Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: the General Linear Model (MLR:GLM/SIG)

MLRV

Volume 22 - Number 1 • Winter 1995

Table of Contents

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: the General Linear Model (MLR:GLM/SIG) through the University of Missouri at St. Louis. *MLRV* abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook* and with the *FAXON* and *READMORE* subscription agencies. MLR:GLM/SIG information and a membership application form can be obtained by writing, FAXing (314-516-5784), Voice Mailing (314-516-5785), or e-MAILing (sspaner@umslvma.umsl.edu) the Executive Secretary. 1995 SIG membership and subscription fees for are: Individual - $10 for one year, $18 for two years; Library/Agency - $20 per year; and Student - $5 for one year. Fee payment should be made payable to the **Multiple Linear Regression SIG** and sent to Steven Spaner, MLR:GLM/SIG Executive Secretary, Department of Behavioral Studies, University of Missouri - St. Louis, 8001 Natural Bridge Road, St Louis, MO 63121-4499.

Department of Educational Psychology
and Special Education
Southern Illinois University at Carbondale
Carbondale, Illinois 62901-4618
Phone: (618)-536-7763
FAX (618)-453-7110
Email: JOHNP@SIU.EDU

MLRV

Multiple Linear Regression Viewpoints

Dear MLRV Subscriber:

We are pleased to provide you with the Winter, 1995 edition of *Viewpoints*. We hope you enjoy this issue and will contribute to future issues of *Viewpoints*. You are invited to submit manuscripts on linear model issues. Theoretical works, applications of the linear model and teaching illustrations are especially sought. If you have any ideas or questions about manuscript possibilities, please contact me.

Manuscripts must be submitted on a 3.5 inch floppy disk. Microsoft Word for the Macintosh, Version 5.1a is used to layout the journal. Please save your manuscript in this format. Manuscripts should conform to the styles standards presented in the *Publication manual of the American Psychological Association* (4th Edition).

I look forward to visiting with you at AERA in New York.

Regressingly Yours,

John T. Pohlmann
Editor

Bias Correction in Risk Assessment When Logistic Regression is used with An Unevenly Proportioned Sample between Risk and Non-Risk Groups

Timothy H. Lee, SPS Payment Systems
Donald T. Searls, University of Northern Colorado

Linear Logistic Regression is a simple but a very powerful tool to assess the likelihood of being in one "category" for an observation with specific independent characteristic values, i.e., when the response variable is dichotomous and the data is replicated, the conditional probability, that an observation belongs to one of the two categories given independent characteristic values, can easily be estimated through Logistic Regression. For various reasons, stratified sampling, sometimes, causes a different sample proportion between the two groups from the population. Many statistical packages allow their users to adjust weights to fix this bias problem as an option in using the Logistic Procedure. The users, however, would experience more computing cost by using the option. In many cases, the purpose of the biased sampling is for computational economy and if the computing cost stays the same, using the biased sample with adjusted weights is not advantageous.

In this study, simple bias correction without using adjusted weights is explained using simulated bankruptcy data. Since the method can be used for any software without adjusting weights, computational economy can be achieved with unbiased results.

1. Introduction

Two group classification techniques are instrumental in many cases of decision making in business, finance, and marketing, etc. For example, when credit grantors extend credit, they need to assess each applicant's credit worthiness or risk for the extension of credit. In marketing analysis, they want to target more potentially responsive populations for direct mailing. These examples are typical cases where the dependent variable is binary, i.e., risk versus non-risk, or response versus non-response. Logistic regression analysis, parametric or nonparametric discriminant function analysis, and neural net are usual candidate tools for such cases. These methods are known to be comparable one to another in terms of classification accuracy. Each of these has merits and demerits depending on the user's point of view such as cost, purpose of analysis, etc. Logistic regression, for many reasons, often has been preferred to other methods, especially to discriminant function analysis. Press and Wilson (1978) made empirical applications to compare logistic regression and discriminant function analysis using breast cancer data and population change data of the U.S. They concluded that Logistic regression outperforms linear discriminant function analysis when the normality assumption is violated. Fienberg (1980), also, mentioned the superiority of logistic regression over discriminant function analysis in case of non-normal populations. In reality, the normality assumption is not easily met, especially in most of the credit or demographic profile data. One of the advantages of using the logistic regression model is that it provides the likelihood of being in one group for an observation given characteristic profile values.

Let E be an event that an observation is from one category and a vector $\underline{x}$ be the characteristic values of the observation. Then, the logistic regression model is

$$p(x) = \Pr\{E \mid \underline{x}\} = 1/[1+ \exp\{ -(\alpha + \underline{\beta}\underline{x})\}],$$

where $(\alpha, \underline{\beta})$ are unknown parameters that are to be estimated from the sample. This model is used to classify an observation into one of the two mutually exclusive categories based on $\underline{x}$.

In actual analysis, the binary dependent variable, usually coded 0 or 1 for event or non-event, is regressed on $\underline{x}$.

1.1 Sample Bias

In many cases of two group classification, the proportion of one group is far smaller than the other.

For instance, the proportion of cancer patients among the population, or the proportion of bankrupt accounts in a portfolio is observed to be very low. In such a case, analysts would rather choose stratified random sampling than simple random sampling. For instance, n observations are taken randomly from the event population, and m observations are taken from the non-event population. The sample ratio between event and non-event in such a sample is quite different from that in the population. For classification purposes, such an uneven proportion shouldn't be a problem, because a classification model developed on an unevenly proportioned sample would work as well as a model developed on an evenly proportioned sample. Such sampling scheme saves sampling cost and in using the data, later on, will bring a reduction of computing cost as well. Immeasurability is, of course, sometimes a cause of the uneven proportion. In this study, we like to consider sample bias in the sense of an uneven or distorted ratio between two mutually exclusive categories.

1.2 Model Bias

If a logistic regression model is derived based on a biased sample, the estimated probability of event given $\mathbf{x}$ would be either underestimated or overestimated even though the model has almost the same classification power as that derived from unbiased data. Let's consider, as a more detailed example, a case when bankruptcy is an event. That is, a model is developed to assess likelihood of bankruptcy given a vector of characteristic values. The risk assessment is biased if the model is developed by using biased data.

2. Analysis of Data

In this study, we used simulated bankruptcy data from Moody's Industrial Manuals 1968-1972 to expand our discussion. The data set has 4 independent variables, x_1 = (cash flow)/(total debt), x_2 = (net income)/(total assets), x_3 = (current assets)/(current liabilities), and x_4 = (current assets)/(net sales). The dependent variable is coded as 0 for bankruptcy and 1, otherwise.

For illustration, let's assume that the proportion of the event (bankruptcy) is 1/50 (=0.02) in a portfolio. A logistic model was derived using a biased development sample which has proportion of event (bankruptcy) 1/3 ($\cong$ 0.33). The parameter estimates on the biased sample were

$$(\alpha', \beta_1', \beta_2', \beta_3', \beta_4')$$

$$= (\ 2.8603, -\ 3.6938, -\ 1.7649, -\ 1.7286,\ \ 0.4760\ ).$$

Figure-1 in the Appendix presents plots between estimated risk versus observed risk for the biased sample. A smooth curve produced by the authors' robust smoother (1990) is superimposed to enhance the visual information. Figure-2 presents the same plots on an unbiased sample which has the same proportion as the population. We can observe that there is, in Figure-1, a strong linear relationship (almost a 45 degree line with some endurable noise) between observed and estimated risks, while, in Figure-2, there is no linearity between the two values and it presents a bias assessment of risk. In most cases, the bias is leaning toward over estimation. That is, when the proportion of the event is very low such as bankruptcy, the sample proportion of the event is usually far higher than the population proportion and may result in an overestimation of risk unless an adjustment is made in the process of estimation.

3. Bias Correction

We can consider two kinds of corrections, i.e., a priori adjustment and a posteri correction.

3.1 A priori Adjustment

One of the easy ways of a priori adjustment is to assign proper weights based on the sampling fraction, $f = n/N$, where, n and N are sample and population sizes, respectively. If, in the case of stratified sampling, f is 0.5 for a stratum, the corresponding sample weight $1/f$ =2 will be assigned in the estimation procedure. This kind of adjustment is allowed, in most of the commercial software, for the price of additional computing cost. To compute estimates of the parameters, Iteratively Reweighted Least Squares (IRLS) or similar methods are used. For example, IRLS for k+1 response categories is used, in SAS, as in the following:

Let $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{(k+1)j})^t$ be a multinomial vector such that

$$Z_{ij} = 1 \quad \text{if } Y_j = i$$

$= 0$ otherwise , for $j = 1, ..., n$

(In two group case, $k = 1$ and Y is a binary response variable)

Let $\mathbf{p}_j = E(\mathbf{Z}_j)$, $\mathbf{V}_j = Cov(\mathbf{Z}_j)$,
and

$\gamma^t = (\alpha_1, \alpha_2, ..., \alpha_k, \beta)$.

And, let $\mathbf{D}_j$ be the matrix of partial derivatives of $\mathbf{p}_j$ with respect to γ. Then, the estimating equation for the regression parameters is

$$\Sigma_j \mathbf{D}_j^t \mathbf{W}_j (\mathbf{Z}_j - \mathbf{p}_j) = 0,$$

where $\mathbf{W}_j = w_j \mathbf{V}_j^-$, w_j is the weight of j-th observation, and $\mathbf{V}_j^-$ is a generalized inverse of $\mathbf{V}_j$. $\mathbf{V}_j^-$ is chosen as the inverse of the diagonal matrix with $\mathbf{p}_j$ as the diagonal. The parameters are estimated iteratively as

$$\gamma'_{m+1} = \gamma'_m + (\Sigma_j \mathbf{D}'^t_j \mathbf{W}'_j \mathbf{D}'_j)^{-1} \Sigma_j \mathbf{D}'^t_j \mathbf{W}'_j (\mathbf{Z}_j - \mathbf{p}'_j)$$

Where $\mathbf{D}'_j$, $\mathbf{W}'_j$, and $\mathbf{p}'_j$ are evaluated values of $\mathbf{D}_j$, $\mathbf{W}_j$, and $\mathbf{p}_j$ at γ'_m.
If the likelihood evaluated at γ'_{m+1} is less than that evaluated at γ'_m, then γ'_{m+1} is recomputed using half the value of the second term of the right hand side.

As was discussed, by assigning proper weights, if it is allowed, or by replicating $1/f$ (to the nearest integer) times, if weighting is not allowed, the sample bias problem in risk assessment can be easily overcome with additional expense.

Our interest, however, is not in a priori adjustment but in a posteri correction. When a model is developed already and the development data is no longer available, or redevelopment causes unexpected inconvenience or cost, posterior correction based on minimal information about the population would be an economical and efficient alternative.

3.2 A posteri Correction

This approach is used to alleviate a biased risk estimation due to an uneven sampling fraction by computing a simple correction factor. For illustration, assume a situation that a probability model is derived using biased data and it is applied in an application data set. The application data is not used for the derivation of the model. We assume, further, that the proportion of the event in the application data will be approximately the same as that of the population. The probability of the event predicted will be biased and it should be corrected. To simplify the discussion, let's define the following:

p: population proportion of events

p': sample proportion of events in a biased data set

ϕ': estimated likelihood of an event for given $\mathbf{x}$ on an application data using the biased model developed on the biased data set

m: number of events observed at ϕ' in the biased data set

n: number of non-events observed at ϕ' in the biased data set

M: total number of events in the biased data set

N: total number of non-events in the biased data set

Further, let:

$f' = m/M$ (Relative frequency of event at ϕ' in the biased data set)

$g' = n/N$ (Relative frequency of non event at ϕ' in the biased data set)

Then, the likelihood of event for an observation estimated on the application data, even though the data is not biased, would be,

$$\phi' \cong m / (m + n)$$

$$= f' * M / (f' * M + g' * N) \quad \text{..... (1)}$$

Since the model was derived on the biased data, ϕ', the conditional probability given characteristic values $\mathbf{x}$, is biased although it is calculated on the application data. It always results in the same likelihood for $\mathbf{x}$ and implies the same likelihood as if it were calculated on the biased sample.

The true likelihood of event at ϕ' can be calculated by,

$$\phi = p * f / [p * f + (1-p) * g] \quad \ldots\ldots\ldots\ldots (2)$$

,where f and g are population relative frequencies for event and non-event, respectively.

The problem is how to estimate (or approximate) ϕ in (2) using ϕ' in (1).
One necessary condition that can be easily proven empirically is that

$$f' \cong f \text{ and } g' \cong g \text{ for any } \phi' \text{ and } p'.$$

From (1), using above condition,

$$[\phi']^{-1} - 1 = (g'/f') * (N/M)$$

$$\cong (g/f) * (N/M) \quad \ldots\ldots\ldots\ldots (3)$$

By multiplying (M/N) * [(1-p)/p] and adding 1 on both sides of (3),

$$\{ (\phi')^{-1} - 1\} * (M/N) * [(1-p)/p] + 1$$

$$\cong [f * p + g * (1-p)] / (f * p) \quad \ldots\ldots\ldots\ldots (4)$$

From (2) and (4), we get,

$$\phi \cong \{ [(1-\phi')/\phi'] * (M/N) * [(1-p)/p] + 1\}^{-1},$$

or by using the fact that (M/N) = [p'/(1-p')], we get

$$\phi \cong \{ [(1-\phi')/\phi'] * [p'/(1-p')] * [(1-p)/p] + 1\}^{-1}$$

The last formula is for bias correction. It shows that the biased likelihood ϕ' can be corrected easily and the only necessary information about the population is the proportion of the event. The formula was applied to the estimated likelihood of event (estimated risk) in Figure - 2 and the corrected risk and observed risk is plotted in Figure - 3. A strong linear relationship is found between the estimated risk and the observed risk, particularly for an observed risk under 20%. This is the region where most of the observed risks occur. This shows that the biased risk is corrected.

4. Discussions

As mentioned above, Logistic regression is a very popular tool in classification analysis. Especially in the two group case such as risk versus non-risk analysis, it is very instrumental in assessing risk level for an observation in a portfolio. An uneven proportion, however, will cause a biased estimation. In business applications, the size of the risk group is usually small compared to the portfolio size. For example, in developing a bankruptcy forecasting model for a portfolio, the number of bankruptcies is very low so all the bankruptcies are taken into the development sample along with a certain number of non-bankruptcies. Even though the resulting model has good separation power when measured by the Kolmogorov-Smirnov test, Apparent Error Rate, or Kull-back Leibler information value, etc., the risk measured by the model would be overly assessed. For worse scenarios, redevelopment of the model is impossible because the original data was purged, or the biased model is installed on the system already and is in production mode. In such cases, a posteri correction is very handy.

Even when the weight option is available in using statistical software, if the weight assigned to one group is too large compared to the other, such as the bankruptcy prediction case, the resulting estimates of risk may not be accurate when round-off error or wrong direction of convergence is cumulated in the process of the iterative reweighted algorithm of the logistic procedure. If such a situation is expected, both the weight assignment and the above correction algorithm can be used for a test.

Acknowledgements

The authors wish to acknowledge Dr. Sam Houston of University of Northern Colorado for his careful review of the paper.

References

Press, S.J. and Wilson, S. (1978), *Choosing between Logistic Regression and Discriminant Analysis*, Journal of the American Statistical Association, Vol. 73, 699 - 705

Fienberg, Stephen (1980), *The Analysis of Cross-Classified Categorical Data*, second ed., The MIT Press.

Lee, Timothy and Searls, Donald (1990), *Two Stage Smoothing of Scatterplots*, ASA Proceedings, Statistical Graphics Sec.

SAS user's Guide, Version 6 (1990), Vol. 2, 1071 - 1126.

Charles W. Therrien (1989), *Decision, Estimation, and Classification: an introduction to pattern recognition and related topics*, John Wiley & Sons.

Richard A. Johnson and Dean W. Wichern (1982), *Applied Multivariate Statistical Analysis*, Prentice-Hall.

Ronald H. Randles and Douglas A. Wolfe (1979), *Theory of Nonparametric Statistics*, John Wiley & Sons.

William G. Cochran (1977), *Sampling Techniques*, third ed., John Wiley & Sons.

Appendix

Figure - 1:
Estimated Risk versus Observed Risk on the biased development data set

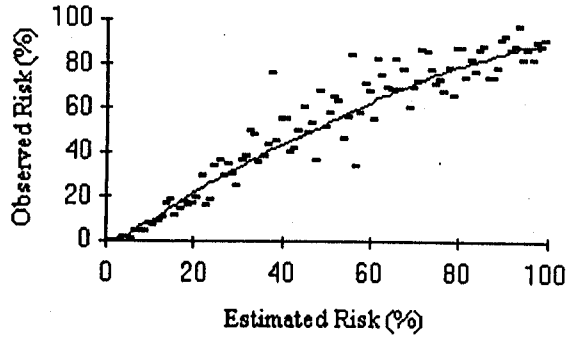


Figure - 2:
Estimated Risk versus Observed Risk on an unbiased application data set

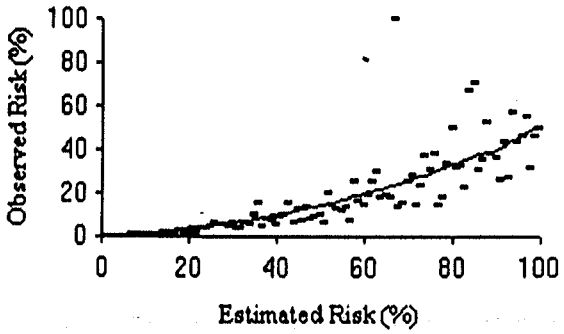
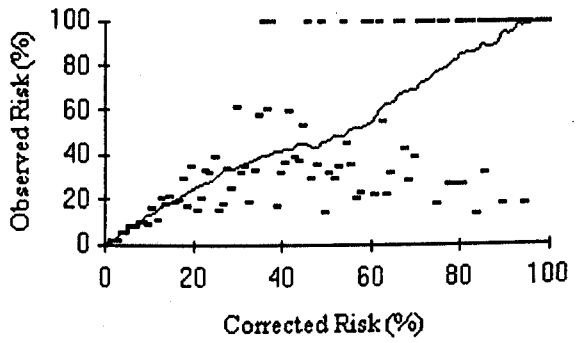


Figure - 3:
Corrected Risk versus Observed Risk on an unbiased application data set



To Path Analyze or Not To Path Analyze: Is There an Alternative Approach

Isadore Newman, The University of Akron
and
Joseph R. Marth, Bluefield College

During the past twenty years there has been a tremendous increase in the frequency of social scientists attempting to investigate phenomena that can not be studied in a laboratory. Since the ideal is to be able to explain complicated relationships in the causal sense, these social scientists have been highly attracted to sophisticated multivariate causal modeling.

Much has been written on the problems of modeling techniques such as path analysis. The concept that any research based upon ex post facto design can not assume causation (post hoc fallacy), that is correlation does not imply causation, has been widely accepted. However, some social scientists are more frequently wondering why not accept causal modeling assumptions? Do the advantages out-weigh the disadvantages? Are the concerns voiced by many statisticians really nitpicking (Cliff, 1983; Daggett & Freedman, 1985; Freedman, 1989; Huber, 1985; Kenny, 1979)?

Purpose

The purpose of this paper is to examine the underlying assumptions of path analysis and to discuss some theoretical concerns. This paper will also suggest an alternative approach that the authors believe to be more robust to the violation of some of the underlying assumptions and still is very effective in testing the overall "goodness of fit" of a theory.

Before beginning, however, a caveat is necessary. There are a number of uses for which researchers employ path analytic procedures that this paper does not deal with. For example, we are not dealing with situations where researchers use path analysis analogous to almost a stepwise model building in which the computer identifies the best fitting models. From a theoretical point of view, this has virtually all of the problems (and maybe even more) of a stepwise regression procedure, and has received much criticism because of its antitheoretical and unstable nature. This paper also does not discuss the use of path analysis for the purpose of determining which alternative models are better. Rather, discussion here is focused on the traditional intent and most conservative approach of path analysis, that of theory testing and model confirmation.

The Assumptions

The underlying premise of path analysis is that if one can meet all of the assumptions, it is justifiable to presume "causation." Therefore, this paper begins with a discussion of these assumptions. The following is a summary of the basic assumptions of path analysis identified by Bollen (1989), Freedman (1987), Dillon and Goldstein (1984), Kenny (1979), and Williams (1978):

1. Requires a theory and nomological net;
2. There is significant relationship between the variable that is assumed to be the cause and the variable that is assumed to be the effect;
3. Causal variable precedes the effect variable in time;
4. Spuriousness has been controlled...all meaningful relationships are included in the model;
5. Variables are additive and no interaction exists;
6. The weights are stable (paths), therefore no multicollinearity;
7. The distribution of residuals are the same no matter what the value of the independent variable;
8. The mean of the residual values is zero;
9. The variance of the residual values is finite;
10. The residuals of each of the variables are independent of all the other variables in the system; and
11. Endogenous variables have at least interval scale properties.

An added concern is that totally different path analytic models can produce a sufficient amount of statistical verification to justify a variety of theoretical explanations for the same variables. Also, there are concerns about the use of latent variables (Cliff, 1983), similar to the concerns of virtually all factor analytic procedures. That is, concern that latent traits, when used, are stable, meaningful, interpretable, and valid. Finally, we should further note that little is known about the effects of heteroscedasticity or autocorrelated disturbances for latent variables (Bollen, 1989).

A discussion of some of the these crucial assumptions and related concerns is presented below, followed by an alternate approach to path analysis should the researcher be unable to meet the assumptions.

The Need for Theory

In path analysis and structural equation modeling (SEM), one builds analytic diagrams that are reflective of the nomological net exposited by the theory it is intended to reflect. Therefore, one of the key underlying assumptions before doing any path analysis or SEM, traditionally, has been the necessity of theory (Bollen, 1989; Borgatta, 1969; Duncan, 1975, 1969; Heise, 1974, 1975, 1977; Williams, 1978). The purpose of theory is to explain and help understand the occurrence of natural phenomena (Kerlinger, 1973). Theory explains the causal effects among and between variables (constructs). Further, since one of the original purposes of path analysis and SEM is to assume "causal" relationships between variables which are frequently, if not always, nonmanipulable (Newman & Newman, 1992; Kerlinger, 1973), one is required to assume causation from correlational-type data. However, this does not mean you can not use path analytic procedures on experimental data. Thusly, theory is an essential component to this process. If one assumes causation which is consistent with a nomological net, one is standing on firmer ground than if one were assuming causation merely because phenomena were correlated.

Happily, when reading research which uses path analysis, there tends to be a much greater explanation of theory and the derivation of its hypotheses, and we strongly support such approaches. This is more likely to require the researcher to know the literature, to know the theory, and to think about the possible logical interrelationships of the variables.

It should also be noted that in the use of path analysis for testing theory, there are goodness of fit indices to help estimate how well the model fits the theoretically predicted relationships. Chi square and the absolute size of the residuals were initially the most frequently used goodness of fit indices. Bentler and Bonnett (1980) and Tanaka and Huba (1985), have developed goodness of fit indices, indicating that they are robust to N size. However, an article by Marsh, Balla, and McDonald (1988) mathematically demonstrates that all of the indices are really dependent to differing degrees on N size.

Time-precedence and Non-spuriousness

Kenny (1979) identified two requirements for path analysis: time precedence and non-spuriousness. These requirements tend to be design concerns in which time precedence indicates that the independent variable, which is the presumed cause of the dependent variable (endogenous variable), logically has to precede the dependent variable. For example, in a causal sense, one would expect IQ to logically precede GPA, but GPA would be less likely to logically precede IQ. Non-spuriousness can be thought of as an underlying assumption of the path analysis design, in that it assumes that the path analytic model contains all of the relevant causal variables.

Interaction

An intriguing aspect for and against the use of path analysis is that, with very few exceptions, little has been said about the issue of interaction. The underlying regression structures of path analysis are analysis of covariance regression models. One of the most important assumptions of analysis of covariance, which can not be violated with impunity, is that there is no significant interaction between the independent variable and the covariates. This means that anyone testing a simple or complex path analytic model which represents a nomological net, is making the assumption, consciously or unconsciously, that there is no interaction. One merely has to think of the social science theories and ask how many of them make that assumption.

In situations where interaction is found, for example between sex and motivation in predicting achievement, one suggested procedure for handling such interactions would have the researcher run separate analyses for males and females. It is likely that a complex path analytic design will have more than one simple first-order interaction. Actually, one would probably expect more than one second-order interaction (which is an interaction between at least two first-order interactions) or third-order interaction (which is an interaction between at least two second-order interactions) to exist in a complex path analytic design. The implications of these interactions for interpretation of path analysis is that researchers will have to consider many subset designs which can become so conditional that they become complex beyond understanding.

For non-linear second-order types of relationships a similar solution has been suggested: that a two-stage least square procedure be incorporated. However, it is interesting to ask individuals who are using path analysis to test a theory if they are in fact assuming that there is neither interaction nor a curvilinear relationship.

To the extent that the path analytic models do not reflect interactions that exist in the theoretical conceptualization, the researcher is actually committing a Type VI Error. That is, there is an inconsistency between the research question of interest and the statistical model which was written to reflect the research question (Newman, Deitchman, Burkholder, Sanders, & Ervin, 1976).

Beta Weight Interpretation

It has been well established in the statistical literature that beta weights are either non-interpretable (Kerlinger & Pedhazer, 1973; McNeil, 1993, 1992,

1991; Ward & Jennings, 1973) or are misleading and should be interpreted with extreme caution. Beta weights are more likely to be interpreted correctly if there is zero multicolinearity between the independent variables. The higher the correlation between these variables, holding everything else constant, the higher the standard deviation and the greater the instability of the weights.

The causal interpretation of a path analytic model needs predictor variables that are low or zero correlated and/or sample sizes that are very large in relation to the number of variables. If the sample sizes are so large, such as the High School and Beyond data set with 58,000 subjects, they can be considered virtual populations. That is, the more subjects per variable, the more stable these weights tend to be. Unfortunately however, when the sample size is very large, traditional tests of significance become virtually meaningless, because any slight difference will be statistically significant. (The proportion of variance accounted for can be considered or the model can be used in a more descriptive manner.)

Some approaches have dealt with the multicolinearity problem by employing measurement models along with statistical models. The measurement model uses a set of indicator variables that are conceptually factor analyzed. These factors, sometimes called latent traits, are assumed to be better measures of the underlying construct than any individual item. These underlying traits are often assumed to be stable or at least more stable than the individual items they are composed of, and therefore are thought to be more reliable and valid. However, one must also keep in mind that these factors are sample specific and may be in turn highly unstable.

Some path analytic users think that using latent traits (factors) decreases or eliminates the multicolinearity problem and reduces measurement error. This is not necessarily the case. For example, if five indicator variables for achievement are factor analyzed and five for ability level, and the ten indicator variables are not factor analyzed together, each set of five items can produce factor solutions that are highly correlated (multicolinear). In addition, five indicator variables may produce three factors when factor analyzed but only the first factor is usually used because this approach assumes the other factors are not meaningful or useful. There may be no justification for such an assumption. Another approach sometimes employed is to only use the first non-rotated factor which maximizes the variance accounted for by that one factor, but also tends to disregard the empirically identified multidimensionality on the construct.

If the sample is virtually a population size or is a population, then the model, even if not causal, can definitely be used descriptively to help explain potential relationships without ever assuming causal effect. There appears to be much less criticism of such an application of path analysis, but there is less interest in using it in this way. For example, while major economic forecasting models that have used path analysis have not held up well (McNees, 1986; Zarnowitz, 1979), they have been found to be useful in a more descriptive sense.

Testing for Underlying Statistical Assumptions

Applied statisticians and sophisticated users of path analysis such as Bollen (1989), Bentler (1987), and Freedman (1985) have pretty much agreed that one should test for certain underlying assumptions and do a pre-analysis of the data related to these assumptions before path-analysis or any statistical treatments are used. Berkane and Bentler (1987) state that BMDP provides a test for multivariate normality, detecting or eliminating outliers for EQS, and Berkane and Bentler (1987) developed a test for homogeneity of kurtosis. In addition, before doing any analysis, one should look at plots of residuals and should always cross validate to establish the stability of the prediction from sample to sample.

Some underlying assumptions are more robust than others. For example, certain assumptions of normality and homogeneity can be violated with virtual impunity if the N is large enough. However, certain assumptions of linearity, no interaction between the independent variables, and no multicolinearity are assumptions to which covariant structural models are highly sensitive (not robust). The question is, how frequently does the literature report the use of these procedures to check underlying assumptions, and why not make this a requirement of the data analysis for publication?

Corrections for Violations of Assumptions

Bollen (1989) and others (Bentler, 1987; Bentler & Dijkstra, 1985; Bentler & Lee, 1983; Freedman, 1985; Johnson, 1984; Joreskog & Sorbon, 1981; Tukey, 1954) have dealt with violations to the assumptions and have suggested solutions. For example, the use of alternate estimators such as General Least Squares (GLS), Unweighted Least Squares (ULS), Elliptical Generalized Least Squares (EGLS), Two-Stage Least Squares (2SLS), Three-Stage Least Squares (3SLS), Instrumental-Variable Estimators (IVE), and Full-Information Maximum Likelihood (FIML) are discussed by Bollen (1989). However, these techniques themselves tend to have assumptions about what the data truly look like in the population. If the researcher is correct about the nature of the distribution of data in the population and s/he picks a statistical procedure that is most appropriate for that distribution, it is obvious that his/her analysis is most likely to produce the most accurate parameter estimates. Unfortunately, however, the researcher frequently does not know what the data look like in the population

and/or is unaware of what is "causing" abnormalities in the distribution. Further, while a statistical technique may allow one to correct for anomalies, the researcher must make the assumption that the anomalies are in fact errors. Otherwise, the very corrections themselves create greater errors than no correction at all. What we are arguing is that statistical corrections for anomalies in the distribution, without considering the causes of the anomalies, is a fatal flaw in the research study. Therefore, one has to be aware of the assumptions one is making about the anomalies when one is making a correction. There is no correction which is a panacea that will replace understanding one's data.

A Simple Alternative Approach to Path Analysis for Testing Theoretical Relationships

The following is a suggested approach that is methodologically much simpler and is more robust to some of the devastating assumptions such as linearity and no interaction that are underlying assumptions of path analysis, and yet has many of the same advantages for testing a nomological net. This approach starts with theory that produces a nomological net, then identifies the logically derived hypothesis to be tested. For example, let's assume that 15 hypotheses are produced from the nomological net. Some can be interactional, repeated measures, time lagged, multiple wave, curvilinear, main effects or direct effects. Let's further assume that 13 of the hypotheses are significant in the predicted direction. One can then get an estimate, by using a Sign test, of how well these hypotheses support the overall theory (nomological net). Depending upon one's productivity and situation specifics, one may choose to do a Sign test on the directions of each individual hypothesis with no concerns for the tests of significance. Or, one can do the Sign test only on the number of significant hypotheses and compare it to the total number of hypotheses. In either case, this nonparametric test can be used to estimate the overall support of the theory. In addition, this test of significance is not dependent upon the N size, but rather on the number of hypotheses generated. It is apparent how this approach can fit well into a meta-analysis. As Pedhazer (1990) and Ward and Jennings (1973) suggest, researchers should keep their analysis simple but well thought out and have hypotheses that are derived from previous research and theory.

The authors believe much path analysis research gets lost in the complexity of the models and the sophistication of the analyses. In cases where more sophisticated analysis may be required, based upon the theory and the derived hypotheses which may infer, for example, underlying latent structures, the suggested approach would be to do:

1. A factor analysis of the variables of interest;
2. A cross validation of the factor structures to estimate stability;
3. A factor regression using the factors as predictor and criterion variables where appropriate; and
4. Cross validation on the regression equations to estimate their stability.

Needless to say, before doing any type of analysis, it is always desirable to first look at your means, standard deviations, frequencies, correlations, and residual plots before proceeding. It is this pre-analysis that helps to identify potential errors in the data, to what degree underlying assumptions have been violated, and if and what data transformations are needed or desirable. We think it is appropriate to end with a quote from Rogosa (1987): "[t]he transition of substantive theory into methods for data collections and analysis is where I think the fertile interaction between statistician and social scientist lies[,] rather than in arguing 'thumbs up' or 'thumbs down' on path analysis" (p. 185).

References

Bentler, P. M. (1987). Structural modeling and the scientific method: Comments on Freedman's critique. *Journal of Educational Statistics, 12*(2), 151-157.

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaish (Ed.), *Multivariate analysis VI,* Amsterdam: North Holland.

Bentler, P. M., & Lee, S. Y. (1983). Covariance structures under polynomial constraints: Application to correlation and alpha-type structural models. *Journal of Educational Statistics,* 8, 207-222.

Berkane, M., & Bentler, P. M. (1987). Distributions of kurtosis, with estimators and tests of homogeneity of kurtosis. *Statistics & Probability Letters,* 5, 201-207.

Blau, P., & Duncan, O. D. (1967). *The American occupational structure.* New York: John Wiley & Sons.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley & Sons.

Borgatta, E. F. (Ed.). (1969). *Sociological Methodology 1969,* San Francisco: Jossey-Bass.

Cliff, N. (1987). Comments on Professor Freedman's paper. *Journal of Educational Statistics, 12*(2), 161-164.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research,* 18, 115-126.

Crano, W. D., & Mellon, P. M. (1978). Causal influence of teachers' expectations on children's academic performance: A cross-lagged panel analysis. *Journal of Educational Psychology, 70,* 39-49.

Daggett, R., & Freedman D. (1985). Econometrics and the law: A case study in the proof of antitrust damages. In L. LeCam & R. Olshen (Eds.), *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer* (Vol. I, pp. 126-175). Belmont, CA: Wadsworth.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications.* New York: John Wiley & Sons.

Duncan, O. D. (1975). *Introduction to structural equations models.* New York: Academic Press.

Duncan, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin,* 72, 177-182.

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement.* New York: Seminar Press.

Freedman, D. (1981). Some pitfalls in large econometric models: A case study. *Journal of Business, 54,* 479-500.

Freedman, D. (1985). Statistics and the scientific method. In W. Mason & S. Fienberg (Eds.), *Cohort analysis in social science research: Beyond the identification problem* (pp. 345-390). New York: Springer.

Fox, J. (1987). Statistical models for nonexperimental data: A comment on Freedman. *Journal of Educational Statistics,* 12(2), 161-164.

Heise, D. R. (Ed.). (1974). *Sociological methodology 1975.* San Francisco: Jossey-Bass.

Heise, D. R. (1975a). *Causal analysis.* New York: Wiley.

Heise, D. R. (Ed.). (1975b). *Sociological methodology 1976.* San Francisco: Jossey-Bass.

Heise, D. R. (Ed.). (1977). *Sociological methodology 1977.* San Francisco: Jossey-Bass.

Huber, P. (1985). Projection pursuit. *Annals of Statistics, 13,* 435-475.

Johnson J., (1984). *Econometric methods.* New York: McGraw Hill.

Joreskog, K. G., & Sorbom, D. (1981). *LISREL - Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: International Educational Services.

Kenny, D. (1979). *Correlation and causation.* New York: Wiley.

Kerlinger, F. N. (1973). *Foundations of behavioral research (2nd edition).* New York: Holt, Rinehart, and Winston.

Marsh, H., Balla J., & McDonald R. P. (1988). Goodness of fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

McNees, S. (1986). Forecasting accuracy of alternate techniques: A comparison of U.S. macroeconomic forecasts. *Journal of Business and Economic Statistics, 4,* 5-24.

McNeil, K. (1991). *Reasons for not interpreting regression coefficients.* Paper presented at the meeting of the Southwest Education Research Association, San Antonio, Texas.

McNeil, K. (1992). Conditions for interpretation of regression weights. *Multiple Linear Regression Viewpoints, 19*(1), 37-43.

McNeil, K. (1993). Cautions and conditions for interpreting weighting coefficients, *Midwestern Educational Researcher, 6*(1), 11-14.

Newman I., Deitchman R., Burkholder J., Sanders R., & Ervin L. (1976). Type VI Error: Inconsistencies between the statistical procedure and the research question. *Multiple Linear Regression Viewpoints, 6*(4), 1-19.

Newman, I., & Newman, C. (1992). *Conceptual statistics for beginners* (2nd edition). Lanham, MD: University Press of America.

Noonan R., & Wold, H. (1983). Evaluating school systems using partial least squares. *Evaluation in Education, 7*, 219-364.

Pedhazer E. J. (1982). *Multiple regression in behavioral research.* New York: Holt, Rinehart, and Winston.

Rogosa, D. R. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88*, 245-258.

Rogosa, D. R. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics, 12(2)*, 185-195.

Simon, H. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association, 49*, 467-479.

Tanka, T. S., & Huba, G. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38*, 197-201.

Tukey, J. W. (1954). Causation, regression, and path analysis. In O. Kempthorne, T.A. Bancroft, J. W. Gowen, & J. L. Lush (Eds.), *Statistics and mathematics in biology.* Ames: Iowa State College Press.

Ward J. H., & Jennings, E. (1973). *Introduction to linear models* Inglewood Cliffs, NJ: Prentice Hall

Williams, J. D. (1978). Path analysis from a regression perspective [Monograph]. *Multiple Linear Regression Viewpoints, 2*(2), 1-81.

Wold, H. (1985). Partial least squares. In J. Kotz & N. L. Johnson (Eds.), *Encyclopedia of social sciences* (Vol. 6). New York: Wiley.

Wold, H. (1987). Response to D. A. Freedman. *Journal of Educational Statistics, 12(2)*, 202-204.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.

Wright, S. (1960). Path coefficients and path regressions. *Biometrics, 16*, 189-202.

Zarnowitz, V. (1979). An analysis of annual and multiperiod quarterly forecasts of aggregated income, output, and price level. *Journal of Business, 52*, 1-34.

The *p*-Problem with Forward Selection Stepwise Regression: Algorithm for Controlling Type I Errors

T. Mark Beasley, St. John's University, New York
Dennis W. Leitner, Southern Illinois University at Carbondale

The use of forward selection stepwise regression has been criticized for both interpretive misunderstandings and statistical aberrations. A major statistical problem with stepwise regression and other procedures that involve multiple significance tests is the inflation of the Type I error rate. Common approaches to control the family-wise error rate (e.g., the Bonferroni and Sidak corrections) are based on the assumptions of independent tests which typically reduce power. Because the presence of correlated predictors is a more realistic situation, other algorithms based on the average correlation in the predictor matrix have been proposed. The present study proposes an algorithm based on the maximum eigenvalue and the determinant of the predictor matrix for controlling the family-wise Type I error rate for multiple, correlated tests in forward selection regression under the complete null hypothesis. A Monte Carlo simulation with 5,000 replications was performed to demonstrate the effectiveness of the proposed algorithm.

Most users of multiple regression techniques in educational research are attempting to reduce a set of k predictor variables in order to report a simplified model. Typically, if a set of k predictors are regressed on a dependent variable, Y, only those predictors that are found statistically significant will be considered substantively valuable. Furthermore, because of the nature of many educational and psychological measurement scales, researchers are less likely to estimate regression coefficients as a way of interpreting substantive findings. Rather, F-ratios or p-values are used in a dichotomous decision process such that the relationship between a predictor and a criterion variable is "significant or not" (e.g., Thompson, 1989b). Furthermore, it is possible to have a statistically significant model (i.e., significant full model R^2) when the component variables are individually nonsignificant in either a zero-order or partial manner. However, educational researchers are not likely to consider such a model in the development of theory. Therefore, the forward selection procedure of stepwise regression became popular among educational researchers because it begins with significance tests of zero-order correlations and proceeds to more complex models.

For several years now, applied statisticians (e.g., Thompson, 1989a; Wilkinson, 1979) have been calling attention to the abuses of stepwise regression in its common use by less statistically sophisticated researchers. But theses and dissertations continue to step (unwisely) across the desks of graduate educators, and articles with many of these same problems continue to appear in print. It is hoped that elaborating these limitations and proposing new methods for using stepwise regression will bring about its more appropriate use. Three statistical procedures are considered under the rubric of stepwise regression: Forward selection; backward elimination; and true stepwise (Draper & Smith, 1981). Specifically, the forward selection procedure forms a model from the set of k dependent variables by first selecting the single best predictor. The second best predictor is then chosen by the criteria of strongest contribution to the prediction of Y, while controlling for the effects of the first predictor entered. Thus, the first step involves k simultaneous tests of zero-order correlations, while the second step involves $(k - 1)$ simultaneous tests of first-order semi-partial correlations (Aitkin, 1974). The process continues so that at each step the variable selected for inclusion significantly increases the prediction of Y (i.e., full model R^2).

The use of the various stepwise regression procedures has been criticized for many interpretative misuses and statistical aberrations. First, researchers often interpret the final solution of a reduced set of g predictors as being the best subset of predictors overall and of that size. Also, there is a tendency to confuse the order of entry and variable importance (Huberty, 1989). Stepwise procedures suffer from the use of the largest partial F as a test of a potential entry variable which is not in the regression model at that stage. The correct null sampling distribution for this test in not the ordinary F distribution, but is a partial F distribution which is very difficult to obtain. (Draper & Smith, 1981). Moreover, researchers often proceed to test each stage in a stepwise regression as if the partial F distribution does not exist and as if the test at that step is the only test that has or will occur. Furthermore, the

degrees of freedom (*df*) used for these tests are often incorrect. For example, in the forward selection procedure the *df*'s used for the first step, a test of zero-order correlations, is (N - 2), while (N - k - 1) would be more appropriate. These considerations, in general, tend to inflate the probability of at least one Type I error (i.e., the probability of forming an erroneous model).

Another interpretative problem arises when two or more predictor variables are highly correlated. In such situations, there is a strong probability that one of the variables will absorb the majority of the other variables' prediction power and therefore cause their exclusion from subsequent models. Not only does a set of correlated predictors lead to potential substantive misinterpretations, it also makes estimating the probability of a Type I error more complex. Thus, due to multiple tests, incorrect *df*s, misunderstood partial F tests, and correlated predictor variables, it is difficult to determine the correct Type I error rate in stepwise regression. To compound these problems, the p-value associated with each variable entered stepwise into a regression equation (except for the final step) is incorrect in many canned statistical packages.

MULTIPLE TESTING AND THE TYPE I ERROR RATE

As with any statistical procedure, two kinds of inferential errors can be made. A Type I error occurs if a variable is selected when the population regression weight is zero. A Type II error occurs when a variable is not selected when it has a non-zero population regression weight. Many educational researchers adopt one of the traditional fixed significance levels (i.e., α = .05 or .01) when evaluating an F-ratio. This significance level determines the *Type I error rate* for each test independently. However, it is rare that educational researchers test a single hypothesis. Several variables and multiple significance tests are common. Thus, a researcher must consider the probability of committing a Type I error when multiple hypotheses are tested (i.e., the *family-wise error rate*).

In the context of post-hoc tests in the analysis of variance (ANOVA), the true family-wise Type I error rate (α_T) for k independent (i.e., orthogonal) tests with the same alpha level (α_i) is defined by the following equation:

$$\alpha_T = 1 - (1 - \alpha_i)^k \ , \qquad (1)$$

assuming the *complete null hypothesis* (i.e., all groups have identical means). Thus, the *family size* of the tests performed is equal to k. In order to return the Type I error to the nominal alpha (α_i), one could adjust α_i by the Sidak method:

$$\alpha_{adj} = 1 - (1 - \alpha_i)^{1/k} \ . \qquad (2)$$

This correction would yield an alpha level smaller than the nominal alpha, but over the course of multiple tests, this adjusted alpha (2) is expected to yield a Type I error rate equivalent to the nominal alpha, α_i.

Similarly, the forward selection method in stepwise regression conducts no less than k simultaneous tests of significance as if multiple tests are not performed. That is, the first predictor is selected by the largest zero-order correlation of all k variables without consideration for the number of tests being conducted. Thus, if an educational researcher using forward selection regression were to commit a Type I error under the complete null hypothesis (i.e., all k zero-order correlations between Y and the predictors were null), it would occur on the first step. That is, when all predictors are not correlated with the dependent variable, testing the maximum of the k zero-order dependent variable-predictor correlations determines the Type I error rate of the forward selection procedure. Thus, assuming independent predictors, the probability of a Type I error on the first step is equal to (1). To adjust α_T so that the Type I error rate returned to the nominal alpha (α_i), one could assume the family size is equal to k and adjust α_i with (2). However, if the k predictors were all perfectly correlated, then the family size would be equal to one (k = 1) and the Type I error rate would equal the nominal alpha (i.e., $\alpha_T = \alpha_i$). In the more realistic situation of correlated predictors, the solution for the correct Type I error rate is considerably more complex and requires the integration of the correlated F distribution (Pope & Webster, 1972). Furthermore, only limited tables of critical values are available (e.g., Games, 1977), while a few Monte Carlo approximations based on averaged correlations have been proposed (i.e., Krishnaiah & Armitage, 1965; Pohlmann, 1979).

For example, Pohlmann (1979) proposed a method based on the average squared correlation in the predictor matrix to control the Type I error in forward selection regression. To elaborate, a value, c, can estimate *family size* and substitute for k in (2) in order to control the *family-wise* Type I error rate. Pohlmann suggested the following function:

$$c = k - (k - 1)\bar{r}_x^2, \qquad (3)$$

where k equals the number of predictors and $\bar{r}_x^2$ equals the averaged squared inter-predictor correlation. Pohlmann also suggested correcting $\bar{r}_x^2$ by using a less biased estimate of the squared correlation based on the McNemar (1969) shrinkage formula. Initially, each squared correlation in the predictor matrix is corrected by:

$$\hat{r}_{ij}^2 = 1 - (1 - r_{ij}^2)\frac{(N-1)}{(N-2)} \ , \qquad (4)$$

where N equals the number of cases and r_{ij}^2 equals the square of the ijth element of the predictor matrix. Then $\bar{r}_x^2$ is calculated by:

$$\bar{r}_x^2 = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \hat{r}_{ij}^2}{k(k-1)/2} \qquad (5)$$

and entered into (3). However, Pohlmann's study simulated cases in which all correlations within the predictor matrix were equal which is an unrealistic expectation. That is, a variety of correlation patterns may yield the same average squared correlation, but it is not likely that the family-wise Type I error rates would be equal for these matrices.

PROPOSED ALGORITHM FOR ESTIMATING FAMILY SIZE

To consider another perspective, however, the appropriate Type I error rate may approach (1) with b orthogonal factors rather than these algorithms based on the average correlation of k predictors. To elaborate, one possible approach to the p-value problem would be to perform a principal component analysis (PCA) on the predictor correlation matrix and extract b orthogonal components. In fact, it can easily be shown that such a linear transformation will not affect the full model R^2. That is, if all k variables and the k components extracted from the predictor matrix are used as separate models to predict a criterion variable, Y, then both models would have the same full model R^2. The expected Type I error rate when using the k orthogonal principal components, however, will equal (1) for the first step of a forward selection stepwise regression. Thus, decomposing the set of k predictors into b orthogonal components and modifying algorithms for correlated predictors may provide a better approximation of the family-wise Type I error rate. Importantly, this indicates a relationship between the transformation matrix and the family-wise Type I error rate. Thus, it is proposed that the maximum eigenvalue (λ_{max}) from an unrotated principal components analysis and the determinant, $|\mathbf{P}|$, of the predictor correlation matrix, $\mathbf{P}$, is related to the proportion of Type I errors on the first-step, which defines the probability of forming an erroneous model under the complete null hypothesis.

The eigenvalues of a correlation matrix, $\mathbf{P}$, are commonly used as indices of the number of factors that underlie a correlation matrix (e.g., Kaiser, 1970). Furthermore, the maximum eigenvalue provides an index for the proportion of variance accounted for by the largest principal component, the average correlation of $\mathbf{P}$, and the number of underlying factors (Tatsuoka, 1988). The determinant of a correlation matrix, $|\mathbf{P}|$ has been used in establishing the independence of variables in PCA (Nagarsenker, 1976). The determinant of the covariance matrix, $|\mathbf{C}|$, gives the generalized variance (Tatsuoka, 1988), and the determinant of the correlation matrix, $|\mathbf{P}|$, is equal to $|\mathbf{C}|$ divided by the determinant of the diagonal variance matrix $|\mathbf{V}|$,

$$|\mathbf{P}| = \frac{|\mathbf{C}|}{|\mathbf{V}|} . \qquad (6)$$

Thus, it follows that the generalized proportion of variance in $\mathbf{P}$, that is the generalized $\mathbf{R}^2$, is equal to:

$$\mathbf{R}^2 = 1 - \frac{|\mathbf{C}|}{|\mathbf{V}|} = 1 - |\mathbf{P}| . \qquad (7)$$

Therefore, in combination λ_{max} and $|\mathbf{P}|$ provide rather unique information about the inter-correlation of the predictor matrix. Specifically in PCA, λ_{max} divided by k gives the proportion of variance in $\mathbf{P}$ accounted for by the first and largest principal component. However, λ_{max} is known to always be greater than one even in random data matrices (Horn, 1965). In fact, when the variables are independent and all off-diagonal elements in $\mathbf{P}$ are zero then $\mathbf{P}$ is an Identity matrix, $\mathbf{I}$, and the expected value of λ_{max} equals one,

$$\lim_{\mathbf{P} \to \mathbf{I}} \lambda_{max} = 1 . \qquad (8)$$

Therefore, subtracting one from λ_{max} and dividing by k would provide a corrected proportion of variance for the largest principal component

$$\frac{\lambda_{max} - 1}{k} \qquad (9)$$

Also, if the variables are independent, then the determinant, $|\mathbf{P}|$, equals one,

$$\lim_{\mathbf{P} \to \mathbf{I}} |\mathbf{P}| = 1 . \qquad (10)$$

Although it is left undefined because such a matrix is not invertable, one can imagine that if all predictor variables were perfectly correlated, then λ_{max} would equal k. That is, the limit of λ_{max} as all the elements of $\mathbf{P}$ approach unity is k:

$$\lim_{\mathbf{P} \to 1} \lambda_{max} = k . \qquad (11)$$

Furthermore, since the product of the eigenvalues must equal the determinant, then under the same conditions specified for (11), the limit of $|\mathbf{P}|$ equals zero as λ_{max} approaches k :

$$\lim_{\mathbf{P} \to 1} |\mathbf{P}| = 0 \quad \text{and} \quad \lim_{\lambda_{max} \to k} |\mathbf{P}| = 0 . \qquad (12)$$

Given conditions (11) and (12), all predictors are perfectly correlated and there is only one "true" variable and the *family size* (denoted as c) should be equal to one, which can be described as:

$$c = k - (k - 1) \quad (13)$$

Thus, $(k - 1)$ multiplied by (9) results in the proportion of $(k - 1)$ that should be subtracted from k; however, c also depends on the correlations in **P** whose generalized estimate comes from $|\mathbf{P}|$. Thus, $(k - 1)$ should be multiplied by (9) and (7). Therefore, c can be estimated by:

$$c = k - \frac{(k - 1)(\lambda_{max} - 1)(1 - |\mathbf{P}|)}{k} . \quad (14)$$

Thus under the conditions set in (8), (10), (11), and (12), as the relationship among the predictor variables approaches perfect multicollinearity, the estimat of family size in (14) approaches one. Also if the k predictors are independent then (14) equals k. Therefore, if a researcher can use k, λ_{max}, and $|\mathbf{P}|$ to estimate the independence of the predictors in **P** with c, then (14) could be substituted for k in equation (2) and used as an estimate of family size to adjust α_T so that it approximates the nominal alpha. Thus in the present study, a Monte Carlo simulation of a forward-selection stepwise procedure with no expected correlation between the dependent variable, Y, and the k predictors was used to estimate the correct Type I error rate (p-values) for k = 2, 3, 4, 5, 7, and 10 correlated variables under various inter-predictor correlation conditions. From these results, the proposed formulation of c (14) was substituted for k in (2) to determine whether it was useful in controlling the Type I error rate.

For comparison purposes, Pohlmann's (1979) algorithm (3) was also used. The Appendix provides numerical examples that demonstrate the differences between the two methods.

METHODS

Simulation Procedure

A Monte Carlo program was written in SAS/IML (SAS Institute, 1990) to simulate the forward selection process of stepwise regression. Initially, the RANNOR fucnction, which provides a pseudo-random clock generated values, was use to generate a normally distributed predictor matrix, **X**, with dimensions of n rows (cases) and k columns (variables). All predictor means were equal to zero and all variances were equal to one. Then by using the fundamental postulate of PCA (Tatsuoka, 1988) and a method described by Kaiser and Dickman (1962), a $k \times k$ matrix of principal component coefficients, **F**, was derived from a prespecified predictor correlation matrix, **P** and pre-multiplied by the transpose of **X** to create a transformed data matrix $\mathbf{Z}_p$ that simulates **P** (see Beasley, 1994):

$$\mathbf{Z}_p = \mathbf{F}\,\mathbf{X}^t \quad (15)$$

Then a normally distributed dependent variable vector, **Y**, was randomly generated and concatenated with the transpose of $\mathbf{Z}_p$ to form the entire data matrix, **M**. Thus, although there was correlation among the k variables in **P**, there was no expected correlation between the predictor variables and Y. This process was replicated 5,000 times. Since an infinite number of correlation matrices can be simulated, various combinations of λ_{max} and $|\mathbf{P}|$ were used for each level of k. Tables 1, 2, and 3 in the Results section reference the values of λ_{max} and $|\mathbf{P}|$ that were imposed on **X**. The number of predictors was manipulated from k = 2, 3, 4, 5, 7, and 10. The number of cases was held constant at a fairly large number of N = 200 in order to avoid extreme shrinkage of R^2 (Harris, 1975) and to reduce the residual error in the transpose of $\mathbf{Z}_p$ as it simulates the predictor correlation matrix, **P**.

Test Procedures

Under conditions of the complete null hypothesis, if an erroneous model is to be formed (i.e., Type I error committed) using a forward selection procedure then it will occur on the first step. Furthermore most packaged stepwise programs (i.e., SAS STEPWISE, SPSS REGRESSION) perform the first entry with $(N - 2)$ dfs. Therefore, the maximum zero-order correlation in the predictor column of **M** was tested. If the calculated $F(1, 180)$ exceeded the critical values for F at the α = .05 level of significance, then it was counted as a Type I error. The number of rejections divided by the 5,000 replications served as empirical p-values and estimates of the true family-wise Type I error rate, α_T. The results of this procedure were used to help estimate family size, c. That is, if α_T and α_i are known then family size, c, can be solved as follows:

$$c = \frac{ln(1 - \alpha_T)}{ln(1 - \alpha_i)} , \quad (16)$$

where ln refers to the Naperian logarithm.

The expected values of k, λ_{max}, and $|\mathbf{P}|$ using the formula described in (14) were regressed on c derived from the simulations and (16) to investigate the goodness of fit. These results were also compared to the results of Pohlmann's (1979) algorithm (3). Furthermore, the effectiveness of (14) in controlling the family-wise Type I error rate was assessed by substituting these estimates of c for k in (2) to set a more stringent α_i in each simulation. These corrected Type I error rates were compared to the nominal alpha of .05..

RESULTS

Using the expected values of k, λ_{max}, and $|\mathbf{P}|$, the family size estimates of c from (14) were regressed on the empirical values of c derived from the proportion of rejections at the $\alpha = .05$ level of significance during the 5,000 replications. Thus, the following model was tested

$$c_{emp} = b_1 k + b_2 \frac{(k - 1)(\lambda_{max} - 1)(1 - |\mathbf{P}|)}{k} \quad (17)$$

with the intercept restricted to zero and the coefficients $b1$ and b_2 restricted to one. The model R^2 with these restriction was 0.9858. Figure 1 (upper panel) shows a scatter plot of this analysis with different elements for each value of k. The model R^2 when using Pohlmann's (1979) algorithm (3) based on averaged squared correlations was 0.9147. A scatter plot of that regression is shown in the lower panel of Figure 1. The diagonals on each panel represent a perfect fit of the expected and empirical values of family size. As can be seen, many more estimates of family size, c, deviate from the perfect fit diagonal for the Pohlmann's average squared correlation estimate of c as compared to the current proposed algorithm. Using a dependent t-test for correlations, the proposed correction (14) was found to be significantly better than Pohlmann's estimate of c, $t(67) = 9.70$, $p < .001$.

Tables 1, 2, and 3 show the expected values for the average squared correlation within the predictor matrix, $\mathbf{P}$, the maximum eigenvalue (λ_{max}), and the determinant of $\mathbf{P}$, $|\mathbf{P}|$. These tables also show the empirical values of the family-wise Type I error rate (Empirical p-values), the estimated value of family size, c from (14), and the corrected p-values after controlling Type I errors with (14).

As can be seen by looking across Tables 1, 2, and 3, when the number of predictors increased from k = 2 to 10 the expected increase in the family-wise Type I error rate also occurred. Also, by examining the first entry for any number of predictors (k), when the average squared correlation of the predictor matrix is zero, the empirical p-values approximate their estimated value from (1). For example, for $k = 4$ independent predictors(i.e., $E(\bar{\rho}^2) = 0$), the expected family size is four. Using (1) the expected family-wise Type I error rate under the complete null hypothesis is 0.1855. In comparison, the simulation in this study estimated the family-wise Type I error rate with an empirical p-value of 0.1870. From (16), the estimated family size is c = 4.0361 (see Table 2, upper panel). One can also see by looking within any Table that as the expected average squared correlation increases the Type I error rate and family size. Yet, some matrices with the same average

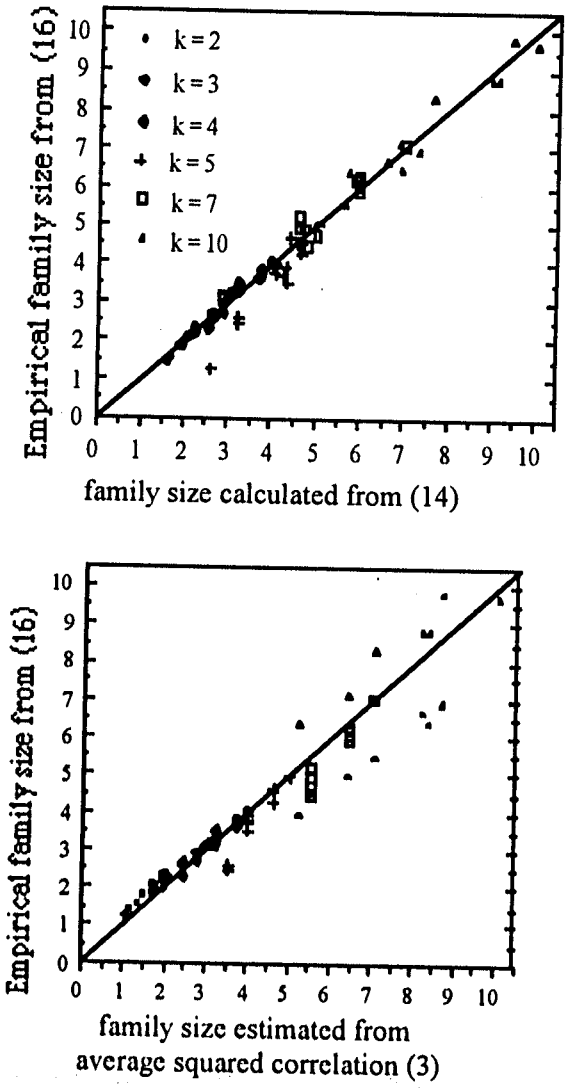


Figure 1. Empirical family size, c, derived from (16) as a function of the estimated family size from (14; upper panel) and from the average squared correlation (3; lower panel).

ρ^2 have different values for λ_{max} and $|\mathbf{P}|$ and more importantly different empirical proportions of Type I errors. This is most notable in Table 3 with $k = 10$. It is important to note that when corrected with (14) the Type I error rates (corrected p-values) are reasonably close to the nominal alpha of .05 regardless of these discrepancies. Corrections based on average squared correlations (i.e., Pohlmann, 1979), however, would correct these discrepant configurations in the same manner. Thus, logically as well as statistically, the correction formula (14) provides a better adjustment for controlling Type I errors for multiple, correlated tests.

In any Monte Carlo study, one must consider the sampling error of the simulation process. Based on the nominal alpha of $\alpha = .05$ and 5,000 replications, the standard error of each estimate is $s_e = .003$, which is used as a general heuristic to evaluate the proposed procedure. Although several corrected *p*-values exceed the + 2 standard error range, most are within the range of acceptability set by Bradley (1978). Explanation for these aberrations for the currently proposed correction may be twofold. One problem may be that some correction for sample size is necessary. Since sample size was held constant in this study, it should not have presented a serious problem. However, this possibility warrants further attention. A second problem is consistent with technical issues involving multicollinearity, in that the use of highly correlated predictor matrices yields extremely small determinants. In this case the accuracy of estimating such small values present a serious computational problem. That is, slight estimation errors can lead to rather large computational errors.

Table 1.

Expected values for the population average r^2, maximum eigenvalue (l_{max}) and determinant ($|\mathbf{P}|$) with empirical Type I error rate, estimated *c* from (16) and corrected Type I error rate (14) for $k = 2$ and 3 predictors.

$E(\bar{\rho}^2)$	$E(\lambda_{max})$	$E(\lvert\mathbf{P}\rvert)$	Empirical *p*-value	*c* (16)	Corrected *p*-value (14)
k = 2					
0.00	1.0000*	1.0000	0.0962	1.9719	0.0506
0.01	1.1000*	0.9900	0.0998	2.0497	0.0518
0.09	1.3000*	0.9100	0.0934	1.9116	0.0496
0.25	1.5000*	0.7500	0.0892	1.8215	0.0510
0.49	1.7000*	0.5100	0.0896	1.8301	0.0532
0.64	1.8000*	0.3600	0.0760	1.5410	0.0454
0.81	1.9000*	0.1900	0.0714	1.4442	0.0438
k = 3					
0.00	1.0000*	1.0000	0.1474	3.1089	0.0502
0.09	1.5695	0.7609	0.1296	2.7061	0.0496
	1.5984	0.7826	0.1394	2.9268	0.0494
	1.6000*	0.7840	0.1300	2.7150	0.0462
0.25	1.8922	0.2910	0.1120	2.3158	0.0452
	1.9860	0.4692	0.1286	2.6837	0.0516
	2.0000*	0.5000	0.1250	2.6033	0.0540
0.49	2.3658	0.0700	0.1052	2.1670	0.0532
	2.3986	0.2531	0.1090	2.2500	0.0504
	2.4000*	0.2160	0.1122	2.3202	0.0492
0.64	2.5885	0.0384	0.0910	1.8601	0.0462
	2.6000*	0.1040	0.1004	2.0627	0.0510

Note. * indicates that all correlations in **P** are equal.

Table 2.
Expected values for the population average r^2, maximum eigenvalue (l_{max}) and determinant ($|\mathbf{P}|$) with empirical Type I error rate, estimated c from (16) and corrected Type I error rate (14) for $k = 4$ and 5 predictors.

$k = 4$					
$E(\bar{\rho}^2)$	$E(\lambda_{max})$	$E(\lvert\mathbf{P}\rvert)$	Empirical p-value	c (16)	Corrected p-value (14)
0.00	1.0000*	1.0000	0.1870	4.0361	0.0494
0.09	1.7926	0.5832	0.1744	3.7363	0.0470
	1.8016	0.5439	0.1706	3.6467	0.0474
	1.8964	0.6481	0.1746	3.7410	0.0478
	1.9000*	0.6517	0.1790	3.8452	0.0552
0.25	2.2670	0.1188	0.1548	3.2788	0.0512
	2.4150	0.1042	0.1508	3.1868	0.0528
	2.4995	0.3019	0.1542	3.2650	0.0582
	2.5000*	0.3125	0.1650	3.5155	0.0572
0.64	3.3696	0.0011	0.1042	2.1453	0.0538
	3.3984	0.0238	0.1106	2.2851	0.0520
	3.4000*	0.0272	0.1104	2.2807	0.0546
0.00	1.0000*	1.0000	0.2252	4.9743	0.0460
0.09	2.0462	0.3866	0.2068	4.5168	0.0516
	2.0558	0.4449	0.2078	4.5414	0.0524
	2.0954	0.3589	0.1970	4.2774	0.0514
	2.1946	0.5221	0.2112	4.6252	0.0486
	2.2000*	0.5282	0.2124	4.6549	0.0508
0.25	2.7652	0.0081	0.1648	3.5109	.0486
	2.8100	0.0768	0.1800	3.8689	0.0556
	2.9094	0.0112	0.1728	3.6985	0.0600
	2.9914	0.1745	0.1758	3.7693	0.0516
	3.0000*	0.1875	0.1818	3.9118	0.0628
0.64	4.1957	0.0039	0.1250	2.6033	0.0560
	4.2000*	0.0067	0.1184	2.4568	0.0598
0.98	4.9600	0.0001	0.0610	1.2271	0.0328

Note. * indicates that all correlations in **P** are equal.

Table 3.
Expected values for the population average r^2, maximum eigenvalue (l_{max}) and determinant ($|\mathbf{P}|$) with empirical Type I error rate, estimated c from (16) and corrected Type I error rate (14) for $k = 7$ and 10 predictors.

$E(\bar{\rho}^2)$	$E(\lambda_{max})$	$E(\lvert\mathbf{P}\rvert)$	Empirical p-value	c (16)	Corrected p-value (14)
$k = 7$					
0.000	1.0000*	1.0000000	0.3024	7.0206	0.0476
0.090	2.3742	0.1097200	0.2664	6.0396	0.0548
	2.5193	0.1764100	0.2708	6.1569	0.0570
	2.5539	0.2311600	0.2738	6.2373	0.0546
	2.7745	0.3024000	0.2720	6.1890	0.0596
	2.8000*	0.3294200	0.2602	5.8755	0.0576
0.250	3.3159	0.0016268	0.2138	4.6896	0.0544
	3.5875	0.0004800	0.2040	4.4481	0.0486
	3.6922	0.0125600	0.2234	4.9291	0.0564
	3.7627	0.0001200	0.2072	4.5266	0.0540
	3.9779	0.0554100	0.2344	5.2072	0.0660
	4.0000*	0.0625000	0.2232	4.9241	0.0558
0.640	5.7938	0.0002600	0.1492	3.1501	0.0586
	5.8000*	0.0003700	0.1466	3.0906	0.0574
$k = 10$					
0.000	1.0000*	1.0000000	0.3920	9:7007	0.0456
0.153	2.3770*	0.5333101	0.3960	9.8289	0.0512
	4.0039	0.0003575	0.2986	6.9147	0.0516
0.187	2.6830*	0.4163279	0.3634	8.8045	0.0512
	4.4190	0.0000643	0.2794	6.3882	0.0514
0.205	2.8450*	O.3608996	0.3636	8.8107	0.0484
	4.7727	0.0000260	0.2900	6.6771	0.0532
0.327	3.9430*	0.1116766	0.3474	8.3206	0.0624
	5.8601	0.0000005	0.2462	5.51004	0.0548
0.400	4.6000*	0.0463574	0.3062	7.12708	0.0558
	6.5298	4.49e-8	0.2268	5.01464	0.0536
0.532	5.7880*	0.0062336	0.2784	6.36115	0.0614
	7.4954	2.84e-9	0.1840	3.96428	0.0550

Note. * indicates that all correlations in **P** are equal.

DISCUSSION

The behavioral science literature is replete with "significant" findings that fail the ultimate test of replication (Pedhazur, 1982; Rosnow & Rosenthal, 1989). One explanation for this conundrum lies in the family-wise Type I error rate that increases when stepwise regression or other multiple testing procedures are used. Faced with the problem of multiple tests that may be correlated, the researcher should take some action to correct the Type I error rate. Possible approaches to this problem include:

a). Prior to performing a stepwise regression, conduct an omnibus test with all potential predictors in the model.

b). When searching for a significant subset of predictors, use stepwise methods with backward elimination

c). When searching for a reduced subset of predictors through stepwise methods, perform a PCA and extract orthogonal components and use (1) to correct the family-wise Type I error rate.

d). In any multiple test situation, use one of several simultaneous inference tests (e.g., Games, 1977; Schafer, 1992; Schafer & Macready, 1975) to control Type I errors.

e). Use the Bonferroni inequality, however, one may over-correct the probability of a Type I error and lose power.

f). Use the algorithm (14) suggested here if multiple, correlated test are being performed.

It should be noted that there are practically an infinite number of configurations a correlation matrix can assume; therefore, there is no way to exhaust those possibilities. Therefore, these findings are limited to the specific correlation matrices simulated. Thus, although extensive replications of this study are needed to assume the generality of these findings, it is not unreasonable to assume that the proposed algorithm (14) will work in other situations.

Although the family-wise Type I error correction suggested here has been framed in terms of the forward selection procedure of stepwise regression, there is no reason for its exclusion from other situations that involve a single dependent variable and multiple tests that are correlated. For example, a set of nonorthogonal contrasts for an ANOVA, although based on coded vectors for means have correlations coefficients associated with them. Therefore, a matrix of correlations among contrasts could be analyzed with (14). In conclusion, the suggested algorithm shows adequate control of the family-wise Type I error rate and is based on more complete information than estimates based simply on the average squared correlation. Yet, in the results the suggested correction sometimes deviated from the nominal alpha. Thus, further investigation will focus on manipulating sample sizes and using a shrinkage correction for the determinant of the predictor correlation matrix.

REFERENCES

Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics, 16*, 221-27.

Beasley, T. M. (1994). CORRMTX: Generating correlated data matrices in SAS/IML, *Applied Psychological Measurement, 18*, 95.

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

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Games, P. A. (1977). An improved table for simultaneous control on *g* contrasts. *Journal of the American Statistical Association, 72*, 531-534.

Harris, R. (1975). *A primer of multivariate statistics*. New York: Academic Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Huberty, C. J. (1989). Problems with stepwise methods - better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 1). Greenwich, CT: JAI Press.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35*, 401-415.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27*, 179-182.

Krishnaiah R. K., & Armitage, R. B. (1965). *Probability integrals of the multivariate **F** distribution, with Tables and applications*. (Report No. ARL 65-236). Wright-Patterson AFB, OH: U. S. Air Force.

McNemar, Q, (1969). *Psychological statistics*. New York: Wiley.

Nagarsenker, B. N. (1976). The distribution of the determinant of correlation matrix useful in principal components analysis. *Communications in Statistics: Simulation and Computation, B5*, 1-13.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. (2nd ed.). New York: Holt, Rinehart, and Winston.

Pohlmann, J. T. (1979). Controlling the Type I error rate in stepwise regression analysis. *Multiple Linear Regression Viewpoints, 10*, 46-60.

Pope, P. T., & Webster, J. T. (1972). The use of an *F*-statistic in stepwise regression procedures. *Technometrics, 14*, 327-340.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

SAS Institute. (1990). *SAS/IML user's guide* (Release 6.04). Cary, NC: Author.

Schafer, W. D. (1992). Simultaneous inference options for statistical decision making. *Measurement and Evaluation in Counseling and Development, 25*, 98-101.

Schafer, W. D., & Macready, G. B. (1975). A modification of the Bonferroni procedure on contrasts which are grouped into internally independent sets. *Biometrics, 31*, 227-228.

Thompson, B. (1989a). Why won't stepwise methods die? Measurement *and Evaluation in Counseling and Development,, 21*, 146-148.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development,, 22*, 2-5.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86*, 168-174.

Some Historical Notes on Statistical Data Analysis

Joe Ward

The historical notes below form a basis for concluding that combining a cell-means prediction-model approach with modern computers can empower data analysts to:

-- Analyze many different-appearing data analysis procedures with one general approach, which reduces the amount of material to be learned.

-- Create models that are more appropriate to the problems of interest, rather than forcing problems into packaged algorithms that may not answer the questions.

-- Reduce the risk of unknowingly obtaining answers from statistical software that are unrelated to the research questions of interest.

-- More easily and correctly specify the computational requirements to the computer.

-- Simplify communicating results of the analyses, since the models are developed from natural language concerns of the researcher.

1951 -

Joe Ward began working at the Air Force Personnel and Training Research Center (AFPTRC) at Lackland Air Force Base to move data analysis from desk calculators to IBM punched card machines. The first task was to implement an iterative algorithm for solving least squares equations that was not sensitive to linearly dependent predictors.

1953 - 1963

Bob Bottenberg and Joe Ward collaborated in enhancing research capabilities at AFPTRC by exploiting the power of Regression Models (Linear Models) made possible through the use of high speed computers. Many experiences combined to bring about a new perspective in research analysis at AFPTRC. While studying at Stanford University, Bottenberg was influenced by Z.W. Birnbaum, Albert H. Bowker, Meyer A. Gershick, George Polya and others. And while attending several Southern Regional Education Board Summer Institutes at the U. of Florida, North Carolina State, and Virginia Polytechnic Institute, Ward had valuable perspectives from association with Richard Anderson, Gertrude Cox, David Duncan, George Nicholson Jr., Lowell Wine and others. Also, of prime importance was the influence of Harry M. Hughes of the Air Force School of Aerospace Medicine.

During the 1950's most of the personnel at AFPTRC were PhD Research Psychologists who had received their statistics education prior to the availability of high speed computers. This meant that techniques of analysis did not involve the use of approaches to analysis that required a large amount of computing. During the late 1950's Bottenberg and Ward developed a Statistics course for personnel at AFPTRC. The plan was to provide a sequence of background concepts that would "eventually" lead to the exploitation of regression models and the computer for analysis. However, the participants were anxious to get on to the highly publicized promises that they would be able to create models appropriate to the research questions of interest and the course contents were adjusted accordingly. Unfortunately, little has changed in many one-semester, required college statistics courses. So much time is spent on the "assumed background prerequisites" that the students are rarely given the opportunity to realize the data analysis capabilities that are readily at their command.

During the AFPTRC course it became apparent that a "Top Down" approach was the way to go for persons who were interested in seeking answers to practical research questions. This implies starting with the problem stated in "natural language" and creating models that fit the problem rather than trying to fit the problem into an easily computable (possibly inappropriate) algorithm. This approach also suggests that concepts be introduced AS NEEDED, rather than spending time on topics which might have been assumed to be prerequisites for creating models to answer questions of interest. In situations where the participants are already indoctrinated with the pre-computer algorithms, it may be useful to relate the regression model approach to the older methods.

This need to empower researchers to create their own models was recognized by Raymond Christal and others at AFPTRC and as a result Bottenberg and Ward were encouraged to develop and document their ideas. This resulted in publication in March, 1963 of "Applied Multiple Linear Regression" by Robert A. Bottenberg and Joe H. Ward, Jr., PRL-TDR-63-6, which is available as AD413-128 from the Clearinghouse for Federal Scientific and Technical Information. For several years after this document was published it was among the highest volume sales from the Clearinghouse.

1964 -

In the summer of 1964 Bottenberg and Ward led a two-week National Science Foundation training session for a group of social sciences university faculty members. This session was directed by Earl Jennings and used the computing and dormitory accommodations at the University of Texas at Austin. The instructional activities focused on the use of regression models and computers in research data analysis. The participants were shown that it was now possible to solve the systems of simultaneous equations that are sometimes required for statistical models. And it wasn't (and still isn't) really necessary to have "equal or proportional n's" that were required BC (Before Computers). Furthermore, even if a researcher has NO OBSERVATIONS in some categories of an "Analysis of Variance" model, the problem can be readily analyzed by stating meaningful hypotheses about the population "cell means" for which there ARE OBSERVATIONS. With model creation skills it may be possible to create a defensible model that produces estimates of population means in cells in which there are no observations and to test hypotheses about the means of those cells.

1967 - 1975

During this period a series of Presessions were led by Bottenberg, Jennings, and Ward at the annual meetings of the American Educational Research Association. These sessions provided an opportunity for practitioners of educational research to become aware of the power of the regression models approach in the computer age. The large number of "graduates" of these Presessions stimulated the creation of the special interest group within AERA, SIG/Multiple Linear Regression. This MLR SIG has an informal publication, "Viewpoints", that provides communication among its members.

1973 -

After many years of teaching about and using regression models and computers, Ward and Jennings collaborated on a book that was to be included in the Prentice-Hall Series in Educational Measurement, Research, and Statistics. Specifically, the book was designed as a supplement to the Gene Glass and Julian Stanley book, "Statistical Methods in Education and Psychology". Englewood Cliffs, NJ: Prentice-Hall, 1970. The book (ILM) by Ward and Jennings was titled "Introduction to Linear Models", Englewood Cliffs, NJ: Prentice-Hall, 1973.

The book was an attempt to provide the reader with fundamental notions that would enable them to create models to answer research questions of interest. The ILM book developed the linear models approach in the traditional sequence presented in the Glass and Stanley book. That sequence was Inferences About the Mean, Difference Between Two Means, One-Factor Analysis of Variance, Two- Factor Analysis of Variance,....

1989 - 1994 >

From 1989-1992 Joe Ward served as a member of the American Statistical Association-National Council of Teachers of Mathematics (ASA-NCTM) Joint Committee on the Curriculum in Statistics and Probability. Ward continues to keep in close contact with the activities of the Committee and continues work with secondary schools through the "Adopt a School" program of the ASA. Ward started working with high school students and teachers in the use of computers in 1958. While the emphasis during those early years was on introducing computers into the secondary schools, Ward took the opportunity to introduce a few high school students to the combined power of regression models and computers. He now works with high school students and teachers in the San Antonio area who wish to enhance their data analysis skills. Ward teams with Laura Niland, a statistics teacher at MacArthur High School and the 1988 Texas Presidential Awardee in Secondary Mathematics, in workshops for high school teachers and students. Ward has taught Problem Solving Using Data Analysis to high school students in the Prefreshman Engineering Program (PREP) of the University of Texas at San Antonio.

Teaching both high school and college students who have had no previous introduction to Data Analysis has lead to the conclusion that a "TOP-DOWN" approach to Data Analysis will allow students to make practical use of their Data Analysis

experiences before they become "turned-off". Notice the use of the term "Data Analysis" in place of "Statistics". The use of a "different name" for the course allows more freedom to start with real-world problems, introduce the use of regression models and computers and apply these techniques to the data analysis requirements. Topics that are frequently taught as prerequisites are introduced when needed in the data analysis process.

1951 - 1994

Ward has interacted with a wide variety of researchers who call themselves by different labels. These include Research Psychologists, Educational Researchers, Operations Researchers, Economists, Statisticians, Computer Scientists, Mathematicians, Sociologists, Management Scientists, Engineers, etc. Fortunately many of these researchers learn -- while on the job -- to create models to fit their problems and to use the computer to "crunch the numbers". However, observations of newly trained researchers and the books used for their training indicate that much time is spent learning the "pre-computer" approaches to data analysis. Those authors that do show the student some examples of the general linear model approach to analysis do little to empower the student to create their own models. It is not clear why many classical texts include so many special computational formulas that were necessary in earlier years. There may be a belief that a learner acquires a stronger degree of understanding if they know how to do the pre-computer arithmetic. Many of the statistical software packages emphasize the use of the computerized versions of the "pre-computer" algorithms. And these packaged programs can occasionally provide answers to uninteresting questions that are different from the hypotheses that the data analyst thought were being tested.

There still remains a great need to develop instructional approaches that will allow researchers to create their own models as required and to use the computer to handle the computational burden. It seems that a good approach to introducing students to model development is to begin with a problem that is of interest to them and to use concepts that are familiar. The use of "averages" of collections of data are a great way to start since learners of all ages have heard the term and have been subjected in school to the use of "averages" as performance indicators. Most learners can talk easily in "natural language" about comparing "averages" among categories (e.g., batting averages, shooting average, etc.). Then these ideas can be expressed in more formal prediction models of forms such as :

DEPENDENT VARIABLE = PREDICTION + ERROR,
DATA = FIT + RESIDUAL,

DATA = MODEL + ERROR, or

$Y = XB + E.$

The important idea is to provide learning experiences that will eventually allow students to create models relevant to the questions of interest. The solutions to these models are now feasible by high-speed computers.

Using Multiple Regression to Develop ANOVA Power Formulae

Dale G. Shaw and David R. McCormack
University of Northern Colorado

Multiple regression may be used to examine the relationship between a single dependent variable and a set of several independent variables. The ANOVA power tables presented by Cohen (1988) can be considered such a data set. In these tables, the power of a balanced one-factor ANOVA design may be considered the dependent variable which is predicted by four independent variables: sample size, alpha level, number of groups, and effect size. Cohen's 66 pages of tables provide 15,526 power values for various combination of values of the four independent variables.

This article presents regression equations fitted to Cohen's ANOVA power tables in an effort to obtain simple yet accurate formulae for estimating the power of an ANOVA design. Simple equations were sought because power analysis is presently receiving limited attention in research planning (Cohen, 1988). Having a simple, easy-to-use formula which estimates a design's power might lead to improved designs. Obviously accurate power estimates are desirable, but the criterion of accuracy is less stringent than might be supposed. The researcher who is planning an ANOVA design does not usually require a power estimate to the nearest percentage point as Cohen's tables provide. For example, if a design's power were estimated to be .88 with a ±.06 margin of error, the experiment could proceed with with reasonable confidence of having high power even though the exact power is unknown.

In contrast to the simplicity criterion which required subjective judgment, the accuracy criterion was quantifiable. We used 2 indicators of accuracy for evaluating and comparing regression models: R^2 (the proportion of variance "explained" by the formulae) and RMSE (the root mean square error). We sought equations with $R^2 > .95$ and RMSE $\leq .03$. Regrettably, these two criteria for desirable formulae, simplicity and accuracy, conflicted with each other. The simplest formulae were not the most accurate and the most accurate formulae were not simple.

Linear Formulae

The first attempt to model ANOVA power was a simple linear model in which the dependent variable was Cohen's (1988) ANOVA power values. The independent variables were α, u, n, and f, described in Table 1. Cohen's tables provided 15,526 power values between .01 and .99, which served as data points on a hyper-surface. As expected, this first model failed to meet the accuracy criterion (R^2 = .4320). General knowledge of power curves as well as inspection of Cohen's tables suggested the surface was curvilinear rather than linear. To accommodate the curvature, and still keep the models composed of fairly simple terms, the predictor set was increased from 4 to 24 variables by including the square, the cube, the square root, the natural logarithm (ln), and the natural logarithm of the natural logarithm (ln(ln)) of each basic predictor. The last of these new predictors was undefined for some data points because ln(ln(1)) is undefined. Therefore, the basic variables were modified: α was multiplied by 1000, f by 100, and u by 10. R^2 values for various models created from the 24 variables did not exceed .95.

The search for a better model progressed by imposing a restriction on the data set. This was justified because a user of the resulting formulae probably would not need accuracy for very high or very low power values. A very low power, whether .10 or .25, indicates the proposed design is probably not worthy of further consideration. On the other hand, a very high power, whether .95 or .99, suggests a design worthy of further consideration. Because a research planner probably needs only limited accuracy at either end of the power range, the data points of these asymptotic tails of the power data (which offer the greatest difficulty in fitting a linear model) were eliminated. The greatest R^2 value of the new models (.9524 using the predictor set ln(f), ln(n), $\sqrt{u}$, $\sqrt{\alpha}$, ln(ln(n)), and u^2) was observed when power was restricted to the interval [.25, .95]. Thus, a decision was made to continue the search for linear formulae using only the reduced data set.

The next step was to increase the set of predictors by including the products of pairs of the 24 predictors so that interactive effects of the predictor variables could be accommodated. For each basic predictor, a set of 6 predictors had already been included, such as f, f^2, f^3, ln(f), ln(ln(f)), and $\sqrt{f}$. When each of the 6

f predictors was paired with each of the 6 n predictors, 36 predictors were possible. When all 4 basic predictor variables were considered, a total of 216 product pairs were added to the former 24 predictors, creating a set of 240 predictors. Regression by the forward, stepwise, and all-possible techniques was employed in search of terms that explained large portions of the variance in p. When the residuals of models based on these predictors were plotted, three somewhat parallel curves were observed. This prompted separation of the data set into three sets, one for each α level. R^2 values greater than .98 were obtained for each α level considered separately. Such formulae marginally satisfied the accuracy criterion but did not meet the simplicity criterion in which one formula incorporating all α levels was desired.

In the interest of simplicity, all terms containing logarithms were eliminated from the model. This reduced the possible predictor set from 240 to 112 predictors. Similar R^2's were attained without the complexity of the logarithmic terms. While marginally acceptable R^2 levels were obtained for specific α levels, the R^2 values obtained for general formulae were not deemed acceptable.

Several recurring predictors were observed in the formulae for the separate α levels, and it was hoped some form of α could be entered as a factor with these predictors to develop formulae that were acceptable for all α levels. Especially encouraging was the pair $f\sqrt{n}$ and f^2n, the second being the square of the first. Because $f\sqrt{n}$ continued to be prominent throughout the experimentation, η (eta) was defined u^2 to simplify future predictor notations. Experimentation with powers of η and α, along with various other terms from the current models, failed at this point to obtain acceptable formulae, however.

As stated earlier, the above models included only power values in the interval [.25, .95]. Continuing the search for formulae which would be simpler and more accurate, another reduction of the data set was tested. Because the user of a k-group ANOVA design is only rarely concerned with a comparison of more than five groups, the data for six or more groups was removed. Because u = k-1, this reduction meant only data points with values of u in the interval [1, 4] were used in the continuing search. This restriction of the data set permitted a model which included all α levels and which attained an R^2 =.9656 for the predictor set η, $f\eta^2$, $\sqrt{u\alpha}$, $u\alpha$, $u^2\alpha^3$, and n^2f^3. Individual models for specific α levels attained R^2 >.99. The best model (for α =.01 using the predictors η, η^2, $\sqrt{u}$, f, f^2, and $f^2\eta$) attained R^2 =.9962.

Many other possibilities were explored in the search for good models. Just as the separation of α levels had been explored, a separation of f levels was tested. Models in which p was replaced with ln(p), exp(p), or a trigonometric function of p were tried. None of these experiments yielded any improvement when compared to those models already reported.

To balance the two criteria, simplicity and accuracy, a compromise was required. Having chosen an accuracy requirement of RMSE < .03, it appeared the best general formula (given here in a factored form) contained six predictors and seven constants:

$$p = -.034\eta^3 + \left(.240 - .720\sqrt{\alpha}\right)\eta^2 + \left(2.178\sqrt{\alpha} + .043u\right)\eta - (.192f + .268) \quad (1)$$

The accuracy of Formula 1 was attained by considering only data points with power in [.25, .90] and u in [1, 4]. While Formula 1 is not as simple as originally hoped, its simplicity was deemed reasonable, considering the magnitude of the problem. The simplest possible linear combination of the four basic predictors would require five constants for the four terms plus an intercept term. That simple model, however, demonstrated very poor accuracy. Formula 1 requires only seven constants and it provides good accuracy, so it may be considered reasonable by potential users.

To reduce the number of terms (and constants) required in the model, consideration was given to re-entering logarithmic predictors into the model. Many combinations were tried using all possible regressions on various predictor sets. The model

$$p = .058 + .149\ln(\alpha) + (.355 + .045u)\eta + .197\ln(n)\sqrt{f} \quad (2)$$

emerged after much experimentation. Formula 2 has only five constants, compared to seven in Formula 1. The four predictors are, however, more complex than the six predictors of Formula 1.

The two formulae presented above appeared comparable in simplicity. To test the accuracy of Formulae 1 and 2, their residuals were analyzed (see Table 2) and the two formulae were again found to be comparable. The residual plots indicated a high degree of accuracy had been attained, but each plot exhibited a curvature which invited further exploration using a cubic function of p. Thus a two stage estimation procedure was considered. Stage One was either of the above two formulae. Stage Two then entered the resulting p into the model $P = b_0 + b_1p + b_2p^2 + b_3p^3$. The new P gave a better power prediction but the improvement was judged too minimal to warrant the application of Stage Two. The first formulae were already less simple than desired and it was felt the application of a second stage formula would probably not be attractive to any user.

Non-Linear Formulae

Because power data is not linear, non-linear models such as $p = b_0\left(1 - \exp\left(b_1 u^{e_1} a^{e_2} n^{e_3} f^{e_4}\right)\right)$ and $p = b_0 + b_1 u^{e_1} a^{e_2} + b_2 n^{e_3} f^{e_4}$ were tested using computer iterations to determine the b and e parameter values which most closely fit the surface. Although many models failed to converge to a set of parameters, the above models did each converge with RMSE < .04. These results were less satisfactory, however, than the results from the linear Formulae 1 and 2 already reported. Finally, the logistic model was considered because its graph approximates a power curve in shape, being asymptotic to zero and one. The logistic model, unlike linear models, might allow use of the full data set and might model ANOVA power well.

The logistic model is based on a sigmoidal curve with an equation similar to $Y = \frac{1}{1+e^{-X}}$. With $P = Y$ and a transformed $p' = x$, $P = \frac{1}{1+e^{p'}}$. Solving, $p' = \ln\left(\frac{1-P}{P}\right)$. Although P is sigmoidal, the transformed p' is linear. This p' was regressed on various sets of predictors using the linear model $p' = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$. The resulting coefficients were then substituted yielding a model similar to $P = \frac{1}{1+e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j}}$.

For initial trials, the five predictors (α, u, n, f, and η) were tested. The standard deviation of the residuals was .0858 with the data set limited to power in the interval [.25, .95] but when the entire data set was allowed, the standard deviation of the residuals was .0768. The logistic model performed as well, if not better, with an unrestricted data set. As experimentation continued, the predictors which had been discovered in the search for linear models were found to be helpful in the search for logistic models.

Although logistic regression produced pleasing power estimates in the asymptotic tales, a disturbing feature of these models was the wide range of the prediction errors indicated by the minimums and maximums of the residuals. Even in the best model, one residual was as large as .39. Through analysis of the data, the source of these extreme residuals was found to be cases of very low n and large u. When the restrictions $n \geq 10$ and u in [1, 4] were placed on the data, similar to the restrictions used in developing the linear models, a better fit was obtained:

$$p = \frac{1}{1 + 2.81\alpha^{-.72} u^{(.31-.27\eta)} \eta e^{[.91f-(2.31+.17u)\eta]}} \quad (3)$$

In Formula 3, all errors were within ± .05 and the standard deviation of the errors was .0150, a very pleasing result when considering the accuracy criterion of acceptable formulae. However, the simplicity criterion was challenged by this model. Any logistic model is by nature complex when compared to a linear model.

The Formulae Compared

The three formulae produced by this study each have features which may be attractive to users. Formula 1 is simple but lengthy. Formula 2 is more compact, but it includes logarithmic terms. Formula 3 is the most accurate, but it is also the most complex. In addition to these basic comparisons, the user might consider the tables of residuals associated with the formulae. Tables 3, 4, and 5 show the standard deviations of the residuals as well as the minimum and maximum residuals under various restrictions of the predictor variables. As an example, a design of five groups (u = 4) and five subjects per group (n = 5) is described in the next to last line of each table. If Formula 3 is chosen, the standard deviation of the residuals is .0125. Assuming normality of the residuals, 95% of the predicted power values would be within ± 1.96(.0125) = ± .0245. For the worst case, the predicted power value could be as much as .0984 too great or .0320 too small. (Power = predicted power + error.)

The superiority of the logistic Formula 3 of Table 5 is obvious, shown by the smaller numbers throughout. In addition, the logistic formula is based on the entire data set with power in [.01, .99]. The linear formulae were developed using only the data with power values in [.25, .90]. Of course, the accuracy of Formula 3 was gained at the expense of simplicity.

The user's choice of one of these three formulae will depend upon the user's desires and purposes. If the user desires the simplest formula, one of the linear formulae (Formula 1 or Formula 2) should be chosen. If greater accuracy is desired, the logistic formula (Formula 3) should be chosen. The user desiring accurate predictions in the tail regions of the model should always choose the logistic formula. A user may use a linear formula several times to test possible models and then, having narrowed the choices, use the logistic formula to make a final model selection. With computer spreadsheets, it is also possible for the user to consider the results of all three formulae simultaneously when proposing various ANOVA designs.

Although the formulae provide good power estimates in most cases, a user never knows whether

the estimate obtained in a particular case is highly accurate or only marginally accurate. Reference to plots of residuals can provide further insight for interpreting the power predictions calculated from the formulae. Figures 2, 3, and 4 show residuals plotted against the predicted power for each of the three formulae. To illustrate, consider Figure 2. Under Formula 1, if the predicted power is .75, the plot shows the residuals vary from -.04 to .06. Thus the actual power of the design is .71 to .81.

Although the user of the formulae may not always have the residual plots available, the formulae can still be used effectively if the user understands the general shape of the residual plots. The user of the linear Formulae 1 and 2 must be aware that the power will be over predicted when power is high, and under predicted when power is low. This is especially clear in Figure 1 where the full data set of power in [.01, .99] is plotted. When such predictions are obtained from the formulae, the user must interpret the results as "high" or "low" power respectively, without stating a specific power value. An example of extreme power predictions is the case of α=.05, n=500, u=1, and f=.4 (prediction = 15.602, error = -14.612). This error results from the dramatic negative effect of the factor η^3 for large values of n. Power for n=500 is expected to be very high, clearly outside of the [.25, .90] power range. Computed power estimates which fall into the range for which the formulae were developed will be reasonable power estimates, but the user is warned that any extreme power predictions of Formulae 1 or 2 should be ignored.

The shape of the residual plot of Formula 3 (Figure 4) is very different from the shape of the plots for the linear formulae, the logistic formula being more accurate in the tail regions than in the central regions. A comparison of the scales of the plots, however, demonstrates that the increased tail accuracy is not at the expense of accuracy in the central regions. Formula 3 meets or exceeds the performance of the other two formulae even in the central regions.

The formulae developed by this study offer a new way to compute the power of ANOVA designs. These formulae resulted from a directed "trial and error" search among those predictors which seemed reasonable. Certainly, the study did not exhaust all possible predictors of power. Thus, other researchers may discover better (simpler and/or more accurate) formulae than those presented here. This may be done with the tool of regression, as used in this study, or by some other method not yet considered.

References

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., & Kane, J. M. (1990). Statistical power analysis for one-way analysis of variance: a computer program. *Behavior Research Methods, Instruments, and Computers, 22,* 271-282.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hilldale, NJ: Lawrence Erlbaum Associates.

Dixon, W. J., & Massey, F. J. (1969). *Introduction to statistical analysis* (3rd ed.). New York: McGraw-Hill.

Donaldson, T. S. 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.

Fox, M. (1956). Charts of the power of the F-test. *Annals of Mathematical Statistics, 27,* 484-497.

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, 237-288.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology (2nd ed.).* Englewood Cliffs, NJ: Prentice-Hall.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *The American Statistician, 43,* 253-260.

Haase, R. F. (1986). A BASIC program to compute statistical power for atypical values of alpha. *Educational and Psychological Measurement, 46,* 629-632.

Lehmer, E. (1944). Inverse tables of probabilities of errors of the second kind. *Annals of Mathematical Statistics, 15,* 388-398.

Lenth, R. V. *PowerPack (Version 2.22)* (no date given) [Computer program]. 3061 Hastings Avenue, Iowa City, IA 52240. (319) 337-8549.

Levin, J. R. (1975). Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement, 12*, 99-108.

McCormack, D. R. (1993). Formulae for estimating the power of one-factor ANOVA designs. (Doctoral dissertation, University of Northern Colorado, Greeley).

Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models.* (2nd ed.). Homewood, IL: Irwin.

NONCDIST *(Quality control and industrial experiments)* (no date given) [Computer program]. Lionheart Press, P. O. Box 379, Alburg, VT 05440. (514) 933-4918.

O'Brien, R. G. (1988). Review of PowerPack. The *American Statistician, 42*, 266-270.

Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central F-distribution. *Biometrika, 38*, 112-130.

Rogers, W. T., & Hopkins, K. D. (1988). Quick power estimates incorporating the joint effects of measurement error and a covariate. *Journal of Experimental Education, 57*, 86-94.

Seber, G. A., & Wild,C. J. (1989). *Nonlinear regression.* New York: Wiley.

Shaw, D. G., & McCormack, D. R. (1992) Estimating the power of a t-test. Unpublished manuscript, University of Northern Colorado, Greeley.

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

Statistical Power Analysis (version 1.0) (no date given) [Computer program]. Lawrence Erlbaum Associates, Inc., 365 Broadway, Hillsdale, NJ 07642. (201) 666-4110.

Tang, P. C. (1938). The power function of the analysis of variance with tables and illustrations of their use. *Statistical Research Memoirs, 2*, 126-149.

Table 1. Independent Variables of the Linear Regression Model

Variable	Description	Values or Range
α	significance level	.01, .05, .10
u = k - 1	numerator degrees of freedom for a k-group ANOVA	1 - 24
n	per group sample size	2 - 1000
f	Cohen's effect size	.05 - .80

Table 2. Comparison of Residuals of Formula 1 and Formula 2

Formula	Mean	Standard Deviation	Minimum	Maximum
1	.0006	.0254	-.1387	.0617
2	.0006	.0251	-.0901	.1080

Table 3. Residuals of Linear Formula 1, p in [.25, .90].

$$p = -.034\eta^3 + \left(.240 - .720\sqrt{\alpha}\right)\eta^2 + \left(2.178\sqrt{\alpha} + .043u\right)\eta - (.192f + .268)$$

u Values	n Values	St Dev	Min	Max
All	All	.1094	-.7782	.0617
[1, 8]	All	.0340	-.2301	.0617
[1, 8]	n≥5	.0313	-.1466	.0617
[1, 8]	n≥10	.0306	-.1414	.0608
[1, 4]	All	.0254	-.1387	.0617
[1, 4]	n≥5	.0242	-.0979	.0617
[1, 4]	n≥10	.0235	-.0897	.0608

Table 4. Residuals of Linear Formula 2, p in [.25, .90].

$$p = .058 + .149\ln(\alpha) + (.355 + .045u)\eta + .197\ln(n)\sqrt{f}$$

u Values	n Values	St Dev	Min	Max
All	All	.1167	-.7375	.1080
[1, 8]	All	.0350	-.2027	.1080
[1, 8]	n≥5	.0338	-.1221	.1080
[1, 8]	n≥10	.0335	-.1221	.1080
[1, 4]	All	.0250	-.0901	.1080
[1, 4]	n≥5	.0247	-.0876	.1080
[1, 4]	n≥10	.0243	-.0876	.1080

Table 5. Residuals of Formula 3, p in [.01, .99].

$$p = \frac{1}{1 + 2.81\alpha^{-.72}u^{(.31-.27\eta)}\eta_e^{[.91f-(2.31+.17u)\eta]}}$$

u Values	n Values	St Dev	Min	Max
All	All	.0445	-.5195	.0320
[1, 8]	All	.0160	-.2156	.0320
[1, 8]	n≥5	.0126	-.0984	.0320
[1, 8]	n≥10	.0115	-.0516	.0295
[1, 4]	All	.0143	-.1276	.0320
[1, 4]	n≥5	.0125	-.0984	.0320
[1, 4]	n≥10	.0113	-.0516	.0295

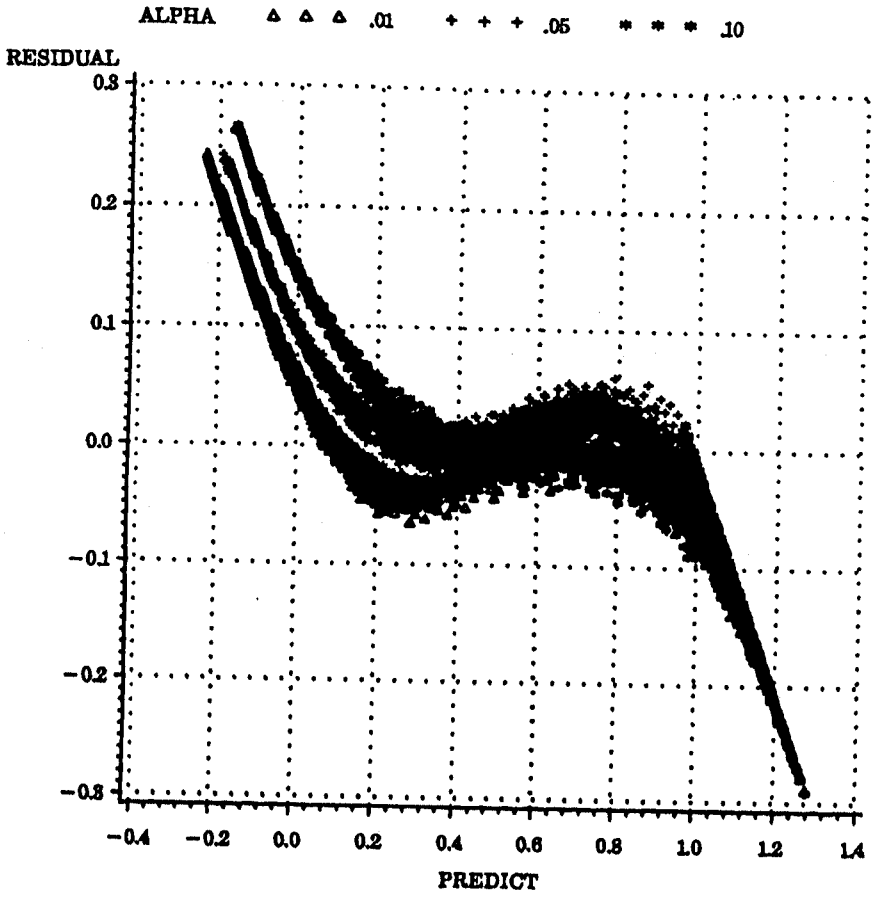


Figure 1. Formula 1, p in [.01, .99], u in [1, 4], n

$$p = -.034\eta^3 + \left(.240 - .720\sqrt{\alpha}\right)\eta^2 + \left(2.178\sqrt{\alpha} + .043u\right)\eta - (.192f + .268)$$

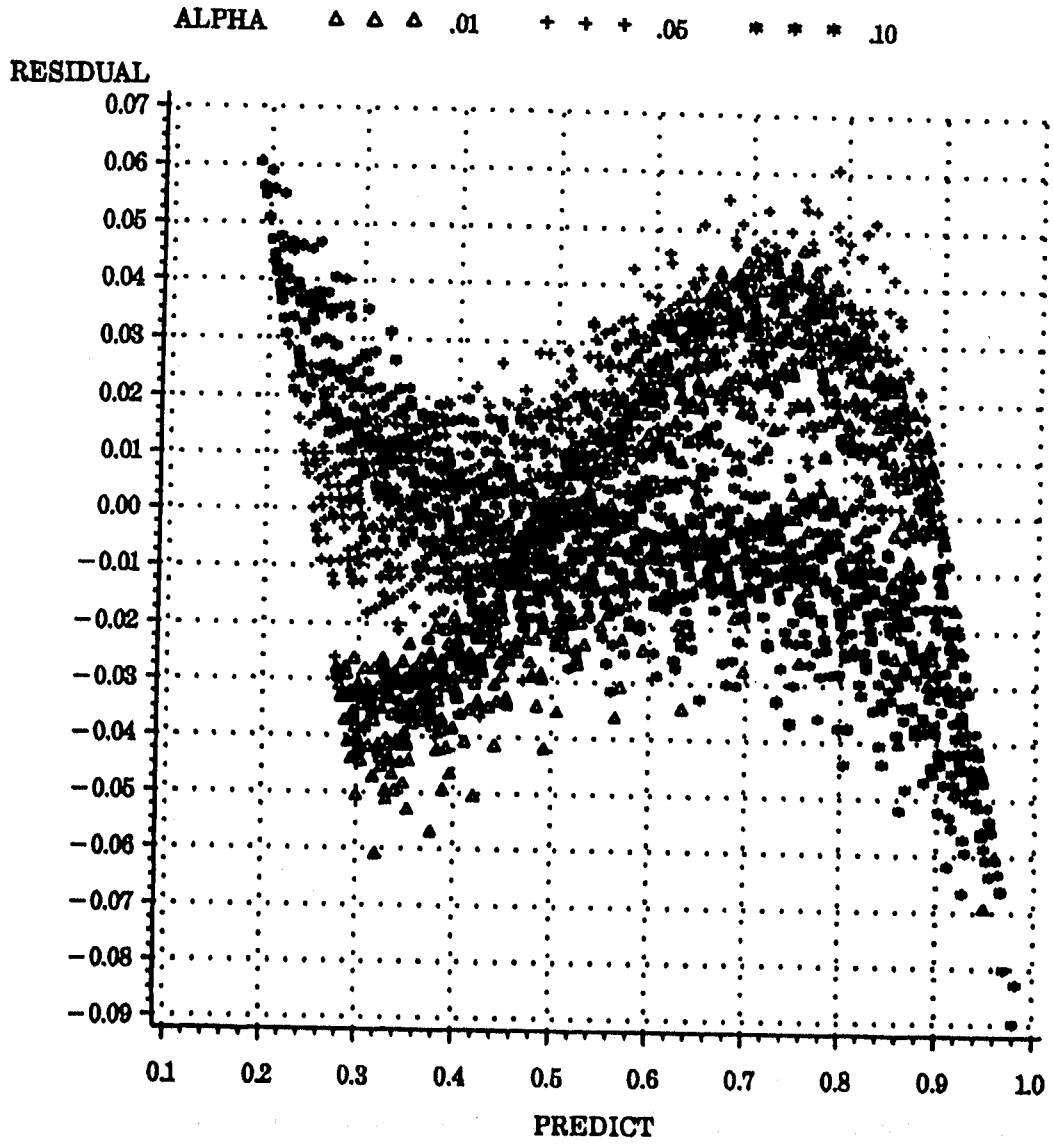


Figure 2. Formula 1, p in [.25, .90], u in [1, 4], n

$$p = -.034\eta^3 + \left(.240 - .720\sqrt{\alpha}\right)\eta^2 + \left(2.178\sqrt{\alpha} + .043u\right)\eta - (.192f + .268)$$

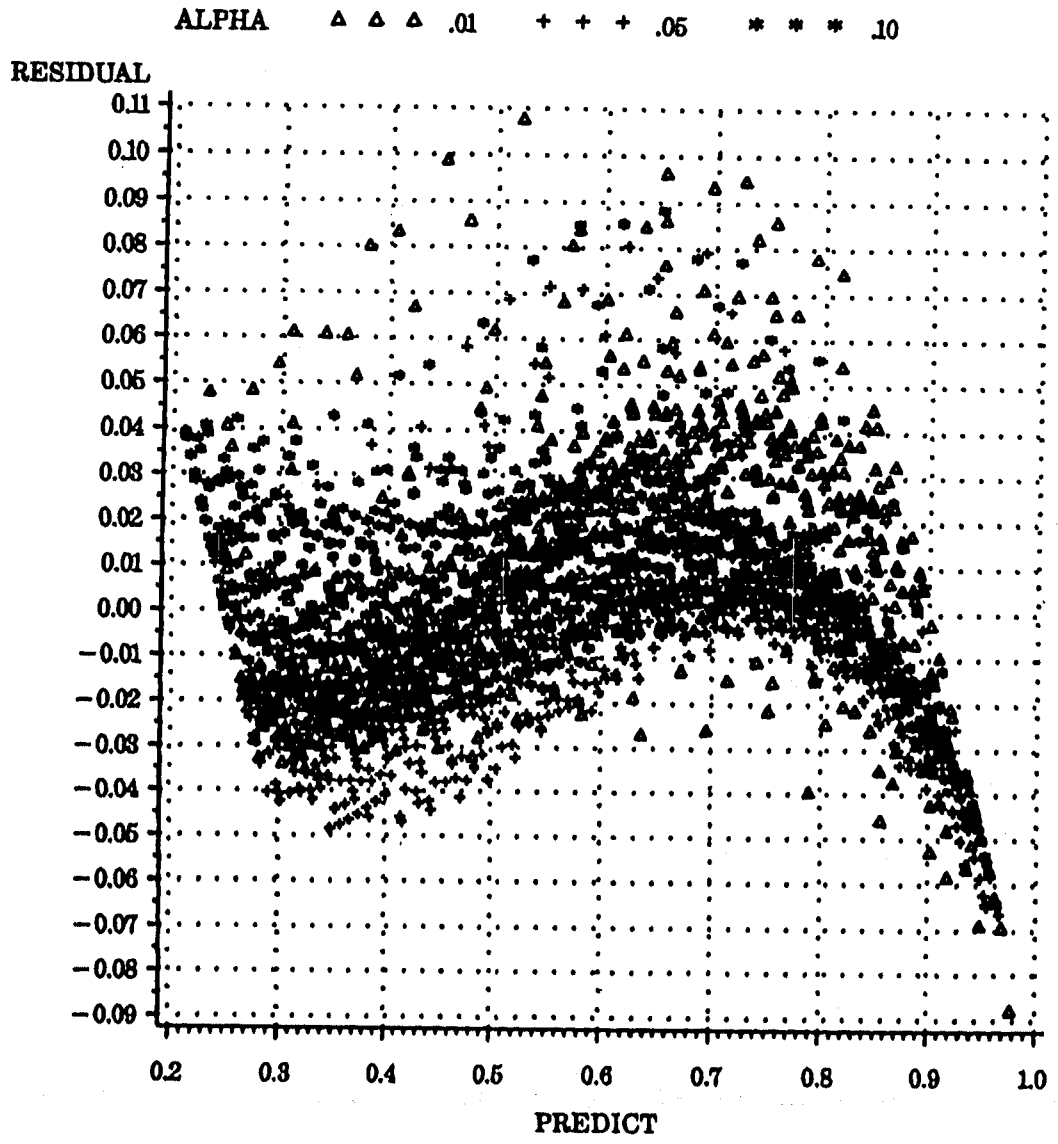


Figure 3. Formula 2, p in [.25, .90], u in [1, 4], n

$$p = .058 + .149\ln(\alpha) + (.355 + .045u)\eta + .197\ln(n)\sqrt{f}$$

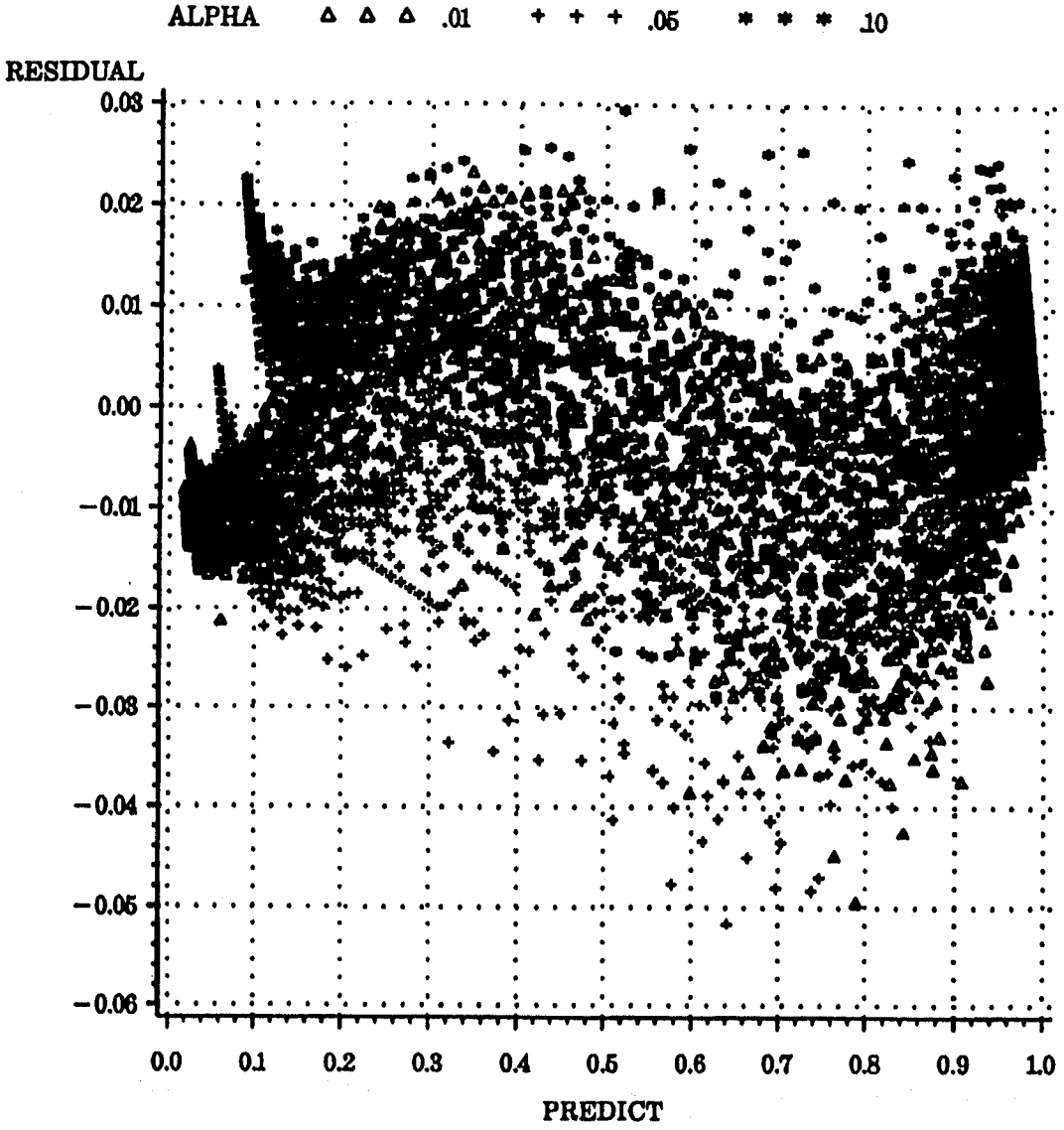


Figure 4. Formula 3, p in [.01, .99], u in [1, 4], n

$$p = \frac{1}{1 + 2.81\alpha^{-.72}u^{(.31-.27\eta)}\eta_e^{[.91f-(2.31+.17u)\eta]}}$$

Comparison of General Linear Model Approaches to Testing Variance Heterogeneity in True and Quasi-Experiments

T. Mark Beasley
St. John's University, New York

Simulation results indicated that when groups were sampled from the same platykurtic population the O'Brien (1981) transformation was preferred except when a positive sample size/variance correlation existed, then the Welch test performed on the O'Brien scores was more powerful. Also consistent with previous research, when grouped data were sampled from the same leptokurtic population the Brown & Forsythe (1974) transformation was preferred for equal sample sizes. The O'Brien test was more powerful with an indirect sample size/variance relationship regardless on distribution shape(e.g., Algina et al., 1989; Olejnik & Algina, 1987, 1988). The study also demonstrated that the Welch test performed on Brown-Forsythe scores was more powerful when a positive sample size/variance correlation existed in leptokurtic data. Furthermore, choosing a test of variance based on an initial test of kurtosis may improve power (Ramsey, 1994). When data were sampled from populations with drastically different shapes (kurtosis), the Type I error rate of most tests was unstable excluding the Hartley *Fmax* test which performed surprisingly well.

The analysis of variance (ANOVA) is one of the most widely used statistical procedures in educational research. Namely, it is the technique of choice for True Experiments in which members sampled from the same population are randomly assigned to treatment conditions. In field research and Quasi-Experiments, comparisons among groups are also of interest; however, there exists the possibility that these groups are sampled from different populations. In either case, behavioral researchers often compare groups with different distributional properties, which may be a result of (a) sampling from different populations or (b) an experimental treatment affecting something other than central location. Thus, as far as analytic procedures are concerned, the distinction between True and Quasi-Experiments becomes ambiguous. For the purposes of this study, a True Experiment is defined as sampling data from a single population and randomly assigning cases to groups. A Quasi-Experiment is defined by separately sampling data from populations which differ in distributional shape (i.e., skew and kurtosis).

One of the most critical conditions for any linear modeling procedure involves the assumption of homoscedasticity across levels of the independent variable. In the ANOVA, it follows that heterogeneous variances may obscure the magnitude of test statistics for comparisons among means. Thus, testing variance equality appropriately is important in checking a vital assumption of the ANOVA. Furthermore, despite the existence of differences in central location, heterogeneous variances may constitute substantive and theoretically valuable results. That is, it may be interesting to know that the responses of two separately sampled populations differ in scale or that an experimental treatment significantly affects response variability.

Traditional tests of variance homogeneity (e.g., Hartley's *Fmax*) can be very simple, calculating the ratio of two sample variances. The *Fmax* test, however, has long been known to be extremely sensitive to deviations in kurtosis (Box, 1953; Scheffe', 1959). Slight departures from normality which involve kurtosis have been shown to make substantial difference in the Type I error rate of the *Fmax* test (Pearson & Please, 1975). For instance, Hartley's *Fmax* test has been shown to be conservative for platykurtic distributions and liberal when distributions have positive kurtosis (Durrand, 1969). Although several tests of variance have been proposed, the *Fmax* remains popular in a variety of applied studies because of its simplicity.

In a simulation study, Conover, Johnson, and Johnson (1981) compared several procedures for testing homogeneous variances and found that most are liberal (i.e., the Type I error rate was considerably larger that the nominal alpha). Thus few tests exist that actually control the Type I error rate. Over the past two decades, robust tests of variances based on applying the ANOVA to transformed scores (e.g., Brown & Forsythe, 1974; O'Brien, 1981) have been proposed. Under conditions of a "True Experiment" and equal sample sizes, these tests have been shown to be powerful in a variety of population distributions (Algina, Olejnik, & Ocanto, 1989; Olejnik & Algina, 1987; Ramsey, 1994; Ramsey & Brailsford, 1990). The Brown-Forsythe (BF) test has been criticized because it has low power for small, odd

sample sizes and only moderate power for platykurtic and normal populations (O'Brien, 1981; Olejnik & Algina, 1987). Under these same conditions, the most common form of the O'Brien (OB) procedure has been shown to be more powerful than BF. Also for unequal sample sizes, OB has been suggested for platykurtic distributions and BF with symmetric and/or leptokurtic distributions (Algina et al., 1989).

Despite these recommendations based on the kurtosis of distributions, a criterion for identifying population shape was not suggested. Ramsey and Brailsford (1990) noted that tests of kurtosis could be used to decide between BF and *Fmax*. Following the suggestions of previous studies, Ramsey (1994) has recently suggested two conditional procedures based on testing kurtosis for each group separately. Ramsey's results confirmed the robustness of OB and BF but indicated that optimal power can be established with the conditional procedure of testing kurtosis to decide between the these tests. However, the power of these conditional procedures has been shown to be dependent on the power of the test of kurtosis. Also, Ramsey's results are limited in the sense that only the conditions of a True Experiment were simulated. That is, the two groups were sampled from the same population. In field research and Quasi-Experiments, comparisons of groups sampled from different populations are often of interest and the suggested conditional procedures have yet to be fully investigated under such conditions.

Olejnik and Algina (1988) found that both OB and BF held the Type I error rate for a limited number of distributions which differed in location and form. The OB tended to be most powerful with equal sample sizes and with an inverse relationship between sample sizes and population variances (i.e., larger sample has the smaller variance). When sample sizes and population variances had a direct relationship (i.e., larger sample has the larger variance), using OB transformed scores as dependent variables and performing the Welch (1951) statistic was the most powerful procedure.

A variety of nonparametric tests of variance are also available; however, they have presented problems with robustness and low power. Two of the better known procedures were proposed by Klotz (1962) and Siegel and Tukey (1960). When data were sampled from a normal population, both tests demonstrated the appropriate Type I error rate (Penfield & Koffler, 1985; Olejnik & Algina, 1985). Also, the Klotz test had power equal to or greater than the power of OB or BF when samples differed in variance only. However, both the Siegel-Tukey and Klotz tests were strongly affected by differences in central location (Moses, 1963). When the sampled distributions share the same asymmetric shape but differ in location, the tests are liberal. Yet, both tests become less powerful as location parameters increase when groups share the same symmetric shape (Olejnik & Algina, 1985). To date, attempts to modify these tests through mean- and median-alignment have not drastically improved their statistical properties (e.g., Conover et al., 1981; Olejnik & Algina, 1988).

Thus the purpose of this study was to investigate the robustness and power of OB, BF, and the use of the Welch statistic on these transformed scores (W_{OB} and W_{BF}, respectively) under conditions of a *True Experiment* (i.e., groups are randomly constructed from the same population) and a *Quasi-Experiment* (i.e., groups are sampled from two different populations). Furthermore, the effectiveness of conditional procedures based on tests of kurtosis (e.g., Ramsey, 1994) was examined. Although tests of variance are themselves of interest, in most educational research differences in central location are to be expected; therefore, nonparametric procedures such as the Siegel-Tukey and Klotz tests were excluded from this study. Under several circumstances, the results were expected to replicate those of Ramsey (1994) and Olejnik and Algina (1987, 1988). Furthermore, the findings of this study should address the issue of the appropriate procedure for testing variances in Quasi-Experiments in which the populations may differ in variance and form and the samples differ in size.

Statistics for Testing Variances

Although many statistical tests for comparing population variances have been developed, only a few of these procedures have demonstrated robustness when populations are nonnormal (i.e., Conover et al., 1981). Of these tests, the general linear model procedures, which involve performing the ANOVA (or some variant) on transformed scores, have shown both robustness and superior power.

Hartley's Fmax test. This test was investigated because of its wide use and known properties when kurtosis deviates form normality. The *Fmax* test is the ratio of the largest to the smallest of *J* variance,

$$Fmax = \frac{s^2_{largest}}{s^2_{smallest}}. \qquad (1)$$

The degrees of freedom are ($n_{largest}$ - 1) for the numerator and ($n_{smallest}$ - 1) for the denominator. Although it is often recommended that the *Fmax* test only be used with approximately equal sample *n*'s, its statistical properties were examined under all condition of this study. Critical values were obtained from the sampling distribution derived by Hartley (1950).

Brown-Forsythe Transformation. To test differences in variances, Levene (1960) proposed using the ANOVA but replacing each score, y_{ij}, of subject *i* within group *j* with the absolute deviation from its respective group mean. Although this procedure is fairly robust, it was found not to be adequately powerful (Conover et al., 1981). Brown and Forsythe (1974)

proposed applying the ANOVA to absolute deviations from respective group medians, m_j, such that:

$$b_{ij} = |y_{ij} - m_j| \quad . \qquad (2)$$

O'Brien Transformation. O'Brien (1979) proposed that the original score, y_{ij}, of subject i in group j be replaced with

$$r_{ij}(w) = \frac{(w + n_j - 2)\, n_j (y_{ij} - \bar{y}_j) - w s_j^2 (n_j - 1)}{(n_j - 1)(n_j - 2)} \qquad (3)$$

where w is a parameter ranging between zero and one and $\bar{y}_j$ equals the mean, s_j^2 equals the variance, and n_j equals the sample size of group j. For most cases, O'Brien (1981) has recommended a value of $w = 0.5$ from which the group means for r in (3) are the variances of each group y: $\bar{r}_j = s_j^2$. The ANOVA is performed on the transformed r values.

Welch Statistic. It is not known whether the OB or BF tests are asymptotically distribution free. Furthermore, because the variance of r is dependent on sample size, O'Brien (1981) suggested using a Welch (1951) approximate degrees of freedom analysis on r values in place of the ANOVA when sample sizes are not equal (WOB). This procedure may also be performed on BF transformed scores (WBF). The Welch statistic is calculated by

$$W = \frac{\sum_{j=1}^{J} c_j (\bar{y}_j - \tilde{y}) / (J - 1)}{1 + \frac{2(J-2)}{J^2 - 1} \sum_{j=1}^{J} (1 - \frac{c_j}{c})^2 / (n_j - 1)} \qquad (4)$$

where J equals the number of groups, $c_j = n_j / s_j^2$, $c_* = \Sigma c_j$, and $\tilde{y} = \sum c_j r_j / c$. The Welch statistic is approximately distributed as F with degrees of freedom equal to $(J - 1)$ and

$$\left[\frac{3}{J^2 - 1} \sum_{j=1}^{J} (1 - \frac{c_j}{c})^2 / (n_j - 1)\right]^{-1} \qquad (5)$$

For $J = 2$ groups, the degrees of freedom in (5) follow the Satterthwaite (1946) formula.

Conditional Tests. Based on the simulation studies of Olejnik & Algina (1987, 1988), BF is preferred for leptokurtic populations, while OB is recommended for normal and platykurtic distributions. To achieve optimal power, Ramsey (1994) proposed two procedures for testing variances that are conditioned on applying a test for kurtosis.

Pearson's traditional sample measure of population kurtosis, γ_2, in group j is $b_2 = m_4 / m_2^2$, where $m_r = \sum (y_{ij} - \bar{y}_j)^r / n_j$. Thus m_2 is the second moment about the mean, the biased sample variance. Although standardized population moments for skewness and kurtosis provide popular significance tests, Ramsey and Ramsey (1993) have supplied a detailed and accurate table of critical values for b_2, which are used to test kurtosis against the null hypothesis (Ho: $\beta_2 = 3$).

For the tests proposed by Ramsey, tests of kurtosis are applied in each of the two samples at the $\alpha = .05$ significance level. A score of -1, 0, or +1 is recorded depending on whether the test of b_2 indicates that the distribution was significantly platykurtic, nonsignificant, or significantly leptokurtic, respectively. Combining scores from the two samples results in a total score, S, ranging from -2 to +2. In a J-group study, S would range from $-J$ to $+J$. The test of kurtosis is taken as identifying the population for the entire experiment as platykurtic if $S \leq -1$, mesokurtic if $S = 0$, and leptokurtic if $S \geq +1$. In one conditional procedure, OB_{BF}, kurtosis is tested and OB is applied if the samples are platykurtic or mesokurtic ($S \leq 0$) and BF if the distributions are significantly leptokurtic ($S \geq +1$). This approach is based on the recommendations of Olejnik and Algina (1987) but does not control the Type I error rate under certain distributional conditions; therefore, Ramsey (1994) suggested another conditional procedure that demonstrated superior power and adequate robustness. This approach, BF_{OB}, involves testing the fourth moment and applying OB with significantly platykurtic distributions ($S \leq -1$) and BF otherwise ($S \geq 0$).

Methods

Consistent with previous studies (e.g., Miller, 1968; Olejnik & Algina, 1987, 1988; Ramsey & Brailsford, 1990), the present investigation was restricted to the two-group case. These studies yielded results congruent with multi-group studies. Furthermore, the restriction to two groups allows more careful consideration of other factors. Since previous studies have indicated that shifts in central location have little to no effect on general linear model tests of

variance, (Beasley & O'Connor, 1995; Olejnik & Algina, 1988), three population variables were manipulated: shape in the form of kurtosis (γ_2), variance in one group, (σ^2); sample size (n_j).

Conditions

Population Kurtosis. Previous studies have indicated that skewness affects the robustness and power of nonparametric tests (Olejnik & Algina, 1988) but only affects the statistical properties of parametric tests in combination with nonnormal kurtosis (Conover et al., 1981; Olejnik & Algina, 1988; Pearson & Please, 1975). The normal and six nonnormal distributions that had no skewness but varied in kurtosis were simulated. They are presented in ascending order from platykurtic to leptokurtic. The first population was extremely platykurtic (XPLT) and continuous with skewness (γ_1) equal to zero and kurtosis (γ_2) equal to -1.80. The second population was also platykurtic (PLAT) and continuous with skewness (γ_1) equal to zero and kurtosis (γ_2) equal to -1.00. It was chosen because it has been used in a variety of other simulation studies (e.g., Olejnik & Algina, 1987, 1988). The third population was slightly platykurtic (SPLT) with γ_1 = 0.0 and γ_2 = -0.50. It was selected as a continuous distribution which closely matches the moments of one of Micceri's (1989) data sets. The fourth population was the normal distribution (NORM) generated with the SAS RANNOR function. The fifth population (LEP1) was selected as a slightly leptokurtic, γ_2 = +1.00, continuous distribution with no skew comparable to the second population (PLAT). The sixth (LEP3) and seventh (XLEP) were selected as highly leptokurtic (γ_2 = +3.00 and +3.75, respectively) with no expected skewness.

Group Size and Variance Ratio Parameters. Equal sample sizes of n_j = 10, 13, and 20 and unequal sample sizes of (10, 20) and (13, 20) were employed. To investigate power, variance ratios of VR = 2.0 and 5.0 were imposed by taking the population from which Group Two was sampled and multiplying it by constant equal to the square root of VR.

Because Olejnik and Algina (1988) found that tests of variance were differentially powerful depending on the relationship between group size and population variance, all conditions were crossed when power was investigated. For example, when the variance ratio was VR = 2.0 and Group One, with n_1 = 13, was sampled from the normal distribution, while Group Two (n_j = 20) was sampled from a platykurtic distribution, an inverse relationship between group size and population variance (negative condition) was imposed. In order to create a positive condition, the sample sizes were reversed so that the larger group had the larger variance. Also because power and robustness may depend on population shape, the seven populations were systematically manipulated as long as conditions did not duplicate (e.g., when investigating Type I error rate for equal sample sizes all population combinations are not necessary). Table 1 shows the sample size conditions for the analyses in this study. Note that two sample size configurations were added to impose positive and negative sample size/variance correlations for investigating power in True Experiments. For Quasi-Experiments, all possible sample size conditions were used since the Type I error rate has been shown to depend on sample size/kurtosis configurations. In examining power, variance constants were imposed on both Group One and Group Two because power has also been shown to depend on the configuration of sample size, kurtosis, and variance (Beasley & O'Connor, 1995; Olejnik & Algina, 1988).

Procedure

The second through seventh populations were generated separately for each group using the RANNOR function is SAS/IML, which provides a clock generated pseudorandom standard normal deviate, z_{ij} (SAS Institute, 1990). Fleishman's (1978) method was used to transform these distributions into non-normal data with specified mean, variance, skewness, and kurtosis values via a polynomial equation of the form,

$$y_{ij} = a + bz_{ij} + cz_{ij}^2 + dz_{ij}^3 \quad . \qquad (6)$$

Since the minimum kurtosis derived by Fleishman is γ_2 = -1.00, the first population (XPLT) was simulated by combining three uniform distributions that varied in central location. A small distribution of 20 cases that centered around 0 and two larger distributions of 990 cases each which centered around -0.75 and 0.75 were concatenated to create this heavy-tailed distribution. Linear transformations were used in order to have the expected variances used in this study. During the simulation procedures, observations were randomly sampled from these distributions during each replication. For each condition elaborated, 5,000 replications were completed. The proportions of rejections at the α = .05 level of significance were used as measures of empirical power and Type I error rate.

Since 5,000 replications were conducted in each condition with α = .05, the standard error is .0031. Thus, any Type I error rate of .0562 or greater exceeded two standard errors and was considered a significant inflation of the Type I error rate. Other less stringent criteria include upper limits of .06 (Cochran, 1954) and .075 (Bradley, 1978). In order to avoid the problems with making multiple comparisons within this study, the standard error of simulation was used as a general heuristic rather than as a statistical test when comparing empirical power estimates. Furthermore, if the

empirical Type I error rate of a test exceeded the nominal alpha by two standard errors, its power was interpreted cautiously. If its Type I error rate exceeded Cochran's limit of .06, its power estimate was not reported.

Results

Simulation Accuracy

When multiplied as in (6), the resulting mean, variance, skewness, and kurtosis of y_j approximate the characteristics of the distribution of interest. It should be noted, however, that the simulated data are not governed by a known mathematical function. Rather, the simulated data represent a distribution with the same skewness and kurtosis as the desired distribution. Table 2 demonstrates the adequacy of the Fleishman simulation method in this study. Values for the mean (μ), variance (σ^2), skew (γ_1) and kurtosis (γ_2) for each group of n_j were taken across 15,000 replications for n_j = 10 and 13 and across 30,000 replications for n_j =20. For all seven populations, μ, σ^2, and, γ_1 were adequately simulated. Furthermore, kurtosis (γ_2) was reasonably simulated for platy- and mesokurtic distributions, especially with n_j =20. For leptokurtic distributions, however, the kurtosis of the group was drastically underestimated which is most likely due to the small sample sizes used.

True Experiments

Type I Error. Table 3 shows the empirical Type I error rate for the seven sampled populations under the conditions of a True Experiment (i.e., both groups drawn from the same population). As would be expected the Hartley's *Fmax* test showed a conservative rejection rate with platykurtic populations (e.g., XPLT, PLAT, & SPLT) but more importantly was liberal when the data were sampled from leptokurtic populations. Furthermore, when sample sizes are unequal, the suspension of the *Fmax* test is often suggested. However, the Type I error rate remained under the nominal alpha of .05 even with unequal samples under the meso- and platykurtic conditions. All other tests, except for WOB which exhibited minor inflation with disparate sample sizes, held the Type I error rate under the nominal alpha.

Power. Tables 4 and 5 show representative results from the comparative power analysis of True Experiments with different populations and variance ratios of *VR* = 2.0 and 5.0, respectively. For all tests, except *Fmax*, it can be seen that heavy-tailed distributions presented a more powerful situation when testing variance heterogeneity. For example, the empirical power estimates were higher when data were sampled from the PLAT population as compared to normally distributed data. Also, higher power estimates were yielded when data were sampled from the normal distribution as compared to the leptokurtic populations(e.g., LEP1 and XLEP, see Tables 4 & 5).

Under the conditions of a normally distributed population, the Hartley's *Fmax* was robust and demonstrated superior power, except when there was a positive relationship between sample size and group variance. In this case, the WOB was most powerful. When the sample size/variance correlation was negative *Fmax* and OB were of similar power.

When data were sampled from the PLAT distribution, the OB transformation and Ramsey's OB_{BF} were the clear choices in low power situations (see Table 4). However, with a variance ratio of *VR* = 5.0, the *Fmax* test was more powerful except when the sample size/variance correlation was positive. Thus, in cases where the smaller group had the smaller variance, WOB was most powerful namely because neither *Fmax* nor the Ramsey's procedures make provisions for such situations. In high power situations (*VR* = 5.0) whether the data were meso- or platykurtic, the *Fmax* was more powerful. However, both *Fmax* and OB are very likely to reject the null hypothesis in such cases. Thus, if the data sampled are platykurtic, one should consider the O'Brien transformation in low power situations. In all conditions with meso- and platykurtic data and a positive relationship between sample size and variance, performing the Welch statistic on O'Brien scores was the most powerful procedure.

When data were sampled from leptokurtic populations, the *Fmax* test was disqualified because it inflated the Type I error rate (see Table 3). Of the remaining tests, BF and BF_{OB} had similar empirical power estimates with small (n_j = 10) equally sized samples. Similarly, as was observed with platykurtic distribution, BF_{OB} was more powerful than BF, which indicates that the Ramsey conditional procedures can provide more power. When sample sizes were unequal and positively related to the group variances, the WBF was more powerful. This finding seems consistent with previous research but has yet to be reported in the literature. When a negative relationship between sample sizes and group variances existed, OB was more powerful regardless of the leptokurtosis of the sampled population. Increasing the group Variance Ratio to *VR* = 5.0 magnified these findings. However, under these more powerful conditions, the power estimates of BF were more competitive and actually exceeded those of OB when there was an inverse sample size/variance relationship. For example in Table 5, when n_1 = 20, σ_1^2 = 1.0, n_2 = 13, and σ_2^2 = 5.0, the power of BF, .5590 was much higher than that of OB, .5110. This indicates that OB is only more powerful under negative sample size/variance conditions in low power situations (i.e., small sample sizes, small differences in variance). That is, if samples are rather large and drawn from leptokurtic populations, the BF and WBF may be better

choices for testing variances. Thus, consistent with previous research, when there is a negative correlation between sample sizes and variances, the advantage in power of OB over BF seems to dissipate with increasing (a) sample sizes for both groups, (b) variance for the smaller group, and/or (c) kurtosis of the sampled population (Olejnik & Algina, 1988; Ramsey, 1994).

Quasi-Experiments

Type I Error. When one group was sampled from a population with extremely negative kurtosis (XPLT), while the second group was sampled from population of varying shapes, the Type I error rates for all tests, except for *Fmax*, were unstable and generally above the nominal alpha of .05 (see Table 6). However, as the extremity of platykurtosis declined, the Type I error rates became more stable for most tests (see Table 7).

When the two groups were sampled from populations with similar positive kurtosis, the results were predictable from the Type I Error results for True Experiments. Table 8 shows that when both groups had positive kurtosis most tests, except for *Fmax*, held the Type I error rate at the nominal alpha of .05. However, WOB showed inflations when the larger group was sampled from a less leptokurtic distribution. These results extended to situations where one group is sampled from a slightly leptokurtic distribution (LEP1) and the other is sampled from a slightly platykurtic distribution (SPLT).

Although the mixture of LEP1 and the normal distribution did not affect the Type I error rate of most tests (see Table 8), when the variance of a normally distributed sample was tested against the variance of data sampled from more leptokurtic populations (e.g., LEP3, XLEP), the Type I Error rate of all tests were affected when sample sizes were unequal (see Table 7). When the normally distributed data were compared to samples from platykurtic populations (PLAT), the Type I error rate was controlled for all tests with equal sample sizes. When sample sizes were not equal, only *Fmax* and BF were consistently robust to these violations to the normality assumption. When the larger group was more platykurtic, OB, WOB, and OB_{BF} tended to inflate the Type I error rate (see Table 7). Thus it would appear that if data are sampled from different populations with similar kurtosis, keeping group sizes approximately equal would be a reasonable step in controlling the Type I error rate.

In some situations where the kurtosis of the sampled distributions differed in sign, the Type I error rate of *Fmax* remained under the nominal alpha of .05. However, when the disparity in kurtosis increased this was not the case. For example, in comparing the variance of data sampled form the extremely platykurtic population (XPLT, γ_2 = -1.80) to the variance of samples from highly leptokurtic distributions (LEP3, γ_2 = 3.00), no test was robust (see Table 6). Thus, it appears that if the kurtosis of distributions differ in sign to the same absolute degree, then the *Fmax* test of variance is robust. This supposition was confirmed in an *ad-hoc* simulation in which the variance of data sampled from the XPLT distribution was tested against the variance of two leptokurtic distributions with population kurtosis values of γ_2 = 1.75 and 2.00. When comparing these variances under the null hypothesis, the Type I error rate of *Fmax* remained under the nominal alpha of .05 while all other tests were not robust.

Power. It should be noted that since Type I error rates for these tests of variance were dependent on the sample size and population kurtosis configuration, power was also dependent on combinations of sample size, population kurtosis, and group variance. In general, when the group with the larger variance was sampled from the heavier-tailed distribution there was more power for the tests of variance. When the more leptokurtic distribution was more variant, a reduction in power was observed. Therefore, results comparing the power of these tests are reported for both situations.

In quasi-experimental situations in which one group was sampled from the extremely platykurtic population, only *Fmax* controlled the Type I error. Therefore, only *Fmax* can be validly used for testing variances when only one group is sampled from an extremely platykurtic population. As this negative kurtosis increased in value and became less extreme, more comparisons were possible.

When one group was sampled from the platykurtic population (PLAT, γ_2 = -1.00) while the other group was normally distributed, all tests of variance held the Type I error rate for equal sample sizes and are comparable (see Table 7). Table 9 shows that under these conditions, OB and OB_{BF} were the most powerful. With unequal sample sizes, OB, WOB, and OB_{BF}, tended to inflate the Type I error rate, and therefore, *Fmax* and BF seem to be the most dependable tests. Furthermore, when there was a positive sample size/variance correlation, WBF was robust and more powerful as long as the disparity in sample sizes was not extreme. With an inverse sample size/variance relationship, *Fmax* is robust and adequately powerful. However, one may consider that OB, WOB, and OB_{BF} only inflated the Type I error rate when the more platykurtic group was larger in size. Thus, under conditions where the sample sizes are equal or the smaller group is more platykurtic, OB and OB_{BF} were more powerful except when the larger (more leptokurtic) sample had the larger variance, in which case, WOB was more powerful.

As with the extremely platykurtic population, the Type I error rate was controlled by *Fmax* when the platykurtic distribution (PLAT, γ_2 = -1.00) was compared to a group sampled from a population with an equal degree of leptokurtosis (LEP1, γ_2 = 1.00); however, no other test was robust (see Table 8). Thus

for a test of variance to be valid when one group is platykurtic, the other group must be either (a) similarly platykurtic, (b) symmetric, or (c) leptokurtic to the same degree. If the sampled distributions are similarly platykurtic, OB or WOB are preferred. If a second group is symmetric in shape then overall, *Fmax* is adequate; however, if the platykurtic distribution has more variance, OB, WOB, BF, WBF may be considered. If the kurtosis of groups differ in sign to the same degree, only *Fmax* is adequate.

Table 9 also shows that when normally distributed scores were compared to data sampled from a leptokurtic distribution with $\gamma_2 = 1.00$, all tests of variance that did not violate the Type I error rate were similarly powerful. Since OB and BF exhibited similar power, one of the conditional procedures may be used to decide which test to perform. That is, BF_{OB} or OB_{BF} can provide more power (Ramsey, 1994). With a positive sample size/variance correlation, the Welch procedures (WOB and WBF) showed more power relative to the other tests. With a negative sample size/variance correlation, OB remained the test of choice. Similar findings extend to situations where one group was leptokurtic ($\gamma_2 = 1.00$) and the other was slightly platykurtic ($\gamma_2 = -0.50$; see Table 10). However, in this situation the Welch procedures were more likely to inflate the Type I error rate, and BF should be considered when the sample size-variance correlation is positive.

As was the case when samples were selected from the same leptokurtic distribution, sampling from different leptokurtic populations demonstrated the superiority of the BF procedure and its variants. For example, Table 10 shows comparative power estimates for the tests of variance when one group was sampled from an extremely leptokurtic population (LEP3, $\gamma_2 = 3.00$) while the other group was less leptokurtic ($\gamma_2 = 1.00$). With equal sample sizes, BF and BF_{OB} were more powerful. As was the case in True Experiments, the high power of BF_{OB} relative to BF indicates the effectiveness of testing kurtosis before applying a test of variance (Ramsey, 1994). When the sample size/variance correlation is positive, WBF was clearly the most powerful procedure, while OB was more powerful with a negative relationship. As with the results for True Experiments, the advantage of OB with an inverse sample size/variance relationship dissipated in high power situations (*VR* =5.0, results not shown).

Discussion

Summary

The results demonstrated that when data were sampled from the same population and randomly assigned (i.e., True Experiments) to equally sized groups, Hartley's *Fmax* test was only robust when the population kurtosis was near or below zero. This confirms the findings of many other studies and establishes the need for analytic alternatives for testing variances when data are nonnormal. When data had a negative kurtosis, the O'Brien (1981) transformation was generally the best choice, while the Brown & Forsythe (1974) transformation was robust and showed superior power for testing variances in leptokurtic data. Also consistent with previous studies, the O'Brien test was generally more powerful when sample sizes and variances were negatively correlated, regardless of the shape of the distribution (Algina et al., 1989; Olejnik & Algina, 1988; Ramsey, 1994). The Welch procedure performed on O'Brien scores was more powerful when the sample size/variance correlation was positive in platykurtic samples (Algina et al., 1989).

Furthermore, under the conditions of a positive sample size/variance correlation in leptokurtic samples, the Welch test applied to Brown-Forsythe scores was robust and demonstrated superior power. Although this finding seems reasonable given previous research, it had yet to be empirically confirmed until this study. The results also demonstrated that choosing a test of variance based on an initial test of kurtosis can increase power (Ramsey, 1994); however, the power of these conditional tests has been shown to be dependent on the power of the test of kurtosis (Beasley & O'Connor, 1995). Thus, if tests of kurtosis are to be used to determine the most powerful and appropriate test of variance to perform, one must be concerned with the power of both tests.

This study also presented many new findings about the statistical properties of testing variances when groups were not sampled from the same population (i.e., Quasi-Experiments). When both groups were sampled from similarly platykurtic or similarly leptokurtic distributions, the results were predictable from the results of True Experiments. However, when one group was extremely platykurtic, only the *Fmax* tests controlled the Type I error rate. Furthermore, if the kurtosis of the groups differed in sign to the same absolute degree, the *Fmax* test was robust.

When the variance of normally distributed data were tested against the variance of data sampled from a platykurtic population, the Type I error rate of many tests were less stable which in turn affected the validity of power estimates and recommendations for use. Most notably, when the larger group was normally distributed with a larger variance and the smaller, platykurtic group was less variable, WOB was robust and more powerful. *Fmax* was preferable when sample sizes were equal or negatively correlated to variances. However, when the larger group was platykurtic, the Type I error rates of OB and WOB were inflated. Thus, when the more platykurtic group had a larger variance, the OB exhibited more power when the sample sizes were equal or inversely related to variance (platykurtic group was smaller in size). For a positive sample size/variance correlation (i.e., larger platykurtic group had larger variance) only BF was robust and adequately powerful.

When the variance of normally distributed data was tested against the variance of data sampled from leptokurtic populations, the Type I error rate of *Fmax* is extremely inflated and the BF is preferred when sample sizes are equal; however, OB showed similar power. When the sample size/variance correlation was positive the Welch test applied to BF scores is generally more powerful, while OB was more powerful when the smaller group had a larger variance despite the leptokurtic shape of one group.

Recommendations

Educational researchers are typically interested in estimating change and differences. However, simply examining shifts in central location does not fully address these issues, all distributional differences should be investigated. Thus testing all moments in the distribution is recommended when comparing groups whether they are intact or randomly constructed. Not only does this approach test the major assumptions for the ANOVA, but it also investigates the issue of whether a treatment condition affected the shape or response variability of a distribution of scores in True Experiments. In this case, tests such as the Kolmogorov-Smirnov test may be used to answer the question "Did the treatment affect the distribution of scores?" If intact groups are compared in central location or if differences in scale of the dependent variable are of interest, a test of variance is needed. The results demonstrate that the shape of the distributions should be examined before choosing a test of variance.

Table 11 shows a summary of these recommendations based on the kurtosis of the distributions and whether the sample sizes are equal, positively correlated, or negatively correlated with the variances. Entries on the diagonal exhibit recommendations for data sampled from the same (i.e., True Experiments) or similar populations. Off-diagonal entries reveal the recommendations for Quasi-Experiments and field research. Since the conditional tests examined are used to select one of these tests of variance (i.e., OB and BF), they are not represented. Furthermore, conditionally choosing the most powerful test based on sample characteristics may capitalize on chance differences in the data and inflate the Type I error rate.

In evaluating the recommendations in Table 11, one should consider that educational data tends to be platykurtic in nature (Micceri, 1989). It should also be noted that the recommendations for situations where the groups are either both leptokurtic or both platykurtic extend to most values of kurtosis. However, one should be aware that for situations in which one group is leptokurtic and the other is platykurtic the recommendations in Table 11 apply only if the kurtosis is of similar absolute value. Thus it is suggested that all relevant tests of variance be performed and agreement among the results assessed. If all tests reject the null hypothesis under conditions in which the Type I error rate is controlled, then the statistical significance is likely to represent a valid result. If there is disagreement among tests, then the consistency of disagreements with empirical findings should be assessed. For example, if equally sized, platykurtic samples are tested for variance heterogeneity and only the O'Brien test rejects the null hypothesis, there is indication of statistical significance because the O'Brien test is robust and most powerful in this situation (see Table 11).

Although this investigation was limited to the two-sample tests, it is believed that these results extend to most multi-group situations. For True Experiments, other studies have shown this to be the case (e.g., Miller, 1968). For Quasi-Experiment recommendations, one should consider the several factors. If the kurtosis values for all groups indicate similar positive or similar negative kurtosis, then the recommendations for leptokurtic and platykurtic groups in Table 11 should be valid. Also if about half of the groups are mesokurtic while the other half are either lepto- or platykurtic, then Table 11 can be used. If the groups are mostly leptokurtic, using the leptokurtic recommendations is advised; however, if the groups are mostly platykurtic, recommendations are more difficult to make. If the groups have drastically different shapes, the results indicated that *Fmax* was the preferred test in the two group situation, but one must consider that the *Fmax* only uses the data of two groups. Thus, if multiple groups are present and the groups with the largest and smallest variances (the values used for *Fmax*) have kurtosis estimates of opposite signs, *Fmax* may be allowable as long as the kurtosis values have approximately the same absolute value.

References

Algina, J., Olejnik, S. F., & Ocanto, R. (1989). Error rates and power estimates forselected two-sample tests of scale. *Journal of Educational Statistics, 14,* 373-384.

Beasley, T. M., & O'Connor, S. A. (1995, April). Testing heterogeneous variances in true and quasi-experiments. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

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

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69,* 364-367.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10*, 417-451.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances with applications to the outer continental shelf bidding data. *Technometrics, 23*, 351-361.

Durrand, A. L. (1969). *Comparative power of various tests of homogeneity of variance.* Unpublished Master's Thesis, University of Colorado, Boulder, Colorado.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Hartley, H. O. (1950). The maximum *F*-ratio as a short-cut test for heterogeneity of variance. *Biometrika, 37*, 308-312.

Klotz, J. (1962). Nonparametric tests for scale. *Annals of Mathematical Statistics, 33*, 495-512.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278-292). Palo Alto, CA: Stanford University Press.

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

Miller, R. G., Jr. (1968). Jackknifing variances. *Annals of Mathematical Statistics, 39*, 567-582.

Moses. L. E. (1963). Rank tests of dispersion. *Annals of Mathematical Statistics, 34*, 973-983.

O'Brien, R. G. (1979). An improved ANOVA method for robust tests of additive models of variance. *Journal of the American Statistical Association, 74*, 877-880.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin, 89*, 570-574.

Olejnik, S. J., & Algina, J. (1985, April). *Power analysis of selected parametric and nonparametric test for heterogeneous variances in non-normal distributions.* Paper presented at the meeting of the American Educational Research Association. Chicago, IL.

Olejnik, S. J., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics, 12*, 45-61.

Olejnik, S. J., & Algina, J. (1988). Test of variance equality when distributions differ in form and location. *Educational and Psychological Measurement, 48*, 317-329.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-242.

Penfield, D. A., & Koffler, S. (1985, April). *A power study of selected nonparametric k-sample tests.* Paper presented at the meeting of the American Educational Research Association. Chicago, IL.

Ramsey, P. H. (1994). Testing variances in psychological and educational research. *Journal of Educational Statistics, 19*, 23-42.

Ramsey, P. H., & Brailsford, E. A. (1990). Robustness and power of tests of variability on two independent groups. *British Journal of Mathematical and Statistical Psychology, 43*, 113-130.

Ramsey, P. H., & Ramsey, P. P. (1993). Updated version of the critical values of the standardized fourth moment. *Journal of Statistical Computation and Simulation, 44*, 231-241.

SAS Institute. (1990). *SAS/IML user's guide* (Release 6.04). Cary, NC: Author.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2*, 110-114.

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

Siegel, S., & Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *American Statistical Association Journal, 55*, 429-445.

Welch, B. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*, 330-336.

Table 1. Summary of conditions analyzed for Sample Size and Population configurations

	True Experiment		Quasi-Experiment		
	Type I	Power	Type I	Power	Power
(n_1, n_2)	Error	$(\sigma_1^2 < \sigma_2^2)$	Error	$(\sigma_1^2 < \sigma_2^2)$	$(\sigma_1^2 > \sigma_2^2)$
(10, 10)	*	*	*	*	*
(13, 13)	*	*	*	*	*
(20, 20)	*	*	*	*	*
(10, 20)	*	+	*	+	-
(13, 20)	*	+	*	+	-
(20, 10)	U	-	*	-	+
(20, 13)	U	-	*	-	+

Note. * indicates the analysis was completed. U indicates the analysis was unnecessary and not completed. - indicates a negative relationship between sample size and variance. + indicates a positive relationship between sample size and variance.

Table 2. Average population parameters across Type I error simulations.

	Population Parameter			
Population	μ	σ^2	γ_1	γ_2
1. **XPLT** $E(\gamma_2) = -1.80$				
$n = 10$	+0.003937	1.009481	-0.001320	-0.964967
$n = 13$	+0.003508	1.011025	-0.006166	-1.205637
$n = 20$	-0.003517	1.006298	-0.010828	-1.447347
2. **PLAT** $E(\gamma_2) = -1.00$				
$n = 10$	+0.005843	1.007704	-0.012320	-0.421073
$n = 13$	+0.005508	1.009243	-0.008166	-0.541785
$n = 20$	+0.013517	1.011201	-0.030828	-0.686890
3. **SPLT** $E(\gamma_2) = -0.50$				
$n = 10$	+0.000090	1.017263	+0.039203	-0.172836
$n = 13$	+0.001756	1.015814	+0.046087	-0.233373
$n = 20$	-0.004001	1.012889	+0.058859	-0.302452
4. **NORM** $E(\gamma_2) = 0.00$				
$n = 10$	-0.002803	1.006276	-0.027654	-0.003053
$n = 13$	-0.001429	1.001976	-0.034611	-0.009850
$n = 20$	+0.000057	1.003101	-0.028943	-0.010843
5. **LEP1** $E(\gamma_2) = +1.00$				
$n = 10$	-0.000713	1.024512	+0.061602	+0.261224
$n = 13$	+0.058517	1.022624	+0.065461	+0.339038
$n = 20$	+0.023376	1.027412	+0.052777	+0.493708
6. **LEP3** $E(\gamma_2) = +3.00$				
$n = 10$	+0.011568	1.028524	+0.032652	+0.553391
$n = 13$	+0.008883	1.012206	+0.035397	+0.765899
$n = 20$	+0.005923	1.020455	+0.213038	+1.218371
7. **XLEP** $E(\gamma_2) = +3.75$				
$n = 10$	+0.010142	1.013801	+0.040116	+0.662791
$n = 13$	+0.005774	1.028358	+0.017173	+0.904860
$n = 20$	+0.004539	0.998350	+0.017142	+1.357478

Table 3. Empirical Type I Error Rate for seven procedures in True Experiments with no differences in central location.

Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.0006	.0248	.0338	.0242	.0292	.0248	.0284
	13, 13	.0018	.0272	.0058	.0262	.0054	.0272	.0272
	20, 20	.0006	.0262	.0186	.0256	.0172	.0262	.0272
	10, 20	.0030	.0372	.0366	.0288	.0316	.0372	.0380
	13, 20	.0026	.0296	.0232	.0288	.0114	.0294	.0308
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0152	.0438	.0344	.0380	.0324	.0444	.0392
	13, 13	.0142	.0464	.0256	.0430	.0248	.0468	.0400
	20, 20	.0082	.0474	.0386	.0456	.0378	.0476	.0444
	10, 20	.0124	.0494	.0384	.0580*	.0436	.0498	.0474
	13, 20	.0110	.0464	.0326	.0510	.0316	.0464	.0424
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0354	.0396	.0330	.0320	.0312	.0400	.0362
	13, 13	.0340	.0416	.0306	.0366	.0294	.0422	.0346
	20, 20	.0306	.0486	.0436	.0458	.0432	.0498	.0456
	10, 20	.0368	.0498	.0458	.0604*	.0526	.0518	.0494
	13, 20	.0332	.0458	.0382	.0472	.0414	.0468	.0406
4. **NORM** (γ_2 = 0.00)								
	10, 10	.0484	.0336	.0402	.0258	.0370	.0352	.0414
	13, 13	.0528	.0384	.0312	.0320	.0274	.0402	.0346
	20, 20	.0460	.0416	.0366	.0392	.0358	.0428	.0368
	10, 20	.0462	.0364	.0392	.0530	.0492	.0380	.0400
	13, 20	.0478	.0422	.0344	.0442	.0356	.0440	.0380
5. **LEP1** (γ_2 = 1.00)								
	10, 10	.0648*	.0352	.0404	.0260	.0368	.0366	.0412
	13, 13	.0648*	.0362	.0324	.0316	.0298	.0376	.0358
	20, 20	.0678*	.0380	.0398	.0352	.0382	.0412	.0388
	10, 20	.0624*	.0368	.0366	.0460	.0456	.0388	.0370
	13, 20	.0664*	.0408	.0328	.0396	.0348	.0416	.0362
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.1466*	.0278	.0328	.0188	.0286	.0294	.0338
	13, 13	.1538*	.0332	.0346	.0278	.0316	.0364	.0360
	20, 20	.1766*	.0292	.0364	.0256	.0348	.0362	.0368
	10, 20	.1476*	.0366	.0358	.0438	.0494	.0420	.0368
	13, 20	.1567*	.0316	.0384	.0302	.0396	.0368	.0392
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1400*	.0240	.0340	.0168	.0278	.0266	.0348
	13, 13	.1496*	.0294	.0338	.0234	.0298	.0326	.0346
	20, 20	.1752*	.0346	0412	.0300	.0380	.0438	.0418
	10, 20	.1502*	.0382	.0372	.0448	.0506	.0416	.0390
	13, 20	.1662*	.0340	.0370	.0344	.0412	.0388	.0384

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of α = .05.

Table 4. Empirical Power for seven procedures in True Experiments with no differences in central location and σ_1^2 = 1.0 and σ_2^2 = 2.0.

Pop.	n_1,n_2	F_{max}	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.0228	.4566	.1842	.4424	.1742	.4558	.4502
	13, 13	.0336	.6310	.0444	.6168	.0420	.6296	.6196
	20, 20	.0856	.8662	.2306	.8608	.2278	.8662	.8630
Pos.	10, 20	.0282	.6502	.2004	.7602	.2024	.6494	.6470
Pos.	13, 20	.0392	.7640	.2044	.8330	.1952	.7642	.7620
Neg.	20, 10	.0546	.7038	.1846	.5224	.2192	.7036	.6964
Neg.	20, 13	.0686	.7544	.0600	.6422	.0302	.7542	.7458
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0984	.1412	.1116	.1188	.1050	.1424	.1262
	13, 13	.1364	.2206	.1348	.2020	.1300	.2204	.1648
	20, 20	.2680	.4086	.2950	.3950	.2910	.4084	.3392
Pos.	10, 20	.1230	.1600	.1716	.3420*	.2480	.1620	.1804
Pos.	13, 20	.1530	.2424	.2144	.3612	.2668	.2442	.2384
Neg.	20, 10	.1898	.3176	.1740	.1184*	.0918	.3176	.2142
Neg.	20, 13	.2150	.3390	.1912	.1952	.1264	.3392	.2390
4. **NORM** (γ_2 = 0.00)								
	10, 10	.1508	.0994	.1070	.0782	.0978	.1030	.1130
	13, 13	.2022	.1454	.1298	.1262	.1226	.1484	.1368
	20, 20	.2850	.2348	.2182	.2190	.2122	.2412	.2250
Pos.	10, 20	.1620	.0750	.1230	.2502	.2172	.0806	.1244
Pos.	13, 20	.2116	.1330	.1732	.2416	.2282	.1406	.1758
Neg.	20, 10	.2530	.2452	.1616	.0650	.0766	.2454	.1734
Neg.	20, 13	.2572	.2382	.1654	.1072	.0952	.2396	.1752
5. **LEP1** (γ_2 = 1.00)								
	10, 10	---	.0686	.0882	.0522	.0782	.0728	.0904
	13, 13	---	.1040	.1114	.0870	.1044	.1118	.1148
	20, 20	---	.1610	.1864	.1470	.1812	.1850	.1888
Pos.	10, 20	---	.0462	.1016	.1934	.2158	.0562	.1002
Pos.	13, 20	---	.0766	.1372	.1710	.1976	.0952	.1376
Neg.	20, 10	---	.1912	.1528	.0400	.0630	.1982	.1580
Neg.	20, 13	---	.1798	.1544	.0668	.0834	.1904	.1576
7. **XLEP** (γ_2 = 3.75)								
	10, 10	---	.0602	.0830	.0444	.0724	.0672	.0848
	13, 13	---	.0830	.1010	.0654	.0900	.0910	.1020
	20, 20	---	.1438	.1776	.1322	.1712	.1764	.1792
Pos.	10, 20	---	.0302	.0916	.1680	.2020	.0438	.0910
Pos.	13, 20	---	.0746	.1376	.1568	.2034	.0984	.1384
Neg.	20, 10	---	.1620	.1336	.0280	.0492	.1688	.1366
Neg.	20, 13	---	.1520	.1386	.0536	.0786	.1668	.1420

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 5. Empirical Power for seven procedures in True Experiments with no differences in central location and $\sigma_1^2 = 1.0$ and $\sigma_2^2 = 5.0$.

Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.6612	.5450	.4874	.4762	.4532	.5476	.5134
	13, 13	.8402	.7594	.6582	.7186	.6332	.7600	.6974
	20, 20	.9724	.9590	.9104	.9526	.9062	.9592	.9406
Pos.	10, 20	.7794	.6856	.7010	.9348	.8618	.6878	.7170
Pos.	13, 20	.9040	.8550	.8138	.9452	.8858	.8554	.8352
Neg.	20, 10	.8782	.8684	.7238	.4664	.4442	.8688	.7788
Neg.	20, 13	.9278	.9158	.8086	.7322	.6380	.9160	.8544
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.6612	.5450	.4874	.4762	.4532	.5476	.5134
	13, 13	.8402	.7594	.6582	.7186	.6332	.7600	.6974
	20, 20	.9724	.9590	.9104	.9526	.9062	.9592	.9406
Pos.	10, 20	.7794	.6856	.7010	.9348*	.8618	.6878	.7170
Pos.	13, 20	.9040	.8550	.8138	.9452	.8858	.8554	.8352
Neg.	20, 10	.8782	.8684	.7238	.4664*	.4442	.8688	.7788
Neg.	20, 13	.9278	.9158	.8086	.7322	.6380	.9160	.8544
4. **NORM** (γ_2 = 0.00)								
	10, 10	.6396	.3616	.4136	.2876	.3772	.3698	.4200
	13, 13	.7690	.5528	.5646	.4924	.5392	.5630	.5726
	20, 20	.9350	.8428	.8330	.8198	.8266	.8594	.8396
Pos.	10, 20	.7264	.3800	.5680	.7796	.7604	.3964	.5654
Pos.	13, 20	.8388	.5860	.6978	.7894	.8008	.6052	.6992
Neg.	20, 10	.8148	.7458	.6470	.2860	.3664	.7538	.6584
Neg.	20, 13	.8648	.7880	.7084	.4954	.5342	.7986	.7224
5. **LEP1** (γ_2 = 1.00)								
	10, 10	---	.2600	.3432	.1984	.3086	.2762	.3446
	13, 13	---	.4088	.4924	.3574	.4640	.4446	.4944
	20, 20	---	.6656	.7536	.6384	.7438	.7454	.7550
Pos.	10, 20	---	.2186	.4674	.6032	.6792	.2690	.4628
Pos.	13, 20	---	.3838	.5948	.6250	.7142	.4524	.5934
Neg.	20, 10	---	.6160	.5600	.1838	.2890	.6372	.5650
Neg.	20, 13	---	.6400	.6362	.3452	.4530	.6852	.6406
6. **LEP3** (γ_2 = 3.00)								
	10, 10	---	.2062	.2990	1590	.2636	.2268	.2986
	13, 13	---	.2948	.4090	.2490	.3866	.3502	.4094
	20, 20	---	.4962	.6662	.4616	.6532	.6556	.6676
Pos.	10, 20	---	.1400	.3666	.4794	.6086	.2082	.3598
Pos.	13, 20	---	.2610	.4902	.4750	.6274	.3678	.4878
Neg.	20, 10	---	.5262	.5166	.1256	.2366	.5696	.5184
Neg.	20, 13	---	.5110	.5590	.2358	.3732	.5888	.5606

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 6. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is platykurtic $\gamma_2 = -1.80$.

Group 2 Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
2. **PLAT** ($\gamma_2 = -1.00$)								
	10, 10	.0092	.0620*	.0484	.0584*	.0460	.0640*	.0632*
	13, 13	.0068	.0642*	.0114	.0616*	.0110	.0654*	.0634*
	20, 20	.0022	.0542	.0476	.0524	.0460	.0550	.0550
	10, 20	.0048	.0282	.0408	.0402	.0658*	.0348	.0364
	13, 20	.0044	.0362	.0160	.0454	.0126	.0388	.0408
	20, 10	.0062	.1178*	.0532	.0808*	.0460	.1180*	.1178*
	20, 13	.0054	.0840*	.0494	.0600*	.0418	.0842*	.0844*
3. **SPLT** ($\gamma_2 = -0.50$)								
	10, 10	.0148	.0734*	.0600*	.0694*	.0568*	.0752*	.0738*
	13, 13	.0116	.0750*	.0160	.0742*	.0150	.0766*	.0742*
	20, 20	.0084	.0674*	.0582*	.0656*	.0566*	.0678*	.0680*
	10, 20	.0082	.0358	.0458	.0462	.0682*	.0420	.0452
	13, 20	.0068	.0468	.0162	.0560	.0122	.0492	.0492
	20, 10	.0096	.1238*	.0660*	.0936*	.0580*	.1240*	.1230*
	20, 13	.0108	.0992*	.0560	.0782*	.0500	.0994*	.0990*
4. **NORM** ($\gamma_2 = 0.00$)								
	10, 10	.0216	.0860*	.0668*	.0832*	.0628*	.0880*	.0858*
	13, 13	.0170	.0856*	.0178	.0830*	.0156	.0874*	.0826*
	20, 20	.0136	.0748*	.0676*	.0738*	.0644*	.0756*	.0786
	10, 20	.0090	.0438	.0464	.0560	.0752*	.0480	.0516
	13, 20	.0122	.0606*	.0202	.0686*	.0132	.0636*	.0610*
	20, 10	.0234	.1530*	.0814*	.1168*	.0780*	.1528*	.1508*
	20, 13	.0156	.1238*	.0752*	.0988*	.0652*	.1244*	.1266*
5. **LEP1** ($\gamma_2 = 1.00$)								
	10, 10	.0324	.1090*	.0712*	.1052*	.0674*	.1106*	.1054*
	13, 13	.0262	.1038*	.0248	.1022*	.0230	.1052*	.0992*
	20, 20	.0252	.1036*	.0872*	.1022*	.0836*	.1040*	.1074*
	10, 20	.0160	.0614*	.0548	.0706*	.0838*	.0642*	.0652*
	13, 20	.0224	.0818*	.0252	.0886*	.0144	.0846*	.0790*
	20, 10	.0460	.1930*	.1034*	.1732*	.1068*	.1934*	.1934*
	20, 13	.0354	.1624*	.1078*	.1440*	.1018*	.1630*	.1626*
6. **LEP3** ($\gamma_2 = 3.00$)								
	10, 10	.0538	.1392*	.0934*	.1354*	.0884*	.1400*	.1366*
	13, 13	.0564*	.1470*	.0398	.1454*	.0366	.1470*	.1380*
	20, 20	.0608*	.1316*	.1172*	.1300*	.1140*	.1316*	.1438*
	10, 20	.0328	.0844*	.0680*	.0888*	.1012*	.0856*	.0856*
	13, 20	.0446	.0990*	.0404	.1036*	.0234	.1006*	.0942*
	20, 10	.0820*	.2430*	.1338*	.2182*	.1518*	.2426*	.2398*
	20, 13	.0804*	.2188*	.1418*	.1890*	.1496*	.2190*	.2178*

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$.

Table 7. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is normally distributed $\gamma_2 = 0$.

Group 2 Pop.	n_1, n_2	*Fmax*	OB	BF	WOE	WBF	OB_{BF}	BF_{OB}
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0362	.0458	.0432	.0420	.0404	.0470	.0484
	13, 13	.0280	.0472	.0340	.0434	.0328	.0478	.0412
	20, 20	.0280	.0532	.0500	.0516	.0492	.0534	.0506
	10, 20	.0304	.0588*	.0536	.0860*	.0668*	.0606*	.0588*
	13, 20	.0304	.0566*	.0484	.0654*	.0536	.0578*	.0534
	20, 10	.0304	.0516	.0390	.0468	.0404	.0532	.0456
	20, 13	.0228	.0422	.0328	.0360	.0282	.0430	.0400
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0446	.0400	.0396	.0312	.0370	.0416	.0412
	13, 13	.0408	.0444	.0312	.0368	.0296	.0454	.0372
	20, 20	.0428	.0482	.0414	.0444	.0412	.0498	.0430
	10, 20	.0442	.0404	.0404	.0670*	.0592	.0416	.0406
	13, 20	.0422	.0474	.0424	.0522	.0460	.0490	.0442
	20, 10	.0424	.0394	.0354	.0454	.0424	.0410	.0380
	20, 13	.0370	.0442	.0310	.0402	.0338	.0454	.0356
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.1048*	.0478	.0500	.0374	.0440	.0496	.0510
	13, 13	.1060*	.0460	.0442	.0370	.0404	.0494	.0464
	20, 20	.1188*	.0536	.0562*	.0488	.0542	.0598*	.0572*
	10, 20	.0960*	.0586*	.0470	.0292	.0366	.0602*	.0492
	13, 20	.1018*	.0564*	.0454	.0340	.0364	.0584*	.0482
	20, 10	.1194*	.0434	.0614*	.0990*	.0926*	.0474	.0624*
	20, 13	.1184*	.0524	.0638*	.0730*	.0746*	.0568*	.0640*
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1076*	.0422	.0494	.0326	.0458	.0446	.0500
	13, 13	.1174*	.0480	.0474	.0406	.0452	.0524	.0510
	20, 20	.1228*	.0564*	.0658*	.0526	.0648*	.0660*	.0650*
	10, 20	.1056*	.0584*	.0468	.0280	.0336	.0614*	.0484
	13, 20	.1166*	.0554	.0510	.0344	.0384	.0606*	.0524
	20, 10	.1272*	.0496	.0680*	.1040*	.0996*	.0550	.0686*
	20, 13	.1376*	.0548	.0676*	.0802*	.0814*	.0594*	.0672*

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$ test.

Table 8. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is leptokurtic $\gamma_2 = 1.00$.

Group 2 Pop.	n_1,n_2	F_{max}	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0406	.0524	.0434	.0436	.0410	.0530	.0492
	13, 13	.0444	.0622*	.0460	.0566*	.0436	.0632*	.0546
	20, 20	.0448	.0684*	.0616*	.0656*	.0606*	.0696*	.0656*
	10, 20	.0538	.0730*	.0720*	.1134*	.0980*	.0756*	.0740*
	13, 20	.0478	.0752*	.0744*	.0954*	.0826*	.0768*	.0756*
	20, 10	.0268	.0520	.0362	.0400	.0336	.0520	.0438
	20, 13	.0350	.0548	.0352	.0402	.0322	.0550	.0446
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0602*	.0436	.0412	.0350	.0368	.0448	.0442
	13, 13	.0548	.0472	.0350	.0394	.0316	.0474	.0404
	20, 20	.0528	.0510	.0464	.0460	.0458	.0534	.0480
	10, 20	.0622*	.0484	.0556	.0922*	.0798*	.0498	.0576*
	13, 20	.0586*	.0468	.0494	.0630*	.0562*	.0492	.0502
	20, 10	.0458	.0546	.0380	.0386	.0410	.0558	.0404
	20, 13	.0436	.0452	.0358	.0354	.0288	.0474	.0394
4. **NORM** (γ_2 = 0.00)								
	10, 10	.0702*	.0322	.0378	.0242	.0348	.0336	.0396
	13, 13	.0702*	.0370	.0302	.0302	.0272	.0382	.0342
	20, 20	.0772*	.0470	.0460	.0424	.0444	.0516	.0492
	10, 20	.0734*	.0444	.0444	.0714*	.0616*	.0466	.0452
	13, 20	.0700*	.0400	.0432	.0514	.0496	.0428	.0444
	20, 10	.0628*	.0436	.0378	.0402	.0420	.0464	.0402
	20, 13	.0670*	.0400	.0342	.0346	.0306	.0422	.0360
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.0858*	.0326	.0370	.0264	.0328	.0336	.0380
	13, 13	.0812*	.0366	.0354	.0316	.0320	.0384	.0382
	20, 20	.0982*	.0472	.0472	.0426	.0460	.0512	.0480
	10, 20	.0858*	.0462	.0392	.0350	.0408	.0496	.0402
	13, 20	.0936*	.0434	.0390	.0304	.0342	.0480	.0404
	20, 10	.1134*	.0384	.0476	.0682*	.0722*	.0424	.0478
	20, 13	.1248*	.0380	.0448	.0522	.0540	.0428	.0456
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1012*	.0352	.0382	.0254	.0338	.0380	.0396
	13, 13	.1078*	.0374	.0360	.0296	.0326	.0400	.0360
	20, 20	.1200*	.0402	.0492	.0360	.0472	.0500	.0494
	10, 20	.0996*	.0492	.0382	.0282	.0342	.0530	.0040
	13, 20	.1110*	.0464	.0438	.0350	.0338	.0526	.0444
	20, 10	.1264*	.0408	.0514	.0704*	.0734*	.0452	.0534
	20, 13	.1210*	.0382	.0496	.0568*	.0584*	.0432	.0502

Note. * indicates the Type I error rate exceed .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$ test.

Table 9. Empirical Power Estimates for seven procedures in Quasi-Experiments with no differences in central location and Group One is normally distributed $\gamma_2 = 0.00$.

								Group Two	
PLAT	γ_2 = -1.00		$\sigma_1^2 = 1.0$			$\sigma_2^2 = 2.0$			
Pop.	n_1, n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}	
	10, 10	.1540	.1724	.1550	.1496	.1462	.1748	.1682	
	13, 13	.1868	.2386	.1862	.2234	.1784	.2418	.2046	
	20, 20	.2956	.3974	.3526	.3888	.3488	.4032	.3672	
Pos.	10, 20	.1684	.2120*	.2358	----	----	----	.2380*	
Pos.	13, 20	.2064	.2746*	.2758	----	.3274	.2812*	.2842	
Neg.	20, 10	.2104	.2906	.2048	.1242	.1248	.2940	.2198	
Neg.	20, 13	.2276	.3222	.2246	.2038	.1542	.3256	.2456	
Group Two **PLAT**	γ_2 = -1.00			$\sigma_1^2 = 2.0$			$\sigma_2^2 = 1.0$		
	10, 10	.1292	.0710	.0722	.0540	.0652	.0722	.0760	
	13, 13	.1810	.1150	.0856	.0964	.0806	.1154	.0960	
	20, 20	.2998	.2264	.1748	.2086	.1706	.2268	.1978	
Neg.	10, 20	.2406	.2520*	.1386	----	----	----	.1762*	
Neg.	13, 20	.2686	.2396*	.1340	----	.0718	.2402*	.1700	
Pos.	20, 10	.1268	.0474	.0930	.2142	.1762	.0512	.0908	
Pos.	20, 10	.1890	.0976	.1232	.2182	.1770	.0992	.1262	
Group Two **LEP1**	γ_2 = 1.00			$\sigma_1^2 = 1.0$			$\sigma_2^2 = 2.0$		
	10, 10	----	.0640	.0798	.0458	.0722	.0670	.0814	
	13, 13	----	.0940	.0890	.0772	.0798	.0968	.0924	
	20, 20	----	.1690	.1690	.1556	.1640	.1844	.1736	
Pos.	10, 20	----	.0404	.0878	.1816	.1820	.0486	.0882	
Pos.	13, 20	----	.0692	.1130	.1618	.1670	.0794	.1130	
Neg.	20, 10	----	.1852	.1238	----	----	.1884	.1314	
Neg.	20, 13	----	.1676	.1224	.0608	.0682	.1732	.1306	
Group Two **LEP1**	γ_2 = 1.00			$\sigma_1^2 = 2.0$			$\sigma_2^2 = 1.0$		
	10, 10	----	.1038	.1148	.0826	.1054	.1088	.1178	
	13, 13	----	.1460	.1420	.1218	.1346	.1514	.1472	
	20, 20	----	.2526	.2646	.2386	.2598	.2690	.2658	
Neg.	10, 20	----	.2320	.1758	.0608	.0818	.2352	.1822	
Neg.	13, 20	----	.2348	.1920	.1058	.1118	.2404	.1984	
Pos.	20, 10	----	.0880	.1654	----	----	.0990	.1622	
Pos.	20, 13	----	.1412	.1998	.2518	.2632	.1574	.2010	

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 10. Empirical Power Estimates for seven procedures in Quasi Experiments with no differences in central location and Group One has positive kurtosis $\gamma_2 = 1.00$.

Group Two **SPLT**		$\gamma_2 = -0.50$			$\sigma_1^2 = 1.0$		$\sigma_2^2 = 2.0$	
Pop.	n_1, n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
	10, 10	---	.1390	.1362	.1102	.1272	.1442	.1418
	13, 13	.2428	.1954	.1782	.1722	.1682	.2012	.1854
	20, 20	.3410	.3082	.3042	.2960	.3006	.3218	.3084
Pos.	10, 20	---	.1278	.1866	----	----	.1382	.1864*
Pos.	13, 20	.2652*	.1878	.2338	----	.2974*	.1994	.2330
Neg.	20, 10	.2680	.2582	.2040	.0844	.1114	.2622	.2086
Neg.	20, 13	.2858	.2706	.2174	.1454	.1462	.2778	.2244
Group Two **SPLT**		$\gamma_2 = -0.50$			$\sigma_1^2 = 2.0$		$\sigma_2^2 = 1.0$	
	10, 10	---	.0588	.0674	.0452	.0602	.0610	.0700
	13, 13	.1740	.0796	.0748	.0638	.0674	.0814	.0782
	20, 20	.2730	.1408	.1286	.1298	.1244	.1486	.1362
Neg.	10, 20	.2242	.1832	.1098	.0350	.0424	.1854	.1218
Neg.	13, 20	.2552	.1682	.1150	.0614	.0574	.1718	.1258
Pos.	20, 10	---	.0318	.0698	----	----	.0364	.0690*
Pos.	20, 13	.2058*	.0694	.1032	----	.1510*	.0768	.1044
Group Two **LEP3**		$\gamma_2 = 3.00$			$\sigma_1^2 = 1.0$		$\sigma_2^2 = 2.0$	
	10, 10	---	.0460	.0630	.0344	.0530	.0496	.0650
	13, 13	---	.0690	.0782	.0558	.0704	.0768	.0812
	20, 20	---	.1048	.1238	.0978	.1196	.1278	.1236
Pos.	10, 20	---	.0272	.0722	.1450	.1654	.0362	.0716
Pos.	13, 20	---	.0444	.0832	.1174	.1386	.0576	.0836
Neg.	20, 10	---	.1388	.1052	----	----	.1460	.1086
Neg.	20, 13	---	.1272	.1062	.0436	.0530	.1386	.1096
Group Two **LEP3**		$\gamma_2 = 3.00$			$\sigma_1^2 = 2.0$		$\sigma_2^2 = 1.0$	
	10, 10	---	.0932	.1174	.0712	.1002	.0982	.1200
	13, 13	---	.1112	.1354	.0920	.1234	.1232	.1374
	20, 20	---	.1898	.2402	.1770	.2346	.2288	.2412
Pos.	20, 10	---	.0732	.1490	.2386	.2654	.0876	.1476
Pos.	20, 13	---	.1020	.1812	.2104	.2450	.1272	.1796
Neg.	10, 20	---	.1962	.1716	----	----	.2102	.1746
Neg.	13, 20	---	.2010	.1876	.0796	.1112	.2150	.1902

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 11. Recommendations based on Variance Ratios and whether the sample sizes are equal, positively correlated, or negatively correlated with the variances for both True and Quasi-Experiments.

Smaller Variance	Larger Variance			
	PLAT	***SPLT***	***NORM***	***LEPT***
PLAT				
Equal	OB	OB, *Fmax*	*Fmax*	*Fmax*
Positive	WOB	WOB	WOB	*Fmax*
Negative	OB	OB	*Fmax*	*Fmax*
SPLT				
Equal	OB, *Fmax*	OB, *Fmax*	*Fmax*, OB	OB, BF
Positive	WOB	WOB	WOB	BF
Negative	OB	OB, *Fmax*	*Fmax*, OB	OB
NORM				
Equal	OB	*Fmax*, OB	*Fmax*	BF, OB
Positive	BF	WOB	WOB	WBF, WOB
Negative	OB	*Fmax*, OB	*Fmax*	OB
LEPT				
Equal	*Fmax*	OB, BF	BF, OB	BF
Positive	*Fmax*	BF	BF, WBF	WBF
Negative	*Fmax*	OB	OB	OB

Note. Entries on the diagonal represent recommendations for True Experiments, while off-diagonal entries are for Quasi-Experiments. PLAT = platykurtic; SPLT = slightly platykurtic; NORM = Normal; LEPT = leptokurtic; OB = O'Brien BF = Brown-Forsythe; W refers to performing Welch procedure

MINUTES

OF THE
ANNUAL MEETING
OF THE
MULTIPLE LINEAR REGRESSION: GENERAL LINEAR MODEL SIG
(New Orleans, LA)

APRIL 19, 1995

Professor Adria Karle-Weiss (Murry State University), SIG Chair, opened the business meeting. The first order of business was the call for nominations of Chair-elect, three replacement Executive Board/Editorial Board members, and Executive Secretary. Executive Secretary, Steve Spaner (University of Missouri - St. Louis), explained that the MLRSIG election procedures call for the election to be held by mail ballot and the business meeting to be a nominating meeting only. It was moved and passed by the members attending to suspend the election by mail ballot rule and to hold the election at the business meeting. Nominations for chair-elect were Professors Randy Schumacher (University of North Texas) and Mark Beasley (St. John's University). Professor Randy Schumacher and was elected Chair-elect for 1996. His term of office will begin following the 1996 business meeting. The nominated Executive Board/Editorial Board replacements were Professors Carolyn Benz (University of Dayton), Mark Beasley , and Jeff Kromrey (University of South Florida). No additional nominations were offered so all were elected by acclamation. Professor Carolyn Benz will complete Professor John Pohlmann's (Southern Illinois University - Carbondale) term (1994 - 98). Pohlmann has taken on the MLRV Editorship thereby sitting on the Board by virtue of office. Professors Mark Beasley and Jeff Kromrey replace Board members Carl Huberty (University of Georgia) and Randy Schumacher and assume the four year terms from 1995 - 99. Finally, Steve Spaner was nominated for Executive Secretary for another three-year term (1995 - 98). No additional nominations were offered so Spaner was elected by acclamation.

Chair Karle-Weiss called upon Steve Spaner to give the treasurers report and the membership update. Spaner reported that the SIG treasury was $1995.56 on 1-1-94, the beginning of the new membership year, and $1917.93 just before the business meeting (3-31-95). Spaner reported that the current paid membership was down from 1994. Spaner attributed the decline to the reduced number of issues of and irregular schedule for the Multiple Linear Regression Viewpoints, the MLRSIG's journal.

(Secretary's note: 1995 membership payment was due at the beginning of the 1995 calendar year; if your mailing label has **94 or earlier** at the end of the first line, **you are unpaid** for the past 1995 MLRSIG membership year as well as now owing for the 1996 MLR:GLM/SIG membership year)

Discussion ensued following the Executive Secretary's report regarding the future of the MLRSIG and the MLRV. It was affirmed that the SIG wanted to continue and wanted to produce the journal. It was suggested, once again, that persons making presentations under the MLRSIG sponsorship at the AERA conference should at least be invited to submit their papers to the MLRV.

It was also pointed out that in the 1970s and 1980s the SIG sponsored presessions on MLR (Joe Ward, Earl Jennings to name a few) and on Path Analysis (Lee Wolfle, John Williams to name a few). These presessions were very successful and provided significant numbers of new members for the SIG. Past chair Keith McNeil and incoming chair Isadore Newman pledged to develop and conduct an AERA presession on MLR/GLM at the 1996 AERA meeting.

Comment was made regarding the name of the SIG (Multiple Linear Regression SIG) and the possible datedness of the term MLR verses, for example, GLM (General Linear Model). After considerable discussion a motion was offered and passed to change the name of the SIG to Multiple Linear Regression: General Linear Model Special Interest Group (MLR:GLM/SIG).

A comment regarding the role and responsibility of SIG Executive/Editorial Board members drew attention to the fact that some journals expect if not require their editorial board members to regularly contribute to their journal. A motion was made and passed to instruct the MLRV editor to send a letter to each MLRV Editorial Board member encouraging them to submit an article for review and publication in MLRV.

Finally, SIG member Sue Trace (California State University - Fresno) moved a resolution thanking the 1994 MLRSIG Chair, Adria Karle-Weiss for her leadership and work for the SIG over the past year, in particular her advocacy for participation and recognition of the female SIG members. The resolution passed unanimously.

The meeting was adjourned to allow members to attend the SIG's social that was to follow.

Respectfully submitted,

Steven D. Spaner, Executive Secretary
Multiple Linear Regression: General Linear Model SIG
Department of Behavioral Studies
University of Missouri - St. Louis
St. Louis, MO 63121-4499
sspaner@umslvma.umsl.edu

Information for Contributors

Multiple Linear Regression Viewpoints (MLRV), a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression (SIG/MLR), is published once or twice per year to facilitate communication among professionals who focus their investigations on the theory, application, or teaching of multiple linear regression models or their extensions. Also, the journal accepts news items of interest to members of the SIG/MLR.

All manuscripts submitted to *MLRV* should conform to the language, style, and format of the *Publication Manual of the American Psychological Association* (4th ed., 1994), available from Order Department, American Psychological Association, P. O. Box 2710, Hyattsville, MD 20784. Three copies of the manuscript, all double spaced (including equations, footnotes, quotes, and references) and accompanied by an abstract of 100 words or less, should be submitted to the editor at the address listed below. Mathematical symbols and Greek letters should be precise and clear and should leave no question as to interpretation. All figures must be camera ready. Manuscripts that do not conform to the above specifications may be returned to the author for style changes before the review process will begin. A submitted manuscript will receive a blind review from at least two members of the editorial board (except occasional invited contributions, letters to the editor, editorials, or news items). Any author identifying information should appear on the title page only. Efforts will be made to keep the review process to a maximum of eight weeks. The final version of an accepted manuscript should be submitted on a 3.5" disk, preferably in Apple Macintosh Microsoft Word, Version 5.1, although other formats might be acceptable. The editor reserves the right to make minor changes to an accepted manuscript in order to facilitate a clear and coherent publication.

Potential authors are encouraged to contact the editor to discuss ideas for contributions or to informally determine whether manuscripts might be appropriate for publication in MLRV. The editor also welcomes suggestions for debates, theme issues, other innovative presentation formats, and general inquiries about the journal. SIG/MLR news items should be sent to the editor as soon as they become available.

Manuscripts and other correspondents with the editor should be addressed to:

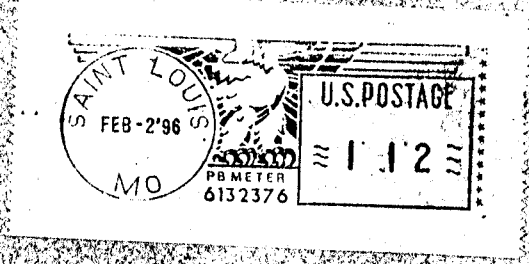
John T. Pohlmann, Editor, *MLRV*
Department of Educational Psychology and Special Education
Southern Illinois Universtiy at Carbondale
Carbondale, Illinois 62901-4618

phone: (618) 536-7763
fax: (618) 453-7110
internet: JOHNP @ SIU.EDU

University of Missouri St. Louis

Postage paid at the University of Missouri at St. Louis, St Louis, MO 63121-4499. POSTMASTER Send address changes to Steven Spaner, MLR:GLM/SIG Executive Secretary, Department of Behavioral Studies, University of Missouri - St. Louis, 8001 Natural Bridge Road, St Louis, MO 63121-4499

School of Education
Department of Behavioral Studies
8001 Natural Bridge Road
St. Louis, Missouri 63121-4499
Telephone: 314-516-5782



695 ROGE 96
ROGERS, BRUCE G.
DEPT OF ED PSYCH
UNIV OF NO IOWA
CEDAR FALLS, IOWA 50614-0607