THE IMPACT OF NESTED TESTING ON EXPERIMENT-WISE TYPE I ERROR RATE

by

JACK SAWILOWSKY

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: EDUCATION EVALUATION AND RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

_____

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ACKNOWLEDGEMENTS

I would like to acknowledge the late Dr. Gail Fahoome, my initial doctoral advisor, for her guidance in helping me to select this dissertation topic. I am grateful to Professor Barry Markman for his support during the transition and in bringing the dissertation to fruition. I am also thankful for the assistance of the members of my doctoral committee: Associate Dean Patrick Bridge, Assistant Dean William Hill, Dr. Julie Smith, and my cognate advisor Dr. Monte Piliawsky. I benefited from conversations with Professor Shlomo Sawilowsky on matters relating to Monte Carlo simulations, and from Professor R. Clifford Blair for correspondence relating to Fortran coding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

When conducting a statistical test the initial risk that must be considered is a Type I error, also known as a false positive. It occurs when “rejecting a null hypothesis when it is true” (Hinkle, Wiersma, & Jurs, 2003, p. 178). The Type I error rate is set by nominal alpha, assuming all underlying conditions of the statistic are met. For example, if nominal $\alpha = 0.05$, then this indicates that the threshold for what constitutes a rare event is set to the odds of less than or equal to 1 in 20, and the probability of a false positive is equal to 5%.

For example, classical parametric statistics – such as the Z, t and F – are based on an underlying theory of probability that equates nominal alpha with the Type I error rate. Therefore, under the truth of the null hypothesis (i.e., no treatment effect), and when the underlying assumptions (i.e., independence, homoscedasticity & normality) are met, a nominal alpha of 0.05 will result in a 5% rejection rate of the null hypothesis.

Consider a selection of random numbers from the Gaussian distribution. Suppose they are subsequently randomly assigned into two groups, and the parametric Student’s t test is conducted on their respective means. Over the course of many replications of this experiment, the long run average rejection rate, even though by definition the random values do not model the presence of a treatment, will be equal to the value set as nominal alpha.

The paradigm of taking the probability of the long run average as the risk for a single hypothesis test is based on the Frequentist approach to statistics. “Neyman throughout his work emphasizes the importance of a probabilistic model of the system

under study and describes frequentist statistics as modelling the phenomenon of the stability of relative frequencies of results of repeated 'trials'" (Mayo & Cox, 2006, p. 79). In contradistinction, "Bayesian statistics is a term applied to the body of inferential techniques that uses Bayes' theorem to combine observation data with personalistic or subjective beliefs" (Press, 2005, p. 1), and therefore, Type I error rates are conceptualized slightly differently. (See Sawilowsky, 2003, for a comparison of both paradigms, along with the Fisherian approach to statistics.)

The Frequentist risk represented by the Type I error only applies if a single statistical test is conducted on the data set. If multiple analyses are conducted the Type I error rate will increase above nominal alpha. This is known as experiment-wise Type I error inflation: the "Experimentwise error rate (α_E) is the probability of making a Type I error rate for the set of all possible comparisons" (Hinkle, Wiersma, & Jurs, 2003, p. 372). Statisticians have considered this problem since the second half of the 20th century and have proposed a variety of solution strategies to handle Type I error inflation, particularly for statistical approaches that invoke multiple procedures.

For example, the question frequently arises as to the reason for caution in conducting both a parametric and a nonparametric test on the same data, and failing to reject the null hypothesis for whichever results are favored. The Type I error inflation makes this approach inappropriate. A solution to this problem is called the maximum test, where critical values are obtained based on conducting both types of tests. See Algina, Blair, and Coombs (1995) and Maggio (2012) for some solutions. This strategy, however, requires special tables of critical values.

According to some viewpoints, there are also statistical layouts that permit a step-down analysis. An example is following a multivariate test (e.g., MANOVA or MANCOVA) with univariate tests. Consider a Hotellings' T^2 which conceptually is an extension of the test of difference in means in the Student's t test to the multivariate case, which is the difference in group centroids. A question that frequently arises following a significant T^2 is if one or the other dependent variable was the greater contributor.

Suppose both a test of reading and mathematics achievement were given following an intervention, and the T^2 test of differences in means between females and males was statistically significant. The step-down univariate test (i.e., Student's t test) on reading by gender, and mathematics by gender, would then be conducted. The statistical literature is not settled on the appropriateness of this approach. The general consensus is if the multivariate test was conducted only to maximize power there is no reason why step-down tests shouldn't be conducted (other than the inflation of Type I errors). However, if the T^2 was conducted because of a multivariate hypothesis with intertwined dependent variables (e.g., self-esteem & self-worth), conducting step-down tests and the concern with experiment-wise Type I error inflation vanishes.

However, there are other layouts that according to all viewpoints require multiple statistical tests. The classical example of this is the one-way analysis of variance. The omnibus F test can be used to determine if there is a difference in means somewhere within the $K \geq 3$ groups. Either a priori or post hoc comparisons must be conducted, however, in order to determine precisely where the difference(s) in means occurred. It is recognized that conducting multiple tests in this application increases the experiment-

wise Type I error rate. In the literature review to follow in Chapter 2, some a priori methods that attempt to prevent this will be discussed. More relevant to the current study are post hoc methods.

Post Hoc Tests: A Resolution to the Inflation Problem

Wilcox (1996) described the most extreme post hoc solution to experiment-wise Type I error inflation:

> The Bonferroni procedure, sometimes called Dunn's Test, provides a simple method of performing two or more tests such that the experimentwise Type I error probability will not exceed α. If you want experimentwise Type I error probability to be at most α, you simply perform paired t-tests, each at the αb = α/C level of significance, where C is the total number of comparisons you plan to perform. (p. 279)

The Bonferroni-Dunn procedure divides alpha by the number of tests to be conducted, to ensure that after all hypothesis tests are computed the total Type I error rate does not exceed nominal α. This method is guaranteed to contain the Type I error rate, but it also guarantees loss of statistical power, because as α decreases, β increases; and as β increases, power decreases (Hinkle, Wiersma, & Jurs, 2003, p. 300). All other multiple comparison procedures are a compromise between the Bonferroni and making no adjustments to control Type I error inflations.

Several procedures have been developed to correct for experiment-wise Type I error rates, especially between the 1940s and 1960s, that are less extreme than the Bonferonni-Dunn approach (e.g., Dunn's, Dunnet's, Fisher's, Scheffé's, Student-Newman-Keuls', & Tukey's tests; see Kirk, 2013, for a comprehensive review). This

topic attracted the attention of researchers worldwide, culminating in the first International Conference on Multiple Comparisons that was held in 1996 in Tel Aviv *(*http://www.mcp-conference.org/1996/*).*

At the conference, Type I error inflations were shown to be pertinent to a variety of research designs and statistical layouts, including the following: interim analysis, sequential analysis, adaptive testing, multivariate contexts, closed stepwise procedures, union-intersection procedures, logically related hypothesis, wavelets, resampling, discrete tests, order statistics, semi-Bayesian methods, Bayesian methods, confidence intervals, inverse problems, simultaneous confidence intervals, global maximizers, multinomial proportions, cross-sectional designs, saturated designs, pre-clinical trials, clinical trials, safety assessment, dose finding, trend tests, multiple endpoint studies, trait loci, transformations, step-up tests, and the Solomon four-group design. The 8th International conference was held in the summer of 2013 (http://www.mcp-conference.org/2013/index.php).

It is clear from the plethora of research designs considered at the International Conference that the inflation of Type I errors should be considered in all research designs, and should never be summarily dismissed. For example, an application of interest not previously considered is the impact of nesting designs on Type I errors.

Nesting

Hierarchical linear modeling (HLM), which is based on testing nested effects, is a popular statistical approach to school-based research. Kreft and De Leeuw (1998) stated “Hierarchical data structures are very common in the social and behavioral sciences… Once you know that hierarchies exist, you see them everywhere” (p. 1).

Kanji (1999) provided a definition of a nested or hierarchical classification as follows:

> In the case of a nested classification, the levels of factor *B* will be said to be nested with the levels of factor *A* if any level of *B* occurs with only a single level of *A*. This means that if *A* has *p* levels, then the *q* levels of *B* will be grouped into *p* mutually exclusive and exhaustive groups, such that the *i*th group of levels of *A* is q_i, i.e. we consider the case where there are $\sum_i q_i$ levels of *B*. (p. 128)

Kanji provided a nested example from education with two factors, which were three teachers each per four schools. The teachers represented level one of the nest, with their school being level two. Two F tests were conducted. One was necessary to determine teacher differences, whereas the other was carried out to determine school differences.

Winer (1971) provided another example of nested factors. Consider a drug trial in which one level of independent variable is the assignment of one of two different drugs. Drug 1 is administered at one group of hospitals, while Drug 2 is administered in another group of hospitals. Hence, Hospitals constitute the second level of the nest. Drug 1 and Drug 2 contribute unique effects within the group of Hospitals. Winer (1971) explained, “Effects which are restricted to a single level of a factor are said to be *nested*

within that factor" (p. 360). The effects of the hospitals were also nested, because they appeared beneath the Drugs factor. Patients at Hospitals 1, 2 and 3 received Drug 1, while patients at Hospitals 4, 5 and 6 received Drug 2. In this case, the classical A×B ANOVA layout is inappropriate, because there is no way to construct the interaction of Hospital by Drug. For that to have occurred, the patients at each hospital needed to have received both drugs. (See Table 1.) Winer (1971) concluded by emphasizing the substantial limitation of nested designs in that they do not permit the testing of an interaction effect.

Table 1

Winer's (1971, p. 359) Hierarchical/Nested Design Example

Drug 1			Drug 2		
Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	Hospital 6
n	*n*	*n*	*n*	*n*	*n*

To ameliorate this limitation, Winer (1971) suggested a two-factor factorial experiment, which is superior to a nested design because the interaction can be measured. This design is illustrated in Table 2. Winer noted this design is not always a viable alternative, due to the requirement of multiple categories within each factor sometimes being studied as relevant factors. As long as this is not the case, however, Winer indicated it is the preferred option when an interaction between factors is of interest and should be tested.

Table 2

Winer's (1971, p. 361) Two-Factor Factorial Experiment Alternative

	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	Hospital 6
Drug 1	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$
Drug 2	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$	$\frac{n}{2}$

Winer (1971) also addressed the possibility of partially-nested designs. The example given was different teaching methods, each taught at different schools (p. 365). Limited to this layout, it is a two-factor factorial experiment. However, if the schools were located in different cities, the layout changes to a partially hierarchical design. The partially-nested design "enables the researcher to eliminate systematic sources of variation associated with differences between cities and differences between schools within cities from the experimental error" (p. 365). There is a cost to pay for invoking the partially-nested layout: "reduced degrees of freedom for experimental error" (p. 365), which reduces the power of the test.

Winer (1971) also expanded "computational procedures for nested factors" (p. 464) for the three factor layout. The same drawback appears with this design in that the nesting sacrifices degrees of freedom. Winer noted with regard to the resulting F ratio, that if the "denominator has relatively few degrees of freedom, the power of the test will be low" (p. 466).

Kreft and De Leeuw (1998) stated that hierarchical modeling tends to address research questions that lack independence and other experimental conditions, which

makes it incompatible with ANCOVA (p. 5). Similarly, Kennedy and Bush (1985) noted "Interaction is not a meaningful consideration when one variable is nested within another" (p. 52). For an interaction effect to be measured, all factors in all levels would need to contain all factors of all other levels. However, nesting is advantageous in order to control for unique effects of a specific level of a nest on another level (e.g., schools on curriculum).

Experiment-wise Type I Error and Nesting

There are also more sophisticated multi-level and longitudinal models based on these basic layouts (Heck, Thomas, & Tabata, 2010). However, there has been little discussion in the literature regarding the impact on the inflation of experiment-wise Type I error rates due to the hierarchical testing of treatment effects. For example, Kanji (1999) did not address the issue of conducting multiple F tests. If each test is set at $\alpha = 0.05$, then in reality there will be an approximate experiment-wise Type I error rate of 0.10. Similarly, Winer's (1971) presentation of the different types of nested designs (2 Factors, Partial, and 3 or more Factors) was not accompanied by a discussion on the experiment-wise Type I error rate.

Marascuillo and Serlin (1988) discussed how the risk of Type I errors are distributed from nested variables.

> In a three-factor design with the inclusion of each of the two-factor interactions and the one three-factor interaction, the total risk of a type I error is
>
> $$\alpha_T \leq \alpha_A + \alpha_B + \alpha_C + \alpha_{AB} + \alpha_{AC} + \alpha_{BC} + \alpha_{ABC}.$$

> If each source of variation is tested at .05, then $\alpha_T \leq 0.35$. If a model is used where some factors are nested in others, then we maintain an overall $\alpha_T \leq 0.35$ by apportioning the risk of a Type I error appropriately. (p. 557)

(Note that α throughout their equation is a misnomer; Marascuillo & Serlin, 1988, were referencing the experiment-wise Type I error rate, which they referred to as the "overall" (p. 557) false positive rate. Hence, they should have used symbols reflecting Type I error instead of alpha.) They noted the summation of alpha's (0.05), when multiplied by the number of effects (7) produces the ceiling of 0.35, the experiment-wise Type I error rate.

Purpose of the Study

The purpose of the study is to use Monte Carlo methods via Fortran to determine if there is an experiment-wise Type I error rate inflation, and if so, what is its magnitude, when testing nested effects common to educational and psychological research. Given Marascuillo and Serlin's (1988) explication of experiment-wise Type I error inflation, the question arises whether nested designs can be used without corrections for this problem.

Research Hypothesis

It is hypothesized that nested designs, despite their currently increasing popularity, are vulnerable to experiment-wise Type I error inflation. If this is found to be the case, a solution strategy will be suggested to control the inflation.

Operational Definitions

Type I error: A Type I error occurs when "rejecting a null hypothesis when it is true" (Hinkle, Wiersma, & Jurs, 2003, p. 178).

Type I error rate: The number of rejections (i.e., false positives) per number of statistical tests conducted. For example, the Type I error rate under the truth of the null hypothesis when nominal $\alpha = 0.05$ and all underlying conditions are met is expected to be 1 out of 20, or 5%.

Experiment-wise Type I error rate: "Experimentwise error rate (α_E) is the probability of making a Type I error rate for the set of all possible comparisons" (Hinkle, Wiersma, & Jurs, 2003, p. 372).

Nested designs: "Effects which are restricted to a single level of a factor are said to be *nested* within that factor" (Winer, 1971, p. 360).

Hierarchical models: "A hierarchy consists of lower-level observations nested within higher level(s)" (Kreft & De Leeuw, 1998, p. 1).

Bonferroni adjustment: Based on the Bonferroni inequality, Dunn (1961) developed an approach to ensure the experiment-wise Type I error rate cannot rise above nominal alpha. It is achieved by dividing the nominal alpha chosen by the number of tests to be conducted. For example, if three tests are to be conducted, each test would have its alpha level set to $\frac{0.05}{3} = 0.01\overline{6}$.

Limitations

The design used in this study will be limited to students' scores obtained in a single nested layout of three teachers per school with four schools. Data will be randomly selected and assigned from a selection of theoretical mathematical distributions and selected real datasets. Moreover, only sample sizes and alpha levels common to social and behavioral sciences will be modeled. Therefore, the results of this study may not be generalizable to other layouts and study conditions.

Chapter 2

Review of the Literature

Saville (1990) argued that Type II errors are of greater concern than Type I errors. The reason is due to the consideration of practical data analysis versus purely theoretical. In theory, hypothesis tests begin with the null assumption that $\mu_1 = \mu_2 \cdot \cdot \cdot = \mu_k$, for *k* groups. However, Saville (1990) stated that in practice, the means rarely ever end up exactly the same. Chance fluctuations in group means can therefore lead researchers to finding false negatives, in an attempt to control for this.

Therefore, Saville (1990) recommended conducting multiple t tests, concluding that "the multiple comparison controversy is resolved if the procedures are thought of as hypothesis generators rather than as methods for simultaneous generation and testing" (p. 179). However, Bradley (1969) gave limits to the extent to which an inflated Type I error is tolerable, which he called the liberal definition of robustness. It is defined as ±0.5α (i.e., 0.025 – 0.075, when α is set to 0.05). In contrast to Saville (1990), most authors are of the opinion that Type I error inflation is of extreme importance.

Types of Type I Error

There are two domains in which the issue of experiment-wise Type I error rate inflation can be an issue: sequential tests and parallel tests.

Sequential (or Serial) tests.

Sequential tests occur in separate phases. For example, there is the recommendation to test for underlying assumptions (homoscedasticity via Levine's test and via Kolmogorov-Smirnov's test), and only on successfully rejecting both proceeding

to conduct a statistical test of effects (e.g., t test). This strategy was recommended in many statistical packages (e.g., SAS, 1990, p. 25; SPSS, 1993, p. 254-255; SYSTAT, 1990, p. 487). However, Sawilowsky (2002) noted “There is a serious problem with this approach that is universally overlooked. The sequential nature of testing for homogeneity of variance as a condition of conducting the independent samples t test leads to an inflation of experiment-wise Type I errors” (p. 466). Sawilowsky (2002) conducted a Monte Carlo study that demonstrated the experiment-wise Type I error rate inflated to almost twice alpha. A possible solution to this is to avoid using a parametric test that requires testing for underlying assumptions when the data are not known to be normally distributed and homogeneous.

In another example, Walton-Braver and Braver (1988) developed a five-test sequence for the analysis of the Solomon four-group design. Sawilowsky and Markman (1988, 1990a) argued the experiment-wise Type I error for the fifth and final test is “dependent upon the Type II error properties of the preliminary four tests” (p. 178). They concluded regarding the serial testing method that “its limitations must be investigated, and we advise using the technique with caution” (p. 178).

Braver and Walton-Braver (1990) responded by reducing their method from a five test sequence to a four test sequence. Whereas this will somewhat alleviate the experiment-wise Type I error rate, as noted by Sawilowsky and Markman’s (1990b) reply, the issue remained unresolved.

Eventually the matter was formally studied in Sawilowsky, Kelley, Blair, and Markman (1994). They noted that Walton-Braver and Braver (1988) had claimed that “in sequential tests, it is difficult to specify a priori the experiment-wise Type I error rate

over the entire sequence" (p. 153). Sawilowsky et al. (1994) responded: "Although it may be difficult to specify a priori the Type I error rate at a particular step in the sequence, it is straightforward to determine empirically the actual Type I error rate through Monte Carlo methods, and therefore these errors should not be ignored" (p. 366). They found the experiment-wise Type I error rate, over the five test sequence, inflated to 0.138 when nominal alpha was set to 0.05.

Parallel tests.

Parallel tests occur when multiple tests are conducted at the same time. For example, in ANOVA, multiple main effects and interactions can all be of interest. There is debate whether to start with the main effects or interactions, and whether to stop or continue after finding significance (see, e.g., Sawilowsky, 2007a, ch. 14). Regardless of the method chosen, all tests are conducted simultaneously. For example, with three main effects, the following seven combinations can be tested for significance: A × B × C, A × B, A × C, B × C, A, B, and C.

There is a commonly held belief by researchers that ANOVA provides weak protection against the inflation of Type I error rates when conducting multiple tests. This is due to the researcher being genuinely interested in multiple hypotheses. It is believed that this interest adequately negates the effect of conducting repeated measures while utilizing the Frequentist approach. It is argued that ANOVA is in contrast to processes such as stepwise regression, in which the researcher does not have prior suspicion or even interest in the various hypotheses being tested.

For example, Kromrey and Dickenson (1995) stated,

> in a two-factor ANOVA, three null hypotheses are tested (one for each main effect and one for the interaction effect), while in a three-factor analysis, seven null hypotheses are tested (three main effects, three first-order interactions, and one second-order interaction), and in a four-factor analysis, fifteen null hypotheses are tested. The effects of multiple testing… in factorial ANOVA has not been undertaken, despite the fact that the problem has been recognized for more than 30 years. (p. 51-52)

Kromrey and Dickenson (1995) conducted a Monte Carlo simulation in which the number of factors (2-4), pattern of effects (null and/or non-null), effect size (small – large), and sample size (5, 10, & 20) were modeled. The simulation was conducted with 5,000 repetitions per experimental condition. In order to safeguard against rival hypotheses affecting the results, the ANOVA F tests were conducted on data sampled from a theoretical normal distribution, thus ensuring internal validity.

Conditioned on a significant omnibus F test, with the two-factor model, the experiment-wise Type I error rate for the null effects were 0.06. With the three-factor model, it was as high as 0.16, and with four factors, it arose to 0.35 for the null effects. These results demonstrated that the issue of experiment-wise Type I error rate applies to the parallel scenario, even in the presence of a known significant non-null effect. (In other words, the weak protection is ineffective in controlling experiment-wise Type I error rate inflation.

Kromrey and Dickenson (1995) resolved this problem by applying the Bonferroni post hoc procedure, as well as other modifications (i.e., Holm and Hochberg). Each hypothesis test was divided by the overall desired alpha level, which prevented

unwanted inflation of the experiment-wise Type I error rate. However, a Google Scholar search indicated that in the 18 years since Kromrey and Dickenson (1995) was published, it has only been cited only 15 times of which 12 were from the applied literature. Based on this, one can conclude the study has had almost no impact on statistical practice. Being interested in multiple effects does not eliminate the inflation of Type I error when conducting multiple tests.

Classical Solutions to Multiple Comparison Inflations

There are a number of classical ways to control experiment-wise Type I errors that improve on the power loss from the Bonferroni-Dunn adjustment, as well as more modern, computer-based approaches. For example, regarding multiple t tests in the context of one-way ANOVA, SPSS (2013, v. 21) provides the following techniques when the underlying assumption of homogeneous variances condition holds:

- LSD. Uses t tests to perform all pairwise comparisons between group means. No adjustment is made to the error rate for multiple comparisons.
- Šídák (sometimes known as Holm-Šídák). Pairwise multiple comparison test based on a t statistic. Šídák adjusts the significance level for multiple comparisons and provides tighter bounds than Bonferroni.
- Scheffé. Performs simultaneous joint pairwise comparisons for all possible pairwise combinations of means. Uses the F sampling

distribution. Can be used to examine all possible linear combinations of group means, not just pairwise comparisons.

- R-E-G-W F. Ryan-Einot-Gabriel-Welsch multiple stepdown procedure based on an F test.
- R-E-G-W Q. Ryan-Einot-Gabriel-Welsch multiple stepdown procedure based on the Studentized range.
- S-N-K (Student-Newman-Keuls). Makes all pairwise comparisons between means using the Studentized range distribution. With equal sample sizes, it also compares pairs of means within homogeneous subsets, using a stepwise procedure. Means are ordered from highest to lowest, and extreme differences are tested first.
- Tukey. Uses the Studentized range statistic to make all of the pairwise comparisons between groups. Sets the experimentwise error rate at the error rate for the collection for all pairwise comparisons.
- Tukey’s b. Uses the Studentized range distribution to make pairwise comparisons between groups. The critical value is the average of the corresponding value for the Tukey’s honestly significant difference test and the Student-Newman-Keuls.
- Duncan. Makes pairwise comparisons using a stepwise order of comparisons identical to the order used by the Student-Newman-Keuls test, but sets a protection level for the error rate for the

collection of tests, rather than an error rate for individual tests. Uses the Studentized range statistic.

- Hochberg's GT2. Multiple comparison and range test that uses the Studentized maximum modulus. Similar to Tukey’s honestly significant difference test.
- Gabriel. Pairwise comparison test that used the Studentized maximum modulus and is generally more powerful than Hochberg's GT2 when the cell sizes are unequal. Gabriel’s test may become liberal when the cell sizes vary greatly.
- Waller-Duncan. Multiple comparison test based on a t statistic; uses a Bayesian approach.
- Dunnett. Pairwise multiple comparison t test that compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. 2-sided tests that the mean at any level (except the control category) of the factor is not equal to that of the control category. < Control tests if the mean at any level of the factor is smaller than that of the control category. > Control tests if the mean at any level of the factor is greater than that of the control category. (SPSS, help/index.jsp?topic=/com.ibm.spss.statistics.help/ idh_onew_post.htm)

When homogeneity of variance cannot be assumed, SPSS offers the following:

- Tamhane's T2. Conservative pairwise comparisons test based on a t test. This test is appropriate when the variances are unequal.
- Dunnett's T3. Pairwise comparison test based on the Studentized maximum modulus. This test is appropriate when the variances are unequal.
- Games-Howell. Pairwise comparison test that is sometimes liberal. This test is appropriate when the variances are unequal.
- Dunnett's C. Pairwise comparison test based on the Studentized range. This test is appropriate when the variances are unequal. (SPSS, help/index.jsp?topic=/com.ibm.spss.statistics.help/idh_onew_post.htm)

Calculating and Estimating Experiment-wise Type I Error Rates

In terms of the false positive rate, there are two main ways to estimate the Type I error rate based on the number of tests being conducted. Bush and Kennedy (1985) defined the first approach as $1 - p^n$, in which "p denotes the probability in a single instance" of "*not* committing an alpha error" (p. 28). Hence, if α is set to 0.05, p = 0.95. Bush and Kennedy (1985, p. 78) illustrated this with the example of three tests conducted on the same data set. The experiment-wise Type I error rate is

$$\begin{aligned} 1 - .95^3 &= 1 - 0.8574 \\ &= 0.1426 \end{aligned}.$$

The second approach is a crude estimation technique based on multiplying nominal alpha by the number of tests conducted. This approach, described by Marascuilo and Serlin (1988, p. 557), would produce an estimated experiment-wise Type I error rate of 3 × 0.05 = 0.15. Note this result is close to the exact value computed above. However, this estimation procedure becomes unusable as the number of tests conducted increases. Indeed, if 21 tests were computed on the same data set, this procedure would produce an estimated experiment-wise Type I error rate of 1.05, which is above the ceiling for *p*.

Table 3 contains a comparison of the two approaches for projecting the experiment-wise Type I error rate with nominal α set to 0.05 for one to 100 multiple tests on the same data set. The fourth column shows the inefficiency (Delta = M&S – K&B) of the Marascuilo and Serlin estimation procedure, which increases as the number of tests conducted increases.

Table 3

Estimated and calculated Type I error rates

# Tests	M&S	K&B	Delta
1	0.0500	0.0500	0.0000
2	0.1000	0.0975	0.0025
3	0.1500	0.1426	0.0074
4	0.2000	0.1855	0.0145
5	0.2500	0.2262	0.0238
6	0.3000	0.2649	0.0351
7	0.3500	0.3017	0.0483

Table 3 continued

Estimated and calculated Type I error rates

# Tests	M&S	K&B	Delta
8	0.4000	0.3366	0.0634
9	0.4500	0.3698	0.0802
10	0.5000	0.4013	0.0987
11	0.5500	0.4312	0.1188
12	0.6000	0.4596	0.1404
13	0.6500	0.4867	0.1633
14	0.7000	0.5123	0.1877
15	0.7500	0.5367	0.2133
16	0.8000	0.5599	0.2401
17	0.8500	0.5819	0.2681
18	0.9000	0.6028	0.2972
19	0.9500	0.6226	0.3274
20	1.0000	0.6415	0.3585
21	1.0500	0.6594	0.3906
22	1.1000	0.6765	0.4235
23	1.1500	0.6926	0.4574
24	1.2000	0.7080	0.4920
25	1.2500	0.7226	0.5274
26	1.3000	0.7365	0.5635
27	1.3500	0.7497	0.6003
28	1.4000	0.7622	0.6378
29	1.4500	0.7741	0.6759
30	1.5000	0.7854	0.7146
31	1.5500	0.7961	0.7539
32	1.6000	0.8063	0.7937
33	1.6500	0.8160	0.8340
34	1.7000	0.8252	0.8748
35	1.7500	0.8339	0.9161
36	1.8000	0.8422	0.9578
37	1.8500	0.8501	0.9999
38	1.9000	0.8576	1.0424
39	1.9500	0.8647	1.0853
40	2.0000	0.8715	1.1285
41	2.0500	0.8779	1.1721
42	2.1000	0.8840	1.2160
43	2.1500	0.8898	1.2602
44	2.2000	0.8953	1.3047

Table 3 continued

Estimated and calculated Type I error rates

# Tests	M&S	K&B	Delta
45	2.2500	0.9006	1.3494
46	2.3000	0.9055	1.3945
47	2.3500	0.9103	1.4397
48	2.4000	0.9147	1.4853
49	2.4500	0.9190	1.5310
50	2.5000	0.9231	1.5769
51	2.5500	0.9269	1.6231
52	2.6000	0.9306	1.6694
53	2.6500	0.9340	1.7160
54	2.7000	0.9373	1.7627
55	2.7500	0.9405	1.8095
56	2.8000	0.9434	1.8566
57	2.8500	0.9463	1.9037
58	2.9000	0.9490	1.9510
59	2.9500	0.9515	1.9985
60	3.0000	0.9539	2.0461
61	3.0500	0.9562	2.0938
62	3.1000	0.9584	2.1416
63	3.1500	0.9605	2.1895
64	3.2000	0.9625	2.2375
65	3.2500	0.9644	2.2856
66	3.3000	0.9661	2.3339
67	3.3500	0.9678	2.3822
68	3.4000	0.9694	2.4306
69	3.4500	0.9710	2.4790
70	3.5000	0.9724	2.5276
71	3.5500	0.9738	2.5762
72	3.6000	0.9751	2.6249
73	3.6500	0.9764	2.6736
74	3.7000	0.9775	2.7225
75	3.7500	0.9787	2.7713
76	3.8000	0.9797	2.8203
77	3.8500	0.9807	2.8693
78	3.9000	0.9817	2.9183
79	3.9500	0.9826	2.9674
80	4.0000	0.9835	3.0165
81	4.0500	0.9843	3.0657

Table 3 continued

Estimated and calculated Type I error rates

# Tests	M&S	K&B	Delta
82	4.1000	0.9851	3.1149
83	4.1500	0.9858	3.1642
84	4.2000	0.9865	3.2135
85	4.2500	0.9872	3.2628
86	4.3000	0.9879	3.3121
87	4.3500	0.9885	3.3615
88	4.4000	0.9890	3.4110
89	4.4500	0.9896	3.4604
90	4.5000	0.9901	3.5099
91	4.5500	0.9906	3.5594
92	4.6000	0.9911	3.6089
93	4.6500	0.9915	3.6585
94	4.7000	0.9919	3.7081
95	4.7500	0.9923	3.7577
96	4.8000	0.9927	3.8073
97	4.8500	0.9931	3.8569
98	4.9000	0.9934	3.9066
99	4.9500	0.9938	3.9562
100	5.0000	0.9941	4.0059

Notes: M&S = Marascuilo and Serlin (1988), K&B = Kennedy and Bush (1985), Delta = M&S – K&B.

Nested Designs

As mentioned above, Kanji (1999) provided an example of the application of the ANOVA F test for a nested design. The example is repeated here both for explication of the calculations and to provide a worked example that will be used to determine the accuracy of the Fortran coding to be developed in this study. In the example, there are four schools with three teachers nested within each school, with the test scores as presented in Table 4.

Kanji (1999) noted the residual, A School factor, and B Teacher nest sums of squares, are respectively

$$S_E^2 = \sum_i \sum_j \sum_k \left(Y_{ijk} - Y_{ijo} \right)^2,$$

$$S_S^2 = \sum_i n_i \left(Y_{i00} - Y_{000} \right)^2, \text{ and}$$

$$S_T^2 = \sum_i \sum_j n_{ij} \left(Y_{ij0} - Y_{i00} \right)^2,$$

where E is the residual, S is the School, T is the Teacher, the test is H_A: $\alpha_i = 0$ for all *i*, and the test for H_B: $\beta_{ij} = 0$ for all *i*, *j*. Supplying the appropriate calculations to the ANOVA table yields Table 5.

With *df* = 8, 60, the critical value for the nominal $\alpha = 0.05$ is 2.10. Because 1.46 < 2.10, the nested effect of differences by Teachers is not statistically significant. However, with *df* = 3, 8, the critical value is 4.07. Because 6.47 > 4.07, the difference between Schools are statistically significant. Note that in this case the test for nested effects was unnecessary. Nevertheless, conducting this test will add to the experiment-wise Type I error rate.

Table 4

Nested design example data from Kanji (1999, p. 129)

	Schools											
	I			II			III			IV		
	Teacher			Teacher			Teacher			Teacher		
	1	2	3	1	2	3	1	2	3	1	2	3
	44	39	39	51	48	44	46	45	43	42	45	39
	41	37	36	49	43	43	43	40	41	39	40	38
	39	35	33	45	42	42	41	38	39	38	37	35
	36	35	31	44	40	39	40	38	37	36	37	35
	35	34	28	40	37	37	36	35	34	34	32	35
	32	30	26	40	34	36	34	34	33	31	32	29
TT	227	210	193	269	244	241	240	230	227	220	223	211
$\bar{X}_T$	37.8	35.0	32.17	44.83	40.67	40.16	40.0	38.33	37.83	36.67	37.17	35.17
ST	630			754			679			654		
$\bar{X}_S$	35.00			41.89			38.72			36.33		

Notes: TT = Teacher total, ST = School total, Grand mean School total = 2,735.

Table 5

Data from the Kanji (1999, p. 130) ANOVA table

	df	SS	Mean Square	F
Schools	3	493.60	164.53	6.47
Teachers within School	8	203.55	25.44	1.46
Pupils within Teachers	60	1047.84	17.46	
Total	71	1744.99		

Chapter 3

Methodology

A Monte Carlo simulation will be conducted via Absoft Pro Fortran v. 11.5.3 (IDE 2.2) Fortran program written for the WIN/TEL platform. It will be executed on a 64 bit Windows Ultimate Eight Core i7 Intel Sandy Bridge 2700k CPU running at 3.5 Ghz, with 16 Gb of RAM.

Design

A two-factor nested layout or hierarchical classification layout will be used. This design assumes errors are normally distributed, and the magnitudes of those errors are independent from either of the two factors. Specifically, the hypothetical layout pertains to an analysis of difference of means between classes taught by different teachers, with teachers in turn being nested within different schools. In this layout, student test scores will be simulated for three teachers (or classrooms) per each of four schools, as noted in the table below.

Table 6

Nested design example data from Kanji (1999, p. 129)

School I			School II			School III			School IV		
Teacher			Teacher			Teacher			Teacher		
1	2	3	4	5	6	7	8	9	10	11	12
n1											
n2											
n3											
…											N

It should be noted in the literature, when nested designs are utilized, they are almost always conducted through the use of multiple ANOVA tests. Others, such as the t test, are generally not found, because rarely are such studies conducted on two schools with two teachers per school (e.g., Kanji, 1999 & Winer, 1971). Therefore, when a nested layout is demonstrated, the ANOVA test is required.

Sampling Plan

A pseudo-random number generator will be used to simulate student test scores. The data will be generated through Roguewave's (2012) subroutine libraries for the theoretical distributions. Data will be simulated to follow the: Gaussian, uniform, exponential, t (*df* = 3), and Chi-squared (*df* = 2) distributions. Variates from the Gaussian (i.e., normal) distribution will be used to demonstrate the veracity of the Fortran coding. Deviates from non-normal distributions are commonly used in Monte Carlo studies to illustrate robustness properties with respect to Type I errors for departure from population normality.

Samples will also be obtained from real data sets (Micceri, 1989) via the Realpops 2.0 subroutine library (Sawilowsky & Fahoome, 2003); Realpops 2.0 is a Fortran 90 updated version of the Fortran 77 subroutine library by Sawilowsky, Blair, and Micceri (1990). (For details on the real data sets, see Micceri, 1989, and Sawilowsky & Blair, 1992.) The real data sets to be sampled will be the smooth symmetric (achievement scores), digit preference (achievement scores), multi-modal lumpy (achievement scores), and extreme asymmetry (psychometric scores).

Sample sizes will be set to n = 2, 10, 30, 45, and 120. Samples of size n = 2 and n = 120 will be selected to represent the theoretical minimum and a reasonable maximum study parameter, as is customarily done in Monte Carlo studies. Samples of size n = 10, 30, and 45 will be selected to represent small, medium and large classrooms, respectively. Under the truth of the null hypothesis (and homoscedasticity as modeled in this study), unbalanced layouts (i.e., unequal sample sizes per teacher or unequal teachers per school) will have no impact on Type I errors and are therefore not modeled. One million repetitions will be executed for each combination of study parameters.

Analysis

The appropriate analysis for the nested design in Table 1 is a series of two F tests. Initially, the F test is conducted to determine if there are teacher differences. Under ideal conditions, the intent is to fail to reject the null hypothesis. This is because it is assumed that the teachers have similar qualifications (e.g., certification, experience) in order to be named the instructor of record.

The more important test is then conducted. This is an F test for effects, which in this case is for the difference in means between schools. When the null hypothesis is false, it means the new curriculum administered in at least one school statistically significantly changed student scores. The F test should reject this null hypothesis.

In the current study, the truth of the null hypothesis is based on the generation of pseudo-random numbers. There will be an expected Type I error rate for each of the

component tests. The experiment-wise Type I error rate will be determined by the sum of those two Type I error rates.

This will be accomplished in two ways. The first is unconditional; meaning the test for effects (i.e., between schools) will be conducted regardless of the results of the test for nesting (i.e., between teachers). The second is conditional; meaning the test for effects will only be conducted if and only if (*iff*) a nesting effect is non-null.

Differentiating between unconditional and conditional testing is advisable if the general purpose for conducting an intervention study is to determine if there is a difference between schools where students did or did not receive an intervention. The impact of teacher differences should be negligible. In other words, the school effect should only be tested when it can be first shown there was no teacher effect.

Alpha Levels

In order to increase generality of results, the F tests invoked in the Monte Carlo simulation will be conducted at both the nominal alpha = 0.05 and 0.01 levels.

Table Template

The results of the Monte Carlo simulation will be presented in the following formats of Table 7 and Table 8. Hence, there will be 2 alpha levels × 9 distributions/data sets × 2 condition statuses = 36 tables of results.

Table 7

Unconditional Type I error rates, α = 0.05, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2			
10			
30			
45			
120			

Notes: n = sample size per cell, Factor = School, Nest = Teacher.

Table 8

Conditional Type I error rates, α = 0.05, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2			
10			
30			
45			
120			

Notes: n = sample size per cell, Factor = School, Nest = Teacher.

Error Isolation

The Monte Carlo is being conducted using parametric or normal theory tests. However, data are also drawn from non-normal distributions. Therefore, the issue arises as to where potential results are originating. If the Type I error rates do inflate, it is important to determine whether these results are due to experiment-wise Type I error inflation or if they are caused by violating the assumption of normality. Typical Type I error rates are listed in Table 9.

By knowing what happens to the Type I error rate with a single statistical test when the assumption of normality is violated, the expected experiment-wise Type I error rate will be double this value (for the unconditional method). For example, under an exponential distribution with sample size n = 15, the Type I error becomes conservative: when nominal alpha is set to 0.05, the actual Type I error rate is 0.04 (Glass, Peckham, & Sanders, 1972). (Although making fewer Type I errors than expected sounds good, the downside is the power is reduced when the test becomes conservative.) Therefore, with unconditional testing of nested effects, the expectation is for 0.04 × 2 = 0.08. In the event of experimental results summing to this value (within sampling error due to the Monte Carlo), experiment-wise inflation will be isolated as the cause for this error. Meanwhile, conditional experiment-wise Type I error rate inflation is more difficult to predict.

Table 9

Expected Type I error rates for normal and selected non-normal data at α = 0.05 and α = 0.01 (Glass, Peckham, & Sanders[1], 1972, p. 250; Sawilowsky and Blair[2], 1982 p. 356-358)

Distribution / Dataset	Resulting alpha (*0.05*)	Resulting alpha (*0.01*)
Normal	0.050	0.010
Exponential[1]	0.040	0.004
Uniform[1]	0.051	0.010
Digit preference[2]	0.050	0.012
Extreme asymmetric[2]	0.047	0.009
Multi-modal lumpy[2]	0.052	0.012
Smooth symmetric[2]	0.050	0.010

Note: These results are for different numbers of repetitions and are based generally on the balanced layout of samples sizes $n_1 = n_2 = 20$. Increasing the number of repetitions and sample sizes will give Type I errors closer to nominal alpha.

Chapter 4

Results

The following sections contain tables showing the results of the study.

Unconditional

The test for the nest and the treatment effect are both conducted in this model of analysis. Although it does not matter which test is conducted first for consistency, the test for the nest was conducted prior to the test of the effect. A series of tabled results are presented, arranged by distribution or dataset type. The entries inside each table represent the Type I error rate for the study conditions.

Gaussian distribution.

The Gaussian distribution is also known as the normal distribution, or bell curve. It is a model commonly used in education. However, a survey by Micceri (1989) indicated that real datasets rarely are symmetric with light tails, two features which are prominent in the bell curve shown in Figure 1 below.

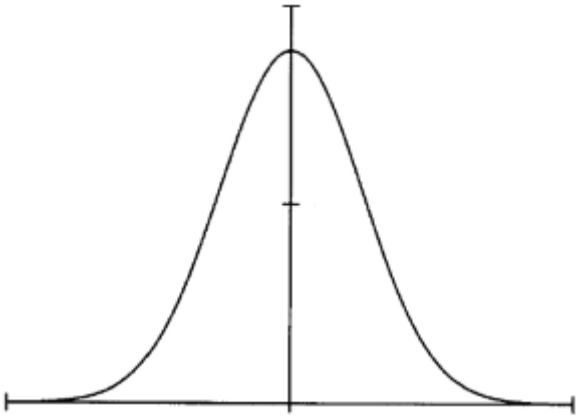

Figure 1: Gaussian distribution (Excerpted from Sawilowsky & Fahoome, 2002, p. 343)

The results for data sampled from this distribution are compiled in Tables 10 and 11, with Table 10 representing results for the nominal alpha = 0.05 and Table 11 representing 0.01.

Table 10

Unconditional Type I error rates, α = 0.05, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.049809	0.050238	0.100047
10	0.050013	0.050271	0.100284
30	0.050134	0.049765	0.099899
45	0.050115	0.050436	0.100551
120	0.050126	0.049639	0.099765

Table 11

Unconditional Type I error rates, α = 0.01, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010093	0.009873	0.019966
10	0.010109	0.010010	0.020119
30	0.010074	0.010117	0.020191
45	0.009909	0.010031	0.019940
120	0.010024	0.009999	0.020023

Chi-squared distribution, 3 degree of freedom.

The chi-squared distribution is “the distribution of the sums of squares” of normal variates (Evans, Hastings, & Peacock, 2000, p. 52). It is a model commonly used in education because it is the referent distribution for the chi-squared statistic which is

used in cross-tabulations. Through straightforward transformations, it is also related to the gamma, F, Poisson, and t distributions.

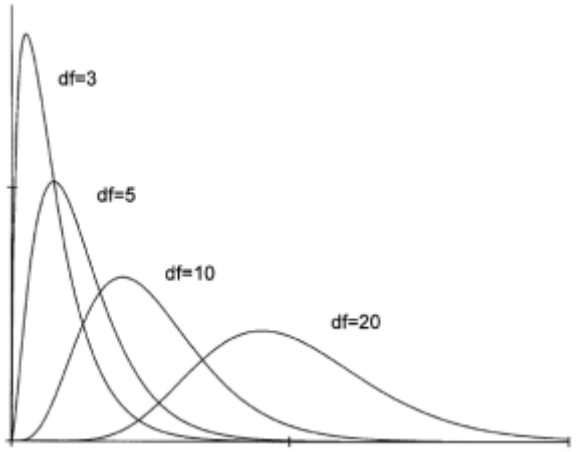

Figure 2: Chi-squared distribution (Excerpted from Sawilowsky & Fahoome, 2002, p. 344)

The results for data sampled from this distribution are compiled in Tables 12 and 13, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 12

Unconditional Type I error rates, α = 0.05, chi-squared (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054072	0.047537	0.101609
10	0.048256	0.049527	0.097783
30	0.049284	0.049878	0.099162
45	0.049365	0.049755	0.099120
120	0.049387	0.050258	0.099645

Table 13

Unconditional Type I error rates, α = 0.01, chi-squared (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.014348	0.010668	0.025016
10	0.009602	0.010208	0.019810
30	0.009574	0.010023	0.019597
45	0.009810	0.010138	0.019948
120	0.009756	0.010142	0.019898

Exponential distribution.

The exponential distribution is commonly used to model growth and decay. It is a special case of the gamma and Weibull distribution. Through various transformations it is related to the uniform and Erlang distributions. This distribution is not only of practical importance but is a long standing theoretical model of note. For example, the difference between two exponential variates is a Laplace variate, and by changing other parameters is related to the Pareto and Gumbel distributions. Other members of the exponential family include Bernoulli, binomial, geometric, Gaussian, logarithmic, Rayleigh and von Mises distributions in the univariate case, and the normal, Dirichlet, Wishart multivariate distributions (Evans, Hastings, & Peacock, 2000, p. 81-82).

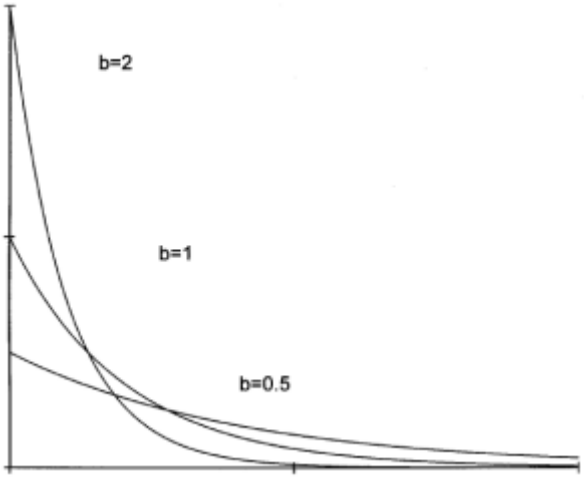

Figure 3: Exponential decay (Excerpted from Sawilowsky & Fahoome, 2002, p. 348)

The results for data sampled from this distribution are compiled in Tables 14 and 15, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 14

Unconditional Type I error rates, α = 0.05, exponential distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.055038	0.046336	0.101374
10	0.047278	0.048773	0.096051
30	0.048832	0.050075	0.098907
45	0.049161	0.049849	0.099010
120	0.049749	0.050005	0.099754

Table 15

Unconditional Type I error rates, α = 0.01, exponential distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.016123	0.010746	0.026869
10	0.009685	0.010227	0.019912
30	0.009812	0.010269	0.020081
45	0.009800	0.009960	0.019760
120	0.010023	0.010070	0.020093

t distribution, 3 degrees of freedom.

The t distribution, developed by William Sealy Gosset (Student, 1908), forms the backdrop for testing the differences between means, one of the most common statistical techniques used in the social and behavioral sciences. In the two sample layout, it is the square root of F. When *df* = 1, it is known as the Cauchy distribution (Evans, Hastings, & Peacock, 2000, p. 182). As $N \rightarrow \infty$, $t \sim Z$.

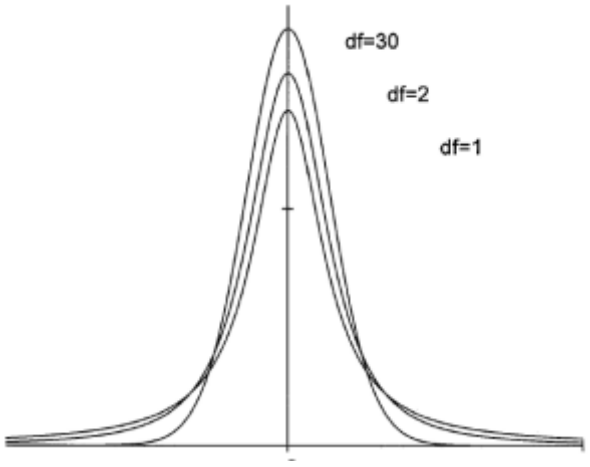

Figure 4: t distribution (Excerpted from Sawilowsky & Fahoome, 2002, p. 345)

The results for data sampled from this distribution are compiled in Tables 16 and 17, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 16

Unconditional Type I error rates, α = 0.05, t (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.041380	0.040655	0.082035
10	0.044071	0.045085	0.089156
30	0.046541	0.047470	0.094011
45	0.047050	0.047384	0.094434
120	0.048256	0.048455	0.096711

Table 17

Unconditional Type I error rates, α = 0.01, t (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.008599	0.007372	0.015971
10	0.007965	0.008549	0.016514
30	0.008581	0.009125	0.017706
45	0.008754	0.009116	0.017870
120	0.009222	0.009480	0.018702

Uniform distribution.

Also known as the rectangular distribution, the uniform distribution forms the basis of Monte Carlo studies, because it is used for the generation of random numbers that are equiprobable. The uniform distribution has both a continuous and discrete form.

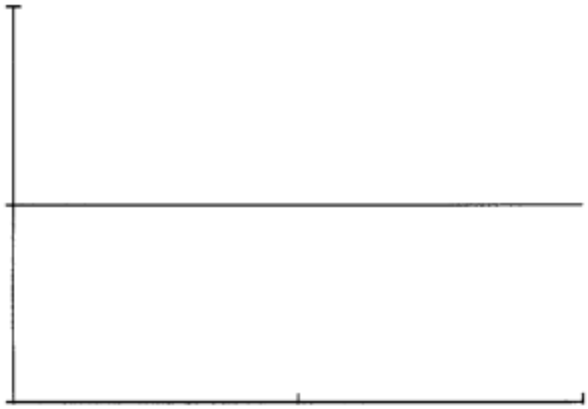

Figure 5: Uniform distribution (Excerpted from Sawilowsky & Fahoome, 2002, p. 342)

The results for data sampled from this distribution are compiled in Tables 18 and 19, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 18

Unconditional Type I error rates, α = 0.05, uniform distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054535	0.052737	0.107272
10	0.050997	0.050217	0.101214
30	0.050176	0.050291	0.100467
45	0.050061	0.049965	0.100026
120	0.050304	0.050057	0.100361

Table 19

Unconditional Type I error rates, α = 0.01, uniform distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.012261	0.011234	0.023495
10	0.010462	0.010010	0.020472
30	0.010214	0.010017	0.020231
45	0.009918	0.010051	0.019969
120	0.010121	0.010119	0.020240

Digit preference dataset.

Tables 20 through 27 repeat the same patterns as Tables 10 – 19 above, except the referent distributions were replaced with the large sample datasets by Micceri (1989), as coded by Sawilowsky, Blair, and Micceri (1990). As can be observed in Figure 6, it is essentially symmetric with light tails, but has certain score prevalences.

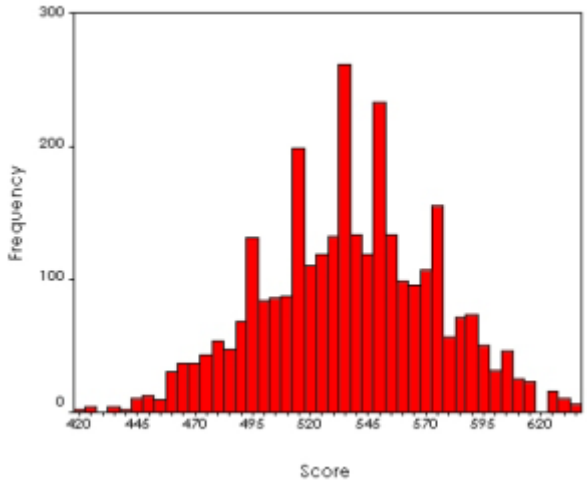

Figure 6: Digit preference dataset by Micceri (1989) (Excerpted from Sawilowsky & Fahoome, 1990, p. 357)

The results for data sampled from this dataset are compiled in Tables 20 and 21, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 20

Unconditional Type I error rates, α = 0.05, digit preference dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.050826	0.050770	0.101596
10	0.050264	0.049959	0.100223
30	0.049751	0.050311	0.100062
45	0.049932	0.049939	0.099871
120	0.050455	0.050024	0.100479

Table 21

Unconditional Type I error rates, α = 0.01, digit preference dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010324	0.010183	0.020507
10	0.010160	0.010139	0.020299
30	0.010059	0.010136	0.020195
45	0.010032	0.009843	0.019875
120	0.010010	0.010163	0.020173

Extreme asymmetric dataset.

This Micceri (1989) dataset is the similar to mathematical exponential distributions. As with the theoretical population, the asymmetric datasets can represent growth (see Figure 7) or decay.

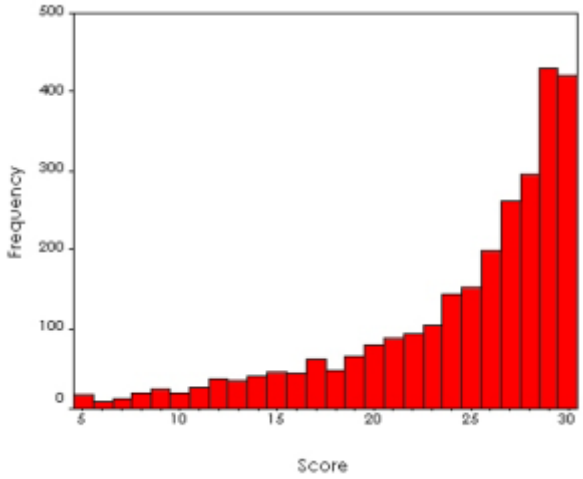

Figure 7: Extreme asymmetric dataset by Micceri (1989) (Excerpted from Sawilowsky & Fahoome, 1990, p. 351)

The results for data sampled from this dataset are compiled in Tables 22 and 23, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 22

Unconditional Type I error rates, α = 0.05, extreme asymmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.064911	0.049738	0.114649
10	0.048979	0.050588	0.099567
30	0.049553	0.050112	0.099665
45	0.049147	0.049971	0.099118
120	0.049837	0.050627	0.100464

Table 23

Unconditional Type I error rates, α = 0.01, extreme asymmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.023892	0.014844	0.038736
10	0.010311	0.010586	0.020897
30	0.009830	0.010206	0.020036
45	0.009966	0.009978	0.019944
120	0.009975	0.010134	0.020109

Multi-modal lumpy dataset.

Theoretical bimodal populations (also known as mixed or contaminated normal) are prevalent, such as the distribution of language scores of students in a school containing native and second language speakers. The Micceri (1989) large sample dataset that depicts this condition is also markedly lumpy, as opposed to smooth and “interesting mathematical functions” (p. 157) in theoretical models.

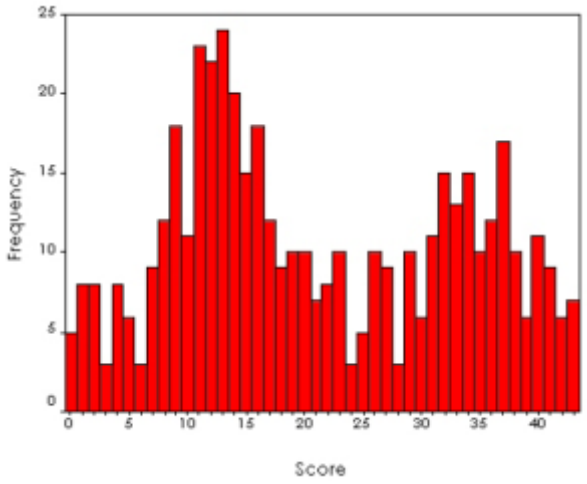

Figure 8: Multi-modal lumpy dataset by Micceri (1989) (Excerpted from Sawilowsky & Fahoome, 1990, p. 354)

The results for data sampled from this dataset are compiled in Tables 24 and 25, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 24

Unconditional Type I error rates, α = 0.05, multi-modal lumpy dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054927	0.053243	0.108170
10	0.058563	0.050270	0.108833
30	0.050003	0.050406	0.100409
45	0.049934	0.049937	0.099871
120	0.050364	0.050075	0.100439

Table 25

Unconditional Type I error rates, α = 0.01, multi-modal lumpy dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.012824	0.011325	0.024149
10	0.013750	0.010146	0.023896
30	0.010275	0.010119	0.020394
45	0.009990	0.009900	0.019890
120	0.009944	0.010086	0.020030

Smooth symmetric dataset.

The smooth symmetric dataset is the best estimate of the Gaussian distribution that is found in nature. It is smoother than most real datasets, and has light tails. However, as indicated by Micceri (1989), none of the datasets of this type found in his survey of social and behavioral science datasets passed the Kolgomorov-Smirnov test for normality.

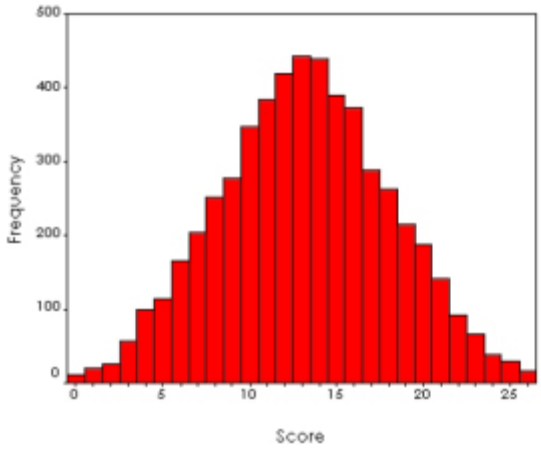

Figure 9: Smooth symmetric Dataset by Micceri (1989) (Excerpted from Sawilowsky & Fahoome, 1990, p. 530)

The results for data sampled from this dataset are compiled in Tables 26 and 27, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 26

Unconditional Type I error rates, α = 0.05, smooth symmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.051114	0.050992	0.102106
10	0.050295	0.049862	0.100157
30	0.049573	0.050358	0.099931
45	0.049793	0.049927	0.099720
120	0.050428	0.050041	0.100469

Table 27

Unconditional Type I error rates, α = 0.01, smooth symmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010500	0.010274	0.020774
10	0.010119	0.010206	0.020325
30	0.010039	0.010199	0.020238
45	0.009953	0.009939	0.019892
120	0.009918	0.010090	0.020008

Conditional

The premise of nested designs is the existence of a confounding factor. Obviously, if it was known that no nest effect exists, there would be no purpose in invoking a nested design. Hence, there is little purpose in conducting a test for effects, conditional on failing to reject the null hypothesis for nest effects. However, when the nested effect is retained, the sums of squares are partitioned to it in an effort to reduce the residual sum of squares used in computing the F ratio for the treatment effect. Therefore, the process is completed with conducting the test of effects. The tabled results in this section depict the Type I errors and experiment-wise inflations when the test of effects is conducted *iff* the nested effect is statistically significant.

Gaussian distribution.

The results for data sampled from this distribution are compiled in Tables 28 and 29, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 28

Conditional Type I error rates, α = 0.05, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.049809	0.001646	0.051455
10	0.050013	0.000080	0.050093
30	0.050134	0.000020	0.050154
45	0.050115	0.000020	0.050135
120	0.050126	0.000020	0.050146

Table 29

Conditional Type I error rates, α = 0.01, Gaussian distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010093	0.000099	0.010192
10	0.010109	0.000000	0.010109
30	0.010074	0.000000	0.010074
45	0.009909	0.000000	0.009909
120	0.010024	0.000000	0.010024

Chi-square distribution, 3 degree of freedom.

The results for data sampled from this distribution are compiled in Tables 30 and 31, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 30

Conditional Type I error rates, α = 0.05, chi-squared (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054072	0.002256	0.056328
10	0.048256	0.000104	0.048360
30	0.049284	0.000000	0.049284
45	0.049365	0.000000	0.049365
120	0.049387	0.000000	0.049387

Table 31

Conditional Type I error rates, α = 0.01, chi-squared (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.014348	0.000070	0.014418
10	0.009602	0.000000	0.009602
30	0.009574	0.000000	0.009574
45	0.009810	0.000000	0.009810
120	0.009756	0.000000	0.009756

Exponential distribution.

The results for data sampled from this distribution are compiled in Tables 32 and 33, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 32

Conditional Type I error rates, α = 0.05, exponential distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.055038	0.002362	0.057400
10	0.047278	0.000021	0.047299
30	0.048832	0.000021	0.048853
45	0.049161	0.000041	0.049202
120	0.049749	0.000000	0.049749

Table 33

Conditional Type I error rates, α = 0.01, exponential distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.016123	0.000062	0.016185
10	0.009685	0.000000	0.009685
30	0.009812	0.000000	0.009812
45	0.009800	0.000000	0.009800
120	0.010023	0.000000	0.010023

t distribution, 3 degrees of freedom.

The results for data sampled from this distribution are compiled in Tables 34 and 35, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 34

Conditional Type I error rates, α = 0.05, t (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.041380	0.001474	0.042854
10	0.044071	0.000023	0.044094
30	0.046541	0.000022	0.046563
45	0.047050	0.000000	0.047050
120	0.048256	0.000000	0.048256

Table 35

Conditional Type I error rates, α = 0.01, t (df = 3) distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.008599	0.000000	0.008599
10	0.007965	0.000000	0.007965
30	0.008581	0.000000	0.008581
45	0.008754	0.000000	0.008754
120	0.009222	0.000000	0.009222

Uniform distribution.

The results for data sampled from this distribution are compiled in Tables 36 and 37, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 36

Conditional Type I error rates, α = 0.05, uniform distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054535	0.002714	0.057249
10	0.050997	0.000059	0.051056
30	0.050176	0.000000	0.050176
45	0.050061	0.000040	0.050101
120	0.050304	0.000000	0.050304

Table 37

Conditional Type I error rates, α = 0.01, uniform distribution, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.012261	0.000082	0.012343
10	0.010462	0.000000	0.010462
30	0.010214	0.000000	0.010214
45	0.009918	0.000000	0.009918
120	0.010121	0.000000	0.010121

Digit preference dataset.

The results for data sampled from this distribution are compiled in Tables 38 and 39, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 38

Conditional Type I error rates, α = 0.05, digit preference dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.050826	0.002086	0.052912
10	0.050264	0.000000	0.050264
30	0.049751	0.000020	0.049771
45	0.049932	0.000000	0.049932
120	0.050455	0.000020	0.050475

Table 39

Conditional Type I error rates, α = 0.01, digit preference dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010324	0.000000	0.010324
10	0.010160	0.000000	0.010160
30	0.010059	0.000000	0.010059
45	0.010032	0.000000	0.010032
120	0.010010	0.000000	0.010010

Extreme asymmetric dataset.

The results for data sampled from this distribution are compiled in Tables 40 and 41, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 40

Conditional Type I error rates, α = 0.05, extreme asymmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.064911	0.003790	0.068701
10	0.048979	0.000041	0.049020
30	0.049553	0.000000	0.049553
45	0.049147	0.000020	0.049167
120	0.049837	0.000000	0.049837

Table 41

Conditional Type I error rates, α = 0.01, extreme asymmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.023892	0.000251	0.024143
10	0.010311	0.000000	0.010311
30	0.009830	0.000000	0.009830
45	0.009966	0.000000	0.009966
120	0.009975	0.000000	0.009975

Multi-modal lumpy dataset.

The results for data sampled from this distribution are compiled in Tables 42 and 43, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 42

Conditional Type I error rates, α = 0.05, multi-modal lumpy dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.054927	0.002968	0.057895
10	0.058563	0.000017	0.058580
30	0.050003	0.000020	0.050023
45	0.049934	0.000020	0.049954
120	0.050364	0.000020	0.050384

Table 43

Conditional Type I error rates, α = 0.01, multi-modal lumpy dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.012824	0.000000	0.012824
10	0.013750	0.000000	0.013750
30	0.010275	0.000000	0.010275
45	0.009990	0.000000	0.009990
120	0.009944	0.000000	0.009944

Smooth symmetric dataset.

The results for data sampled from this distribution are compiled in Tables 44 and 45, with the first table representing results for the nominal alpha = 0.05 and the second table representing 0.01.

Table 44

Conditional Type I error rates, α = 0.05, smooth symmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.051114	0.002015	0.053129
10	0.050295	0.000000	0.050295
30	0.049573	0.000040	0.049613
45	0.049793	0.000000	0.049793
120	0.050428	0.000000	0.050428

Table 45

Conditional Type I error rates, α = 0.01, smooth symmetric dataset, repetitions = 1,000,000

n	Nest	Factor	Experiment-wise
2	0.010500	0.000000	0.010500
10	0.010119	0.000000	0.010119
30	0.010039	0.000000	0.010039
45	0.009953	0.000000	0.009953
120	0.009918	0.000000	0.009918

Chapter 5

Discussion

Kreft and De Leeuw (1998) noted,

> Hierarchical data structures are very common in the social and behavioral sciences… Once you know that hierarchies exist, you see them everywhere… Examples include students nested within schools, employees nested within firms, or repeated measurements nested within persons. (p. 1)

Similarly, Gonzales (2009) indicated when the "factors are not crossed… we cannot use the machinery of the factorial analysis of variance" (p. 313). The proposed solution is to turn to nested designs which are "now a major area of research in social science statistics" (p. 314). Gonzales (2009) concluded: "Multilevel modeling techniques permit simultaneous modeling of all the levels that are accounted for in the design" (p. 315).

Unfortunately, the observations of Kreft and De Leeuw and Gonzales overlook the impact of conducting statistical tests in a hierarchical model in general and in nested designs in particular. Gonzales (2009) attempted to forestall the impact of multiple testing with the rhetorical question, "Aren't we capitalizing on chance by making so many comparisons?" (p. 336). The first answer given was to make nested designs analogous to factorial ANOVA where there appears to be no concern in the statistical literature over the inflation of Type I error in testing main effects and interactions. However, as noted by Kromrey and Dickenson (1995), and discussed at length in Chapter 2, this provides no safe haven from experiment-wise Type I error inflation.

The second argument advanced by Gonzales to preclude issues of multiple testing in nested designs was, "Replication is the best way to deal with concerns about

multiple tests and inflated Type I error rates" (p. 337). However, Sawilowsky (2007b) demonstrated in a Monte Carlo experiment that "replicating the same poor design has little chance of contributing accurate evidence for or against the effectiveness of a treatment, or for quantifying the magnitude of its effectiveness if it exists" (p. 221-222).

The third argument advanced by Gonzales (2009) was to apply a correction such as the Bonferroni-Dunn technique (p. 285). This is precisely the solution strategy previously proposed by Kromrey and Dickenson (1995). However, such methods always result in a reduction of statistical power and should be used as a last resort.

Indeed, despite offering these three solution strategies, Gonzales (2009) concluded that experiment-wise Type I error rate inflation was something that researchers need not take seriously. However, to his credit, Gonzales' final word on this issue was "We admit that we are in the minority among methodologists on this particular point" (p. 285).

Hence, the purpose of this study was to explicate the impact of simple nesting designs on experiment-wise Type I error rates via a Monte Carlo exercise. Study parameters included popular population distributions and vetted large datasets to generate samples using common sample sizes and alpha levels for the single nested layout of three teachers per school with four schools. The tests for the nest and effect were conducted unconditionally and conditionally.

As predicted by theory (Marascuillo & Serlin, 1988), the results in Tables 10-27 demonstrate that conducting a series of two statistical tests unconditionally, regardless of the nature of those tests, produces an experiment-wise Type I error rate of approximately twice nominal alpha. Tables 46-47 contain a compilation of those results.

Table 46

Summary of average Type I Error rates for various distributions/datasets, unconditional, alpha = 0.05

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.050039	0.050070	0.100109
Chi-square (df=3)	0.050073	0.049391	0.099464
Exponential	0.050012	0.049008	0.099019
t (df=3)	0.045460	0.045810	0.091269
Uniform	0.051215	0.050653	0.101868
Digit preference	0.050246	0.050201	0.100446
Extreme asymmetric	0.052485	0.050207	0.102693
Multi-modal lumpy	0.052758	0.050786	0.103544
Smooth symmetric	0.050241	0.050236	0.100477

Table 47

Summary of average Type I Error rates for various distributions/datasets, unconditional, alpha = 0.01

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.010042	0.010006	0.020048
Chi-square (df=3)	0.010618	0.010236	0.020854
Exponential	0.011089	0.010254	0.021343
t (df=3)	0.008624	0.008728	0.017353
Uniform	0.010595	0.010286	0.020881
Digit preference	0.010117	0.010093	0.020210
Extreme asymmetric	0.012795	0.011150	0.023944
Multi-modal lumpy	0.011357	0.010315	0.021672
Smooth symmetric	0.010106	0.010142	0.020247

In Tables 48-49, the Type I error rates are averaged as in the previous two tables, except the test for the factor (i.e., School) is conducted conditionally subsequent to a significant test of the nesting effect. In order to understand these results, consider Bradley's (1978) definition for two levels of robustness. The conservative definition is

met when the Type I error rate is within the bounded interval [0.5α - 1.5α] inclusive, and the liberal definition is met when the Type I error rate is within the bounded interval [0.9α - 1.1α] inclusive. The results for the factor (School) are ultra-conservative, falling far below 0.025 when the test is conducted at the 0.05 nominal alpha level, and below .005 when the test is conducted at the 0.01 nominal alpha level. In addition, the impact of being ultra conservative means the test for the factor (School) greatly lacks statistical power.

Table 48

Summary of average Type I Error rates for various distributions/datasets, conditional, alpha = 0.05

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.050039	*0.000357*	0.050397
Chi-square (df=3)	0.050073	*0.000472*	0.050545
Exponential	0.050012	*0.000489*	0.050500
t (df=3)	0.045460	*0.000304*	0.045763
Uniform	0.051215	*0.000563*	0.051777
Digit preference	0.050246	*0.000425*	0.050671
Extreme asymmetric	0.052485	*0.000770*	0.053256
Multi-modal lumpy	0.052758	*0.000609*	0.053367
Smooth symmetric	0.050241	*0.000411*	0.050652

Note: Values in italics are nonrobust according to Bradley's (1978) liberal definition.

Table 49

Summary of average Type I Error rates for various distributions/datasets, conditional, alpha = 0.01

Distribution/Dataset	Nest (Teacher)	Factor (School)	Experiment-wise
Normal	0.010042	*0.000020*	0.010062
Chi-square (df=3)	0.010618	*0.000014*	0.010632
Exponential	0.011089	*0.000012*	0.011101
t (df=3)	0.008624	*0.000000*	0.008624
Uniform	0.010595	*0.000016*	0.010612
Digit preference	0.010117	*0.000000*	0.010117
Extreme asymmetric	0.012795	*0.000050*	0.012845
Multi-modal lumpy	0.011357	*0.000000*	0.011357
Smooth symmetric	0.010106	*0.000000*	0.010106

Note: Values in italics are non-robust according to Bradley's (1978) liberal definition.

Statistical Power Projections

As previously noted, conducting the test of the Factor (i.e., School) conditionally will create a lack of statistical power due to the ultra-conservative nature of being the second in sequence in a series of two tests. Although it is beyond the scope of the current study to conduct a full-scale power spectrum analysis, in an attempt to explain the impact on statistical power, a treatment alternative of shift in location parameter was introduced.

The study parameters for this brief power study included setting nominal alpha to 0.05. Data were sampled from the Gaussian distribution, the sample size was set at n = 2, and both unconditional and conditional testing were conducted. The treatment was modeled by the addition of a constant equal to 0.5σ, where σ = 1 when the referent distribution is normal to create an effect size of Cohen's d = 0.5. The magnitude of this effect size is considered moderate (Cohen, 1988).

The treatment conditions were set in two studies as follows. For Study 1, an effect size of 0.5 was added to a single teacher per school. This created a difference among the twelve teachers, while leaving the schools equal. For Study 2, all teachers in a single school were simulated to receive the treatment, creating a difference between both the teachers and the schools. Due to the layout of nested designs, in this case with teachers contained within the school where they work, it is impossible to simulate a change between schools only. The results are compiled in Table 50.

Table 50

Statistical power projection, normal distribution, alpha = 0.05, n = 2

				Power			
	Study parameters			Unconditional		Conditional	
Recipient	Alpha	ES Teacher	ES School	Teacher	School	Teacher	School
Teacher	0.05	0.5	0.0	0.194	0.054	0.194	0.011
Teacher and School	0.05	S1 = 0.5	S2-4 = 0.0	0.121	0.114	0.121	0.089

Notes: ES = effect size in standard deviations, S1 = School 1, S2-4 = Schools 2, 3 & 4.

As noted, with the given study parameters, the unconditional and conditional power for the test of the nest effect (Teacher) was 0.194. In the unconditional layout, the expected Type I error rate of approximately 0.05 was obtained, however, in the conditional, the Type I error rate was ultra-conservative at 0.011. The loss in power becomes apparent in Study 2. Although the power was approximately the same for the

treatment effect (0.121 and 0.114, respectively) for the conditional layout, the power obtained for the effect (school) was reduced to from 0.141 to 0.089, which is a severe loss in power of approximately 22%.

Conclusion

Prior to drawing a conclusion in resolving the issue of the impact of nesting on the inflation of experiment-wise Type I error rates, it should be mentioned that there are potentially other statistical techniques that could have been incorporated, such as the nonparametric Kruskal-Wallis and the rank transform tests. Neither test is a solution for the inflation of experiment-wise Type I errors, but it is not known if either would help recover some of the lost power. However, because neither the Kruskal-Wallis nor the rank transform tests have been developed specifically for nested layouts, they were not incorporated in the study.

As Kromrey and Dickenson (1995) showed, the testing of multiple effects in a layout can be safely carried out via invoking a Bonferroni-Dunn or similar technique. However, as it stands, the statistical power available to the testing of the treatment effect conditional on a significant nested effect is already severely reduced due to the procedure being ultra-conservative. The use of Bonferroni-Dunn or related methods will only further reduce statistical power.

Heck, Thomas, and Tabata (2010) noted more sophisticated nested designs “are rapidly growing in their popularity and use” (p. 320), which will only exacerbate the issues outlined in this study. In conclusion, researchers should heavily weigh the trade-

offs of experiment-wise Type I error inflation for unconditional and statistical power loss for conditional nested designs.

REFERENCES

Algina, J., Blair, R. C., & Coombs, W. T. (1995). A maximum test for scale: Type I error rates and power. *Journal of Educational and Behavioral Statistics*, *20*(1), 27-39.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Braver, M. C., & Walton-Braver, S. L. (1988). Statistical Treatment of the Solomon Four-Group Design: A Meta-Analytic Approach. *Psychological Bulletin*, *104*(1), 150-154.

Braver, S. L., & Walton Braver, M. C. (1990). Meta-analysis for Solomon four-group designs reconsidered: A reply to Sawilowsky and Markman. *Perceptual and Motor Skills*, *71*, 321-322.

Cohen, J. (1988). *Power analysis for the behavioral sciences*, second edition. Mahwah, NJ: Erlbaum.

Dunn, O.J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52-64.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical distributions*, third edition. New York, NY: Wiley.

Fletcher, H. J., Daw, H., & Young, J. (1989). Controlling multiple F test errors with an overall F test. *Journal of Applied Behavioral Science*, *25*, 101-108.

Glass, G. V., Peckham, P. D., Sanders, and J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42(*3), 237-288.

Gonales, R. (2009). *Data analysis for experimental design*. New York, NY: Guilford Press.

Heck, R. H., Thomas, S. L, & Tabata, L. N. (2010). *Multilevel and longitudinal modeling with IBM SPSS*. New York, NY: Routledge/Tailor & Francis.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences*. New York, NY: Houghton Mifflin.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Kanji, G. K. (1999). *100 statistical tests*. London, UK: Sage.

Kennedy, J. J. & Bush, A. J. (1985). *An introduction to the design and analysis of experiments in behavioral research*. Lanham, MD: University Press of America.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences*. (4th Ed.). Thousand Oaks, CA: Sage.

Kreft, I. & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

Kromrey, J. D., & Dickenson, W. B. (1995). The use of an overall f test to control type I error rates in factorial analyses of variance: Limitations and better strategies analyses of variance: limitations and better strategies. *Journal of Applied Behavioral Science*, *31*, 51-64.

Maggio, S. (2012). The t-Wilcoxon maximum test: Critical values and illustration. Unpublished doctoral dissertation. Detroit, MI: Wayne State University.

Marascuilo, L. A. & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York, NY: W. H. Freedman and Company.

Mayo, D. G., & Cox, D. R. (2006). *Frequentist statistics as a theory of inductive inference: Lecture notes-monograph series*. The Second Erich L. Lehmann Symposium, *49*. Institute of Mathematical Statistics, http://www.jstor.org/stable/4356393, p. 77-97.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105* (1), 156-166.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263-1282.

Press, S. J. (2005). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. (2nd. Ed.) Mineola, NY: Dover.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*, p. 1350-1353.

SAS (1990). *SAS/STAT user's guide, Vol. 1*. (4th ed.) Cary, NC: SAS Institute.

Savage, L. J. (1962). *The foundations of statistical inference*. Methuen & Co. Ltd., London, UK.

Sawilowsky, S. S. (June 23, 1996). *Controlling experiment-wise Type I error in the Solomon four-group design*. Proceedings of the International Conference On Multiple Comparisons. Tel Aviv, Israel.

Sawilowsky, S. S. (2002). The probable difference between two means when $\sigma_1 \neq \sigma_2$. *Journal of Modern Applied Statistical Methods*, *1*(2), 461-472.

Sawilowsky, S. S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods*, *2*(2), 467-474.

Sawilowsky, S. S. (2007a). ANOVA: Effect sizes, interaction vs. main effects, and a modified ANOVA table. In S. Sawilowsky (Ed.) *Real data analysis*. American Educational Research Association Educational Statisticians. Washington, DC: InfoAge Publishing.

Sawilowsky, S. S. (2007b). ANCOVA and quasi-experimental design: The Legacy of Campbell and Stanley. In S. Sawilowsky (Ed.) Real data analysis. American Educational Research Association Educational Statisticians. Washington, DC: InfoAge Publishing, p. 213-238.

Sawilowsky, S. S., & Fahoome, G. F. (2003). *Statistics via Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.

Sawilowsky, S. S., & Markman, B. (1988). *Another look at the power of meta-analysis in the Solomon four-group design*. ED 316 556.

Sawilowsky, S. S., & Markman, B. S. (1990a). Another look at the power of meta-analysis in the Solomon four-group design. Perceptual and Motor Skills, 71, 177-178.

Sawilowsky, S. S., & Markman, B. S. (1990b). Rejoinder to Braver and Walton Braver. *Perceptual and Motor Skills*, *71*, 424-426.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). REALPOPS.LIB: a PC Fortran library of eight real distributions in psychology and education. *Psychometrika*, *55*, 729.

Sawilowsky, S. S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. *Journal of Experimental Education*, *62*, 361-376.

SPSS (1993). *SPSS for Windows: Base system user's guide release 6.0*. Chicago: SPSS.

SPSS (2013). *IBM SPSS User's Manual, v. 21*. Reference #7024972, Modified date: 2012-08-14,

Student. (1908). The probable error of a mean. *Biometrika*, *6* (1), 1–25. doi:10.1093/biomet/6.1.1.

SYSTAT (1990). *SYSTAT: The system for statistics*: Evanston, IL: SYSTAT.

Walton-Braver, M. C., & Braver S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, *104*, 150-154.

Wilcox, R. R. (1996). *Statistics for the social sciences*. London, UK: Academic Press.

Winer, B. J. (1962). *Statistical principles in experimental design*. (2nd. Ed.) NY, NY: McGraw-Hill.

ABSTRACT

THE IMPACT OF NESTED TESTING ON EXPERIMENT-WISE TYPE I ERROR RATE

by

JACK SAWILOWSKY

May 2014

Advisor: Professor Barry S. Markman

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

Keywords: Experiment-wise Type I error inflation, Nested testing, Monte Carlo Simulation, Hierarchical linear modeling, Bonferroni-Dunn

When conducting a statistical test the initial risk that must be considered is a Type I error, also known as a false positive. The Type I error rate is set by nominal alpha, assuming all underlying conditions of the statistic are met. Experiment-wise Type I error inflation occurs when multiple tests are conducted overall for a single experiment. There is a growing trend in the social and behavioral sciences utilizing nested designs. A Monte Carlo study was conducted using a two layer design. Five theoretical distributions and four real datasets taken from Micceri (1989) were used, each with five different samples sizes and conducted with nominal alpha set to 0.05 and 0.01. These were conducted both unconditionally and conditionally. All permutations were executed for 1,000,000 repetitions. It was found that when conducted unconditionally, the experiment-wise Type I error rate increases from alpha = 0.05 to 0.10 and 0.01 increases to 0.02. Conditionally, it is extremely unlikely to ever find results for the factor, as it requires a statistically significant nest as a precursor, which leads to extremely reduced power. Hence, caution should be used when interpreting nested designs.

AUTOBIOGRAPHICAL STATEMENT

Jack Sawilowsky was born May 5th, 1987, in Clearwater, FL. His family moved north to Metro Detroit, MI, when he was approximately three months old. He has lived in Michigan ever since.

Jack received his Bachelor's of Science degree in 2009 from Michigan State University, East Lansing, MI. He majored in Interdisciplinary Studies in Social Science – Human Resources, with a cognate in Psychology and certified specialization in Business. He took a vacation approximately three hours long before beginning his Masters degree the same day. He majored in Education Evaluation and Research – Quantitative track, at Wayne State University, Detroit, MI. He plans to immediately proceed through the doctoral program in EER on a similar timeline. His Masters cognate is in Political Science, while his doctoral cognate will be in Business.

Jack has taught graduate level courses in Evaluation and Measurement since 2010 at WSU. He has also tutored students in the field. For the past few years, he has been consulting as an evaluator for federally funded grants, including the National Science Foundation and Edge Foundation. His work encompasses research design, survey methodology, test development and psychometrics, and data analysis. He published an entry in the *International Encyclopedia of Statistical Science*, which was nominated by the Republic of Serbia for the 2011 Nobel Peace Prize.

Jack has considerable international work experience. This includes an internship at the International Institute for Counter-Terrorism, Herzliya, Israel, and a Sponsorship Assistant position at the Institution of Engineering and Technology in London, UK, and the Michael Faraday House in Stevenage, UK.