, and statistical power (Mizumoto & Takeuchi, 2011). Among these elements, effect size is particularly salient as it quantifies the magnitude of the difference, thereby providing meaningful context to the statistical significance indicated by p-values. Indeed, while p-values inform researchers of the likelihood that observed results are due to chance (Larson-Hall, 2016), they do not convey the size of the effect or the practical significance of the findings. The introduction of effect size offers a more practical understanding of research outcomes.

For instance, one may conduct a statistical test and obtain a significant p-value (e.g., $p < .05$), suggesting that there is a statistically significant difference between groups. However, the same analysis might report a small effect size, indicating that the actual difference between the groups is small in a practical sense. Thus, relying solely on p-values can lead researchers to overestimate the importance of their findings, as significant p-values do not necessarily equate to substantial effect sizes.

This distinction is important, particularly in educational research, where the implications of findings can influence teaching practices and policy decisions. By reporting effect sizes alongside p-values, researchers can provide a more comprehensive view of their results, detailing not only whether a difference exists but also how meaningful the results are in practical senses.

Moreover, the relationship between sample size and effect size calls for attention. Effect size remains unaffected by sample size; thus, a study with a small sample might obtain a large effect size indicative of a practically significant difference, while a large sample might reveal a small effect that has limited implications for practice (Cumming, 2012). In addition, effect size also affects statistical power analysis. Researchers can determine the appropriate sample size necessary for their studies based on these calculations, thus planning effectively to avoid Type I and Type II errors³. The value of power analysis is heightened when investigators strive to detect meaningful changes and ensure that their studies are adequately equipped to test their hypotheses.

With this in mind, understanding of both statistical testing and effect size is indispensable for researchers within the field of TESOL and EFL. Effect size enhances the interpretation of research

³ In Type I error, a researcher concludes that there is an effect for some treatment, but the truth is that there is actually no effect; Type II error means that a researcher is not able to detect the effect in the sample "when it actually does in the population" (Larson-Hall, 2016, p. 156).

findings as well as complements the information provided by p-values. Researchers should prioritize the reporting of both p-values and effect sizes to help provide better understanding of their results in educational research.

2.2 Effect size reporting in the TESOL and EFL fields

The TESOL and EFL fields have not seen extensive studies focused specifically on effect size reporting practices. Among a small number of the studies, Plonsky's (2013) analysis revealed an effect size reporting rate of 25% when examining over 600 articles from both *Language Learning* and *Studies in Second Language Acquisition*. Similarly, Lindstromberg's (2016) findings, which posited that 49% of the 96 (quasi)experimental studies examined in 90 articles spanning 19 volumes from 1997 to 2015 in *Language Teaching Research* included effect size reporting. Even though detailed coding concerning papers that utilized multiple statistical approaches remained insufficiently addressed and both studies did not clarify how such papers were coded (Wei et al., 2019), such studies emphasize the necessity for a dedicated focus on effect size reporting to establish clearer standards and expectations in journals.

ANOVA, in particular, is one of the most frequently used methods in the major TESOL and EFL academic journals (Wei et al., 2019). Plonsky and Oswald (2014) states that the effect size reporting practices have dramatically increased. In fact, reporting effect sizes for ANOVA seems to have normalized in the past five years. However, as mentioned, reporting effect sizes in post-ANOVA multiple comparison analyses is insufficiently practiced. At this point, though the survey was focused on educational technology, the similar condition was pointed out by Kido and Ikeda (2022). In their research, which examined effect size reports published in a 10-year period in the journal of Japan Journal of Educational Technology, the analysis of 188 studies that conducted multiple comparisons, only four provided a description of effect sizes. While many of these multiple comparisons were employed as post-hoc tests within the framework of ANOVA, effect sizes were reported in 41 out of 204 analyses in ANOVA contexts, whereas they were scarcely mentioned in multiple comparisons. This indicates that even in studies where effect sizes were reported in the context of ANOVA, the same studies often failed to report them for multiple comparisons. One possible reason for this discrepancy is that, in the case of multiple comparisons, documenting the effect sizes for various pairwise differences might be perceived as redundant.

In a similar vein, within the Japanese TESOL or EFL community, the reporting of effect sizes in post-ANOVA multiple comparisons analyses has traditionally been underexplored. This study aims to fill this gap by a two-fold examination. First, the author conducted an examination of effect size reporting in multiple comparison analyses published in the JACET Journal over the past five years (2020-2024). The JACET Journal is a well-known journal that has led the English language education in Japan. Given the potential for increased awareness of reporting standards in recent years, it is essential to investigate whether there has been an improvement in effect size reporting among the papers published in this leading journal. As the second phase, this paper includes a simulation study employing the author's prior research data and the imaginary data to demonstrate how variations in sample size can significantly affect p-value outcomes, further contributing to our understanding of

rigorous statistical practices in this field.

3. Research Areas

1. In the first phase, the study conducts an examination of effect size reporting practices in post-ANOVA multiple comparison analyses published in the JACET Journal over the past five years (2020-2024) to assess how reporting standards might have evolved in the Japanese TESOL context.

2. In the second phase, a simulation study utilizes the author's prior research data and the imaginary data to illustrate the impact of sample size on effect size estimates, emphasizing the importance of rigorous statistical considerations.

4. Study 1

4.1 Sampling

To investigate effect size reporting practices in the field of TESOL, the study involved an examination of 43 articles published in the JACET Journal between 2020 and 2024, specifically examining the effect sizes related to significant differences in post-ANOVA multiple comparisons across these papers. This selection period was chosen to provide a comprehensive view of research output within a notable timeframe, allowing the researchers to assess the trends in effect size reporting, particularly concerning the ANOVA and the multiple comparison statistical method.

4.2 Coding

The coding process was conducted following a systematic approach. The author counted the total number of ANOVA executions along with the corresponding instances of effect size reporting for the multiple comparisons. To enhance reliability, the author and another researcher independently coded the data using the same criteria. After coding, the results were compared, and any discrepancies were discussed and resolved through consensus. Although no formal reliability metric was calculated, the collaborative process helped ensure consistency in coding. The key criteria of the coding are as follows: both paired and independent samples were included in the count; however, when a single article contained multiple ANOVA executions, it was counted as one to avoid duplicative calculations (Wei et al., 2019). Moreover, if multiple ANOVA executions were conducted within an article and at least one multiple comparison analysis was performed, that was likewise counted as one. Instances where degrees of freedom (df) or sample size (n) were missing were excluded from the analysis, as well as cases where discrepancies existed between the reported information in the papers and the actual df and n values. Also, in this study, other variants such as analysis of covariance (ANCOVA) and multivariate analysis of variance (MANOVA) were not covered.

4.3 Data Analysis

After completing the coding process, the author summarized the effect size reporting, presenting frequencies to highlight the overall trends in the journal. This comprehensive data analysis demonstrates the current condition of effect size reporting for post-ANOVA multiple comparison analysis within the TESOL and EFL research in Japan.

5. Study 2

In Study 2, the author conducted a simulation analysis utilizing data from his previous study on vocabulary learning devices conducted in 2019 (Miyazaki, 2019). This prior research had investigated the effectiveness of vocabulary learning devices by measuring participant progress across three language tests, aimed at determining whether significant differences existed based on proficiency levels. To achieve this, a cluster analysis had been employed to categorize participants into three proficiency level groups. However, one group gained 12 participants, which was notably small in sample size and could have potentially skewed the results.

Despite the small sample size in one of the proficiency groups, the post-ANOVA multiple comparisons had revealed significant differences in two pairs of comparisons. This prompted the author to simulate varying sample sizes to examine how they might affect the outcome of significance testing. By increasing the sample size up to 100, the author explored whether the previously observed significant differences would remain robust or change, thereby highlighting the influence of sample size on significance reporting. However, multiple comparison analysis revealed significant differences for the same pairs in a sample of 100 participants. This result led to calculating Cohen's effect sizes, and to perform a post-ANOVA multiple comparison analysis using the same sample sizes (12 and 100 subjects) on another imaginary data set.

The simulation follows a methodology utilized in Mizumoto and Takeuchi's (2011) study, where two data sets are aligned in terms of means and standard deviations. The simulation manipulates sample sizes to investigate the resultant effect sizes and significance levels. By comparing the two data sets, the present study seeks to clarify the critical role sample size plays in determining statistical significance.

6. Results and Discussion

6.1 Results and Discussion in Study 1

This examination analyzes the reporting of effect sizes in multiple comparison analyses conducted in articles published in the JACET Journal from 2020 to 2024. Over the examined period, various authors employed ANOVA methodologies with varying practices regarding effect size reporting, particularly in the context of multiple comparisons.

In the 2020 issue (Volume 64), three articles utilized ANOVA. one employed a 2×2 ANOVA, which did not have the effect size for the primary ANOVA, nor were the multiple comparisons conducted. This omission might indicate that multiple comparisons were either deemed unnecessary due to the study's

design or that the results did not warrant further analysis. Another study similarly conducted a two-way ANOVA but did not report the effect size; instead, showing confidence intervals in the figure. In contrast, the other study executed a one-way ANOVA, with the effect size explicitly reported. Besides, the multiple comparisons utilized Cohen's d as a measure. This choice stemmed from conducting t-tests.

The 2021 issue (Volume 65) featured one study that properly employed a one-way ANOVA with effect size reported. Notably, this study also provided confidence intervals for multiple comparisons, enhancing the clarity and reliability of the statistical conclusions drawn. The 2022 issue (Volume 66) saw a year without any ANOVA-based studies.

Moving to the 2023 issue (Volume 67), three articles implemented ANOVA methodologies. Two studies utilized a two-way ANOVA, reported effect sizes, and included confidence intervals post-multiple comparisons. One of the studies continued to build upon previous findings by conducting a further investigation employing a 2×3 ANOVA. Consistent with earlier methodologies, an effect size was reported alongside confidence intervals following multiple comparisons. Reporting confidence intervals enhances the interpretability and reliability of their findings. In contrast, another study reported effect sizes within the ANOVA analysis only where significant differences existed, omitting them for non-significant outcomes and failing to report effect sizes or confidential intervals regarding multiple comparisons.

Lastly, the 2024 issue (Volume 68) featured two studies employing ANOVA. One study employed a 2×3 ANOVA. In the following multiple comparisons, effect size was not reported. Another study, which utilized a two-way ANOVA, also reported effect sizes but did not conduct any multiple comparisons. Instead, the study employed Wilcoxon's signed rank tests in the place of multiple comparisons.

Overall, while there is a gradual trend toward reporting practices concerning effect sizes in ANOVA in the articles published in the JACET Journal, awareness of practices in effect size reporting for post-ANOVA multiple comparison remains an area for further discussion about applicability of findings derived from ANOVA methodologies in the field of TESOL and EFL.

Table 1.
Summary of the Examination of Effect Size Reporting in ANOVA and Multiple Comparison

Year	Number of ANOVA	Effect Size for ANOVA	Multiple Comparison	Effect Size for Multiple Comparison	Confidence Intervals for Multiple Comparison	Others
2020	3	2	2	1	1	0
2021	1	1	1	0	1	0
2022	0	0	0	0	0	0
2023	3	3	3	0	1	0
2024	2	2	1	0	0	1

Note: ANCOVA and MANOVA were not counted in the data.

6.2 Results and Discussion in Study 2

The analysis of the simulation study utilizing prior research data obtained results regarding the

influence of sample size on the interpretation of statistical significance. In a prior study conducted in 2019, the author employed cluster analysis to measure the English proficiency progress of participants, categorizing them into three distinct groups. Notably, one of these groups was composed of only 12 participants. Despite the small sample size, significant differences emerged in two pairs during the multiple comparison analyses (Table 2). As this analysis was a repeated-measured format, immediate effects were recorded following the usage of specific vocabulary devices in both the first and second testing sessions; however, the results of the delayed post-test indicated no significant differences, suggesting a short effect rather than a long-lasting impact on the participants' proficiency.

Table 2.
Multiple Comparisons of the Three Test Results for the Group

Groups	Pair	r	Nominal level	t	p	sig.
(Group)	2 – 1	3	.02	4.02	.00	s.
	2 – 3	2	.03	0.17	.87	n.s.
	3 – 1	2	.03	3.85	.00	s.

Note: MSe=12.99, df=22, *p < .05

To further investigate the robustness of these findings, the author expanded the original data set to include a total of 100 participants, while maintaining the mean scores and standard deviations consistent with the initial analysis. Subsequent post-ANOVA multiple comparison analyses obtained results mirroring those of the original data set. For further information, Cohen's effect sizes were computed, revealing a medium effect between pairs one and two ($d = .46$), as well as pairs one and three ($d = .45$), while pairs two and three exhibited smaller effect sizes ($d = .02$). This suggested that there was a moderate effect attributable to device usage between the first and second assessments based on Cohen's guidelines (1988).

In a subsequent phase of analysis, the author created imaginary data with the same sample size and parameters as the previous data sets, matching each piece of data to the first decimal place using the Statistical Package for Social Science (SPSS 26.0) (Figure 1).

Figure 1. Descriptive Statistics in SPSS for the Imaginary Data Sets

Descriptive Statistics				Descriptive Statistics			
	Mean	Std. Deviation	N		Mean	Std. Deviation	N
VAR00001	48.33	21.25	12	VAR00001	48.33	21.27	100
VAR00002	54.17	19.29	12	VAR00002	54.17	19.29	100
VAR00003	52.08	21.05	12	VAR00003	52.08	21.07	100

Utilizing the same approach for post-ANOVA multiple comparison analyses, the author found significant differences in the newly constructed data set. The ANOVA results indicated significant

differences for both data sets: ($F(2, 22) = 13.28, p < .05, \eta^2 = .55$) for the 12-member group; ($F(2, 198) = 89.93, p < .05, \eta^2 = .48$) for the 100-member group, leading to Bonferroni adjustments for multiple comparisons. Contrastingly, the pair two and three for the 12-member group showed no significance (Figure 2), while the 100-member group revealed differences (Figure 3).

Figure 2. Multiple Comparison Analysis in SPSS for Three Tests in the 12-member Group

Pairwise Comparisons						
Measure: MEASURE_1						
(I) factor1	(J) factor1	Mean Difference	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
		(I-J)			Lower Bound	Upper Bound
1	2	-5.833*	1.486	.007	-10.025	-1.641
	3	-3.750*	1.088	.016	-6.818	-.682
2	1	5.833*	1.486	.007	1.641	10.025
	3	2.083	.743	.052	-.013	4.179
3	1	3.750*	1.088	.016	.682	6.818
	2	-2.083	.743	.052	-4.179	.013

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Figure 3. Multiple Comparison Analysis in SPSS for Three Tests in the 100-member Group

Pairwise Comparisons						
Measure: MEASURE_1						
(I) factor1	(J) factor1	Mean Difference	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
		(I-J)			Lower Bound	Upper Bound
1	2	-5.830*	.559	.000	-7.192	-4.468
	3	-3.750*	.356	.000	-4.577	-2.843
2	1	5.830*	.559	.000	4.468	7.192
	3	2.086*	.376	.000	1.204	3.036
3	1	3.750*	.356	.000	2.843	4.577
	2	-2.086*	.376	.000	-3.036	-1.204

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Across both data sets, effect sizes remained stable: $d = .29$ for pair one and two, $d = .10$ for pair two and three, and $d = .18$ for pair one and three. Judging from this effect size in Cohen's guidelines (1988), it could be fair to think that none of the three tests showed much effect. However, judging only the p -values, some pairs were judged to have an effect. The size of the confidence interval is also useful, but the effect sizes also provide a detailed indication of the meaning of the significant level.

7. Conclusion and Future Implications

From the examination of JACET Journal, as noted by Plonsky and Oswald (2014), there has been a trend towards increased reporting of effect sizes within ANOVA in Japanese TESOL and EFL journals. The simulation study clearly demonstrates the crucial relationship between sample size, significance, and effect sizes, even in the context of multiple comparisons. The findings emphasize the necessity for researchers to report effect sizes alongside p-values to fully convey the practical implications of their findings.

The lack of reporting could be seen as redundant, as argued by Kido and Ikeda (2019), but it is essential to recognize that, as evidenced by the simulations conducted in this study, the magnitude of effects related to multiple comparisons can significantly influence research outcomes. Therefore, it is probable for researchers to explore ways to report effect sizes, or their implications related to multiple comparisons more consistently.

Looking ahead, a pressing challenge for researchers lies in the question of whether confidence intervals can effectively serve as a substitute for effect size reporting in the context of multiple comparison analyses. Although several studies have begun to explore the display of confidence intervals, it remains unclear whether this practice can adequately compensate for the absence of direct effect size calculations. Furthermore, the issue of calculating effect sizes through t-tests—especially when significant differences emerge across two or more pairs in multiple comparisons—raises concerns about associating these calculations with Type I error rates.

Moreover, given the limitations of SPSS in detecting effect sizes related to multiple comparisons, researchers should consider the potential benefits of utilizing alternative statistical software, such as R or jamovi⁴. These tools may not only facilitate more sophisticated analyses but also enhance the robustness of findings in educational research (Ponce-Renova, 2022). By adopting these advancements, researchers can contribute to a more comprehensive understanding of effect size reporting practices in the field.

References

- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
- Hirano, A., & Yokokawa, H. (2024). Effects of experience-based learning on Japanese EFL learners' relative clause processing: Evidence from syntactic priming by self-paced reading procedure.

⁴ *jamovi* is a statistical software, which provides a complete suite of analyses for the social sciences; t-tests, ANOVAs, correlation and regression, non-parametric tests, contingency tables, reliability and factor analysis.

JACET Journal, 68, 9-26.

- Ito, A. (2023). Interpretation of VP ellipsis sentence by Japanese learners of English as a Foreign Language. *JACET Journal*, 67, 127-145.
- Kido, K. & Ikeda, M. (2022). Treatment of effect size when conducting significance tests in the field of educational technology. *Japan Journal of Educational Technology*, 46(3), 597-587.
- Konishi, M. (2021). The gap between enjoyment as EFL learners and anxiety as pre-service teachers with eTandem online international video chat exchanges. *JACET Journal*, 65, 89-106.
- Konishi, M. (2023). Adaptation to a new communication mode for an online international exchange through video chatting. *JACET Journal*, 67, 33-50.
- Kovalyova, A. (2023). Usability of AI-powered chatbot applications for EFL conversation practice. *JACET Journal*, 67, 51-68.
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. Routledge.
- Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research*, 20, 741-768.
- Mikami, Y. (2024). Roles of text relationships and a highlighting task on cross-textual connections across multiple texts. *JACET Journal*, 68, 53-69.
- Miyasako, N. (2020). Does CBI of English language teacher education in cooperative learning after university students' perceptions of cooperative learning and motivation relevant to English language teacher education? *JACET Journal*, 64, 95-115.
- Miyazaki, K. (2019). The effect of online vocabulary learning on passive and active vocabulary use at a range of proficiency levels. *Journal of Pan-Pacific Association of Applied Linguistics*, 23 (2), 85-108.
- Mizumoto, A., & Takeuchi, O. (水本篤・竹内理) (2011) . 「効果量と検定力分析入門—統計的検定を正しく使うために—」 『より良い外国語教育のための方法—外国語教育メディア学会 (LET) 関西支部メソドロジー研究部会 2010 年 度報告論集—』 48-74.
- Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, 35, 655-687.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 Research. *Language Learning*, 64, 878-912.
- Ponce-Renova, H. F. (2022). Comparing effect sizes and their confidence intervals: A primer on equivalence testing in educational research. *Journal of New Approaches in Educational Research*, 11(2), 209-225.
- Ueno, M., & Yamane, S. (2020). Effects of L1-pivotal approach on L2 pronunciation instruction: Teaching English /r/ and /l/ to Japanese EFL Students. *JACET Journal*, 64, 171-186.
- Ushiro, Y., Hosoda, M., Kamimura, K., Ogiso, T., & Sasaki, Y. (2020). EFL readers' processing difficulty with understanding different situational dimensions of narrative texts: An eye-tracking study. *JACET Journal*, 64, 187-203.
- Wei, R., Hu, Y., & Xiong, J. (2019). Effect size reporting practices in applied linguistics research: A study of one major journal. *Sage Open*, 9(2), 1-11.