Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 4

5-1-2017

Experiment-wise Type I Error Rates in Nested (Hierarchical) Study Designs

Jack Sawilowsky
Citigroup, jsitm585@gmail.com

Barry Markman
Wayne State University, barry.markman@wayne.edu

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 52-68.
doi: 10.22237/jmasm/1493596980

ISSN 1538 − 9472

Experiment-wise Type I Error Rates in Nested (Hierarchical) Study Designs

Jack Sawilowsky
Citigroup
Tampa, FL

Barry Markman
Wayne State University
Detroit, MI

When conducting a statistical test one of the initial risks 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 sample sizes and conducted with nominal alpha set to 0.05 and 0.01. These were conducted both unconditionally and conditionally. All permutations were performed 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.

Keywords: Experiment-wise Type I error inflation, nested testing, Monte Carlo simulation, hierarchical linear modeling, Bonferroni-Dunn

Type I Error

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

Jack Sawilowsky, Ph.D,. is a Vice President – Senior Data Scientist. Email him at: jsitm585@gmail.com. Barry Markman is a Professor of Educational Psychology. Email him at: barry.markman@wayne.edu.

The 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 et al., 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.

Type I error inflation can arise in many statistical procedures. In some circumstances, such as the one-way independent samples ANOVA layout, there is a storied history of the development of a priori and post-hoc corrections to the *F* test to ameliorate this problem. Unfortunately, the experiment-wise inflation problem does surface in certain seemingly innocuous layouts, and results are often presented without recognizing the need for adjustment.

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 and self-worth), conducting step-down tests and the concern with experiment-wise Type I error inflation vanishes.

There are, however, 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 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.

Sequential (or Serial) Tests

Sequential tests occur in separate phases. For example, there is the recommendation to test for underlying assumptions (e.g., homoscedasticity via Levine's test and normality via Kolmogorov-Smirnov's test), and only after failing to reject both proceeding to conduct a statistical test of effects (such as the *t*-test). This strategy was recommended in many statistical packages (e.g., Statistical Analysis Systems Institute, Inc., 1990, p. 25; Norušis, 1993, pp. 254-255; Wilkinson, 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, and using a nonparametric alternative in its place.

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. However, 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. (pp. 51-52)

They 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, and 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 rose 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.

Post-Hoc Tests: A Resolution to the Type I Error 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 $\alpha b = \alpha/C$ level of significance, where *C* is the total number of comparisons you plan to perform. (pp. 279-280)

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 alpha. 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 et al., 2003, p. 300). All other multiple comparison procedures are a compromise between the Bonferroni and making no adjustments to control Type I error inflations.

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)

Winer (1971) explained, "Effects which are restricted to a single level of a factor are said to be nested within that factor" (p. 360). Winer emphasized the substantial limitation of nested designs in that they do not permit the testing of an interaction effect.

As an example of a nested design, consider a teacher within school layout. Kanji (1999) decomposed the three components (A School factor, B Teacher factor, Residual) nested sums of squares as

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

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

$$S_{\text{E}}^2 = \sum_i \sum_j \sum_k \left(Y_{ijk} - Y_{ij0}\right)^2$$

Table 1. 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
$\bar{X}_T$	37.80	35.00	32.17	44.83	40.67	40.16	40.00	38.33	37.83	36.67	37.17	35.17
ST	630			754			679			654		
$\bar{X}_S$	35			41.89			38.72			36.33		

Note: TT = Teacher total, ST = School total, $\bar{X}_T$ = Teacher mean, $\bar{X}_S$ = School mean, Grand mean School total = 2,735

Table 2. 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		

where S is the School, T is the Teacher, and E is the residual, where H_A: $\alpha_i = 0$ for all i and H_B: $\beta_{ij} = 0$ for all i, j. The data for the example are compiled in Table 1, and the traditional ANOVA table is presented in Table 2.

Hierarchical Modeling

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).

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 following the results obtained in Table 2 above. 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.

Methodology

Design

A two-factor nested layout or hierarchical classification layout was used. This design assumed errors would be normally distributed, with the magnitudes of

those errors being independent from either of the two factors. Specifically, the hypothetical layout pertained 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 were simulated for three teachers (or classrooms) per each of four schools, as noted in the table below.

Nested designs 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 found in the literature, generally the ANOVA test is required.

Sampling Plan

A pseudo-random number generator was used to simulate student test scores. The data were generated through Roguewave's (2012) subroutine libraries for the theoretical distributions. Data were simulated to follow the Gaussian, uniform, exponential, *t* (df = 3), and Chi-squared (df = 2) distributions. Variates from the Gaussian (i.e., normal) distribution were 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 were also 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 and Blair (1992). The real data sets to be sampled were the smooth symmetric (achievement scores), digit preference (achievement scores), multi-modal lumpy (achievement scores), and extreme asymmetry (psychometric scores).

Sample sizes were set to n = 2, 10, 30, 45, and 120. Samples of size n = 2 and n = 120 were 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 were 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) have no impact on Type I errors and are therefore not modeled. One million repetitions were executed for each combination of study parameters.

Table 3. Expected Type I error rates for normal and selected non-normal data at $\alpha = 0.05$ and $\alpha = 0.01$

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: [1]Glass, Peckham, and Sanders (1972, p. 250); [2]Sawilowsky and Blair (1992, pp. 356-358); 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

Analysis

The appropriate analysis for the nested design in Table 1 above is a series of two *F* tests. Initially, the *F* test was 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 was 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 was false, it meant 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 was 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 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.

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

Error Isolation

The Monte Carlo simulation was conducted using parametric or normal theory tests. However, data were 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 3.

Results

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.

As predicted by theory (Marascuilo & Serlin, 1988), the results in Tables 4 and 5 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 4 and 5 contain a compilation of those results.

In Tables 6 and 7, 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 (1968) definition for two levels of robustness. The conservative definition is met when the Type I error rate is within the bounded interval $[0.5\alpha, 1.5\alpha]$ inclusive, and the liberal definition is met when the Type I error rate is within the bounded interval $[0.9\alpha, 1.1\alpha]$ 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 0.005 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 4. 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 5. 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

Table 6. 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 (1968) liberal definition

Table 7. 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 nonrobust according to Bradley's (1968) 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 $\alpha = 0.05$. 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 $\sigma = 1$ 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 8.

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%.

Restating and expanding on Kreft and De Leeuw (1998):

> 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 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 overlooked 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 earlier in this article, this provides no safe haven from experiment-wise Type I error inflation.

Table 8. Statistical power projections, normal distribution, $\alpha = 0.05$, $n = 2$

				Power			
	Study Parameters			Unconditional		Conditional	
Recipeint	a	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

Note: ES = effect size in standard deviations, S1 = School 1, S2-4 = Schools 2, 3, and 4

The second argument advanced by Gonzales (2009) 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" (pp. 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 for four schools. The tests for the nest and effect were conducted unconditionally and conditionally.

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. When the same issue arose in analyzing the Solomon four-group design (Sawilowsky & Markman, 1990a, b; Sawilowsky, Kelley, Blair, & Markman, 1994), a solution based on an asymmetric Bonferroni-Dunn (i.e., disproportionate allocation of nominal alpha to constituent tests) was proposed by Sawilowsky (1996).

Nevertheless, Heck et al. (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. Hence, researchers should heavily weigh the trade-offs of experiment-wise Type I error inflation for unconditional and statistical power loss for conditional nested designs before utilizing them.

References

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). *Power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.2307/1169991

Gonzales, 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/Taylor & Francis. doi: 10.4324/9780203855263

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

Kanji, G. K. (1999). *100 statistical tests*. London, UK: Sage. doi: 10.4135/9781849208499

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.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage. doi: 10.4135/9781849209366

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*(1), 51-64. doi: 10.1177/0021886395311006

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Norušis, M. J. (1993). *SPSS for Windows: Base system user's guide, release 6.0*. Chicago, IL: SPSS Inc.

Sawilowsky, S. S. (1996, June 23). *Controlling experiment-wise Type I error in the Solomon four-group design*. Presented at the 1st 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. doi: 10.22237/jmasm/1036109940

Sawilowsky, S. S. (2007a). ANOVA: Effect sizes, interaction vs. main effects, and a modified ANOVA table. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 191-212). Washington, DC: InfoAge Publishing.

Sawilowsky, S. S. (2007b). ANCOVA and quasi-experimental design: The legacy of Campbell and Stanley. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 213-238). Washington, DC: InfoAge Publishing.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360. doi: 10.1037/0033-2909.111.2.352

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*(4), 729.

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

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*(4), 361-376. doi: 10.1080/00220973.1994.9944140

Sawilowsky, S. S., & Markman, B. (1988). Another look at the power of meta-analysis in the Solomon four-group design. Retrieved from ERIC database. (ED316556)

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*(1), 177-178. doi: 10.2466/pms.1990.71.1.177

Sawilowsky, S. S., & Markman, B. S. (1990b). Rejoinder to Braver and Walton Braver. *Perceptual and Motor Skills, 71*(2), 424-426. doi: 10.2466/pms.1990.71.2.424

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

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

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

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York, NY: McGraw-Hill.