Journal of Methods and Measurement in the Social Sciences,
Vol. 8, No. 2, 58-71, 2017

Chi-square and *F* Ratio: Which should be used when?

Richard L. Gorsuch
UniMult inc.

Curtis Lehmann
Azusa Pacific University

Approximations for Chi-square and *F* distributions can both be computed to provide a *p*-value, or probability of Type I error, to evaluate statistical significance. Although Chi-square has been used traditionally for tests of count data and nominal or categorical criterion variables (such as contingency tables) and *F* ratios for tests of non-nominal or continuous criterion variables (such as regression and analysis of variance), we demonstrate that either statistic can be applied in both situations. We used data simulation studies to examine when one statistic may be more accurate than the other for estimating Type I error rates across different types of analysis (count data/contingencies, dichotomous, and non-nominal) and across sample sizes (*N*s) ranging from 20 to 160 (using 25,000 replications for simulating *p*-value derived from either Chi-squares or *F*-ratios). Our results showed that those derived from *F* ratios were generally closer to nominal Type I error rates than those derived from Chi-squares. The *p*-values derived from *F* ratios were more consistent for contingency table count data than those derived from Chi-squares. The smaller than 100 the *N* was, the more discrepant *p*-values derived from Chi-squares were from the nominal *p*-value. Only when the *N* was greater than 80 did the *p*-values from Chi-square tests become as accurate as those derived from *F* ratios in reproducing the nominal *p*-values. Thus, there was no evidence of any need for special treatment of dichotomous dependent variables. The most accurate and/or consistent *p*'s were derived from *F* ratios. We conclude that Chi-square should be replaced generally with the *F* ratio as the statistic of choice and that the Chi-square test should only be taught as history.

Key words: Chi-square, *F* ratio, teaching statistics, simulation study

For general linear model (GLM) frequentist statistical procedures, there is a choice between a Chi-square and an *F* ratio for testing significance. But when it is best to use one or the other is relatively unaddressed in the literature. The fact that any *p* is only an estimate is well illustrated by Boos and Stefanski (2011). Based on their analysis of the precision of *p*s and the implications for decisions about reproducibility, they suggest that the accuracy is such that an asterisk system based on the number of leading zeros in the *p* is all that is warranted. However, they did not take into account that the *p* may be computed from a Chi-square or from an *F* ratio, both of which only give an approximation to the correct *p*. Do the conclusions apply to both these approximations? As their conclusions do not take into account the approximation process, the conclusions reached by Boos and Stefanksi (2011) may be optimistic.

There has been an occasional study evaluating *p* estimates. For example, Larntz (1978) evaluated Chi-square and found it to give more accurate contingency Type I error rates than maximum likelihood or Friedman-Tukey statistic. Richardson (1990) ran simulations for the .05 Type I probability, comparing variations on Chi-square such as Upton's Chi-square and Yate's correction. His results showed that the usual Chi-square *without* corrections was appropriate but observed .05 Type I error rates between .04 and .06. However, to our knowledge no prior study has tested both Chi-square and *F* Ratio against each other. The purpose of this paper is to compare *p*'s computed by Chi-square and *F* using Monte Carlo simulations for standard general linear model (GLM) applications, such as phi and eta coefficients, multiple correlation, and ANOVA.

The *F* ratio is calculated as

$$F = ((r^2/df1)/((1-r^2)/df2), \quad (1)$$

where *F* is Fisher's *F* ratio and *r* is the Pearson product moment correlation coefficient, the phi /eta coefficient, or the multiple correlation coefficient (Guilford & Frutcher, 1978) . Degrees of freedom for correlations (including phi and multiple correlation) are:

$$df\,1 = df\text{x} * df\text{y} \quad (2)$$

$$df\,2 = s(N - df\text{x} - 1 - df\text{y} + s) \quad (3)$$

where *df*x is the number of degrees of freedom for the Xs (predictors and independent variables), that is, the number of non-nominal Xs and the number of categories minus 1 for nominal predictors, and *df*y is the number of degrees of freedom for the Y variable(s). *N* is the number of cases and *s* is the minimum of *df*x and *df*y.

Perhaps not as well-known but equally well established is the computation of Chi-square using a similar general formula. Wherry (1984) notes the Chi-square statistic is calculated as:

$$\text{Chi-square} = r^2 * N \quad (4)$$

where *r* is the phi coefficient for a two way contingency table, the Pearson correlation for bivariate analyses, the multiple correlation for multivariable/multicategory independent variables, and eta for nominal dependent variables, thus including multiple regression analysis and ANVOA as well as contingency tables. The degrees of freedom are:

$$df = df\text{x} * df\,\text{y} \quad (5)$$

where *df*x is the number of categories of the X minus 1 for each nominal variable and the number of non-nominal Xs, and *df*y is the same for the Ys.

This Chi-square has been most widely presented as the test of the phi coefficient, computed from a pair of dichotomous variables (e.g., Guilford & Frutcher, 1978). With an appropriate change in degrees of freedom, it can be used with an eta for a multiple category nominal dependent variable as well as ANOVA and regression.

Both the *F* and Chi-square formulae can both be applied to any data set. Thus Knapp (1978) noted that Chi-square can be replaced by *F* tests in analyzing, for example, contingency tables. Both formulae can also be used with multivariate analyses with Pillai's lambda (the multivariate generalization of *r*) (Haase, 2011). Given that *F* and Chi-square can both be computed from the same effect size, it is perhaps not surprising that Glass et al. (1972) note that both have the same statistical assumptions.

For both the *F* ratio and Chi-square, other formulae were developed to ease hand calculations, such as those calculated directly from the cells of the contingency (cross tab) table or those to compute the Chi-square from the roots of a matrix in canonical analysis. Note that with the GLM it makes no difference whether the conditions are labeled contingencies, ANOVA, or regression. Thus the same hypotheses and conclusions apply regardless of the traditional jargon used.

Both *F* ratios and Chi-squares are compared to the *F* and Chi-square distributions, respectively, for an estimate of *p*, the probability of a Type I error. For example, consider a 2 X 2 contingency table with 1 *df* and equal 50/50 splits on both variables. Table 1 contains the Chi-square, *F*, and their *p*'s for 3 effect sizes:

Table 1
Sample Significance Tests Using Chi-square and F Ratio for Selected Effect Sizes

		Chi-square		*F* ratio
r	Chi-square	*p*-value	*F* ratio	*p*-value
0.1	1	0.68	0.99	0.68
0.2	4	0.046	4.08	0.038
0.3	9	0.0026	9.69	0.0034

Note. $N = 100$, $df = 1$ for Chi-square; $df = 1/98$ for F; r is the correlation/phi coefficient.

The probabilities of Table 1 are similar but not identical. For these small correlations they are within rounding error when rounded to the first non-zero digit.

While the analyses and results below obviously apply to null hypothesis significance testing (NHST), all are based in the central Chi-square or *F* distributions (with *t* distributions being a special case of the latter). Therefore, the conclusion above that all analyses could use either the Chi-square or *F* distribution applies equally to CIs (with the non-central distribution where appropriate).

Criteria. The question of whether a Chi-square or an *F* ratio gives a more accurate Type I error can be evaluated by two criteria. First, which one averages closest to the actual Type I error rate? The more accurate *p*'s would be those closer to, that is, less discrepant from, the nominal Type I error rate.

However, accuracy by itself is insufficient. It may be that one is more consistent than the other. If the one with the mean closest to the actual rate is more inconsistent, it would produce more errors than the one with a slightly less discrepant mean but which was more consistent.

Consistency can be measured as the squared discrepancy of the observed *p's* from the alpha. Hence both the mean discrepancy and the squared discrepancy of the approximate *p*'s are needed to judge the performance of the approximate *p*'s from Chi-squares and *F* ratios.

Dichotomous Variables. As nominal categories are represented by dichotomous variables, there may be cause for concern because of statements in many recent texts suggesting that log linear analysis be used with a dichotomous dependent variable. While those critiques are primarily based on regression analysis producing estimates outside the range of 0.0 to 1.0, it seems worthwhile to compare the effectiveness of both Chi-square and *F* ratios for dichotomous variables.

Count data. In addition to the questions of general adequacy and adequacy for dichotomous variables, another question can be asked based on the traditional applications of Chi-square and *F*. Chi-square is most closely linked to contingency table analysis – often referred to as count data -- whereas *F* ratios are most closely linked to continuous data such as regression or continuous Ys such as ANOVA. While, as the formula above suggest, either is equally appropriate for both types of data and, as Glass et al. (1972) note, the assumptions are the same, given the historical association one might predict that Chi-square would function better for count data than *F* ratios.

N. A final question can be addressed by these simulations: the degree to which the accuracy and consistency are affected by the *N*. There is one apparent difference between Chi-square and *F*. Chi-square uses only *df1*, the degrees of freedom related to the number of degrees of freedom for the variables, whereas *F* ratio also uses *df*2, which is based on the total sample size. However, Chi-square also accounts for the *N*, but it does so by incorporating it into the Chi-square formula rather than having a *df2*. If

one statistic accounts for the *N* in a more appropriate way, there should be a variation in the estimated *p*'s which varies as a function of the *N*.

As statistical formulas such as the standard error generally use the square root of *N*, we would expect that to be the best basis for checking for the adequacy of Chi-square and *F* across *N*. The relationship could be linear or curvilinear.

METHOD

Simulations were run with a variety of variables and analyses across selected sample sizes. The data were created using the common pseudo-random number generator found in programs such as Excel. The conditions included were: contingency tables (2x2 with .5:.5 probabilities, 2x3 with .5:.33 probabilities, 2x5 with .5:.20 probabilities (without the *N*s that could not give equal *n*s across the cells)), an equal *n* two-level ANOVA with a normally distributed continuous Y (dependent or outcome) and with a dichotomous Y, and a multiple regression with both a continuous and a dichotomous Y.

Each condition was run multiple times, once for each *N*. The following sample sizes were tested: 20, 24, 28, 32, 36, 40, 44, 48, 56, 64, 72, 80, 100, and 160 as these could give the desired splits for the contingency tables. The variables in the data sets were selected for different analyses as noted below. For each condition and sample, 25,000 replications were run.

For each replication, an effect size was calculated. The Type 1 *p*s of these effect sizes were then estimated using the Chi-square and *F* formulas given above. The mean observed probabilities were compared against the expected outcomes at designed nominal alpha levels of .05, .01, and .005 by subtracting the alpha from the observed *p* to obtain the observed discrepancy. The data were then analyzed for the two criteria: the observed discrepancy of *p* and the squared discrepancy from the nominal alpha.

The analyses addressed the questions of general accuracy, dichotomous Ys, count/ contingency table data, overall tests (that is, the total overall effect of main effects and the interaction of the ANOVA and the multiple correlation of the multiple regression) and the effect of *N*. As statistical formulae use the square root of *N*, that was tested for both linear and curvilinear relationships with the discrepancies.

Given that a number of analyses were computed, multivariate omnibus and family-wide tests were computed when appropriate; if those were significant, then protected post hoc tests were computed for separate effects. The analyses below suggested *F* is better than Chi-square and so the former is used for *p*'s computed by the analyses.

ANALYSES AND RESULTS

The following analyses are for main effects using significance tests for those cases, with the *p* set at .01. However, multivariate omnibus and family-wide tests were also computed, were all significant at $p < .0001$ except as noted below, and supported the results in the tables presented. Interactions were also tested and, when significant, were generally an order of magnitude smaller than the results below and were consistent with the presented results. Each datum was based on a simulation sample of 25,000 which appears to have led to the identification of differences sufficiently small so they would not affect the conclusions from the usual uses of *p*s. Hence, the main effects are presented here.

Overall Accuracy. An overall examination of the accuracy of Chi-square and *F* distributions was computed for an initial comparison across all conditions. Table 2A gives the average discrepancy from *p* for Chi-square and for *F* for nominal Type 1 error rates of .05, .01, and .005 (with the SDs). The differences between the discrepancies, calculated as the Chi-square discrepancy minus the *F* discrepancy, are also given. These were tested against the expected discrepancy values of 0 using Hotelling's T^2. The Chi-square and *F* squared discrepancies are presented in Table 2B. The table also contains tests of the differences between Chi-square and *F*, computed by calculating the discrepancy (or squared discrepancy) for Chi-square minus the discrepancy (or squared discrepancy) for *F*.

Table 2
Overall Accuracy and Consistency: Chi-square and F Ratio Discrepancies from Nominal Alphas

	Mean *(SD)*					
alpha	α =.05		α =.01		α =.001	
A. Raw discrepancy						
Chi-square	.0011***	(-.0021)	-.0009***	(-.001)	-.0008***	(-.0007)
F ratio	.0	(-.002)	.0002*	(-.0008)	.0002**	(-.0006)
Difference	.0011***	(-.0017)	-.0011***	(-.001)	-.0010***	(-.0008)
B. Squared discrepancy						
Chi-square	0.0056	(-.0102)	.0018	(-.0023)	.0012	(-.0015)
F ratio	.0039	(-.0089)	.0007	(-.0011)	.0004	(-.0007)
Difference	.0018**	(-.0061)	.0011***	(-.0023)	.0009***	(-.0015)

Note. Significance was tested with Hotelling T2 tests against an expected discrepancy of 0.0 with N = 165. In B. tests were only computed on the last row as tests of the first two rows are not meaningful.
* $p < .01$, ** $p < .001$, *** $p < .0001$.

The first conclusion is that both gave, as indicated by the more careful presentations in the literature, approximations to their distributions,

being accurate to the first non-zero digit for Chi-square or to three decimal places for *F* but no more.

Chi-square produced significantly higher *p*s at the .05 level but produced significantly lower *p*s at the .01 and .001 levels. The general accuracy of mean results suggest *F* is in general the more accurate (the first criterion).

Consistency was computed for all the samples as the squared discrepancy from the nominal alpha level. The results in Table 2B include the difference between the two tests. Significance tests are only given for the last row of B. as the tests of the first two rows are just tests of whether there is variation which, of course, there is; presenting these tests would imply that they could be interpreted for the purposes of this paper but only the tests of the last row are relevant (the tests were run; of course, they were all significant at *p* less than .0001). For all three alpha levels, *F* had lower variability (squared discrepancy) than Chi-square and so better meets the second criterion of being consistent. The results were such that both *F* and Chi-square are accurate within rounding to the first non-zero digit.

Dichotomous variables. There were 43 results for dichotomous variables which may be considered Ys (dependent variables). Table 3 contains the Chi-square and *F* mean discrepancies and squared discrepancies for the three alpha levels for dichotomous variables to evaluate whether Chi-square or *F* may be more appropriate.

Table 3
Dichotomous Data: Chi-square and F Ratio Discrepancies from Nominal Alpha

	Mean *(SD)*		
alpha	α =.05	α =.01	α =.001
A. Raw discrepancy			
Chi-square	.0011*** (.0021)	-.0009*** (.0010)	-.0008*** (.0007)
F ratio	.0000 (.0020)	.0002* (.0008)	.0002** (.0006)
Difference	.0011*** (.0017)	-.0011*** (.0010)	-.0010*** (.0008)
B. Squared discrepancy			
Chi-square	.0056 (.0102)	.0018 (.0023)	.0012 (.0015)
F ratio	.0039 (.0089)	.0007 (.0011)	.0004 (.0007)
Difference	.0018** (.0061)	.0011*** (.0023)	.0009*** (.0015)

Note. Significance was tested with Hotelling T2 tests against an expected discrepancy of 0.0 with N = 165. In B. tests were only computed on the last row as tests of the first two rows are not meaningful.
* $p < .01$, ** $p < .001$, *** $p < .0001$.

In Table 3, the *F* ratio discrepancies are closer to 0 than those for Chi-square and have lower standard deviations and squared

discrepancies. Thus the criteria of closer mean accuracy and greater consistency of the approximate *p*'s are better met for dichotomous conditions by *F* ratio than by Chi-square. These results are similar to the results in Table 2. A comparison of the two tables suggests that there is no need of special tests just because the Y is dichotomous. (Significance tests between dichotomous and non-dichotomous conditions found the discrepancies and squared discrepancies to be insignificantly different.)

Count data. Is Chi-square more appropriate for count data (contingency tables) than *F*? While the test for accuracy and consistency for dichotomies is a partial subset set of this question, the contingency tables differ in having both independent and dependent variables which are count data. Table 4 contains the results for the contingency tables discrepancy and squared discrepancy means and standard deviations using Chi-square and *F* for each of the alpha levels. (Results with Yate's correction are not reported because it produced poorer results than the uncorrected Chi-squares.)

Table 4
Count Data: Chi-square and F Ratio Discrepancies from Nominal Alphas

	Mean *(SD)*		
alpha	α =.05	α =.01	α =.001
A. Raw discrepancy			
Chi-square	-.0026*** (-.0021)	-.0030*** (-.0018)	-.0020*** (-.0012)
F ratio	.0000 (-.0013)	.0002 (-.0007)	.0001 (-.0005)
Difference	-.0026*** (-.0018)	-.0031*** (-.0017)	-.0021*** (-.0011)
B. Squared discrepancy			
Chi-square	.011 (-.0157)	.012 (-.0139)	.0052 (-.0054)
F ratio	.0016 (-.0015)	.0005 (-.0006)	.0002 (-.0003)
Difference	.0094** (-.0152)	.0115*** (-.0138)	.0050*** (-.0054)

Note. Significance was tested with Hotelling's T2 tests against the expected alpha level with N = 165. * $p < .01$, ** $p < .001$, *** $p < .0001$.

In B, the overall test of the differences between squared discrepancies from Chi-square and *F* was not significant, Hotelling's $T^2 = 3.71$, $F(3, 36) = 1.17$, $p = .3$, so individual significance tests were not performed. However a simple binominal test on the Chi-square and *F* differed at the .06 level. In B. tests were only computed on the last row as the first two rows are not meaningful.

As predicted by the history of usage, Table 4A shows the mean discrepancies to be lower for Chi-square than *F* but the differences are significant primarily because they bracketed the nominal alpha. If this were the only criterion examined, Chi-square would be favored for count data. But note that the standard deviations were higher, suggesting that

more extreme discrepancies from the nominal Type I error can occur with Chi-square.

Table 4B shows the Chi-square squared discrepancies to be higher than those for *F*. *F* gives lower squared discrepancy with a lower standard deviation than does Chi-square. While Chi-square's raw discrepancy indicates adequacy, this analysis found Chi-square *p*s may have more extreme "misses" at each level of the nominal alpha than did *p*s from *F*. That would generally be considered evidence of a greater number of misinterpretations if Chi-square is used instead of *F*. The SDs and squared discrepancies are all lower for *F*, a result almost significant (p = .06) by a simple binominal test. Hence the results suggest *F* to be more consistent than Chi-square for count data.

When rounded to the first non-zero digit, both Chi-square and *F* results would give values within round error of the nominal alpha level.

Omnibus Tests. While the above apply to main effects for ANOVA and individual regression, do they apply when the tests are omnibus or overall tests which include all predictors for ANOVA and for multiple regression? Across all conditions which had multiple IVs, the *p* discrepancies from the Chi-squares and *F*'s were compared. Table 5A contains the mean discrepancies (and SDs) and Table 5B contains the squared discrepancies (and SDs).

Table 5
Omnibus Variables: Chi-square and F Ratio Discrepancies from Nominal Alpha

	Mean *(SD)*		
alpha	α =.05	α =.01	α =.001
A. Raw discrepancy			
Chi-square	-.0026*** (-.0021)	-.0030*** (-.0018)	-.0020*** (-.0012)
F ratio	.0000 (-.0013)	.0002 (-.0007)	.0001 (-.0005)
Difference	-.0026*** (-.0018)	-.0031*** (-.0017)	-.0021*** (-.0011)
B. Squared discrepancy			
Chi-square	.011 (-.0157)	.012 (-.0139)	.0052 (-.0054)
F ratio	.0016 (-.0015)	.0005 (-.0006)	.0002 (-.0003)
Difference	.0094** (-.0152)	.0115*** (-.0138)	.0050*** (-.0054)

Note. Significance was testing with Hotelling T^2 tests against an expected discrepancy of 0.00 with N = 165. In B., tests were performed only of the differences between Chi-square and *F*; the squared discrepancies were re-scaled for ease of presentation. * $p < .01$; ** $p < .001$; *** $p < .0001$.

Table 5 shows that *F* gives results closer to the nominal Type I error rate and to have lower squared discrepancies than Chi-square. This is the same conclusion as for the individual tests given in Tables 2 and 3. (No significance tests comparing omnibus to single tests was computed as the

former are a function of the latter, thus violating the independence of observation assumption of such tests.)

N. To explore the relationship of *N* to *F* and Chi-square *p*s, analyses were computed with the discrepancies and squared discrepancies as the Ys and the square root of *N* as the X (square root was used as this is the form of *N* that appears in formulas for estimating *p*). Linear, quadratic, and cubic tests were computed to check for curvilinearity. The cubic tests were all insignificant ($p > .01$) and are not reported.

Table 6 contains the results for each analysis using *N* in linear and quadratic forms to the *p* discrepancies. Table 6A contains the results for discrepancies and 6B for the squared discrepancies. The Chi-square results showed it to be clearly influenced by *N* whether the criterion was accuracy (discrepancy) or consistency (square discrepancy). The plotted curves consistently showed accuracy and consistency to increase with *N* until the *N* was 60 to 80, after which they leveled off. The approximate discrepancies for *N* of 25 were .0025, -.0017, and - .0016 for the .05, .01, and .001 nominal alphas. For *N* of 100, they were .0005, -.0004, and -.0004. These latter values are accurate to three decimal places.

Table 6
Discrepancies as a Function of N

Nominal alpha	α =.05			α =.01			α =.001		
	R	Line	Quad	*R*	Line	Quad	*R*	Line	Quad
A. Raw discrepancy									
Chi-square	.32[c]	-.28[c]	.16	.47[c]	.44[c]	-.16	.60[c]	.54[c]	-.25[c]
F	NOT SIGNIFICANT (lambda = .03, $p = .64$)								
B. Squared discrepancy									
Chi-square	.24[a]	-.24[b]	.06	.53[c]	-.44[c]	.28[c]	.64[c]	-.54[c]	.35[c]
F	NOT SIGNIFICANT (lambda = .04, $p = .38$)								

Note. $df1 = 1$ but 2 for *R*, $df2 = 160$. Significance: A < .01, b < .001, c < . 0001. *R* is the multiple correlation; Line gives the linear *r*, and Quad gives quadratic *r*.

The discrepancies and squared discrepancies for the *p*s from *F*s were unrelated to *N*. They were equally accurate for *N*s from 20 to 160. In this regard, *F* is more consistent than Chi-square. As Table 2 shows, *F*'s discrepancies are as low at 20 as Chi-square's are at 100. Hence, in regard to *N*, the conclusion is that *F* is as or more accurate than Chi-square at all levels of *N*.

As with the prior analyses, these analyses found only small differences between the *p*s from Chi-square and from *F*.

DISCUSSION

The first conclusion we draw from the analyses is that both Chi-square and *F* are reasonably accurate given their common usage for *p's* from .05 to .005 when the null hypothesis is true. If this is found to be consistent with other relevant analyses, the common practice of reporting more than the first non-zero digit needs to be up dated. This is a more optimistic conclusion than Boos and Stefanski (2011) as it suggests the first non-zero digit may be accurate in addition to the number of zeros before that digit (although the issue of reproducibility which influenced Boos and Stefanski conclusions is not directly addressed by our data). The differences between Chi-square and *F p*s were within rounding error for the first non-zero digit and hence both can be considered equally accurate for their common usage.

The phrase "within rounding error" suggests a .05 result may be reported as .04 to .06, and a .005 reported as .004 to .006. However, an alternative explanation consistent with these data may be that these *p*s are accurate to three decimal places. The latter is suggested by the fact that while the discrepancies are the same order of magnitude across the three nominal alpha levels in Tables 2 to 6, the squared discrepancies -- with the exception of Table 3 -- decreased one order of magnitude between the .05 and the .005 levels, suggesting the rounding error interpretation.

The reason for limitations in the accuracy may, in addition to *p*'s intrinsic limitations noted by Boos and Stefano (2011), lay in the fact that Chi-square and *F* are approximations or l in the fact that computational procedures are limited in accuracy due to internal rounding or truncation. It appears that it is more likely to be the former as the improvement in computation since 1990 has not lead to an increase in accuracy as Gorsuch (1991) reported the same level of accuracy from spot-checks using early micro-computers.

The variations between Chi-square and *F* are small and probably negligible given the rule of accuracy to the first non-zero digit. The *p's* from either procedure are adequate for most work. This conclusion applies to count and dichotomous variables as well as non-count variables, and generally across *N*s from 20 to 160; when differences were found, they were small and specific to a limited set of conditions. The differences are clearly smaller than those induced by variations in reliability and validity of the measures (Gorsuch & Lehmann, 2010).

The results indicate that the *F* ratio provides an estimation of probability that is slightly better than Chi-square overall. The conclusion includes dichotomous data whether they be independent variables (Xs) or dependent variable (Ys). It also applies to multiple *df*1 uses such as omnibus tests whether an overall test for ANOVA or the test for a multiple

correlation. The only place that Chi-square discrepancies were slightly less than those for *F*, the SDs and squared discrepancies were greater than for *F* suggesting that Chi-square may produce more large, potentially misleading, differences from the nominal *p* than *F*. In addition, Chi-square was less consistent in the sense that is was less accurate with moderate to small *N*s. These results suggest that the *F* ratio may be the better procedure to use in all analyses.

Given the general accuracy of both, they could both be continued to be used. But why have both? Given our results and those of Boos and Stefanski (2011), the differences between Chi-square and *F* are negligible to slightly in favor of *F*. Continuing to use both Chi-square and *F* means both must be taught, a practice that violates the rule of elegance so prized in mathematics and contributes to "mathematistry" (Little, 2013). Teaching one instead of two ways to estimate *p* is in keeping with parsimony. And as a general procedure, *F* is easier to present, as the ratio of two estimates of variability is readily generalizable to all situations. Teaching only one method of computing *p* also means that some other important point can be made in the time that would otherwise be taken for teaching Chi-square (it would still need to be mentioned in a history section to enable reading of the literature and understanding results from computer programs that have not yet been brought up to date or are inferior in the computational procedures they use (Keeling & Pavur, 2011)).

Perhaps these results are not surprising considering one historical fact. Fisher was well acquainted with Chi-square before he introduced the *F* ratio.

The current study is limited to tests of the null hypothesis, the most common use of *p*s. There is also the need to evaluate the *p*s for data drawn from data sets containing moderate to strong effects where results from Chi-square and *F* may show greater differences than when there are no effects.

Multivariate usage of *F* and Chi-square is not examined in this paper. However, the same formulae apply with eta/lambda being the multivariate generalization of *r*. Hence, the current results are expected to generalize to multivariate tests.

While the *N*s of the current study ranged as low as 20, the results may differ with even smaller *N*s. Larntz (1978) evaluated Type I error rates for Chi-square small samples and found the accuracy to be, for the .05 level, from .035 to .049 and, for the .01 level, from .0036 to .0092. It appeared that larger Ns (e.g., 24 and 32) give more accurate *p*'s than smaller Ns (e.g., 8 and 12). While our smallest *N* was only 20, this study found Chi-square to give lower *p's* for our smaller *N*s which is consistent with Larntz. The results of Larntz and the present study are more optimistic than Boos and Stefanski (2011) who suggested that the accuracy of *p*s is only to the

order of magnitude (as in the asterisk system, * for .01, ** for .001, and *** for .0001).

It should be noted that these results are with standard approximation procedures for Chi-square and for *F*. Better procedures may well be developed and impact the differences between these two distributions (Keeling & Pavur, 2011). However, note that the tables show *p*s from both distributions are accurate enough for decision making in most situations. Other changes in a study, such as in measurement or sampling, can be expected to impact the effect sizes and accuracy of *ps* to a greater degree than just the second non-zero digit (Gorsuch & Lehmann, 2010). In a sense, the major conclusion from this study is that, for practical purposes, both Chi-square and *F* are equally accurate for all but the most exacting purposes. In that case we would still propose using only one of the two to simplify teaching, and the most accurate of the two and easiest to describe is *F*.

This study has demonstrated several key ideas that are not widely known. The first is that *F* ratios can be calculated not only with ANOVA and regression but also with contingency tables, and provides adequate probability values for them all. This is done by computing an effect size -- correlation/multiple correlation/phi coefficient/eta (the several labels identify special cases of the multivariate lambda) -- which is used to compute *F*. The findings clearly demonstrate that the Chi-square statistic is slightly inferior to the *F* ratio in instances where the *F* ratio is commonly utilized and probably less consistent for count data where Chi-square is commonly used. There is no need to teach Chi-square; instead introduce the effect size and use *F* to test even contingency tables. Doing so would reduce "mathematistry", which results from ignoring parsimony (Little, 2013).

Acknowledgements. Richard L. Gorsuch was Senior Professor at the School of Psychology at Fuller Theological Seminary and CEO of UniMult inc. until his passing in February 2016; Curtis Lehmann is Assistant Professor at Azusa Pacific University, and was a graduate student at the Fuller School of Psychology when this work was done (email: clehmann@apu.edu). An earlier draft of this paper with Steve Brown was presented at the 2014 annual meeting of the American Psychological Association, Washington, DC.

References

Boos, D. D., and Stefanski, L. A. (2011). *P*-value precision and reproducibility. *The American Statistician, 65*:4, 213-221, DOI: 10.1198/tas.2011.10129

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

Gorsuch, R. L. (1991). *UniMult Guide*. Altadena, CA: UniMult.

Gorsuch, R. L., and Lehmann, C. S. (2010). Correlation coefficients: Mean bias and confidence interval distortion. *Journal of Methods and Measurement in the Social Sciences, 1*, 52-65.

Guilford, J., and Frutcher, B. (1978). *Fundamental Statistics in Psychology and Education*. London: McGraw-Hill International.

Haase, R. (2011). *Multivariate General Linear Models*. Thousand Oaks, CA: Sage Publications.

Keeling, K. B., and Pavur, R. J. (2011). Statistical accuracy of spreadsheet software. *The American Statistician*, *65*:4, 265-273, DOI: 10.1198/tas.2011.09076

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410–416.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-square goodness-of-fit statistics. *Journal of the American Statistical Association, 73*, 253- 263.

Little, R. (2013). In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association, 108:502*, 359-369.

Richardson, J. T. E. (1990). Variants of chi-square for 2 x 2 contingency tables. *British Journal of Mathematical and Statistical Psychology*, *43*, 309-326.

Wherry, R. J. (1984). *Contributions to Correlational Analysis*. Albany, NY: State University of New York Press.