Spring 3-25-2022

A Reflection of Real Time Educational, Within-Subjects Data in Diverse Classrooms: One-Way ANOVA as a Solution to Data with Nonignorable Missingness and Skew

Amilynn Campa
acampa2@mail.stmarytx.edu

A Reflection of Real Time Educational, Within-Subjects Data in Diverse Classrooms:

One-Way ANOVA as a Solution to Data with Nonignorable Missingness and Skew

by

Amilynn N. Campa

HONORS THESIS

Presented In Partial Fulfillment of the Requirements for
Graduation from the Honors Program of
St. Mary's University
San Antonio, Texas

Approved by:

Rick Sperling Digitally signed by Rick Sperling
Date: 2022.03.04 09:20:59 -06'00'

Dr. Rick Sperling
Professor of Psychology

Camille A Langston Digitally signed by Camille A Langston
Date: 2022.03.25 12:53:17 -05'00'

Dr. Camille Langston
Director, Honors Program

November 13, 2020

Abstract

Schools across the United States and throughout the world administer tests to students to evaluate their academic performance. In many instances, however, especially in classrooms with higher populations of racial/ethnic minorities and low SES students, there are often missing scores, attributed to higher rates of absenteeism among these demographics (Callahan, 2019; Friedman-Krauss & Raver, 2015; Evans, 2004). When evaluating within-subjects, longitudinal data, many will utilize a pre/post significance test such as a paired-samples *t*-test or a Repeated-Measures ANOVA, however, due to the assumptions of these tests, missing data has posed a problem and requires data manipulation tactics that may distort data representation. Unfortunately, data misrepresentation may disproportionately affect students of the described demographic. Prior studies have explored the possibility of utilizing independent-samples *t*-tests and One-Way ANOVAs as a method of significance testing and have found that in analyzing data containing missingness, these tests yield less-biased results with higher amounts of statistical power. The current paper continues on this path, exploring the extent to which One-Way ANOVAs exhibits results with higher statistical power as it relates to mean difference values in skewed data containing missingness, as compared to Repeated-Measures ANOVAs. Although only simulated data was used, it was found that One-Way ANOVAs outperformed Repeated-Measures ANOVAs and as evidenced by results with lower rates of type I and type II error.

Keywords: statistical power, missingness, skew

A Reflection of Real Time Educational, Within-Subjects Data in Diverse Classrooms:

One-Way ANOVA as a Solution to Data with Nonignorable Missingness and Skew

Literature Review

Lack of Quality Education, Absenteeism, and Other Adversities

Students of lower socio-economic status (SES) often attend lower quality schools with less qualified teachers and fewer resources (Friedman-Krauss & Raver, 2015; Evans, 2004). Additionally, schools that are considered to be “high poverty” institutions tend to have higher populations of racial minorities in their student body, as well as less funding, less well-qualified teachers, and poorer quality education, leading to a lack in academic success (Friedman-Krauss & Raver, 2015; Evans, 2004; Peske & Haycock, 2006; Orfield & Lee, 2005). Along with lower quality schools with fewer resources, frequent mobility and stress in low-income households (due to issues such as violence, familial disruption and separation: all occurring at higher rates in this demographic), are found to cause poor concentration in the school environment (Friedman-Krauss & Raver, 2015; Evans, 2004). Students experiencing poverty not only suffer in their schooling, but they also suffer as a result of the “multidimensional” poverty they experience. These dimensions include low household income, limited education, no health insurance, low income areas of living, and unemployment. Additionally, these components of multidimensional poverty are found most prevalently amongst racial minorities (Reeves et al, 2020).

Students of racial/ethnic minorities and low-SES experience higher rates of school mobility and absenteeism as compared to other students in the nation (Friedman-Krauss & Raver, 2015; Evans, 2004). In fact, research shows that black, Hispanic, and low-income students tend to have higher mobility rates than White and Asian students (Welsh, 2016). Several national datasets show that “children who live at or below the poverty line in America change residences more than twice as often than children who do not live in poverty" (Evans, 2004). For example, primarily representing low-income, racial/ethnic-minority students, the Chicago Longitudinal Study indicated that (excluding promotional transitions) 73% of children moved at least one time between kindergarten and seventh grade, and 21% moved at least three times (Friedman-Krauss et al, 2015). Studies have confirmed that the repercussions of this frequent mobility have a more substantial effect on low-SES and racial minority students as well (Friedman-Krauss & Raver, 2015; Evans, 2004). Additionally, studies show that there is a significant and strong correlation between schools' demographics, achievement levels and mobility. Racial, economic, and achievement segregation is a current and concerning problem in public education, as each of these components are interrelated (Welsh, 2016).

Researchers examining this phenomenon through a 5-year, longitudinal study of racial and SES minority students from preschool to 4th grade suggest that school mobility and academic success are inversely related, due to reasons such as cognitive dysregulation, often resulting from the stress of transitioning schools. This cognitive dysregulation is characterized by lack of attention and poor working memory skills, both of which are imperative for academic success. Specifically, it was found that the frequency of school mobility was a key indicator of math achievement in fourth grade, contributing to the finding that students of racial-minority and low SES tend to perform poorly in their academics when compared to their peers (Friedman-Krauss & Raver, 2015).

Testing and the Shortcomings of Current Data Analysis Methods

In many schools across the United States, tests to evaluate student performance in academics are administered throughout the year to examine the effectiveness of various teaching methods. These scores can give teachers and administrators an estimate of student success as it relates to particular teaching interventions (Callahan, 2019). In within-subject designs such as this, results are commonly modeled and tested for significance in paired-samples *t*-tests and repeated measures ANOVAs (RMAs) (Callahan, 2019), as these allow for score comparison between groups, prior to and after a manipulation. Because the scores are collected from the same subject at different time points, a fraction of the total variation can be explained by the fact that results are coming from the same subject. Statistical power in both RMAs and paired-samples *t*-tests, therefore, is partially derived from the correlation between a subject both prior to and after a manipulation (Callahan, 2020). One-Way ANOVAs (OWAs) and independent-samples *t*-tests are also utilized for significance testing. However, rather than scores being grouped as in a pre-test, post-test scenario (two or more scores for the same subject), OWAs and independent-samples *t*-tests compare separate subject groups and evaluate data as if the scores were collected from different subjects, taken at only one point in time.

Because of the high rates of absenteeism in populations of low-income, racial/ethnic minority students, due to school mobility or any other reason, there is a higher risk of their scores not being collected when an exam is administered. Since scores are evaluated for each subject at multiple time points in RMAs and paired-samples *t*-tests, missingness presents a problem in that at one time point or another, there is no score to draw a relationship between the pre and post-test. In data analysis, this invokes the need to “fill the gaps” that these missing scores present. Compensating for missing data usually involves listwise deletion or imputation methods which can pose problems, such as an increase in type I error rate, if the missingness falls within a certain criterion. In OWAs, however, there is little to no issue with missingness as the scores are not paired but rather treated as individual values from different subjects. Hence, there is no need for any forms of case substitution or deletion.

As mentioned previously, teachers will commonly utilize listwise deletion (LD) and imputation methods, involving either the deletion of scores or the imputation of values (such as the mean value) in place of the missing scores in paired-samples t-tests or RMAs (Callahan, 2019). While these methods may appear to solve the issue presented by missing data, it has been shown that on the contrary, they increase type I error rate and decrease statistical power due to the high likelihood that the missing scores that are being deleted or replaced by estimates are mostly comprised of low-SES and racial minority students, who as examined are at an academic disadvantage and perform at lower levels as compared to their non-minority and higher-SES peers (Friedman-Krauss & Raver, 2015). As reported in Callahan's study (2019), the only case in which listwise deletion would not affect statistical power would be if both the sample was of a sufficient size and if the missingness was completely at random (MCAR for short). In order for missingness in the data to be MCAR, the missingness must be truly random, unrelated to participants or any other variable in the experiment, known or unknown. This implies there must be no relation to any variables such as race or socio-economic status, whatsoever (Callahan, 2019). However, when considering the high probability that many of the missing scores happen to belong to this particular demographic, due to reasons such as school mobility among others, it can be inferred that the missingness is not random at all, it is non-random missingness (MNAR), also referred to as nonignorable missingness (Callahan, 2019; Yang, 2015). This missingness can usually be found in the tail ends of distributions, representing the lower-performing areas of the student sample. Especially in schools where there are larger proportions of children of low SES. the risk of data-analysis conclusions being overall nonrepresentative of the whole population is apparent, as the missingness in these cases correlates with variables like race and socioeconomic status (Callahan, 2020).

Application of Preceding Studies

In a Monte Carlo simulation study, distributions containing missingness were tested for significance under RMA modeling (requiring use of methods such as LD and imputation) as well as OWA modeling (not requiring manipulations such as imputation or LD). It was found that in all conditions of varying correlation, mean difference, and sample size, OWA produced results of lower type I error rate and higher statistical power than the RMA models with LD (Aquino Aguilar, 2017). However, this examination was limited in that it only accounted for scenarios of random missingness (MAR), missingness that is accounted for in the researcher's analysis as it is attributed to observed data rather than unknown factors. The study neglected the testing of conditions with MNAR, the type of missingness more commonly found in distributions from described classrooms (Callahan, 2019).

Following this discrepancy, researchers through methods of a Monte Carlo simulation study set to examine the extent to which OWA produced results with higher statistical power as compared to RMAs (using LD and imputation techniques) in distributions containing simulated MNAR in the tail end of the distribution instead of MAR, more accurately reflecting the type of missing data found in the distributions from real-time, educational data (Callahan, 2019). In this study, varying conditions of missingness and skew were again utilized to examine how each manipulation affected or did not affect values of statistical power and type I error in both OWAs and RMAs. It was found that OWAs outperformed RMAs in all conditions except in instances of low missingness and high levels of correlation. However, these manipulations were applied to distributions lacking the levels of skew that are found in classroom test-score data (Callahan, 2019). As proposed by the study, further research involving "real data", or at least simulated data that more accurately reflects what is collected in classrooms with students of low-SES and racial minority needs to be done to provide evidence of the superiority of OWAs in statistical analysis of distributions containing MNAR (Callahan, 2019).

The Presence of Skew and its Importance

Positive skew, a common component of distributions that are produced in school settings as a result of test score data, provides an impediment to significance testing (Callahan, 2020). In studies prior to this one, the extent to which OWAs and independent sample t-tests had greater statistical power and lower type I error rates were only examined in normal distributions (Callahan, 2020). “Grading on the curve” is a common practice used to adjust scores that are skewed to fit a normal distribution, as it is assumed that test scores should fit a normal distribution: few students performing exceptionally well and few performing terribly, with most students falling between the extremities. This method is inadequate for accounting for skew because “it leads to artificial grade fluctuation, biases results, and can make grades depend on chance rather than student ability or preparedness” (Kulick et al, 2008; Callahan, 2020). The presence of skew and non-normality in distributions presents the risk of yielding biased parameter estimates and results, due to a violation of the F-test assumptions. These assumptions include that the outcome variable is normally and independently distributed with equal variances among groups. Through a Monte Carlo simulation study, researchers explored the robustness of the F-test, or One-way Analysis of Variance, in the context of non-normality, a violation of these assumptions, and type I error. Although previous research as described in the literature review has investigated the robustness of the F-test and concluded that:

> The F-test was valid provided that the deviation from normality was not extreme and the number of degrees of freedom apportioned to the residual variation was not too small (Blanca et al, 2017, p.554)

Researchers in this study aimed to analyze its validity according to Bradley's criterion, stating that:

> a statistical test is considered robust if the empirical type I error rate is between .025 and .075 for a nominal alpha level of .05 (p.552).

The study was conducted with the following manipulations: equal and unequal group sample sizes; group sample size and total sample size; coefficient of sample size variation; shape of the distribution and equal or unequal shapes of the group distributions; and pairing of group size with the degree of contamination in the distribution. It was found that in 100% of cases, regardless of the degree of deviation from a normal distribution, sample size, balanced or unbalanced cells, equal or unequal distribution in the groups, and in degrees of skewness and kurtosis ranging from -1 to 1, the F-test was robust according to Bradley's criterion (Blanca et al, 2017). By the evidence provided in this study and others, we expect that OWAs remains robust with skewed data containing MNAR, producing lower rates of type I error, as well as higher statistical power, especially as compared to RMAs, in the analysis of distributions with non-normality.

Independent vs. Paired-Samples *t*-tests Leading to the Current Study

In an attempt to reflect the reality that test score data collected in classrooms is typically skewed, researchers (Callahan, 2020) utilized a Monte Carlo simulation study to determine the source of low statistical power produced in paired-samples *t*-tests analyzing distributions with missingness (requiring LD or imputation) and compare results to those of independent-samples *t*-tests (not requiring LD or imputation). Specifically, independent *t*-test modeling was applied to skewed distributions with missingness to determine if the model presented a lesser likelihood of incorrectly failing to reject the null hypothesis, and therefore, exhibiting greater statistical power, as compared to a paired-samples *t*-test. Using 72 sets of 1000 samples with varying correlations, skews, and mean differences with nonrandom missingness, distributions were tested for significance in both *t*-tests. For most of the simulations, independent samples *t*-tests produced results with greater statistical power, with the greatest limitation being that they also produced a slightly inflated type I error rate. It was found that power was positively correlated with mean difference and skew (Callahan, 2020).

In conjunction with the study by Callahan (2020), the current study aims to determine the extent to which OWAs have greater statistical power in distributions containing varying conditions of MNAR and skew as a function of mean difference. It is predicted that as mean difference increases, OWAs will produce results with significantly greater levels of statistical power as compared to RMAs using LD. Within conditions of a mean difference of 0.0, type I error rates produced by both the OWA and RMA with LD will also be compared and assessed as a function of skew and MNAR levels.

Method

Sample

To simulate student scores in a pretest, posttest, and follow-up test in schools with higher levels of poverty, a Monte Carlo simulation study using the R statistical software program produced distributions with varying levels of missingness (0, 0.2, 0.3, 0.4), skew (0.0, 0.2, 0.4, 0.6), and mean difference (0, 0.25, 0.50, 0.75, 1). Each of these manipulations represents a component of real-time educational data; for example, skew is typically present in testing data, due to scores not falling into an even distribution. Sample size was set to 30, representing a classroom size ($n=30$), and correlation was set at 0.2, as there is usually a correlation present. In conjunction with the independent *t*-test study, (Callahan, 2020), nonignorable missingness

(MNAR) was introduced to each by creating a biserial correlation between missingness and the value of scores on the pretest, such that the lower a student scored on the pretest, the more likely it was that the student would be absent for the posttest and follow-up test.

Procedure

OWA. The OWAs were performed by classifying the pretest, posttest, and follow-up tests as separate groups rather than longitudinal data; essentially ignoring missing values and assuming the sample sizes in each group were different.

RMA. Due to the need for equal sample sizes across groups, a missing score in the RMA analysis required either the deletion or insertion of values in place of the missing scores across a longitudinal study.

In my analysis of the produced data, type I error rates and type II error rates of the Repeated-Measures ANOVA (requiring listwise deletion) were compared to One-Way ANOVA (not requiring listwise deletion) across degrees of nonrandom missingness, mean difference, and varying levels of skew. To evaluate resulting statistical power as a function of these modifications, tables displaying number of null hypothesis rejects out of 1000 times were produced. Specifically, the R statistical software program produced 12 tables displaying the number of times the null hypothesis was rejected (*y*) in relation to 5 different mean difference values (*x*) including 0.0, 0.25, 0.50, 0.75, and 1.0, as well as 4 degrees of skew (0, 0.5, 1, 1.5) and 3 varying levels of missingness (0.2, 0.3, and 0.4) for both OWAs and RMAs.

Values produced in relation to a mean difference of zero were used to calculated type I error for all degrees of skew and missingness, as a mean difference of 0.0 would indicate that there was no significant relationship at the population level. Therefore, if the null hypothesis was rejected at any point, it would be considered a type I error. Type II error was calculated for mean differences 0.25 through 1.0 in described increments. Since each of these mean differences indicated a significant relationship between related groups, all failures to reject the null hypothesis were considered a type II error and were determined by subtracting the produced value on the table (displaying number of rejects) from 1,000: the number of times the test was run for each condition.

Results

Type I Error

As shown in Figure 1, in all tested conditions of skew and MNAR, data with a mean difference of 0.0 had lower rates of type I error when tested with the OWA as compared to the RMA with LD. Results were most profound in conditions of low skew (0.5) and high missingness (0.4) and were less profound in conditions of low skew (0.5) and missingness (0.2), as well as in conditions of both high skew (1.5) and high missingness (0.4).

Type II Error and Statistical Power

Type II error was calculated for mean differences 0.25 through 1.0, however, it should be noted that in conditions with mean difference values of 0.75 and 1.0, the null hypothesis was correctly rejected every time (1000/1000 times), and hence, produced no type II errors.

Type II error rates in conditions of a 0.25 mean difference are displayed in Table 2. As shown by the winner frequency of 0.75, the RMA generally outperformed the OWA. There were exceptions of the OWA outperforming the RMA with LD in cases skew of 0.0 and higher levels of missingness (ranging from 0.3 to 0.4), as well as in conditions with higher levels of both skew (1.5) and missingness (0.4).

Values produced with from 0.50 mean difference conditions, as shown in Table 2, indicate that the OWA instead outperformed the RMA with LD. There were exceptions including two ties in conditions of skew at 0.0 and MNAR ranging from 0.3 to 0.4: where both the RMA with LD and OWA produced equivalent type II error rates.

Discussion

As schools continue to administer examinations to determine student performance, missingness will always pose a risk that may affect the data analysis outcomes. Since this risk is more highly associated with schools that have higher populations of students of racial/ethnic minorities and lower SES, there could be further implications through inaccurate data analysis that lead to detrimental effects, such as lack of intervention, for these disadvantaged populations.

Within conditions of a 0.0 mean difference, the OWA as a method of analyzing skewed data with MNAR has shown to produce results with significantly less risk of type I error. The most profound effect was found in tests of low skew and high missingness. This finding points to an advantage of utilizing the OWA rather than the RMA with LD in terms of avoiding the detection of a significant relationship when there is not one present.

In testing data with a mean difference of 0.25, the RMA with LD generally outperformed the OWA with a winning frequency of approximately 66.7%. Notably, however, in conditions of high skew and MNAR, the OWA was superior in producing lower type II error rates. Contrary to these results, in conditions of 0.50 mean difference, the OWA outperformed the RMA with LD with a winning frequency of 75%. Only in the case of high skew and medium levels of MNAR did the RMA with LD outperform the OWA by 0.1%. The two tests tied in type II error rates in conditions with skews of zero and medium to high levels of MNAR. As mean difference value increased, the OWA overall outperformed the RMA as a method of analysis in terms of producing findings with lower rates type II error, and hence, greater levels of statistical power. In the 0.25 mean difference test, the power derived by the OWA is not as profound as in the 0.50

mean difference category. It is predicted that this finding was related to the small effect size between groups in the 0.25 mean difference category.

By these results, it can be concluded that in longitudinal experiments containing data with skew and MNAR, there is a particular mean difference in which a OWA as a method of analysis provides more statistically powerful results than a RMA with LD. Results indicated that this value may be between 0.25 and 0.50, and the particular mean difference in which the OWA begins to outperform the RMA with LD should be further examined. Because this study was conducted using simulated data, it is unknown if experimental results could potentially be significant if provided real-time, educational data. Furthermore, the current study only tested manipulations of skew and missingness, whereas other data components such as correlation and sample size were not; provided the conduction of these tests, a more thoughtful conclusion can be made regarding the superiority of the OWA as a method of evaluating skewed data with MNAR.

If methods of utilizing OWAs for testing skewed longitudinal data with MNAR and higher mean differences for significance are implemented at a larger scale, teachers can more accurately evaluate the effectiveness of their teaching interventions. This endeavor may involve the development of a “package”, or function, in statistical software programs like R to help teachers and other educational professionals utilize the correct significance testing method depending on their collected data. Increased accessibility and understanding of these important statistical functions can improve the overall usage and outcome of these tests. The OWA has the potential to more accurately reflect the adversities experienced by those in racial/ethnic minority and low SES communities. This in turn, will hopefully improve the construction of teaching interventions, school programs, and help lead to productive policy changes in the school setting.

References

Aquino Aguilar, G. A. (2017). The relative advantage of one-way ANOVA to repeated measures ANOVA with high student turnover. McNair Scholars Research Journal, 10(1)

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option?. Psicothema, 29(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Callahan, S., Castro, M., Walls, M., (2020, April). Closing the evaluation gap: One-way ANOVA as a solution to non-random missingness in diverse classrooms. Poster presented at the annual Teachers College Winter Roundtable Teaching to Transform: Mentoring and Supervision in Psychology and Education, New York, NY.

Evans, G. W. (2004). The environment of childhood poverty. American Psychologist, 59(2), 77–92. doi:10.1037/0003066x.59.2.77f1065011d342%40sessionmgr103&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#AN=EJ913412&db=eric

Friedman-Krauss, A. H., & Raver, C. C. (2015). Does school mobility place elementary school children at risk for lower math achievement? The mediating role of cognitive dysregulation. Developmental Psychology, 51(12), 1725-1739. doi:10.1037/a0039795

Kulick, G., & Wright, R. (2008). The impact of grading on the curve: A simulation analysis. International Journal for the Scholarship of Teaching and Learning, 2(2). doi:10.20429/ijsotl.2008.020205

Orfield, G., & Lee, C. (2005). Why segregation matters: Poverty and educational inequality. The Civil Rights Project. Retrieved from https://www.civilrightsproject.ucla.edu/research/k-

12-education/integration-and-diversity/why-segregation-matters-poverty-and-educational-inequality/orfield-why-segregation-matters-2005.pdf

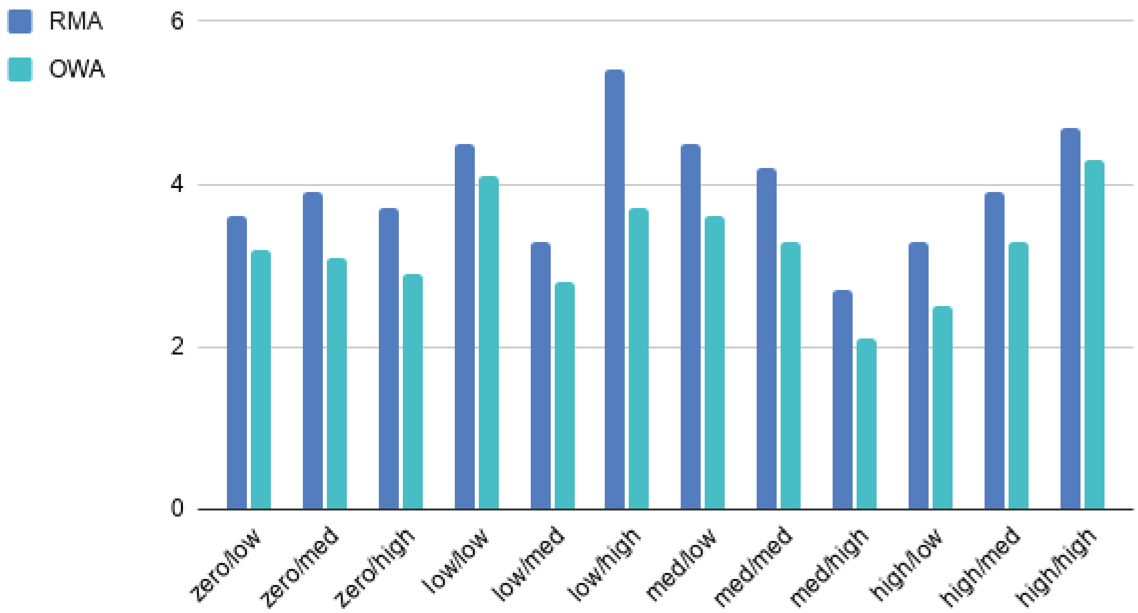
Peske, H. G., & Haycock, K. (2006). Teaching inequality: How poor and minority students are shortchanged on teacher quality. Washington , DC: The Education Trust. Retrieved from https://files.eric.ed.gov/fulltext/ED494820.pdf

Reeves, R. V., Kneebone, E., & Rodriguez, E. (2016). Five Evils: Multidimensional Poverty and Race in America. Retrieved from Brookings Institution website: https://www.brookings.edu/wp-content/uploads/2016/06/ReevesKneeboneRodrigue_MultidimensionalPoverty_FullPaper.pdf

Welsh, R.O. (2016). Student Mobility, Segregation, and Achievement Gaps: Evidence From Clark County, Nevada. *Urban Education, 53*(1), 55-85. doi:10.1177/0042085916660349

Figure 1

Histogram depicting type I error rates of Repeated-Measures ANOVA and One-Way ANOVA



Note. The graph depicts type I error rates in conditions with a mean difference of 0.0, n=30, correlation set at 0.1, and varying levels of skew and MNAR. Each category along the x-axis represents the level of skew and MNAR for that particular test and are positioned as follows: "skew/MNAR". For the first position of skew, "zero" represents a skew of 0, while "low" is equivalent to a skew of 0.5. "Med" and "high" represent the two greater levels of tested skew: 1.0 and 1.5, respectively. The second position indicates level of MNAR, where "low" equals 0.2, "med" equals 0.3, and "high" represents a missingness of 0.4.

Table 1

Type II Error Rates of 0.25 Mean Difference (%), n=30, correlation=0.1

Skew	Missingness	RMA	OWA
0.0	0.2	**50.0***	51.5
0.0	0.3	48.3	**48.2***
0.0	0.4	47.8	**47.3***
0.5	0.2	**48.7***	49.7
0.5	0.3	**48.5***	50.2
0.5	0.4	**48.2***	49.9
1.0	0.2	**48.3***	49.7
1.0	0.3	50.2	50.2
1.0	0.4	**50.9***	51.9
1.5	0.2	**50.2***	50.6
1.5	0.3	**44.9***	51.9
1.5	0.4	51.6	**51.1***

Note. * indicates preferred result

Table 2

Type II Error Rates of 0.50 Mean Difference (%), n=30, correlation=0.1

Skew	Missingness	RMA	OWA
0.0	0.2	1.1	**0.9***
0.0	0.3	0.6	0.6
0.0	0.4	2.0	2.0
0.5	0.2	0.6	**0.5***
0.5	0.3	0.5	**0.4***
0.5	0.4	0.7	**0.4***
1.0	0.2	0.5	**0.2***
1.0	0.3	0.4	**0.1***
1.0	0.4	0.7	**0.1***
1.5	0.2	0.3	**0.0***
1.5	0.3	**0.0***	0.1
1.5	0.4	0.3	**0.1***

Note. * indicates preferred result