Journal of Educational Statistics
Winter 1992, Vol. 17, No. 4, pp. 315–339

Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases

Michael R. Harwell
Elaine N. Rubinstein
William S. Hayes
Corley C. Olds
University of Pittsburgh

Key words: *meta-analysis, Monte Carlo, ANOVA*

Meta-analytic methods were used to integrate the findings of a sample of Monte Carlo studies of the robustness of the F *test in the one- and two-factor fixed effects ANOVA models. Monte Carlo results for the Welch (1947) and Kruskal-Wallis (Kruskal & Wallis, 1952) tests were also analyzed. The meta-analytic results provided strong support for the robustness of the Type I error rate of the* F *test when certain assumptions were violated. The* F *test also showed excellent power properties. However, the Type I error rate of the* F *test was sensitive to unequal variances, even when sample sizes were equal. The error rate of the Welch test was insensitive to unequal variances when the population distribution was normal, but nonnormal distributions tended to inflate its error rate and to depress its power. Meta-analytic and exact statistical theory results were used to summarize the effects of assumption violations for the tests.*

An ongoing concern of quantitative methodologists is the validity of inferences from statistical tests performed on data that may violate underlying assumptions. Monte Carlo (MC) studies are used to identify tests that are insensitive to assumption violations (i.e., those Type I error rate and power properties that are not deleteriously affected).

The effective use of MC studies is hampered by the lack of an overarching theory to guide their interpretation and their impressionistic nature. Harwell (1992) suggested that these deficiencies could be addressed by applying meta-analytic methods as conceptualized by Glass (1976) to quantitatively integrate the results of MC studies for a statistical test. The resulting empirical network of studies provides a summary of the effects of assumption violations, which should assist educational and psychological researchers in

selecting the "best" test (e.g., one that minimizes Type I errors and maximizes power).

The meta-analytic framework suggested in Harwell (1992) was used to integrate the findings of MC studies of the F, Welch (W), and Kruskal-Wallis (KW) tests in the single-factor ANOVA model and of the F test in the two-factor model. Independent and normally distributed scores that share a common variance are the assumptions that underly the F test (Marascuilo & Serlin, 1988, p. 477). As demonstrated by exact statistical theory (Walsh, 1947) and empirical work (Harwell, 1991; Smith & Lewis, 1980), violating the assumption of independent scores directly affects the Type I error rate (α) and power of the F test. Monte Carlo studies have focused on the less well understood effects of violating the assumptions of equal variances and normality. The W test requires independent and normally distributed scores but does not require equal population variances (Welch, 1947); the KW test requires independent scores sharing a common variance but does not require normality (Pratt, 1964).

Results from exact statistical theory for the F test are discussed first. The framework illustrated in Cooper (1982) is then used to report the results of the meta-analysis: problem formulation, data collection, data evaluation, data analysis and interpretation, and presentation of results. Meta-analytic results are combined with results for exact statistical theory to summarize the effects of assumption violations for the tests.

Effects of Nonnormality on F *(Assuming Equal Population Variances)*

The effect of nonnormality on α was investigated by Gayen (1949, 1950), who assumed that j independent populations shared the same nonnormal distribution. Gayen showed that α was directly related to the skewness and kurtosis of a distribution. Other things being equal, the closer these indexes were to the skewness and kurtosis of a normal distribution, the less α was affected. The results of Scheffé (1959, chap. 10) suggested that skewness had a greater effect on α than kurtosis. Tiku (1964) considered the case where j population distributions were nonnormal but not necessarily the same. Other things being equal, distributions with skewness values in different directions had a greater effect on α than distributions with skewness values in the same direction.

The effects of nonnormality on the power of the F test have been investigated by David and Johnson (1951), Srivastava (1959), and Tiku (1971). Their results indicate that, for the conditions studied, mild departures from normality have little effect on the power of the F test. Moderately nonnormal distributions (e.g., skewness values larger than +.5) lead to power values that peak around .8 and then decline. Negative kurtosis tends to reduce power.

Effects of Unequal Variances on F *(Assuming Normality)*

The work of Hsu (cited in Scheffé, 1959, p. 353), Box (1954), and Horsnell (1953) demonstrated that negatively pairing unequal sample sizes and variances (e.g., small samples paired with large variances) produces an inflated α rate; positive pairings (e.g., small samples paired with small variances) produce conservative α rates. Box and Scheffé (chap. 10) also suggested that equal sample sizes mitigate the role of unequal variances on α, especially for two groups. However, Box found that three or more groups and relatively modest variance ratios—for example, 1:1:1:1:3—produced an α of .074 for a nominal level of significance of .05 for equal sample sizes. Similar results were reported by Gronow (1951), Horsnell, and Ramsey (1980).

The effect of unequal variances on the power of the *F* test has been investigated by Box (1954), Horsnell (1953), and Gronow (1951). Exact calculations of power have not been carried out for cases when variances are unequal, but several suggestions have been made for computing approximate power values. These include using the arithmetic average of the unequal variances and generating a pooled average by weighting each variance by one less the number of scores in that group. Box used both kinds of weighted variances in defining a bias index. Horsnell used the arithmetic average of the unequal variances to compute approximate power values. On the whole, it is difficult to characterize general conclusions about power using exact statistical theory because of the difficulty of "finding a suitable way of defining the noncentrality parameter in the case of inequality of variance" (Scheffé, 1959, p. 361). Glass, Peckham, and Sanders (1972) provide a thorough discussion of this problem.

Where possible, meta-analytic results will be compared to results for exact statistical theory. The credibility of the meta-analysis depends largely on the extent to which these results are consistent with those for exact statistical theory.

Meta-Analysis of Monte Carlo Studies for the Single-Factor, Fixed Effects ANOVA Model

Problem Formulation

The primary research question was: What data conditions are associated with deviations from theoretical α and power levels for the *F*, W, and KW tests in the single-factor ANOVA model? Meta-analytic techniques were used to summarize the available MC literature by quantitatively modeling variation in the empirical α and power values as a function of study characteristics.

Data Collection

A population of MC studies for the F, W, and KW tests in the single-factor, fixed effects ANOVA model was identified by searching the *Educational Resources Information Center* data base, *Dissertation Abstracts International,* and the *Current Index to Statistics.* Key words used to locate relevant studies were: ANOVA, distribution-free, F test, heterogeneity, Kruskal-Wallis, Monte Carlo, nonnormality, nonparametric, power, ranks, robustness, simulation, t test, Type I error, Wilcoxon, and Welch. References in periodicals such as the *Journal of Educational Statistics* and *Communications in Statistics—Simulation and Computation* also proved to be rich sources of literature.

Searching large data bases does not ensure that all relevant studies will be identified. For example, MC results reported in unpublished technical reports and master's theses are likely to be underepresented or missed completely in our sample. Under these conditions, the MC studies used in the meta-analysis may differ from those not included. However, the relatively homogeneous nature of MC studies makes it likely that the potentially nonrandom sample of MC studies is representative of the population of studies that could have been done (Harwell, 1992).

The literature search yielded 31 journal articles and three dissertations which were accessible. Six of the articles did not report empirical Type I error rates or power values, instead choosing to report statistics like the percentage of empirical Type I error rates exceeding a certain cutoff. These studies were excluded from the meta-analysis. Twenty-six of the remaining 28 studies investigated the F test; 6 investigated the W test, and 11 examined the KW test. The relatively small number of accessible studies led to the decision to use every available study in the meta-analysis. The articles and dissertations used in the meta-analysis are listed in the references.

The 28 MC studies were initially screened for methodological flaws. Twenty-five of the studies were published in refereed journals, which may provide some protection against studies of poor quality. In addition, each study was examined for inconsistent or unusual procedures and results using the following criteria: (a) data generation method (e.g., random number generator used), (b) evidence of the success of the data generation (e.g., skewness and kurtosis statistics computed for the simulated data), and (c) pattern of empirical α and power values when underlying assumptions of the test were satisfied. No irregularities were noted, and all 28 studies were used in the meta-analysis.

Empirical Type I error rates and power values served as outcome variables for the meta-analysis. Only results associated with a significance level of .05 were coded. Various characteristics of the MC studies were coded as explanatory variables and are listed in the appendix. Information about population distributions was captured by coding skewness and kurtosis values

(Hastings & Peacock, 1975). In some cases, information about skewness and kurtosis was not reported (e.g., for a mixed normal distribution).

The selection of ranges for coding patterns of sample sizes and variances was guided by conditions reported in the sample of studies. Three noncentrality structures were coded: All means differed (i.e., $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_j$); one mean differed from the others (e.g., $\mu_1 \neq \mu_2 = \cdots = \mu_j$), and two means differed from the others (e.g., $\mu_1 = \mu_2 \neq \mu_3 \neq \cdots \neq \mu_j$). The noncentrality parameter was also coded (see Kirk, 1968, p. 107). When variances differed, the arithmetic average of the unequal variances was used in computing the noncentrality parameter. The number of MC samples was coded because this variable was related to the magnitude of sampling error for the empirical proportions of rejections. The amount of information provided about the data generation was coded using a 3-point scale (1 = little, 2 = moderate, 3 = substantial).

Data Evaluation

A three-phase training process was developed to ensure that the characteristics of each MC study were accurately coded. In the first phase, we reviewed two of the 28 MC studies. The structure of one of the studies was relatively simple, and the other was more complex. Coding forms were completed for each study by each author, and the completed coding forms were compared. Disagreements over particular characteristics of a MC study—for example, how nonnormal distribution were modeled—were resolved by group consensus. Information from this training phase was used to modify the coding forms.

In the next phase, eight of the 28 MC studies were equally divided among two teams of coders, each made up of two of the authors of this article. The members of a team independently reviewed and coded each article assigned to them using the modified coding forms. Members of a team then compared their results and attempted to resolve discrepancies among themselves. On the whole, the percent of agreement among reviewers was nearly 100%. The remaining MC studies were then coded.

In the third phase, the coded MC data were entered into a computer data file. In some studies, several noncentrality parameters and structures were examined for a given set of conditions. For example, a study might yield one Type I error rate for a set of conditions and three power values, corresponding to different noncentrality parameters, for the same conditions. The three power values were coded and were assumed to be associated with the same Type I error rate.

Two strategies were used to detect and correct data entry errors. First, a computer printout of the entire data file was scanned in order to detect obvious errors—for example, error rates falling outside an expected range. Second, the agreement between the information in a study and the corre-

sponding data entered in the data file was reviewed by at least two of the authors for each of the 28 studies.

Data Analysis and Interpretation

Summary information for quantitative variables that were coded is reported in Table 1; Table 2 reports frequencies associated with qualitative variables that were coded. The small number of Type I error cases for the Cauchy and mixed normal distributions precluded these distributions from being included in the Type I error analyses; however, power analyses were conducted for the mixed normal distribution. Table 3 presents the average Type I error rates and power values by study.

Statistical Analyses

To construct and evaluate explanatory models, fixed effects regression models were fitted to the Type I error rates and power values (see Hedges & Olkin, 1985, pp. 168–174). The population regression models were of the form:

$$\pi_k = \beta_0 + X_{k1}\beta_1 + X_{k2}\beta_2 + \cdots + X_{kT}\beta_T. \quad (1)$$

In Equation 1, π_k is a population proportion representing an effect magnitude that depends on a set of T fixed predictor variables X_{kT}; β_0 is an intercept, and β_T is a regression coefficient that captures the relationship between the T(th) predictor and π_k (Harwell, 1992). The Type I error rates and power values served as the π_k, and the coded characteristics of the MC studies served as the X_{kT}.

A test of the relationship between a set of T predictors and the π_k was performed using the weighted sum of squares due to regression statistic Q_R (see Harwell, 1992, for an explanation of the weights). A test of model misspecification (i.e., whether all of the explanatory variables needed to explain variation in the outcomes are in the model) was performed using the Q_E statistic. All statistical tests used a significance level of .05. Listwise deletion of missing data reduced the number of cases used in some analyses. The SPSSX (SPSS, 1988) computer program was used to perform the analyses, which are reported in Tables 4–6. The Type I error and power cases are discussed separately.[1]

Type I Error Case

Preliminary analyses indicated that the contribution of the variable representing trends in sample sizes (NTREND) was negligible, and it was dropped from subsequent models. The contribution of the amount of information provided about the data generation (DATAGEN) was nonnegligible, and it was used in the regression models. However, DATAGEN's

TABLE 1
Summary statistics for quantitative variables for the sample of Monte Carlo studies

Variable	*k*	Average	*SD*	Min	Max
F test					
TYPEI	1210	.063	.036	.004	.333
POWER1	944	.48	.32	.026	1.0
POWER2	277	.70	.20	.038	.987
POWER3	16	.69	.04	.609	.737
TOTALN	1210	111	154	8	750
SKEW	1099	.27	.87	0	6.2
KURT	1099	2.02	11.6	−3.75	111
MCSAMPLES1	1210	22658	41294	400	154556
MCSAMPLES2	1422	4861	4211	400	15000
Welch test					
TYPEI	245	.055	.013	.035	.134
POWER1	84	.47	.26	.10	.95
POWER2	61	.54	.25	.106	.987
POWER3	21	.79	.15	.447	.999
TOTALN	245	44	24	8	200
SKEW	245	.12	.48	0	2
KURT	245	.74	1.97	0	6
MCSAMPLES1	245	3592	3696	1000	10000
MCSAMPLES2	216	5458	4629	1000	15000
Kruskal-Wallis test					
TYPEI	242	.057	.051	.005	.381
POWER1	410	.60	.32	.032	1.0
POWER2	32	.65	.10	.503	.838
POWER3	16	.69	.02	.643	.723
TOTALN	242	51	64	9	750
SKEW	147	.48	.84	0	2
KURT	147	1.44	2.72	−3.75	6
MCSAMPLES1	242	1072	841	400	5000
MCSAMPLES2	482	1385	1147	400	5000

Note. The variables are defined in the appendix. *k* = number of cases, *SD* = standard deviation, Min = minimum, Max = maximum.

contribution to the regression equations as reflected in standardized regression coefficients was quite modest.

The results of the explanatory models for the Type I error (TYPEI) case are reported in Table 4. Examination of the residuals of each of the models indicated no gross departures from normality. All of the Q_R statistics in Table 4 were significant ($p < .001$) and were typically quite large. Similarly,

TABLE 2
Summary statistics for qualitative variables for the sample of Monte Carlo studies

Variable	Frequency	Percent	Variable	Frequency	Percent
Distribution	1697	100.0	VARRATIO	1697	100.0
Normal	1160	68.3	Equal	653	38.3
Uniform	23	1.4	>1 and ≤2	134	7.9
Double exponential	18	1.1	>2 and ≤3	70	4.1
Log-normal	18	1.1	>3 and ≤5	234	13.8
Cauchy	12	.7	>5 and ≤8	208	12.2
Exponential	116	6.8	>8	398	23.4
Logistic	8	.5			
t	30	1.8			
Mixed normal	28	1.6			
Other	284	16.7			

NGRPS	1697	100.0	Pairing	1697	100.0
2	486	28.6	Positive	213	12.5
3	330	19.4	Negative	215	12.7
4	490	28.9	Not applicable (equal σ_j^2/N_j)	1269	74.8
5	127	7.5			
6	47	2.8			
7	103	6.1			
8	11	.6			
9	103	6.1			
NRATIO	1697	100.0	DATAGEN	1697	100.0
Equal	1086	64	Low	320	18.9
>1 and ≤1.25	42	2.5	Average	669	39.4
>1.25 and ≤1.5	63	3.7	High	708	41.7
>1.5 and ≤1.75	48	2.8			
>1.75 and ≤2	32	1.9			
>2 and ≤3	250	14.7			
>3 and ≤5	176	10.4			

TABLE 3
Average TYPE1 and POWER1 (all means differ) values by study

Study	TYPE1	POWER1	Study	TYPE1	POWER1
1	.041 (48)	.55 (144)	15	.041 (16)	.52 (96)
2	.081 (62)	*	16	.044 (27)	.55 (189)
3	.053 (27)	.59 (7)	17	.062 (96)	.70 (96)
4	.042 (16)	*	18	.052 (12)	.44 (48)
5	.066 (412)	*	19	.040 (12)	.50 (72)
6	.066 (54)	*	20	*	.47 (59)
7	.068 (90)	.90 (90)	21	.050 (5)	*
8	.065 (90)	.25 (90)	22+	.065 (14)	*
9	.066 (80)	*	23	.057 (20)	*
10	.051 (60)	*	24	.049 (12)	.59 (24)
11	.047 (10)	.58 (75)	25	.076 (36)	*
12	.045 (24)	.57 (144)	26+	.058 (168)	*
13	.045 (162)	.37 (162)	27+	.081 (108)	.30 (108)
14	.050 (18)	.57 (18)	28	.045 (18)	.07 (16)

Note. * = did not examine Type I error rate or power, + = dissertation, values in parentheses represent the number of cases. There is no correspondence between the study identification number and the MC studies.

the Q_E statistics (not reported) were significant ($p < .001$) for each of the models. Despite the misspecification of the models, the squared multiple correlation R^2, adjusted for the number of predictors, appeared to be a useful index of the explanatory power of a model.

Model 1

Model 1 investigated the relationship between TYPEI and the predictor variables skewness (SKEW), kurtosis (KURT), number of groups (NGRPS), total sample size (TOTALN), ratio of largest to smallest sample sizes (NRATIO), DATAGEN, ratio of largest to smallest variances (VARRATIO), and their two-factor interactions:

$$\textit{Model 1a.}\quad \text{TYPEI} = \beta_0 + \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + \text{NGRPS}\ \beta_3 + \text{TOTALN}\ \beta_4 + \text{NRATIO}\ \beta_5 + \text{DATAGEN}\ \beta_6 + \text{VARRATIO}\ \beta_7.$$

$$\textit{Model 1b.}\quad \text{TYPEI} = \beta_0 + \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + \text{NGRPS}\ \beta_3 + \text{TOTALN}\ \beta_4 + \text{NRATIO}\ \beta_5 + \text{DATAGEN}\ \beta_6 + \text{VARRATIO}\ \beta_7 + 21 \text{ two-variable-at-a-time interaction predictors.} \qquad (2)$$

The R^2 of .26 for the F test for Model 1a in Table 4 suggests a moderate

TABLE 4
Summary of Type I error results by test

Model	*F* test				Welch test				Kruskal-Wallis test			
	T	k	Q_R	R^2	T	k	Q_R	R^2	T	k	Q_R	R^2
1a	7	1099	38228	.26	7	245	1075	.50	7	147	458	.02
1b	28	1099	44902	.30	21	245	1237	.56	26	147	1186	.01
2a	5	1099	37708	.26	5	245	601	.27	5	147	391	.03
2b	7	1099	38228	.26	7	245	1075	.50	7	147	458	.03
3a	7	251	4298	.05	7	133	771	.54	6	43	627	.00
3b	8	251	40916	.72	8	133	891	.62	7	43	2938	.46
3c	7	126	1445	.56	7	66	229	.58	6	20	406	.46
3d	7	125	6268	.38	7	67	622	.71	6	23	806	.19
4a	5	534	4808	.06	5	61	191	.53	5	22	150	.07
4b	6	534	23710	.34	6	61	202	.56	6	22	159	.03

Note. T = number of predictors, Q_R = the weighted sum of squares due to regression statistic, k = number of cases, R^2 = the squared multiple correlation adjusted for the number of predictors. Every Q_R statistic was significant at $p < .001$.

correlation between Type I error rates and the set of predictors, and a model with modest explanatory power. An examination of the estimated regression coefficients and their standard errors (corrected following Hedges & Olkin, 1985, p. 174) suggests that VARRATIO plays a key role. Among the seven variables, VARRATIO had the largest standardized regression coefficient (.4). With the exception of KURT, all of the associated regression coefficients were significantly different from zero (an arbitrary criterion of $p < .01$ was used in testing the regression coefficients).

The W and KW tests produced somewhat different results for Model 1a. The correlation between error rates and the set of predictors was quite weak for the KW test (i.e., error rates of the KW test were robust). On the other hand, there was a strong correlation between error rates and the set of predictors for the W test, and all of the associated regression coefficients were significant.

These results suggest that the Type I error rate of the KW test was insensitive to the predictors in Model 1a, that the error rate of the *F* test was moderately affected, and that the error rate of the W test was quite sensitive to these predictors. The analysis involving interaction predictors suggests that the three tests are no more sensitive to multiple assumption violation than they would be to the effects of each violation separately.

Model 2

Model 2 investigated the effect of the shape of a population distribution, as captured with skewness and kurtosis indexes, on Type I error rates:

$$\textit{Model 2a.}\quad \text{TYPEI} = \beta_0 + \text{NGRPS}\ \beta_1 + \text{TOTALN}\ \beta_2 + \text{NRATIO}\ \beta_3 + \text{DATAGEN}\ \beta_4 + \text{VARRATIO}\ \beta_5. \tag{3}$$

Model 2b included the predictors in Model 2a plus predictors representing skewness (SKEW) and kurtosis (KURT). The results for the *F* and KW tests indicate that the shape of a population distribution has little to do with explaining variation in Type I error rates. The standardized regression coefficient for the *F* test for SKEW was significant, but the coefficient for KURT was not, suggesting that skewness has a greater effect on Type I error rates than kurtosis. This is consistent with the results reported by Scheffé (1959, chap. 10). On the whole, the Model 2 results support the perception that the Type I error rates of the *F* and KW tests are relatively insensitive to the shape of a population distribution.

The effect of skewness and kurtosis on the W test was quite different. The difference in R^2s (.23), and the size of the (significant) standardized regression coefficients for SKEW and KURT (.46 and −.43, respectively), suggests that the W test was more sensitive to the shape of a population distribution than the *F* or KW tests. The average error rates for the three

distributions used with the W test (normal, t, exponential) were, based on 215, 15, and 15 cases: .053, .043, and .082, respectively. Although the number of cases was quite small, these results suggest that distinctly nonnormal distributions like the exponential have an inflationary effect on the Type I error rate of the W test.

Pearson's (1929) observation that equal sample sizes mitigate the effect of nonnormal distributions was investigated by repeating Model 2 using data associated with equal, and, separately, unequal samples. The R^2 for equal samples for Model 2b for the F test was .15, while the R^2 for the unequal samples case was .21, supporting Pearson's observation. These results are also consistent with the work of Box (1954).

Model 3

The relationship between Type I eror rates and unequal sample sizes paired with unequal variances (PAIRING) was investigated in Models 3a–3d. Model 3a was identical to Model 1a, while 3b investigated the effect of adding PAIRING. The known sensitivity of the F test to paired samples and variances implies that the meta-analysis should detect a strong relationship between the set of predictor variables (including PAIRING) and error rates. To reiterate, Model 3 used MC data that were associated with a positive or negative pairing of sample sizes and variances.

The results for the F test suggest a very strong relationship between Type I error rates and sample size and variance pairings. These results are consistent with the work of Box (1954) and Horsnell (1953). A similar pattern was observed for the KW test (NGRPS was dropped as a predictor because all of the cases were based on four groups). The W test was less sensitive to PAIRING, which was expected because this test does not require equal variances.

Models 3c and 3d provided specific evidence about the role of sample size and variance pairings. Model 3c investigated the relationship between error rates and the predictors in 3a for MC data in which samples and variances were positively paired (e.g., smaller samples paired with smaller variances), and Model 3d investigated this relationship when samples and variances were negatively paired (e.g., smaller samples paired with larger variances).

The results in Table 4 for the F test suggest a strong relationship between error rates and the explanatory models for a positive pairing of sample size and variances and a weaker relationship for negative pairings. The regression coefficients for VARRATIO were significant for both Models 3c and 3d. The average error rates for the F test for the five unequal variance conditions of VARRATIO for the positively paired condition were, going from smaller to larger variance ratios: .034 (15), .03 (19), .027 (34), .032 (6), and .029 (52), respectively. (Values in parentheses represent the number of cases.) For the negative pairing case, the average error rates across the

unequal variance conditions of VARRATIO for F were: .082 (15), .10 (19), .12 (34), .14 (6), and .18 (51), respectively. These results are consistent with the work of Box (1954), Horsnell (1953), and Ramsey (1980).

The pattern of results for the KW test for positive pairings was similar to that of the F test. However, there was a much smaller correlation between error rates and the predictors in Model 3d for negatively paired samples and variances than there was for F.

The similarity of Model 3c results for the W test versus those of Model 1a supports the relative insensitivity of this test to unequal variances. All of the associated regression coefficients for Model 3c were significant except the coefficient for VARRATIO. For positively paired samples and variances, the average error rates for the W test for the unequal variance categories of VARRATIO were: .056 (5), .05 (9), .046 (12), .051 (6), and .053 (34), respectively. The W test was more sensitive to negatively paired samples and variances. The average error rates for this case were: .054 (5), .051 (9), .057 (12), .051 (6), and .064 (35), respectively.

Model 4

Model 4 investigated the relationship between Type I error rates and unequal variances when sample sizes were equal. The models were

$$\textit{Model 4a.}\quad \text{TYPEI} = \beta_0 + \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + \text{NGRPS}\ \beta_3 + \text{TOTALN}\ \beta_4 + \text{DATAGEN}\ \beta_5.$$

$$\textit{Model 4b.}\quad \text{TYPEI} = \beta_0 + \text{SKEW}\ \beta_1 + \text{KURT}\ \beta_2 + \text{NGRPS}\ \beta_3 + \text{TOTALN}\ \beta_4 + \text{DATAGEN}\ \beta_5 + \text{VARRATIO}\ \beta_6. \tag{4}$$

The results for the F test suggest a moderately strong relationship between variance inequality and error rates when sample sizes are equal. In addition, the standardized regression coefficient for VARRATIO (.53) in Model 4b was significant. The average error rates for VARRATIO when samples were equal are reported in Table 5. VARRATIO values greater than 2 produced inflated error rates, a pattern that was exacerbated as the variance ratio increased. The average error rates for large ratios are consistent with the results of Box (1954).

On the whole, the results of Model 4 suggest that equal sample sizes provided little protection against inflated error rates for unequal variances. The two-group case appeared to be less sensitive to unequal variances than the three- or more groups case, as predicted by the work of Box (1954) and Scheffé (1959, chap. 10).

Power Case

Power values equal to one were recoded to .99 for the regression analyses because the weights for the regression models involved the product of

TABLE 5
Average Type I error rates by test and variance ratios for equal samples

VARRATIO	Equal	1–2	2–3	3–5	5–8	>8
F	.047 (193)	.053 (87)	.066 (2)	.062 (105)	.07 (166)	.076 (68)
W	.052 (22)	.052 (5)	.048 (6)	.051 (9)	.051 (6)	.056 (35)
KW	.045 (59)			.104 (9)	.054 (4)	.067 (7)
**F*	.046 (114)	.051 (2)	.053 (6)	.057 (13)		.064 (15)
*W	.050 (5)		.048 (6)	.051 (6)		.054 (8)
*KW	.047 (39)	.059 (2)				.076 (3)

Note. Refer to the coding scheme given in the appendix for a definition of VARRATIO. Blanks indicate no data. *F* = *F* test, W = Welch, KW = Kruskal-Wallis, * = 2 groups. Values in parentheses represent the number of cases.

power × (1 − power) in the denominator. Failure to recode these values would result in these cases being dropped from the regression models.

TYPEI and the noncentrality (NONCEN) were entered first into each of the regression models, permitting the comparison of power results of tests with varying TYPEI and NONCEN values. With the exception of Models 3c–3d, Models 1–4 were repeated for POWER1 (all means differ); available POWER2 data (one mean differs from the others) permitted Model 2 to be analyzed for the *F* test only; and none of the models could be analyzed for POWER3 (two means differ from the others) due to lack of data.

The POWER1 results in Table 6 indicate greater explanatory power for the *F* test than for the Type I error case. This occurs because of the close relationship between power and noncentrality parameters. In fact, NONCEN accounted for at least 80% of the variation in POWER1 in the *F* test analyses. All of the regression coefficients for Model 1a for the *F* test were significant, with NONCEN producing the largest standardized coefficient (.87). The results of Model 1b indicate that combinations of assumption violations had little effect on the power of the *F* test. These conclusions extend to a sample of mixed normal cases for the *F* test for which Model 1 (less SKEW and KURT) was analyzed. The power of the W and KW tests was not as highly dependent on NONCEN and was more sensitive to combinations of assumption violations (especially W).

The power of the *F* test was insensitive to SKEW and KURT, which is generally consistent with the results of David and Johnson (1951), Srivastava (1959), and Tiku (1971). The KW test was also insensitive to SKEW and KURT. The power of the W test was slightly affected by SKEW and somewhat more affected by KURT. The average POWER1 values for the W test for a normal, *t*, and exponential distribution were, based on 38, 15, and 15 cases: .52, .26, and .28, respectively.

The power of the tests was relatively insensitive to the effects of the

TABLE 6
Summary of POWER1 (all means differ) results

	F test				Welch test				Kruskal-Wallis test			
Model	T	k	Q_R	R^2	T	k	Q_R	R^2	T	k	Q_R	R^2
1a	9	662	8584329	.80	8	68	63015	.28	9	206	98060	.24
1b	29	662	8810350	.81	18	68	135558	.70	26	206	192656	.47
2a	7	662	8565624	.79	6	68	55526	.25	7	206	93282	.23
2b	9	662	8584329	.80	8	68	63015	.28	9	206	98060	.24
3a	9	94	226676	.93	8	35	58339	.93	6	26	4489	.91
3b	10	94	226884	.93	9	35	59019	.94	7	26	4515	.92
4a	7	73	109047	.86	6	18	5632	.26	7	17	5247	.82
4b	8	73	109485	.86	7	18	6250	.29	8	17	5248	.80

Note. T = number of predictors, Q_R = weighted sum of squares due to regression statistic, k = number of cases, R^2 = the squared multiple correlation adjusted for the number of predictors. Some predictors were dropped because of a lack of variability. Every Q_R statistic was significant at $p < .001$.

PAIRING variable in Model 3. The effect of unequal variances when sample sizes were equal was negligible for the F test, which is consistent with the approximate results of Horsnell (1953). The power of the W test for Model 4 was slightly affected by unequal variances, with a standardized coefficient for VARRATIO of .30, while the power of the KW test was strongly correlated with the predictors in Model 4a but was not affected by unequal variances.

The lack of POWER2 data meant that only Model 2 could be analyzed for the F test. The results (not reported) were similar to those in Table 6, although the R^2 values were slightly larger.

Meta-Analysis of Monte Carlo Studies for the Two-Factor ANOVA Model

An immediate question arising from the meta-analytic results in the single-factor case is their generalizability to the multifactor case. It would be surprising indeed if meta-analytic results differed for these two cases, since the assumptions of the F test are the same in the single- and multifactor models. To investigate this empirically, a meta-analysis was performed on a sample of MC studies for the two-factor ANOVA model. Only the F test was investigated because of the absence of any well-documented alternative. The two-factor results are summarized below.

The research question for the two-factor model was the same as that for the single-factor case. In collecting studies for the two-factor case, the same literature sources used in the single-factor case were searched using the key words: ANOVA, factorial, Monte Carlo, simulation, and heterogeneous variances. The literature search yielded seven accessible studies, which are listed in the references.

The sample of MC studies showed the following characteristics: Type I error rate and power information was provided for (at most) three F tests (two main effects, one interaction); all cell sample sizes and variances were equal, and the distributions studied were fewer in number than in the single-factor case. The coding scheme was modified as follows: The explanatory variables NGRPS, VARRATIO, and PAIRING were dropped; Type I error rates and power values were coded for each of the three effects when data were available; the distinctions among noncentrality structures were ignored, and a single power variable was coded. Several of the studies focused on the test of the interaction effect and ignored tests of the main effects. Only Models 1 and 2 in the single-factor case could be analyzed, and, on the whole, inferences for the two-factor model are more restricted than those associated with the single-factor model.

The average Type I error rates for the row, column, and interaction effects were: .049 (69), .048 (54), and .05 (269), respectively. The average power values for the row, column, and interaction effects were: .45 (20), .55 (33), and .59 (282), respectively. Seven distributions were represented: normal,

uniform, double exponential, log-normal, Cauchy, exponential, and mixed normal. Approximately 80% of the data were sampled from a normal distribution.

The meta-analytic results indicated that the Type I error rate of the F test was relatively insensitive to the effects of the explanatory variables DATAGEN, TOTALN, SKEW, and KURT and to multiple assumption violations (Model 1). It was also insensitive to the shape of a population distribution (Model 2). The power results for Model 1 produced a weaker relationship between power and the set of explanatory variables (including TYPEI and NONCEN) than they did for the single-factor case. The two-factor power results were not differentially insensitive to multiple assumption violations or to the shape of a population distribution. On the whole, the empirical Type I error and power results are less sensitive to the explanatory variables in Models 1 and 2 than they are to those in the single-factor case. This is probably due to the fact that the bulk of the two-factor data was from a normal distribution.

Summary of the Effects of Assumption Violations for the Three Tests

The results of the meta-analysis were combined with exact statistical theory results to summarize the effects of assumption violations on the Type I error rate and power of the F, W, and KW tests in Table 7. The format of Table 16 in Glass et al. (1972), in which conclusions regarding the Type I error rate and power of the single-factor F test were presented as a function of various assumption violations for equal and unequal sample sizes, was used in Table 7. Models 1 and 2 for the single-factor case were repeated for equal and unequal sample sizes. These results were similar to those presented in Tables 4 and 6 in which no distinction was made between equal and unequal samples.

The conclusions regarding power in Table 7 were based on the POWER1 results. The available evidence for POWER2 suggests that conclusions regarding the two noncentrality structures are quite similar. Inferences about the effects of skewness and kurytosis in Table 7 can probably be extended to the F test for the two-factor case.

Two results in Table 7 deserve special attention. One is that equal sample sizes provided little protection against inflated error rates for the F and KW tests when variances were unequal, even when sample sizes were equal. Whether the inflation is serious enough to consider an alternative test which is insensitive to unequal variances (e.g., the W test) is properly left to individual researchers. Table 5 provides researchers with some idea of the amount of Type I error inflation that can be expected for various variance ratios. The information in Table 5 may also address the concern of Glass et al. (1972, pp. 244–245) that "the conventional conclusion that heterogeneous variances are not important when n's are equal seems to have

TABLE 7
Summary of consequences of violations of assumptions for ANOVA

Type of violation	Equal samples		Unequal samples	
	Effect on α	Effect on power	Effect on α	Effect on power
Nonindependence of scores	Nonindependence of scores affects both α and power of the *F*, W, and KW tests regardless of whether samples are equal or unequal			
Nonnormality	Negligible effect for the *F* and KW tests; moderate inflation for W test (especially skewness)	Negligible effect for *F*; slight depressive effect for KW; moderate depressive effect for W test (kurtosis more than skewness)	Negligible effect for KW test; slight inflation for *F* test (skewness more than kurtosis); more substantial inflation for W test (skewness more than kurtosis)	Negligible effect for *F* test; modest depressive effect for KW; pronounced depressive effect for W test
Unequal variances	Modest inflation for *F* test that increases with increasing variance ratios; KW test more erratic; effect for W test is modest up to ratios of 8:1	Negligible effect for *F* and W tests; modest depressive effect for KW test	The *F* test is seriously affected, and KW test somewhat less so; positive pairings produce conservative αs; negative pairings produce inflated αs; slight inflation for W test	Negligible effect for *F* and KW tests; slight inflation for W test
Nonnormality unequal variances	Modest inflation for *F*; KW more erratic; moderate inflating effect for W test that depends on distribution and variance ratio	Modest effect for all three tests that depends on distribution and variance ratio	Negligible effect for *F* and KW tests; moderate inflation for W test that depends on distribution and variance ratio	Negligible effect for *F* and KW tests; slight depressive effect for W test that depends on distribution and variance ratio

Note. W = Welch test, KW = Kruskal-Wallis test.

boundary conditions like all other conclusions in this area, and the boundary conditions may not have been sufficiently probed."

Another important result in Table 7 is the apparent sensitivity of the Welch test to distinctly nonnormal score distributions, regardless of whether sample sizes were equal. There was ample evidence of the effectiveness of the Welch test in controlling Type I error rates near the nominal value when variances were unequal (regardless of whether samples were equal or unequal) and when distributions were normal or nearly normal. However, unconditional recommendations for the use of this test must be tempered because of its apparent sensitivity to distinctly nonnormal distributions which, based on a small number of cases, appear to produce inflated error rates and depressed power values. The magnitude of these effects for the W test across a range of nonnormal distributions has not been documented.

Conclusion

The application of quantitative methods of research synthesis to quantitatively integrate Monte Carlo results shows promise for helping researchers to select the best statistical test for their data. The resulting empirical framework of Monte Carlo studies of a statistical test provides a summary of the effects of assumption violations on the Type I error rate and power. Such summaries also permit the validity of previous statistical analyses employing the test to be evaluated.

The meta-analytic results support the perception that the Type I error rate of the F test for the single-factor fixed effects model is robust across a variety of conditions. The credibility of these findings is enhanced by their frequent agreement with exact statistical theory. The meta-analytic results provided evidence that researchers should not rely on equal sample sizes to neutralize the effects of unequal variances on the F test or its nonparametric counterpart, the Kruskal-Wallis test. Under these conditions, the likely result is an inflated Type I error rate. The meta-analysis also indicated that the Welch test is a viable alternative if the distribution is nearly normal. The summary of the behavior of the F test that was presented in Glass et al. (1972) was updated to include the meta-analytic results.

Note

[1] On the suggestion of an anonymous reviewer, the analyses reported in Tables 4–6 were repeated using the weighted logistic regression procedure in the SAS (SAS Institute, 1989) computer program. All of the associated log-likelihood test statistics were significant at $p < .001$, and the pattern among the estimated slopes for the two sets of analyses was similar.

APPENDIX
Coding scheme

(1) Distribution (γ_1 = skewness, γ_2 = kurtosis)
  Normal ($\gamma_1 = 0$, $\gamma_2 = 0$)
  Uniform ($\gamma_1 = 0$, $\gamma_2 = -1.2$)
  Double exponential ($\gamma_1 = 0$, $\gamma_2 = 3$)
  Log-normal (γ_1, γ_2 depend on the parameters used)
  Cauchy ($\gamma_1 = 0$, γ_2 undefined)
  Exponential ($\gamma_1 = 2$, $\gamma_2 = 6$)
  Logistic ($\gamma_1 = 0$, $\gamma_2 = 4.8$)
  t ($\gamma_1 = 0$, $\gamma_2 = \nu/(\nu - 4)$, ν = error degrees of freedom)
  Mixed normal (γ_1, γ_2 depend on the parameters used)
  Other [includes the binomial, Poisson, and Erlang distributions]
(2) Skewness (SKEW)
(3) Kurtosis (KURT)
(4) Number of groups (NGRPS)
(5) Total sample size (TOTALN)
(6) Type I error rate (TYPEI)
(7) Number of Monte Carlo samples for Type I error rate (MCSAMPLES1)
(8) Amount of information provided about the data generation (DATAGEN)
(9) Sample size trend (NTREND)
  1 = steady increase
  2 = steady decrease
  3 = not applicable (equal samples and/or two groups)
(10) Ratio of largest to smallest sample sizes (NRATIO)
  1 = 1 (sample sizes equal)
  2 = >1 and ≤1.25
  3 = >1.25 and ≤1.5
  4 = >1.5 and ≤1.75
  5 = >1.75 and ≤2.0
  6 = >2.0 and ≤3
  7 = >3 and ≤5
  8 = >5
(11) Ratio of largest to smallest variances (VARRATIO)
  1 = 1 (all variances equal)
  2 = >1 and ≤2
  3 = >2 and ≤3
  4 = >3 and ≤5
  5 = >5 and ≤8
  6 = >8
(12) Pairing of sample size and variance (PAIRING)
  1 = positively correlated (e.g., small variances paired with small sample sizes)
  2 = negatively correlated (e.g., small variances paired with large sample sizes)
  3 = not applicable
(13) Power for noncentrality structure in which all means differ (POWER1)

(continued p. 336)

APPENDIX (Continued)

(14) Power for noncentrality structure in which one mean differs from the others (POWER2)
(15) Power for noncentrality structure in which two means differ from the others (POWER3)
(16) Noncentrality parameter (NONCEN)
(17) Number of Monte Carlo samples for power (MCSAMPLES2)

References

* Bishop, T. A. (1976). Heteroscedastic ANOVA, MANOVA, and multiple-comparisons (Doctoral dissertation, Ohio State University, 1976). *Dissertation Abstracts International 37,* 3822B.

* Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology, 33,* 114–120.

** Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics—Simulation and Computation, 16,* 1133–1145.

* Boehnke, K. (1984). F- and H-test assumptions revisited. *Educational and Psychological Measurement, 44,* 609–617.

* Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57,* 49–64.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems I: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290–302.

* Budescu, D. V., & Appelbaum, M. I. (1981). Variance stabilizing transformations and the power of the *F* test. *Journal of Educational Statistics, 6,* 55–74.

* Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 1,* 207–214.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52,* 291–302.

David, F. N., & Johnson, N. L. (1951). The effect of nonnormality on the power function of the F-test in the analysis of variance. *Biometrika, 38,* 43–57.

* Dijkstra, J. B., & Werter, P. S. P. J. (1981). Testing the equality of several means when the population variances are unequal. *Communications in Statistics—Simulation and Computation, 10,* 557–569.

* Donaldson, T. S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. *Journal of the American Statistical Association, 63,* 660–676.

* Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34,* 789–799.

* Games, P. A., & Lucas, P. A. (1966). The analysis of variance of independent

groups on non-normal and normally transformed data. *Educational and Psychological Measurement, 26,* 311–327.

Gayen, A. K. (1949). The distribution of 'Student' t in the random samples of any size drawn from non-normal universes. *Biometrika, 36,* 353–369.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika, 37,* 236–255.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

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

** Groggel, D. J. (1987). A Monte Carlo study of rank tests for block designs. *Communications in Statistics—Simulation and Computation, 16,* 601–620.

Gronow, D. G. C. (1951). Test for the significance of the difference between means in two normal populations having unequal variances. *Biometrika, 38,* 252–256.

** Harwell, M. R. (1991). Completely randomized factorial analysis of variance using ranks. *British Journal of Mathematical and Statistical Psychology, 44,* 383–401.

Harwell, M. R. (1991). Using randomization tests when errors are unequally correlated. *Computational Statistics and Data Analysis, 11,* 75–85.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297–313.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical distributions*. London: Butterworths.

* Havlicek, L. L., & Peterson, N. L. (1974). Robustness of the t test: A guide for researchers on effect of violations of assumptions. *Psychological Reports, 34,* 1095-1114.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Harcourt, Brace, Jovanovich.

Horsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika, 40,* 128–136.

** Iman, R. L., & Conover, W. J. (1976). *A comparison of several rank tests for the two-way layout* (Tech. Rep. No. SAND76-0631). Los Alamos, NM: Sandia Laboratories.

** Kanji, G. K. (1976). Effect of nonnormality on the power in analysis of variance: A simulation study. *International Journal of Educational and Science Technology, 7,* 155–160.

* Kanji, G. K. (1976). The study of robustness of power in the analysis of variance. *International Journal of Mathematical Education in Science Technology, 7,* 401–407.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.

* Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education, 43,* 61–69.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47,* 583–621.

** Lemmer, H. H. (1980). Some empirical results on the two-way analysis of variance by ranks. *Communications in Statistics—Simulation and Computation, 14,* 1427–1438.

* Levine, D. W., & Dunlap, W. P. (1982). Power of the *F* test with skewed data: should one transform or not? *Psychological Bulletin, 92,* 272–280.

* Lin, L. I., & Sanford, R. L. (1983). The robustness of the likelihood ratio test, the nonparametric rank sum test, and F-ratio tests when the populations are from the negative binomial family. *Communications in Statistics—Simulation and Computation, 12,* 523–539.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavior sciences.* New York: Freeman.

* McSweeney, M., & Penfield, D. (1969). The normal scores test for the c-sample problem. *British Journal of Mathematical and Statistical Psychology, 22,* 177–192.

* Nath, R., & Duran, B. S. (1981). The rank transform in the two-sample location problem. *Communications in Statistics—Simulation and Computation, 10,* 383–394.

* Neave, H. R., & Granger, W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in means. *Technometrics, 10,* 509–532.

* Norton, D. W. (1952). An empirical investigation of the effects of nonnormality and heterogeneity upon the *F*-test of analysis of variance. Unpublished doctoral dissertation, University of Iowa, Iowa City.

* Olejnik, S. (1984). Conditional ANOVA for mean differences when population variances are unknown. *Journal of Experimental Education, 53,* 141–148.

** Patel, K. M., & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association, 68,* 615–620.

Pearson, E. G. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika, 21,* 259–286.

* Penfield, D. A., & Koffler, S. L. (1985, March). *A power study of selected nonparametric* K*-sample tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association, 59,* 665–680.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of student's *t* test with unequal variances. *Journal of Educational Statistics, 5,* 337–349.

* Randolph, E., Robey, R., & Barcikowski, R. (1990, April). *Type I error of the ANOVA revisited using power analysis criteria.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

*Rasmussen, J. L. (1985). An evaluation of parametric and non-parametric tests on modified and non-modified data. *British Journal of Mathematical and Statistical Psychology, 39,* 213–220.

* Rasmussen, J. L. (1985). The power of student's t and Wilcoxon statistics. *Evaluation Review, 9,* 505–510.

* Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust of variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal, 14,* 493–498.

SAS Institute. (1989). *SAS/STAT user's guide* (4th ed., Vol. 2). Cary, NC: Author.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Smith, J. H., & Lewis, T. O. (1980). Determining the effects of intraclass correlation on factorial experiments. *Communications in Statistics—Simulation and Computation, 9,* 1353–1364.

SPSS. (1988). *SPSSX user's guide* (3rd ed.). Chicago: Author.

Srivastava, A. B. L. (1959). Effects of non-normality on the power of the analysis of variance test. *Biometrika, 46,* 114–122.

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. *Biometrika, 51,* 83–95.

Tiku, M. L. (1971). Power function of the F-test under non-normal situations. *Journal of the American Statistical Association, 66,* 913–916.

* Tomarkin, A., & Serlin, R. C. (1986). A comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychologial Bulletin, 99,* 90–99.

** Toothaker, L. E., & Chang, H. (1980). On "The analysis of ranked data derived from completely randomized factorial designs." *Journal of Educational Statistics, 5,* 169–176.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics, 18,* 88–96.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika, 34,* 28–35.

* Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics—Simulation and Computation, 15,* 933–943.

* Zimmerman, D. W. (1987). Comparative power of student t test and Mann-Whitney U test for unequal sample sizes and variances. *Journal of Experimental Education, 55,* 171–174.

---

* Single-factor case studies used in the meta-analysis.

** Two-factor case studies used in the meta-analysis.

Authors

MICHAEL R. HARWELL is Associate Professor, Department of Psychology in Education, University of Pittsburgh, 5B27 Forbes Quad, Pittsburgh, PA 15260. He specializes in meta-analysis and nonparametric statistics.

ELAINE N. RUBINSTEIN is PhD Candidate, University of Pittsburgh, 5B27 Forbes Quad, Pittsburgh, PA 15260. She specializes in research methodology.

WILLIAM S. HAYES is PhD Candidate, University of Pittsburgh, 5B27 Forbes Quad, Pittsburgh, PA 15260. He specializes in research methodology.

CORLEY C. OLDS is PhD Candidate, University of Pittsburgh, 5B27 Forbes Quad, Pittsburgh, PA 15260. She specializes in research methodology.