UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

INTERNAL AND EXTERNAL VALIDITY: TWO FACES OF THE SAME COIN

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfilment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

CHRISTOPHER S. CHASTEEN
Norman, Oklahoma
2009

INTERNAL AND EXTERNAL VALIDITY: TWO FACES OF THE SAME COIN

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Jorge Mendoza, Chair

Dr. Larry Toothaker

Dr. Michael Mumford

Dr. Eric Day

Dr. Terry Pace

Acknowledgments

I would like to express my sincere thanks to several people without whom this dissertation would not have been possible. First is Jorge Mendoza, whose guidance and support have proven invaluable throughout my entire graduate training. Second, I would also like to acknowledge the contributions of my committee members, Drs. Michael Mumford, Larry Toothaker, Eric Day and Terry Pace.

On a personal note, this dissertation and my success in my graduate training would also have not been possible without the support and patience of my wife, Sandy. I would graciously like to thank her and note that I would not be where I am today without her.

Table of Contents

Acknowledgements iv
List of Tables vi
List of Figures vii
Abstract viii
Introduction 1
A Selection Situation 7
Selection and Individual Differences 9
Missing Data as Missing Subjects 12
Appropriate Analysis 14
Interactions between a Covariate and a Treatment Factor 15
Selection Scenarios as Degrees of Restriction 16
Simulations 19
The Linear Models 20
ANOVA/Restricted ANOVA 21
ANCOVA1/Restricted ANCOVA1 21
ANCOVA2/Restricted ANCOVA2 22
ANCOVA1 Example 22
ANCOVA2 Example (Treatment by selection variable interaction) 24
Summary of Models 26
Results 28
Mean Square Error (MSE) 28
Treatment Effect Size 33
Block Effect 39
Discussion 43
ANOVA Questions 43
Conclusions 45
References 49
Appendix A 52
Selection Process and Unconditional Normality 52
Appendix B 57
Selection Process and Conditional Normality 59

List of Tables

1 Summary of the Simulation Parameters 20

2 Summary of the Data Generation Models and Descriptors Used in the Results 27

3 Mean Square Error Values for the Treatment Effect Condition 29

4 MSE Values with Covariate Treatment Interaction Present and Shift not Present....30

5 Power of ANCOVA2 Model to Detect Interaction Effect.....31

6 Mean Squares for Covariate/Treatment Interaction Effect.....32

7 Coefficient Estimates of the Interaction Effect.....32

8 Partial Omega-Squared Values for the Treatment Effect without Treatment Covariate Interaction and without Shift.....34

9 Power to Detect Treatment Effect without Shift or Interaction.....34

10 Power to Detect Treatment Effect When Shift is Present.....35

11 Partial Omega-Squared Values for Treatment Effect with Shift Present 36

12 Partial Omega Squared Values for the Treatment Effect with Treatment-Covariate Interaction 37

13 Treatment Mean Squares 38

14 Treatment Mean Differences with Interaction Present 39

15 Power for Block Effect with No Distribution Shift 40

16 Probability of Type I Error for Block Effect with Distribution Shift 41

17 Partial Omega-Square of 'Shift Effect' When Block Effect is Null.....41

18 Power for Block Effect with Distribution Shift 42

19 Power to Detect Distribution Shift with ANCOVA 43

List of Figures

1 An example of a simple theoretical research paradigm 8

2 An example of an organizational research paradigm 9

3 Two populations with relative cut point and block mean difference 17

4 Two populations with absolute cut point and no initial mean difference 18

5 Two populations with absolute cut point and an initial block mean difference 18

1A o(x,) with a=0 and b=.3 53

2A o(x,) with a=1 and b=.3 54

3A Selected Distribution with *s*(*x*, 0, 3) 55

4A Selected Distribution with with *s*(*x*, 1, 3) 56

5A Overlay of functions 56

Abstract

Internal and external validities have been traditionally presented as two different topics. The present paper brings them together by showing how the two are related when the samples have been previously selected on a covariate. Selected samples are common in the social sciences. We show in a computer simulation that the ANOVA fails both in terms of external and internal validities when performed in a selected sample. On the other hand, when we use an ANCOVA in a randomized experiment the ANCOVA retains both internal validity as well as external validity. The simulation was used to compare the performance of ANOVA and ANCOVA in random and nonrandom (selected) samples. When the covariates that were related to the selection process were included in the ANCOVA, the ANCOVA outperformed the ANOVA in many of the situations studied. When a covariate-by-treatment interaction was present, the treatment effects were overestimated by the ANOVAs and underestimated by the ANCOVAs that incorporated the interaction. In addition, with selected samples, the covariate-by-treatment interaction was hard to detect due to the lack of power. Researchers working with selected samples should conduct an ANCOVA using covariates that are related to both selection and the dependent variable.

Introduction

Campbell in 1957 famously articulated the distinction between internal and external validity. He then quickly noted that "they are to some extent incompatible (pg. 297)." This incompatibility between internal validity, focusing on the causality of the effect, and external validity, the generalizability of the effect, has led to considerable study and discussion. One area of discussion has been with regard to the external validity of laboratory studies and internal validity of field studies (e.g. Dipboye & Flanagan, 1979; Dobbins, Lane & Steiner, 1988; Gordon, Slade & Schmitt, 1986). This present paper presents these two validities as two sides of the same coin; that both are dependent on conditional relationships in the data.

An example used to demonstrate these concepts through this paper will be the validities of the use of college students in the psychological laboratory. It is generally assumed that laboratory studies are more internally but less externally valid (e.g. Lucas, 2003; Gordon, and Slade & Schmitt, 1986). However, Dipboye and Flanagan (1979) point out that the samples in both cases can be equally homogenous, limiting the assessment and generalization of effects in either case. This homogeneity can occur because elements associated with quasi-experimental situations, such as non-random selection, often occur prior to random assignment in an experimental design. This limitation is the standard critique of the generalizability of using homogeneous groups such as college students in psychological research.

The current study is focused on alternate analysis methods that may be useful to increase the generalizability of the results of randomized experiments. In both the field and in the laboratory, selection can impact internal and external validity and occur at

multiple stages of a study. In addition, selection can lead to violations of assumptions, such as heterogeneity of variances and normality. While random selection provides the best conditions for inferences, in most cases random selection is rarely, if ever, possible. With nonrandom (selected) samples, the inferences to broader populations are conditional on the characteristics of the sample (Rosenbaum & Rubin, 1983). Understanding the limitations of samples and implications of selection is crucial for both internal and external validities. The issue then changes to from 'if students are representative' of the general population, to 'what degree are they representative', given the characteristics that led to their participation, and the data that is missing for a complete analysis. The degree of representativeness is a function of the various forms of selection that serve as causes of the missing data. While the trade-offs of heterogeneous versus homogenous samples have been discussed in the literature (e.g., Cook, Campbell & Paracchio, 1990; Calder, Phillips & Tybout, 1982; Lynch, 1983), the current paper expands on these concepts by showing how both external and internal validities are conditional processes. Conditional analysis can, in some cases, be used as a more appropriate analysis than some traditionally employed models. While the approach presented here is not necessarily a traditional one in selection research, McCourt (1999) has found that methods of validation in practice often diverge from traditional academic models.

Researchers in both basic and applied fields in the social sciences are often interested in going beyond simple associations to determine the causes and effects of behavioral phenomena. Cause and relationship assessment is not only necessary for theory advancement but for the development of effective interventions (clinical, educational, and professional). The establishment of causal relationships, however, can be significantly

more complicated (and costly) than identifying associations. There are several approaches for establishing causal relationships, varying in method and scope.

One of the most useful structures for developing the ideas of causality is known as the Rubin Causal Model (RCM) (Holland, 1986). Viewing the concept of causal effects from a statistical framework, Rubin extended the logic of randomized experimentation into the formal RCM model. The RCM model represents the causal relationships in randomized experiments and has even been used to extend causal inference to non-randomized designs (Rosenbaum & Rubin, 1983).

The RCM starts by postulating that the strongest causal inference is drawn when a single unit is subject to both the treatment and non-treatment (control) conditions and the unit is identical when exposed to each condition. As the subject is identical in all possible aspects, an observed difference on the outcome measure can be due only to the treatment effect. The effect of *t* is given as the difference between the control and the treatment

$$Y_t(u) - Y_c(u), \tag{1}$$

where the treatment (*t*) causes the difference on unit (*u*) relative to control (*c*).

This difference serves as a mythical logical starting point, as it is impossible to have exactly identical cases to expose one to treatment and the other to control conditions (it is not even approachable in the behavioral sciences). Also, it is not possible to simultaneously subject a single unit to two (or more) treatment conditions. Holland (1986) refers to this impossibility as the "fundamental problem of causal inference" and suggests two possible solutions; the "scientific solution" and the "statistical solution."

The scientific solution utilizes various homogeneity or invariance assumptions -- assumptions that are often plausible but not testable. One such assumption is unit homogeneity, where treatment differences between different units that *appear* identical in all relevant aspects are assessed (Holland, 1986). Other assumptions include causal transience and temporal stability. Causal transience is where the effect of exposure to one condition does not affect the impact of the next condition. Temporal stability is where the effects of the treatment (and measure) are independent of the time on which they are administered. These assumptions address the nature of establishing causality within a single experiment and directly mirror the assumptions typically posed when making generalizations from experiments where non-random sampling has occurred.

Rather than relying on a series of assumptions of the characteristics of subjects, an alternate "statistical solution" (Holland, 1986) extends the logic of exposing identical units to exposing statistically equivalent groups to treatments. As the ideal causal model at the individual (unit) level exists only in theory, the next logical step is to move up in terms of aggregation and consider causal effects averaged across subjects. By studying groups of individuals who are, on average, equivalent on all variables (measured and unmeasured) average causal effects can be measured as

$$E(Y_t - Y_c) = T \tag{2}$$

which, following the rules of expectation, can also be expressed as

$$E(Y_t) - E(Y_c) = T. \tag{3}$$

That is, causal effects can be measured as a difference between a treated population and a control population (at least on average).

In practice not all members of a population are measured, so these measurements are actually conditional values; conditional given selection into a particular treatment condition. If the assumption of independence is met, then the conditional probability is equal to the unconditional, that is

$$E(Y_t) = E(Y_t|S = t) \tag{4}$$

and

$$E(Y_c) = E(Y_c|S = c) . \tag{5}$$

If the selection function *S* is a random process, then *S* will not be related statistically to Y_t and Y_c. Because the selection process (or any other variable, on average) is not related to the dependent measures, then only the treatment is responsible for the effect.

When *S* is not a random process, the observed difference represents a difference in conditional values. The RCM has been extended to include nonrandomized (nonexperimental) designs by using observed covariates to establish a conditional independence assumption. The assignment to a treatment condition, *t*, given a set of covariates, *Z* can be represented by

$$f(x) = P(S = t|Z). \tag{6}$$

In a randomized experiment, *f*(*x*) is specified and not dependent on *Z*. In nonexperimental designs, *f*(*x*) is most likely unknown, however may be estimated from the observed data using, say, a logit model. Extending this logic of conditional independence in group assignment to conditional analyses for generalizable findings is a key point developed through this paper.

While issues of random assignment to treatment conditions and experimental control lay in the domain of internal validity issues, researchers are also concerned with issues of

external validity. The process of research in the social sciences is typically a two-step process. The investigation or intervention is planned based on theory and sound design principles. Strong causal relationships are established with an experimental design, with strong internal validity. In some situations because environmental, ethical or design restrictions exist that do not permit true experimentation, the focus turns to the assessment of associations. Once a relationship is observed in a particular group or situation, it is then typically of interest to generalize the results beyond the groups studied. Research studies are generally conducted not to confine the conclusions to the participants themselves, but to investigate how a phenomena would impact people with varying conditions and characteristics. The utility of theory lies in the simplification of complex realities. If a theory must be qualified for each subset, then the parsimony of the theory is reduced. However, the generalizability of findings is difficult to quantify in many research situations.

The method for obtaining samples statistically determines the limits of the generalizability of the research findings beyond the groups studied. Ideally, a research study clearly defines the target population and obtains a random sample from that population. In such cases, the studied units are, on average, independent on all variables, including the dependent measure of interest. For a variety of practical reasons, such a protocol is very often not possible in applied research. Yet researchers generalize their findings (explicitly or implicitly) to a broader group than those studied. In actuality, many samples are drawn from conveniently available pools, and their characteristics are often unknown. Under these conditions, the appropriateness of generalizations is impossible to gauge.

In the field of psychology, this problem manifests when research is conducted on the best available sample – for academic researchers, the most available source of subjects is often the undergraduate subject pool. The subject pool is a self selected group of potential participants. They are self- and organizationally-selected on known and unknown sets of nonrandom selection processes. This selection process has led to research and conjecture on the applicability of findings based on college students to other populations (Gordon, Slade, & Schmitt, 1986; Farber, 1952). While this discussion has focused on the use of college students in applied research areas, the problem is a more general one. Dipboye and Flanagan (1979) point out even a field study could have a very narrowly defined (selected) study group that may impair generalizations, despite the fact that it was conducted 'in the field.' Psychology is not the only field that has had such exchanges; for example, the use of college students in consumer research has also been subject of debate (Calder, Phillips & Tybout, 1983).

A Selection Situation

Study participants can differ from nonparticipants in infinite ways and ultimately only the theory itself can provide initial guidance on whether certain background variables are relevant or not (Lucas, 2003; Lynch, 1982). Whether in a university or an organizational setting, selection has already occurred on any number of variables, leading conceptually to missing subjects. If a researcher is interested in generalizing beyond the current subject pool with their particular idiosyncrasies, then considering the nature of the sample can potentially provide further insight. These concepts directly map back to the RCM; a random assignment process, in essence, represents an unconditional causal relationship

(Equation 3). When assignment is non-random, the relationship (and inference) is conditional (Equation 6) on the covariates that may or may not be measurable.

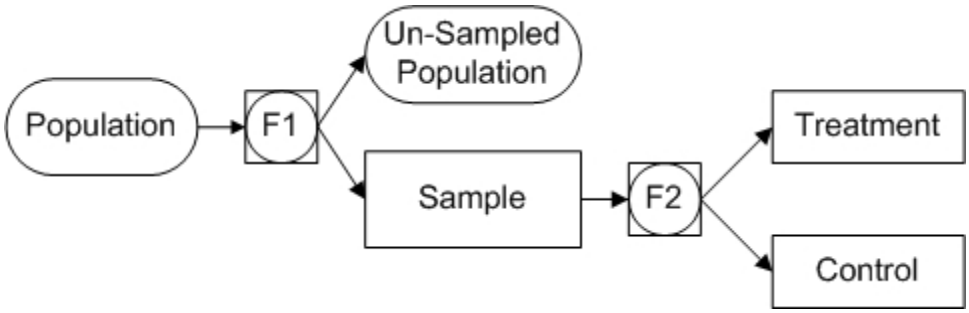

Figure 1. An example of a simple theoretical research paradigm.

Figure 1 depicts a simplified, familiar research design. In the optimal design both selection processes (*F1* and *F2*) are random. If *F2* is random, then the treatment effect for groups is unbiased. If *F1* is random, then effects observed in the sample generalize to the population. As discussed, if *F2* is not random, it is possible to establish independence conditionally by observing covariates related to *F2*. Likewise, if *F1* is not random, it is possible to use covariate information related to *F1* to improve generalizations.

Figure 2 presents a research paradigm more typical of actual research scenarios, both academic and field studies. Even here the design is a simplification of likely additional multiple selection functions (implicit and explicit) operating in real research situations. In Figure 2, *F2* remains the assignment-to-condition function as it was in Figure 1 and it is still possible to establish an unbiased treatment effect if *F2* is independent of *Y* (or the strongly ignorable assumption holds, i.e., sufficient conditional independence). Now, however, here there is an additional sample selection function, *F1a*. Working back toward the general population, the first nonrandom function serves

as the 'block' to simple generalization, that is, generalization without conditioning. Note that if *F1b* is not random, results may not easily generalize back to the subject pool itself, at least not without additional information on the selection variables.

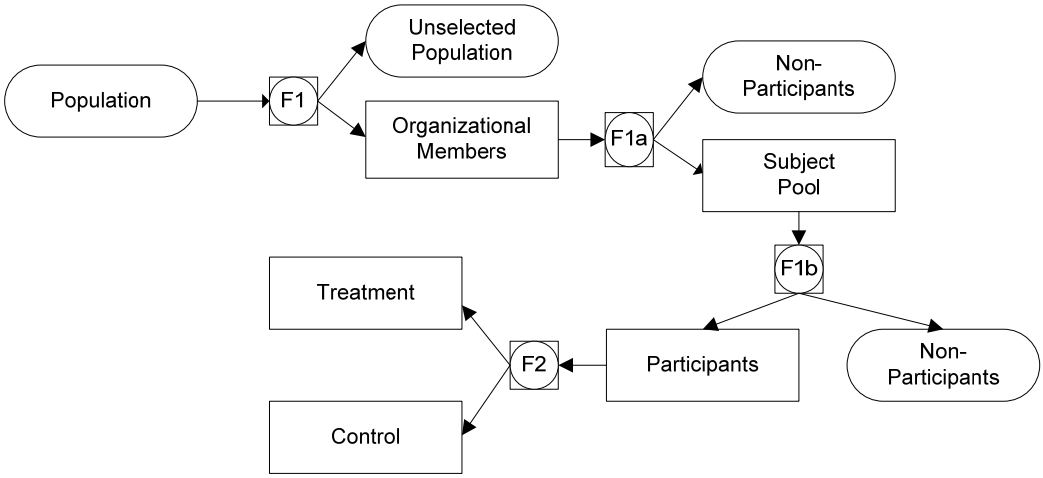

Figure 2. An example of an organizational research paradigm.

Selection and Individual Differences

The process whereby a particular individual becomes a member of a particular organization and, ultimately, a member of a subject pool is a function of a vast complexity of factors and relationships between the factors. For a university student, these could include relevant variables such as prior education, abilities, interests, experience, age, health, SES, and geography. Other, more serendipitous, variables may also play a role, including parental preference or affiliation, performance of athletic teams and a composite overall 'feels right' variable. While these are clearly not all the potentially relevant variables (which may be different for many of the participants), the structure of the factors of the decision process is less of a concern than creating a simple

model of the probability of explicit and implicit selection. Despite all the potentially relevant factors, many selection scenarios can be represented by a series of conditional decisions (probabilities). First, the potential participant (applicant, R) chooses to apply to an organization. Then, the organization makes a choice to accept the applicant (S|R); if accepted, the applicant makes a final choice to join the organization (A|R, S). Because they are a proper subset, the probability of being present in an organization can be represented as a product of conditional probabilities

$$P(o) = P(R)*P(S|R)*P(A|R, S). \quad (7)$$

The probability of being a member of a particular organization is the product of the probability of applying to that organization P(R), times the probability of being selected by the organization given application P(S|R), times the probability of accepting the invitation to join the organization given application and selection. This also highlights that the organization has its own set of criteria used to select applicants and these two sets (subject and organization) work in concert to ultimately determine the composition of the subject or incumbent pool. McCourt (1999) presents this contingency perspective as an alternative that may account for the incongruence between selection research and organizational selection practice.

Prior to selection, we consider each individual *I* to have a set of inherent or acquired characteristics or covariates $z_1, z_2, \ldots, z_q, x_1, x_2, \ldots, x_p$ and a dependent variable of interest y,

$$I(z_{1,}\, z_{2,\,\ldots}\,,\, z_{q,}x_1,\, x_2,\, \ldots,\, x_p,\, y) = I(\mathbf{Z, X, Y}). \quad (8)$$

Here the individual is characterized by two sets of individual differences, *Z* and *X*. (This could have been one or more of these sets. To simplify matters, the focus here is only on one: *X*). When researchers select or exclude an individual on the basis of one of these covariates, they affect the distribution of these individual differences in our sample. They may find themselves working with a sample of individuals with a particular set of characteristics that is different from those of the general population, thus affecting the marginal distributions of these variables. We may have, for example, a sample of teenagers or a sample of senior citizens. This is not generally a problem as long as our generalizations are just confided to say teenagers. The more restrictive or complex the selection process is, the more potential for restriction in the range of the covariates exists. If we want to make our generalizations to a broader set or population, we must account for those variables that are affected by selection.

When sampling has not been random, the argument for valid generalizations has traditionally been made from a theoretical perspective and not from a statistical one (e.g. Maxwell & Delany, 2000). A statistical approach would require taking the perspective that generalization is similar to missing-data or missing-subject problem. Consider that our usual experimental subjects, college students, are a subset of the general population that has been restricted on a particular set of variables through the process of selection. The selection is not random in that a college student self selects in applying, the institution selects to admit, and then the student decides whether or not to attend (a process mirrored in the business environment). Identifying or estimating all of the

selection processes involved in the final decision could be very difficult, if not impossible, in many situations. If we can identify the relevant variables involved in the selection process then it is possible to address the generalization issue from the perspective of missing information by using concepts from the RCM and missing data theory. Looking to modern missing data methods gives us a useful perspective to address external validity and internal validity in nonrandom samples from the perspective of conditional probability.

Missing Data as Missing Subjects

Missing data and statistically appropriate methods for accounting for missing data have received considerable attention in recent years (e.g., see Psychological Methods Special Section in Dec 2001, Shafer & Graham, 2002; Horton & Kleinman, 2007). By conceptualizing problems of external validity as cases of missing data (or missing subjects), we can take advantage of the theory and, in some cases, the techniques. One of the primary distinctions to make between the missing data methods is to address why and/or how the data are missing. If the data are missing due to a random process, the data are considered missing completely at random (MCAR). This is equivalent, in terms of sampling, to a randomly drawn sample. By the characteristics of random sampling, MCAR data is generalizable to the sampled population. MCAR also holds if the cause of missingness is unrelated to the dependent variable in the study, that is

$$P\left(Y_{missing}\middle|Y\right) = P(Y_{missing}). \tag{9}$$

If the data are missing due to a (selection) process that is related to a (measured) covariate, then the data can be considered missing at random (MAR). In the MAR condition, the missingness is random conditional on the covariate. In a selection scenario where all applicants take a selection test, but only the selected applicants become incumbents and are subsequently measured on performance, the missing data on the performance measure would be MAR (Mendoza, Mumford, Bard & Ang, 2004; Chasteen & Mendoza, 2003). This would be equivalent to screening the applicant pool on a covariate. The equation, in this case, being

$$P\left(Y_{missing}\middle|Y,X\right) = P(Y_{missing}|X)\ . \tag{10}$$

where the covariate *X* is always observed.

If the missing data are related to the dependent variable that is missing, then the data are missing not at random (MNAR). This is equivalent to the external validity scenario where both the covariate and dependent variable information is missing for the larger population, in this case,

$$P\left(Y_{missing}\middle|Y,X\right) \neq P(Y_{missing}|X). \tag{11}$$

In selection scenarios that would otherwise be MAR, with restricted information on *X* also related to missingness, it is likely necessary to assume the data are, to some degree, MNAR and some account of the selection function should be made. Missing data methods, such as multiple imputation, are flexible in that they can use the available

covariate information to recreate the distribution of the missing information. These methods do assume that the MAR condition has been met. If the data are MNAR, then the model of the data itself must take into account the missingness mechanism. In terms of external validity, this means extending the traditional analysis to account for the missingness mechanism. For experimental designs after nonrandom selection, this entails extending a traditional model (such as an ANOVA model) and incorporating effects into the model that are a function of selection, including conditional factors and potential interactions.

Appropriate Analyses

While there are several important conditions to consider, the primary concern is that of the effect of organizational selection on the generalizability of the results of a randomized experiment. Since a typical initial analysis in an experimental design is an ANOVA, this analysis method will be compared to a method not traditionally employed in randomized experiments, ANCOVA. The hypothesis is that controlling for group differences between selected and non-selected participants allows for results to be conditionally generalized back to the pool of potential participants prior to any organizational selection.

The composition of the selection function as a conditional probability has been described. However, to examine the effect of selection, it is not necessary to model each stage of selection; but what is necessary is to include the relevant covariates and degree of restriction imposed by selection. With the selection scenarios usually encountered by researchers, it is not possible to measure information on the unselected participants. This lack of information has two effects; first, it prevents an estimation of the selection

function, eliminating a more straightforward analysis, such as a propensity score analysis. Second, it can impose range restriction on the covariate(s). In some cases, however, the covariate information may be fully available, such as having applicant data for unselected applicants. It is likely that a combination of restricted and unrestricted covariate information can be utilized (unrestricted covariates enables the use of missing data methods to reduce the impact of restriction in range of related restricted covariates). In this study, the more likely condition of only having restricted covariate information available will be explored under several degrees of selection.

Interactions between a Covariate and a Treatment Factor

In the presence of a covariate by treatment interaction, we can only speak of mean differences conditionally upon a particular value of the measured covariate, assuming here that the covariate is collected before the experiment and is not affected by the treatment. Failing to test for an interaction between the treatment and the selection covariate could yield an inappropriate model and biased results. Clearly, if the potentially relevant background variables are not measured, testing for these interactions is not possible. Any selection that importantly restricts the range of the covariates may impede the exploration of slope differences and, accordingly, group differences. A similar problem exists when using two covariates that have been restricted by the selection process to predict the dependent variable through their interaction. While the ANOVA in such a case would be employing the wrong model across the range of the covariate, generalizing from the ANCOVA would be extrapolating into areas where the model is not known to hold. The restriction in range can, in addition, impact the normality of the variables, which additionally may impact the results. Normality was

evaluated analytically and it was determined that range restrictions in the scenarios of this study would have to be very severe to impact the normality of the distributions to a degree that would likely exceed the robustness of the statistics. These results are presented in Appendix A.

Selection Scenarios as Degrees of Restriction

The present simulation study focuses on several selection scenarios likely to be encountered in practice. Figure 3 depicts a simple selection scenario. In this scenario, 50% of Population A (the upper ellipse) is above the cut point (represented by the solid vertical line). Likewise 50% of Population B is above the cut point. Drawing 50% of the research sample from the restricted portion of Population A and 50% from the restricted portion of Population B will result in characteristics that will not be representative of the larger populations. In this case, the variances will be restricted and the means will be affected. Likewise, the restriction will result in a distribution that will be somewhat non-normal. All three of these effects will be exacerbated by the degree of restriction imposed on the populations. Figure 3 presents the situation where there is a block effect, as there is a difference in the means of the populations prior to range restriction due to selection (represented by the horizontal dashed lines). In this case, the size of the block effect remains constant after selection, while the unadjusted means increase (represented by the upward shift in the horizontal lines from the dashed lines to the solid lines).

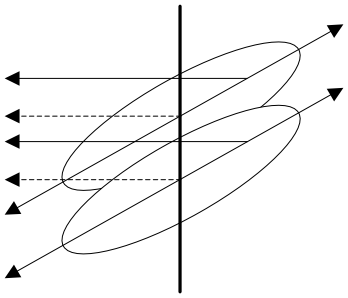

Figure 3. Two populations with relative cut point and block mean difference.

Figure 4 presents a slightly more complex situation. Here only 30% of Population A is above the cut point, while 70% of Population B is above the cut point. Again, the variances will be restricted, though Population A is affected to a much greater extent than Population B. In this scenario, there originally was no difference in the population means (again represented by the horizontal dashed lines) however restriction has now created a mean difference. This type of range restriction is an absolute cut point; the same cut point is applied to both distributions. For example, in an educational situation using, say, males and females as the blocking variable, only applicants who score above a set score on a math test are available for the study, despite the fact that males and females may have different population means on the test. A relative cut point would only, say, select the top 10% from each block. Obviously the relative cut point would be much more difficult to implement, as it would require information on the distribution as a whole prior to selection. Interestingly, when the means of the covariate of the two groups are equal, the relative and absolute cut points are the same (see Figure 3).

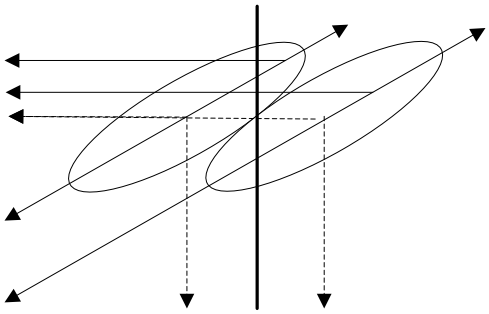

Figure 4. Two populations with absolute cut point and no initial mean difference.

Figure 5 presents the case where a mean difference in the block effect exists prior to selection, but that mean difference is reduced due to selection. In this scenario, both a group mean difference and a covariate mean difference existed prior to selection.

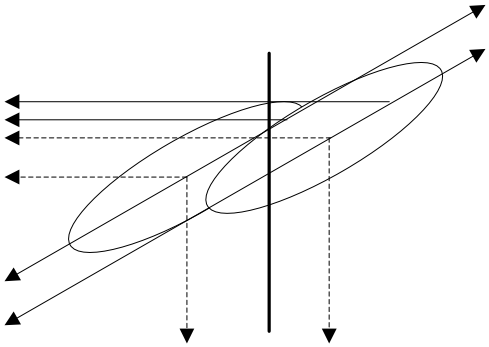

Figure 5. Two populations with absolute cut point and an initial block mean difference.

The present study will consider the effects of these two scenarios. The first is the more straightforward case (though likely the less prevalent case) where both groups have exactly equal means on the covariate measure (Figure 3). The second scenario is represented in Figures 4 and 5, where the populations have different means on the covariate (this will be referred to in the study as the "shifted" distribution for simplicity

and clarity). Both conditions will be subject to various block and treatment effects and treatment by covariate interactions.

Simulations

A series of simulations will be used to compare analysis strategies. For each of the two scenarios (shift/no shift), a variety of combinations of effects and models will be studied. First, three levels of effect size for the block effect (without a treatment effect) will be investigated. Second, three levels of the treatment effect will be modeled. Finally, three levels of treatment effect will be modeled with a covariate by treatment interaction present. All conditions will be run for three levels of a covariate/dependent variable correlation. The result is a 2 x (3+6) x 3 condition design. Each condition will be subject to non-selected (unrestricted) and selected (restricted) situations and analyzed using three models. The result is 54 conditions under two selection situations (restricted/non-restricted) analyzed with three models for a total of 324 cells. The 54 conditions were treated as populations and simulated 50,000 times.

Corresponding to effect sizes proposed by Cohen (1988), three conditions of block and treatment effect magnitudes were studied. Sample sizes and power to detect these effect sizes are presented in Table 1. Originally, the design was to study conditions where power was set to 0.8 for all effects (so that the effect would be detected as significant 80% of the time). However, the sample size necessary to detect the 'small' effect was N = 966 for the treatment effect, which, in the examples presented for psychological studies, is likely too large for a typical study from a research pool. For this effect size, power was reduced to 0.5. While this means that the true effect could be detected as significant only 50% of the time, this, unfortunately, may be more

representative of the actual state of psychological studies, at least for small effects (Rossi, 1990; Maxwell, 2004).

The correlation between the covariate and dependent variable was determined by the small, medium and large effect sizes for correlations presented in Cohen (1988). These combinations of effect sizes, power and sample sizes are presented in Table 1.

For cases with a shift in the distribution, the mean of the second distribution was shifted 1.0488. The mean of the first distribution, A, was zero with a variance of 1. The mean of the second distribution, B, was 1.0488 with a variance of 1. The cutpoint was set at 0.5244. This cut the distributions so that 30% of distribution A was above the cutpoint and 70% of distribution B was above the cutpoint.

Table 1

Summary of the Simulation Parameters

Effect Size	f	ρ	Power	Sample Size	
				Block	Treatment
Small	0.10	0.10	0.50	386	498
Medium	0.25	0.30	0.80	128	156
Large	0.40	0.50	0.80	52	64

The Linear Models

When a linear model is defined in a subpopulation, we must take into consideration whether a factor is a treatment (manipulated) or blocked factor. For example, if we are looking at existing differences among levels of a blocked factor (males vs. females), the groups could be unequally restricted, as they are in Figure 4. In contrast to the

experimental factor in which subjects are randomized into treatments, in the non-experimental factor the groups are likely to differ along the covariates (selection variables). Throughout the discussion we will assume that the treatment (blocked or induced) π_j is not dependent on *X*, and if it is dependent, the dependence is manifested in the product $\pi_j x$.

ANOVA/Restricted ANOVA. All conditions and restricted/unrestricted combinations were analyzed with three statistical models. The first was a standard two-way ANOVA model representing the standard analysis tool for use with randomized experiments. The difference between the results of the ANOVA model with a random sample (Random ANOVA) and a selected sample (Restricted ANOVA) is a key comparison in the study. The model with α_j as the block effect, π_k as the treatment and $\alpha\pi_{j,k}$ as the interaction term is given as

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + e_{i,j,k}. \quad (12)$$

ANCOVA1/Restricted ANCOVA1. The second analysis model studied was the standard ANCOVA model presented in Equation 13. This model will be referred to as the ANCOVA1 model in the present study to distinguish it from the ANCOVA2 model to be presented next. There are several important comparisons to be made with this analysis model. First comparison is to determine if the restricted ANCOVA1 (Restricted ANCOVA1) results are representative of the unrestricted ANCOVA1 (Random ANCOVA1). In addition, the consideration of whether the restricted ANCOVA1 is an

improvement over the restricted ANOVA in terms of representativeness of the unrestricted ANOVA will be important. The model is given as

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + \beta x_{i,j,k} + e_{i,j,k}. \quad (13)$$

ANCOVA2/Restricted ANCOVA2. The final model to be analyzed is an ANCOVA model with a treatment by covariate interaction term presented in Equation 14. This model, referred to as the ANCOVA2 model, is important for investigating the ability of an ANCOVA model in detecting the presence and effect size of this interaction, and is given by

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + \beta x_{i,j,k} + \beta_i \pi_k x_{i,j,k} + e_{i,j,k}. \quad (14)$$

ANCOVA1 example. Consider an ANCOVA (without a block effect α_j) where a non-interactive model holds in the unrestricted population,

$$y_{i,k} = \mu + \pi_k + \beta x_{i,k} + e_{i,k}. \quad (15)$$

Next, consider performing an ANOVA to compare groups using a sample that has being selected (explicitly or implicitly) using *x*. Although the ANOVA is often used by researchers to compare existing groups, computing this analysis on an *x*-selected sample is not appropriate. To understand why, we take the expected value of *y* at each (ANOVA) cell. Because we are working with an *x*-selected sample (superscript '*x*'

indicates selection on *x*), we take the expectation of *y* over the selected conditional subpopulation and find that the expected value of the cell mean is given by

$$\mu^{x}_{y_k} = \mu^{x}_{y} + \pi_k + \beta\mu^{x}_{x_k} \ . \tag{16}$$

The model holds in the subpopulation, because π_k is not dependent on *x* or *e*. However, when there is no random assignment to treatment, the groups may differ along the *x* variable. The difference would then be

$$\mu^{x}_{y_k} - \mu^{x}_{y_{k'}} = (\mu^{x}_{y} + \pi_k + \beta\mu^{x}_{x_k}) - (\mu^{x}_{y_o} + \pi_{k'} + \beta\mu^{x}_{x_{k'}}) = (\pi_k - \pi_{k'}) + \beta(\mu^{x}_{x_k} - \mu^{x}_{x_{k'}}) \ . \tag{17}$$

We can see that the difference between the two *y* means reflects the effect as well as the differences between the restricted means along the *x* variable. The ANOVA is inappropriate in that it ascribes differences between the cell means to the treatment effect, missing the differences along the *x* variable.

In addition, we can see that the *MSE* in the ANOVA under multivariate normality (in the general population) underestimates the variability in *Y* (refer to Appendix B). Note that had we been able to randomize individuals into treatment as one would do on an experimental situation, the expected value of the differences between the two *x* means would be zero. In this case, the cell means in the ANOVA would have reflected only treatment differences. However, the *MSE* would still be based on the restricted marginal *y*-distribution yielding a biased *MSE*, and normality may be questionable. Because the *MSE* underestimated the variability in *Y*, failing to reject the Null Hypothesis in the subpopulation implies failing to reject in the general population, if normality has not been

violated too badly. However, rejection of the Null in the subpopulation does not guarantee its rejection on the general population.

When we are dealing with a selected sample, the ANCOVA is the appropriate analysis assuming that we covary on the selection variables and that these variables are correlated with *Y*. The ANCOVA controlling for *X* is the appropriate design to assess treatment differences. The elegance of the ANCOVA is that it allows us to use a "biased" sample to make conditional inferences about the general population.

ANCOVA2 Example (Treatment by selection variable interaction). Consider an ANCOVA (without block effect α_j) with a fixed treatment factor and the covariate *x* model, under this model,

$$y_{i,k} = \mu_y^x + \pi_k + \beta x_{i,k} + \beta_i \pi_k x_{i,k} + e_{i,k} \ , \tag{18}$$

we have two important results. When the subjects can be assigned at random to the levels of the treatment, each cell has the same expected value on *X*; that is,

$$E(\mu_{x,k}^x - \mu_{x,k'}^x) = 0 \, . \tag{19}$$

Also, the expected value of a cell mean is

$$E(\bar{y}_k^x) = \mu_k^x + \pi_k + \beta \mu_{x,k}^x + \beta_i \pi_k \mu_{x,k}^x \, . \tag{20}$$

It follows that the expected value of the difference between two cell means is

$$\mu_{y,k}^x - \mu_{y,k'}^x = (\pi_k - \pi_{k'}) + \beta(\mu_{x,k}^x - \mu_{x,k'}^x) + \beta_i(\pi_k \mu_{x,k}^x - \pi_{k'} \mu_{x,k'}^x) \, , \tag{21}$$

When dealing with a treatment factor, randomization reduces the difference to

$$\mu_{y,k}^{x} - \mu_{y,k'}^{x} = (\pi_k - \pi_{k'}) + 0 + \beta_i \mu_x^{x} (\pi_k - \pi_{k'}) . \quad (22)$$

Then either under Equation 21 or 22, the ANOVA would incorrectly assign to treatment effect the differences between means. On the other hand, an ANCOVA with *X* as a covariate would correctly identify the treatment effect. However, the treatment effect would depend on the value of *X*. If the full range of *X* is not represented in the selected sample then the results may not generalize to the general population. If the interaction is present the researcher would have to go the general population to extend her results, especially if the interaction is disordinal. This is because we can only speak of mean differences conditionally upon a particular *x** value. We are assuming here that *X* is collected before the experiment and that *X* is not affected by the treatment.

Unfortunately, an ANCOVA that fails to test the interaction between the treatment and the covariate would yield an inappropriate conclusion. The ANCOVA model would have to include the interaction term to be appropriate. Here testing for the interaction term is equivalent to the test of slope homogeneity. When the slopes and intercepts are significantly different, the treatment difference must be assessed conditionally on a specific value of *x**. Any selection that importantly affects the range of *x* may impede the exploration of group differences. Because the treatment differences depend on *x** any statement about the magnitude of the effect would have to be conditional on *x**. A disordinal interaction would gravely impact on the generalizability of our results.

While the performance of the methods have been theoretically established, of particular applied interest is the impact of effect/sample size, degree of selection/restriction, type of restriction, and, as mentioned, presence of interactions on the

performance of the methods. While the simulations cannot cover all potential research conditions, conditions likely to be found in practice were investigated using selection ratios previously described as present in practice (Alexander, Alliger, Carson & Cronshaw, 1989; Chasteen, & Mendoza, 2005) and effect sizes suggested by Cohen (1988). The number of iterations for the simulations was determined by reviewing cumulative mean plots. It was determined that convergence would occur prior to 50,000 iterations, but the simulation was run out to 50,000 for each of the 54 study cells.

Summary of Models

It is important to draw a distinction between the data generation models and the analysis models in the simulation. Table 2 describes the data generation models for the various conditions. Note that *x'* indicates the case where the means of the covariates differ, creating a shift in the two block distributions (Figure 4). Magnitudes of the effects are abbreviated in the table as "S" for small, "M" for medium and "L" for large effects (as presented in Table 1).

The analysis models where the standard models a researcher may initially employ when analyzing the results of a study. The ANOVA model applied to both the restricted and unrestricted data, (as presented in Equation 12) is

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + e_{i,j,k}. \quad (23)$$

The ANCOVA analysis model, referred to in the study as the ANCOVA1 model is the standard two-way ANCOVA model,

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + \beta x_{i,j,k} + e_{i,j,k}. \quad (24)$$

The ANCOVA2 model employed for data analysis is

$$y_{i,j,k} = \mu + \alpha_j + \pi_k + \alpha\pi_{j,k} + \beta x_{i,j,k} + \beta_i \pi_k x_{i,k} + e_{i,j,k} \quad . \tag{25}$$

Table 2

Summary of the Data Generation Models and Descriptors Used in the Results

Effect				Data Generation Model
Block	Treatment	$\beta_i\pi_k x_{i,k}$	Shift	
S				$y_{i,j} = \mu + \alpha_j + \beta x_{i,j}$
M				$y_{i,j} = \mu + \alpha_j + \beta x_{i,j}$
L				$y_{i,j} = \mu + \alpha_j + \beta x_{i,j}$
S			Y	$y_{i,j} = \mu + \alpha_j + \beta x'_{i,j}$
M			Y	$y_{i,j} = \mu + \alpha_j + \beta x'_{i,j}$
L			Y	$y_{i,j} = \mu + \alpha_j + \beta x'_{i,j}$
	S			$y_{i,j} = \mu + \pi_j + \beta x_{i,j}$
	M			$y_{i,j} = \mu + \pi_j + \beta x_{i,j}$
	L			$y_{i,j} = \mu + \pi_j + \beta x_{i,j}$
	S		Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j}$
	M		Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j}$
	L		Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j}$
	S	Y		$y_{i,j} = \mu + \pi_j + \beta x_{i,j} + \beta_i \pi_k x_{i,k}$
	M	Y		$y_{i,j} = \mu + \pi_j + \beta x_{i,j} + \beta_i \pi_k x_{i,k}$
	L	Y		$y_{i,j} = \mu + \pi_j + \beta x_{i,j} + \beta_i \pi_k x_{i,k}$
	S	Y	Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j} + \beta_i \pi_k x'_{i,k}$
	M	Y	Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j} + \beta_i \pi_k x'_{i,k}$
	L	Y	Y	$y_{i,j} = \mu + \pi_j + \beta x'_{i,j} + \beta_i \pi_k x'_{i,k}$

Notes: A "Y" indicates the presence of a given effect.
S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Blanks in Effects column indicate that effect is not present.

Results

The goal of the study was to consider the effect of range restriction on the ANOVA and ANCOVA techniques. While there were many possible combinations of comparisons, several important ones are highlighted. The first comparison looked at the differences between a restricted ANOVA and a (hypothetical) unrestricted one, providing a sense of the robustness of the ANOVA to range restriction. The second comparison assessed the differences between an ANCOVA computed on a restricted sample versus an ANOVA computed on a random sample, providing a sense of an "ANCOVA" fix. The third comparison looked at the differences between the restricted ANCOVA and a random one, showing us the disadvantages if any of the random ANOVA over our fix. These comparisons were made for the three selection scenarios presented in Figures 3, 4 and 5.

These comparisons were made along several statistical criteria. Initially the mean square error (*MSE*) results were compared. Analyses of the block and treatment effects provided details on how range restriction affects these two different types of effects. Finally, the power of the ANCOVA2 to detect a covariate by dependent variable interaction was explored.

Mean Square Error (MSE)

Table 3 gives us an indication of how the increase in homogeneity of the restricted sample when there was no interaction between the covariate and the treatment, led to a decrease in the ANOVA *MSE*. The reduction in error associated with the restricted ANOVA was inversely related to the magnitude of the correlation between the covariate and the dependent variable. Note that in the restricted ANOVA the *MSE* decreased from

0.99 to 0.84 as the correlation increased; in the random ANOVA it was always 1.00. As Table 3 highlights, the reduction in error associated with the ANCOVAs was also a function of the correlation. The reduction for the random and the restricted ANCOVAs was always one minus the squared correlation. Most importantly, the error of the restricted ANCOVAs was always equivalent to the unrestricted, showing that, indeed, the conditional analysis provided the correct *MSE* values with the restricted sample. In addition, the reduction in error for the ANCOVAs was always greater than or equal to the reduction associated with the restricted ANOVA.

Results for the *MSE* values with the shift but no interaction were very similar to the results that have been presented in Table 3 (without the shift or interaction). The only difference with the shift is that the restricted ANOVA did not have quite the same degree of reduction in error (the *MSE* values with shift were 0.99, 0.95 and 0.86 versus 0.99, 0.94 and 0.84 without). The *MSE* values for all remaining models were not affected.

Table 3

Mean Square Error Values for the Treatment Effect Condition

Effect	Model					
Covariate	ANOVA		ANCOVA1		ANCOVA2	
Dependent Variable Correlation	Random Case	Restricted Case	Random Case	Restricted Case	Random Case	Restricted Case
0.1	1.00	0.99	0.99	0.99	0.99	0.99
0.3	1.00	0.94	0.91	0.91	0.91	0.91
0.5	1.00	0.84	0.75	0.75	0.75	0.75

Notes: Treatment Effect is “small”. Other effects are not present.

Table 4

MSE Values with Covariate Treatment Interaction Present and Shift not Present

Effect	Model					
	ANOVA		ANCOVA1		ANCOVA2	
Treatment	Random Case	Restricted Case	Random Case	Restricted Case	Random Case	Restricted Case
S	1.01	0.95	0.92	0.91	0.91	0.91
M	1.06	0.97	0.97	0.93	0.91	0.91
L	1.17	1.00	1.06	0.97	0.91	0.91

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate/dependent variable correlation is .3. Block effect is null.

When focusing on the *MSE* values with the covariate/treatment interaction present, the correct model is the ANCOVA2. Table 4 shows that the *MSE* values for the random and restricted ANCOVA2 model remained the same as when the interaction is not present (at 0.91). In addition, Table 4 shows there was an increase in error associated the other analyses. The random ANOVA had a 17% increase in *MSE* values with a large treatment effect and the restricted ANOVA a 19% increase. The random case ANCOVA1 showed a 41% increase in *MSE* values with a large treatment effect and the restricted ANCOVA1 a 29% increase. When the model did not include the interaction term and the interaction was present in the data, both random and restricted ANOVA and ANCOVA1 models' *MSE* values were nearly always overestimated. The introduction of the shift provided results very similar to those without the shift. The interaction had a much greater impact on the *MSE* results than did the shift.

Despite the ANCOVA2 model being correct in the presence of the covariate/treatment interaction, there was some difficulty in the ability to detect the interaction. Both the power to detect the interaction and the effect size were reduced for the restricted ANCOVA2. Table 5 demonstrates this reduction in power. When the

treatment effect was large, power dropped from 0.76 to 0.35 when the sample was restricted. This highlights the difficulty for a researcher to correctly identify the ANCOVA2 as the appropriate model when working with selected samples. The restriction increased the standard errors of both the estimates of $\beta x_{i,j,k}$ and$\beta_i \pi_k x_{i,k}$, reducing the power of the restricted analysis.

The probability of Type I error was not affected by restriction. Table 6 shows that the average interaction mean square values when the interaction was not present (the top six rows) were nearly always equal to one minus the correlation squared (.91, which was also the *MSE* values). Despite the reduced effected sizes for the restricted ANCOVA2 for the interaction effect, the model did produce coefficient estimates that were accurate (Table 7).

Table 5

Power of ANCOVA2 Model to Detect Interaction Effect

Effect	Model	
	ANCOVA2	
Treatment	Random Case	Restricted Case
S	0.53	0.22
M	0.81	0.38
L	0.76	0.35

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate/Dependent variable correlation = .3.

Table 6

Mean Squares for Covariate/Treatment Interaction Effect

Effect			Model	
			ANCOVA2	
Treatment	$\beta_i \pi_k x_{i,k}$	Shift	Random Case	Selected Case
S	No	No	0.91	0.91
M	No	No	0.91	0.91
L	No	No	0.91	0.92
S	No	Yes	0.92	0.91
M	No	Yes	0.91	0.91
L	No	Yes	0.91	0.91
S	Yes	No	3.38	1.79
M	Yes	No	5.53	2.59
L	Yes	No	5.30	2.49
S	Yes	Yes	3.36	1.95
M	Yes	Yes	5.53	2.87
L	Yes	Yes	5.30	2.83

Note: Covariate/Dependent variable correlation = .3.

Table 7

Coefficient Estimates of the Interaction Effect

Effect	Parameter	Model	
		ANCOVA2	
Treatment		Random Case	Restricted Case
S	-0.12	-0.12	-0.12
M	-0.31	-0.31	-0.31
L	-0.49	-0.49	-0.49

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate dependent variable correlation = .3.

Treatment Effect Size

Ultimately researchers are interested in obtaining representative effect sizes of the treatment effect. Results of the simulation indicate that without a distribution shift (Figure 3) or covariate/treatment interaction that five of the six models reported very close to the nominal effect size. The ANCOVA2 with a restricted sample returned reduced effect sizes (Table 8, last column). The restriction in range on the covariate ultimately resulted in the reduced treatment effect for the ANCOVA2. Table 8 presents the effect sizes for the case with a covariate/dependent variable correlation of 0.3; results for the other two correlations were very similar.

The power of the restricted ANOVA to detect the treatment effect was higher than the random ANOVA without the shift or interaction (4% higher with a small treatment effect (Table 9)). This is due to the increase in homogeneity of the sample following restriction. Only the ANCOVA1s were not affected by restriction. The random ANCOVA2 reported similar power as the ANCOVA1s (two of three conditions were equal), but the restricted ANCOVA2 reported substantially lower power (60% less with a small treatment effect).

Table 8

Partial Omega-Squared Values for the Treatment Effect without Treatment Covariate Interaction and without Shift

Effect	Parameter	Model					
Treatment		ANOVA		ANCOVA1		ANCOVA2	
		Random Case	Restricted Case	Random Case	Restricted Case	Random Case	Restricted Case
S	0.01	0.01	0.01	0.01	0.01	0.01	0
M	0.06	0.06	0.06	0.06	0.06	0.06	0.02
L	0.13	0.13	0.14	0.13	0.13	0.13	0.05

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large).
Covariate dependent variable correlation = .3.

Table 9

Power to Detect Treatment Effect without Shift or Interaction

Effect	ANOVA Parameter	Model					
Treatment		ANOVA		ANCOVA1		ANCOVA2	
		Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	0.50	0.50	0.52	0.54	0.54	0.54	0.22
M	0.80	0.80	0.82	0.83	0.83	0.83	0.39
L	0.80	0.80	0.82	0.83	0.83	0.81	0.37

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large).
Covariate dependent variable correlation = .3.

When the shift in the distribution was introduced, the power of the restricted samples to detect the treatment effect was reduced. Table 10 shows the power reduction for a small treatment effect dropped from 0.50 for the random case ANOVA to 0.45 for the restricted ANOVA. Power for the restricted ANCOVA1 showed a 15% reduction, despite being the correct model. The restriction increased the standard error of the estimate of β. leading to the reduction in power of the ANCOVA1. The ANCOVA2

showed the greatest decrease among the models, 203%, as both β parameter estimates had restriction affecting them.

Table 10

Power to Detect Treatment Effect When Shift is Present

Effect	ANOVA Parameter	Model					
		ANOVA		ANCOVA1		ANCOVA2	
Treatment		Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	0.50	0.50	0.45	0.54	0.47	0.43	0.13
M	0.80	0.80	0.74	0.83	0.76	0.72	0.21
L	0.80	0.80	0.76	0.83	0.77	0.70	0.20

Notes: A "Y" indicates the presence of a given effect. S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate dependent variable correlation = .3.

Despite the drops in power for the restricted ANOVA and ANCOVA1, restriction had little effect on the treatment effect sizes when the shift (without interaction) was present (Table 11). The ANCOVA2 model did have reduced effect sizes, especially the restricted ANCOVA2. Overall, for ANOVA and ANCOVA1, the primary effect of the shift on the treatment effect was a reduction in power, the effect size estimates remained nearly at the same magnitude as without a shift. The ANCOVA2 did not perform as well, with both lowered power and effect size estimates.

Table 11

Partial Omega-Squared Values for Treatment Effect with Shift Present

Effect	Parameter	Model					
		ANOVA		ANCOVA1		ANCOVA2	
Treatment		Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	0.01	0.01	0.01	0.01	0.01	0.01	0.00
M	0.06	0.06	0.05	0.06	0.05	0.04	0.01
L	0.13	0.13	0.12	0.13	0.12	0.10	0.02

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large).
Covariate/Dependent variable correlation = .3.

Adding the covariate by treatment interaction led to some very interesting results under range restriction, showing that the treatment effect can't be accurately estimated by any of the models. These results are reported in Table 12. The effect size estimates for the restricted ANOVA and restricted ANCOVA1 cases were about 2.8 times their unrestricted counterparts when the covariate/treatment interaction was present and the shift was not (items in bold). In contrast, both the restricted ANCOVA2 model and its random counterpart showed underestimation. This underestimation was more pronounced in the restricted ANCOVA2 case. Here the ANCOVA2 model is accounting for the interaction, but the restriction in range of the covariate leads to effect size reduction.

The bottom three rows of Table 12 show that the impact of a covariate/treatment interaction combined with a distribution shift can even affect the unrestricted cases (we also saw this in Table 4 with the *MSE* values). In this situation, the effect sizes for the unrestricted ANOVA and ANCOVA1 models were about twice that of when the shift was not present. The restricted ANOVA and ANCOVA1 models showed a compounding

effect, where the effect sizes were about 3.8 times larger, on average, than the original values. Again, the ANCOVA2 models showed reduction, which was now apparent in the unrestricted case. Despite the underestimation associated with the ANCOVA2 model, the random ANCOVA2 model performed the best (while erring on the conservative side), and is the correct model in this case. The bottom six rows of Table 13 show how these effects are reflected in the mean square values. In Table 13, the top three rows for the random cases show the baselines for the respective models for the rest of the table.

Table 12

Partial Omega Squared Values for the Treatment Effect with Treatment-Covariate Interaction

Effect		Parameter	Model					
			ANOVA		ANCOVA1		ANCOVA2	
Treatment	Shift		Random Case	Restricted Case	Random Case	Restricted Case	Random Case	Restricted Case
S	No	0.01	0.01	**0.03**	0.01	**0.03**	0.01	0.00
M	No	0.06	0.06	**0.17**	0.05	**0.17**	0.05	0.02
L	No	0.13	0.12	**0.33**	0.11	**0.32**	0.11	0.03
S	Yes	0.01	0.02	0.05	0.02	0.05	0.01	0.00
M	Yes	0.06	0.12	0.21	0.12	0.21	0.04	0.01
L	Yes	0.13	0.23	0.37	0.22	0.37	0.07	0.01

Notes: A "Y" indicates the presence of a given effect. S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate/dependent variable correlation = .3.

Table 13

Treatment Mean Squares

Effect			Model					
			ANOVA		ANCOVA1		ANCOVA2	
Treatment	$\beta_t \pi_k x_{t,k}$	Shift	Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	No	No	3.50	3.42	3.40	3.39	3.39	1.81
M	No	No	5.89	5.81	5.77	5.75	5.70	2.62
L	No	No	6.01	6.01	5.84	5.89	5.67	2.59
S	No	**Yes**	3.49	3.03	3.40	3.00	2.85	1.36
M	No	**Yes**	5.86	5.01	5.74	4.96	4.64	1.76
L	No	**Yes**	6.03	5.28	5.85	5.19	4.62	1.75
S	**Yes**	No	3.50	8.99	3.41	8.93	3.39	1.80
M	**Yes**	No	5.95	16.70	5.82	16.50	5.70	2.61
L	**Yes**	No	6.19	17.28	6.03	16.84	5.69	2.58
S	**Yes**	**Yes**	6.79	12.50	6.69	12.44	2.85	1.36
M	**Yes**	**Yes**	12.39	23.52	12.24	23.31	4.65	1.76
L	**Yes**	**Yes**	12.83	24.98	12.55	24.48	4.63	1.75

Note: Block effect is null. Treatment by covariate correlation = 0.3

While the ANCOVA2, especially the restricted ANCOVA2, underestimated the treatment effect, they both provided accurate estimates of the mean differences between the treatments when the interaction was present, whether or not the shift was also present. Table 14 demonstrates that with interaction, the restricted ANOVA and ANCOVA1 overestimated the mean differences with and without the shift present. This reflects the previously noted problem with the effect size estimates. With both the shift and interaction together, the random ANOVA reported a mean difference over twice that of the actual parameter (-1.00 versus -.49). It is important to note that the restricted

ANCOVA2 may have reduced effect sizes for the treatment effect, but did provide accurate coefficient estimates when the interaction was present.

Table 14

Treatment Mean Differences with Interaction Present

Effect		Parameter	Model					
			ANOVA		ANCOVA1		ANCOVA2	
Treatment	Shift		Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	No	-0.12	-0.12	-0.22	-0.12	-0.22	-0.12	-0.12
M	No	-0.31	-0.31	-0.55	-0.31	-0.55	-0.31	-0.31
L	No	-0.49	-0.49	-0.88	-0.49	-0.87	-0.49	-0.49
S	Yes	-0.12	-0.25	-0.31	-0.25	-0.31	-0.12	-0.12
M	Yes	-0.31	-0.63	-0.78	-0.63	-0.78	-0.31	-0.31
L	Yes	-0.49	-1.00	-1.25	-1.00	-1.24	-0.49	-0.49

Note: Block effect is null. Treatment by covariate correlation = 0.3

Block Effect

The block effect sizes and power, in general, were not affected when there was not a shift in the distributions due to the difference in the means of the covariates (scenario corresponding to Figure 3). As an example, the results for the power to detect the block effect with no shift or interaction were very similar to the results of the power to detect the treatment effect under the same conditions (treatment results were presented in Table 9). The main difference between the treatment results and the block results was that the restricted ANCOVA2 suffered no loss of power for detection of the block effect, as it did with the treatment effect. The results for power for detection of the block effect are presented in Table 15.

Table 15

Power for Block Effect with No Distribution Shift

Effect	ANOVA Parameter	Model					
		ANOVA		ANCOVA1		ANCOVA2	
Block		Random Case	Selected Case	Random Case	Selected Case	Random Case	Selected Case
S	0.5	0.50	0.52	0.54	0.54	0.54	0.54
M	0.8	0.80	0.82	0.83	0.83	0.83	0.83
L	0.8	0.80	0.83	0.83	0.83	0.81	0.81

Notes: S/M/L indicates the magnitude of the effect ("S"=Small, "M"=Medium, "L"=Large). Covariate/Dependent variable correlation = .3.

When there is a shift due to a mean difference on the covariate, the situations correspond to those presented in Figures 4 and 5. Figure 4 suggests that restriction will create a block effect when one did not exist prior to selection. Table 16 reports the probability of Type I error for the two ANOVA cases for the block effect. The shift in distributions did create an increase in the probability of Type I error for the restricted case (0.05 for the random ANOVA versus 0.53 for the restricted ANOVA when the correlation was 0.5). The probability of Type I error for the restricted ANOVA decreased as the correlation decreased, though it did not quite reach 0.05, even when the correlation was equal to 0.1 (it did reach 0.07). Figure 4 reflects a condition where the test of the block effect with the restricted ANOVA is biased when there is covariate /treatment correlation. As the ANCOVAs estimate the conditional means, they can detect the shift in the distributions but not the block effect. Therefore there is no unbiased test of the block effect when the sample is restricted.

Table 16

Probability of Type I Error for Block Effect with Distribution Shift

Effect	Parameter	Model	
Covariate/ Dependent Correlation		ANOVA	
		Random Case	Restricted Case
.1	0.05	0.05	0.07
.3	0.05	0.05	0.21
.5	0.05	0.05	0.53

Notes: Magnitude of the treatment effect is Medium.

Table 17 shows the partial omega square values of the random and restricted ANCOVAs. The effect sizes increased as the magnitude of the covariate/dependent variable correlation increased. This is because if the correlation is zero and there is no block effect, there is no difference in the adjusted means (represented by the difference in diagonal lines in Figure 4). As the correlation increased to 0.3, the partial omega-squares were between 'small' and 'medium' (0.02). At a correlation of 0.5, the effect sizes increased to between 0.05 and 0.06. The ANCOVAs are all 'shifting' in terms of adjusted mean differences as the correlation rises above 0.1.

Table 17

Partial Omega-Square of 'Shift Effect' When Block Effect is Null

Effect	Model			
Covariate/ Dependent Correlation	ANCOVA1		ANCOVA2	
	Random Case	Restricted Case	Random Case	Restricted Case
.1	0	0	0	0
.3	0.02	0.02	0.02	0.02
.5	0.05	0.06	0.05	0.06

Notes: Magnitude of the treatment effect is Medium.

Table 18 shows that there is a reduction in the power to detect an existing block effect with a restricted ANOVA when the means of the covariate are not equal (when the correlation is 0.5, the restricted ANOVA had power 81% less than the random ANOVA). This can be inferred from Figure 5, where the restriction creates horizontal lines that are closer together than the horizontal lines associated with the unrestricted distributions. It is clear that both the increase in the probability of Type I error (Table 16) and decrease in power (Table 18) were influenced by the magnitude of the covariate/dependent variable correlation, with a higher correlation having a greater negative impact.

The ANCOVA models in the Figure 5 scenario are again looking at the presence of the shift, however now one distribution is also 'moved' up because of the existing block effect. In this situation, the greatest difference in the adjusted means occurs when the correlations are at their smallest level. As the correlations increase, the distributions 'line up' so that the adjusted means are closer together. We see this effect in Table 19. Of particular importance is that the effects are the same regardless of whether range restriction is present or not.

Table 18

Power for Block Effect with Distribution Shift

Effect	Parameter	Model	
Covariate/ Dependent Correlation		ANOVA Random Case	ANOVA Selected Case
.1	0.80	0.80	0.60
.3	0.80	0.80	0.35
.5	0.80	0.80	0.15

Note: Magnitude of the block effect is Medium.

Table 19

Power to Detect Distribution Shift with ANCOVA

Effect	Model			
Covariate/	ANCOVA1		ANCOVA2	
Dependent Correlation	Random Case	Selected Case	Random Case	Selected Case
.1	0.50	0.50	0.49	0.49
.3	0.16	0.16	0.16	0.16
.5	0.05	0.05	0.05	0.05

Note: Magnitude of the block effect is Medium.

Discussion

ANOVA Questions

The results presented here leave researchers with questions about the appropriate use of ANOVA. Although strictly speaking the ANOVA is not correct with selected samples if the selection variable is related to dependent variable *Y,* it may be acceptable if the correlation between the selection variable and *Y* is not above 0.1 when there is a mean difference on the covariate. The ANOVA is especially troublesome when we were testing the blocked factor as opposed to a treatment factor. It stands to reason that if the blocking is performed using anchors based on the full population distribution as opposed to only the restricted distribution, the results are less likely to be biased. On the other hand, if it is not important that the test results generalize to the broader population then the selected ANOVA is quite appropriate (assuming that selection has not badly affected the normality of *Y*). Some have justified the selected ANOVA by claiming that if one fails to reject the Null Hypothesis with homogenous subjects then one is less likely to reject it with heterogeneous subjects (Calder, Phillips & Tybout, 1982). The problem with this argument is that if you reject the null with the homogenous subjects, then you

do not know whether you would reject it with heterogeneous subjects. You in essence would have to re-run the analysis for any generalizations. In addition there is a bias issue; the test for the block effect will be confounded with the shift. There is also a further weakness with this line of argument in situations where we look at the treatment effect; the possibility of interaction of the covariate and dependent variable can bias the results. If the range of the covariate has been so restricted that investigating the interaction is not possible, we are less likely to detect an interaction between the treatment and the covariate (what Lynch, 1982, called background factors). As we have seen here, this is not only an ANOVA problem, but it is also a problem with the ANCOVA. In general, selection will restrict the background factors by making the subjects more homogeneous. Because interactions affect our ability to interpret the main effects it is important that we identify them when present. This is almost an impossible task in a highly selected sample. Not having adequate variability in the background factors is germane to the experiment, and a lack there of diminishes our ability to identify an interaction effect.

The issue of external validity in the ANCOVA becomes one of homogeneity of slopes. We have seen that this is a valid approach as long as there is no interaction between treatment and the selected variable. If there is an interaction, it may be difficult to detect, and using an incorrect model can lead to overly conservative results. If we fail to detect the interaction, the use an incorrect model (an ANOVA or ANCOVA1) can lead to overestimates. If the slopes are not equal over the treatment conditions, then the results do not generalize across the covariate. Any treatment recommendations would have to take into account the covariate. However, if the slopes are equal, then we would

be able to generalize our results across levels of the covariate. The test of homogeneity of slopes can be conducted in the selected sample if there is "sufficient" range in the covariate; severe restriction will make the detection impossible.

Conclusions

The present paper began by pointing out the problem often encountered in the social sciences with being unable to obtain a truly random sample. Under this situation we have seen that the usual ANOVA can be misleading. Instead of the ANOVA, we have proposed employing an ANCOVA. In a randomized experiment in which the restricted variable is included in the model, the ANCOVA will lead to more generalizable and less biased results. Without an interaction, the effect size estimates for randomized treatment effects for the restricted ANCOVA were similar to those if the random ANOVA, providing evidence of internal validity. Without interaction, the effect size estimates of the restricted ANCOVA for the randomized treatment and block effects were similar to those of the random ANCOVA, demonstrating external validity.

In organizational settings, academic and applied, selection can occur in a variety of ways. The selection may be complex involving direct and indirect methods and occurring at multiple levels. The current study has provided some interesting observations and practical implications for researchers in spite of the fact that it focused on a simple case of direct restriction with a single covariate. We have seen that when there is no 'shift' in the covariate, yielding the same means across the two groups, as depicted in Figure 3, all the methods performed well in terms of probability of Type I error, power and estimating effect size of the treatment and block effects when there was no interaction between the covariate and treatment. Selection creates a more

homogeneous sample of subjects, impacting external validity. In the case of Scenario 3 the external validity is affected, but the tests results are not biased. However, it is unlikely that this situation occurs in practice, when separate naturally occurring distributions have the same means on a selection covariate. It is much more likely that the means would vary to some degree across groups, implying the scenarios in Figures 4 and particularly 5, which has both shift and block effects.

In the situations corresponding to Figures 4 and 5, the only technique that correctly assessed the block effect was the random ANOVA. The best case for the restricted ANOVA was when the selection variable had a very low correlation with the dependent variable. Even when the correlation was as low 0.1, the probability of Type I error for the block effect in the restricted ANOVA was 0.07 (when it should have been 0.05) and power was .60 (when it should have been .80). The restricted ANCOVA, on the other hand, provided estimates that were nearly equivalent to the unrestricted ANCOVA (in terms of both power and probability of Type I error). A researcher can generally trust the results of a restricted ANCOVA but not those of a restricted ANOVA. Thus in scenarios in both Figures 4 and 5, there were not only generalizability issues but also effect biases.

When there was an interaction between the covariate and the treatment, both the ANOVA and ANCOVA models performed poorly in the restricted sample. In fact, the *MSE* values when the interaction term was not appropriately included in the model were elevated by up to 17%. On the other hand, when the interaction term was inappropriately included in the model and not present in the data, the treatment effects were underestimated by around 63%. Furthermore, there is a difficulty in determining whether the interaction is present. In restricted samples the power to detect the interaction was

very low. It is important for researchers, when testing for interaction in a restricted space, to recognize that it is very difficult to detect.

Recommendations

There are a couple of strategies that can be employed when one is interested in generalizing results obtained in a restricted sample. Bringing in covariates related to selection into an ANCOVA provides similar or better results than using the traditional ANOVA. However, we have seen that when a covariate by treatment interaction is present along with a 'shift', the treatment effects are generally overestimated with the ANOVA. Unfortunately, detecting the interaction can be difficult, because the tests lack power, suggesting that testing the interaction using a more liberal test may be warranted. Using the ANCOVA with the interaction in the model provided conservative estimates of treatment effects.

Researchers employing the ANCOVA should pick covariates for an analysis that area related to the dependent variable as well as the selection process. The approach presented here requires the researcher to step back and consider the nature of the selection to identify covariates related to both selection and the dependent variable. That is, the identification of the covariate must take into consideration both theory and the variables responsible for selection to be able to generalize our results beyond the sample itself.

Internal and external validity have traditionally been presented in the social sciences as two separate and distinct concepts (e.g. Kirk, 1995). However, this is not the case, simply because both rely on conditional characteristics of the data to draw valid conclusions. We have seen in our results that both the block and treatment effects are

biased and lack power in selected situations thus affecting both internal and external validity. In light of the RCM, missing data theory (selection), and the results of this study it becomes evident that the selection process can have significant impact on the generalizability of the study and accounting for variables related to selection is likely to improve its generalizability.

References

Alexander, R. A., Alliger, G. M., Carson, K. P., & Cronshaw, S. F. (1989). Empirical distributions of range restricted SDx in validity studies. *Journal of Applied Psychology*, *74*(2), 253-258.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research, 9*(3), 240-244.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1983). Beyond external validity. *Journal of Consumer Research*, *10*(1), 112-114.

Chasteen, C. S., & Mendoza, J. L. (2003, April). *Restriction in range issues in validation designs: Modern tools for old problems*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Chasteen, C. S., & Mendoza, J. L. (2005). *Range restriction as a missing data problem: A comparison of ranges restriction corrections in the multiple hurdle validation design.* Unpublished Manuscript, University of Oklahoma, Norman.

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

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi Experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 1, pp. 491-576). Palo Alto, CA: Consulting Psychologists Press, Inc.

Dipboye, R. L., & Flanagan, M. F. (1979). Research settings in industrial and

organizational psychology: Are findings in the field more generalizable than in the laboratory. *American Psychologist*, *34*(2), 141-150.

Dobbins, G. H., Lane, I. M. & Steiner, D. D. (1988). A note on the role of laboratory methodologies in applied behavioural research: Don't throw out the baby with the bath water. *Journal of Organizational Behavior*, *9*(3), 281-286.

Farber, M. L. (1952). The college student as laboratory animal. *American Psychologist*, *7*(3), 102.

Gordon, M. E., Slade, L. A. & Schmitt, N. (1986). The "Science of the Sophomore" revisited: From conjecture to empiricism. *The Academy of Management Review*, *11*(1), 191-207.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 79-90.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 945-960.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed). Pacific Grove, CA: Brooks/Cole Publishing Company.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, *21*(3), 236-253.

Lynch, J. G. (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, *9*(3), 225-239.

Lynch, J. G. (1983). The role of external validity in theoretical research. *Journal of Consumer Research*, *10*(1), 109-111.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences and remedies. *Psychological Methods*, *9*(2), 147-163.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing Experiments and Analyzing Data: A Model Comparison Approach* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McCourt, W. (1999). Paradigms and their development: The psychometric paradigm of personnel selection as a case study of paradigm diversity and consensus. *Organizational Studies*, 20(6), 1011-1033.

Mendoza, J. L., Bard, B., Mumford, M., & Ang, S. (2004). Criterion-related validity in multiple-hurdle designs: estimation and bias. *Organizational Research Methods*, *7*(4), 418-441.

Rosenbaum, P. R., & Rubin, D. R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rossi, J. S. (1990). Statistical power of psychological research: What heave we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646-656.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Appendix A

An issue that must be addressed is the impact of nonnormality. When the samples have been shaped by a selection process, the samples can no longer be considered to be random samples from the (entire) normal population. The selection itself can create a variety of distributions that can impact analyses and the validity of their assumptions.

To examine this question, it is important to consider who is typically available to participate in an experiment involving a subject pool. Furthermore, to simplify this example bivariate normality for the *X* and *Y* variable in the general population is assumed. The subject pool consists of students who applied, were selected by the organization and accepted the offer (and have not quit). Next, assuming that selection (acceptance and the other factors) was a function of *X*, it is relatively easy to show that selection on *X* affects the joint distribution of *X* and *Y* rendering the selected population nonnormal.

Selection Process and Unconditional Normality

Although generally robust to violation of the normality assumption, the ANOVA assumes normality for the validity of the *F*-ratios. It is well know that the $F = MSB/MSE$ ratio in the ANOVA follows the *F*-distribution in random samples from a normal distribution. The question that we must examine is whether the *F*-ratios in the ANOVA follows the *F*-distribution when the samples have been influenced by a selection process; that is, when the samples can no longer be considered to be random from the entire (normal) population.

To examine the question, we examine the individuals available for the experiment. To simplify our discussion we will assume a bivariate situation with variables *X* and *Y*. A

similar argument could be made with three variables or more. Next assume that $P(o)$ is a function of X and that this probability follows the cumulative two-parameter logistic distribution. Clearly, other probability distributions could be assumed. Then $P(o)$ is

$$o(x) = \frac{1}{1 + e^{-(x-a)/b}} \quad . \tag{1A}$$

By carefully selecting a and b, we can obtain a variety of shapes for $o(x)$ that could be used to describe an applicant-selection-acceptance process. Figure 1A gives an example of a selection curve. This curve describes a selection process in which most of the "selection" occurs to the right of -1, with the probability of selection increasing as we go from -1 to 1.

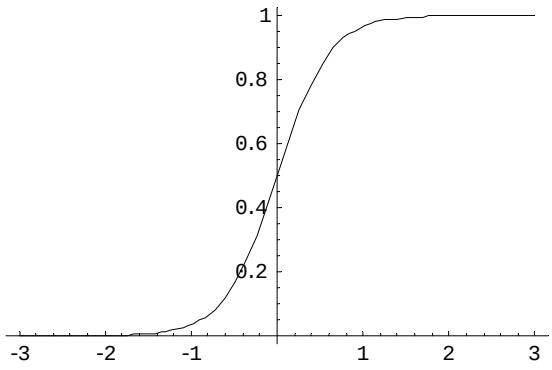

Figure 1A. o(x,) with a=0 and b=.3

(If an organization has sufficient data this function could be estimated using a Poisson or Logistic regression. Once estimated the results could be used to calibrate the selected

distribution.) Figure 2A shows a similar selection process shows that in this case most of the selection takes place to the right of zero.

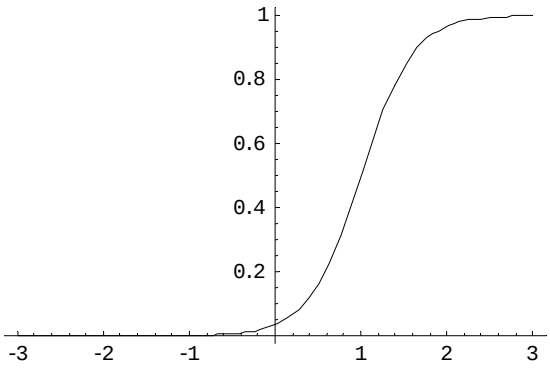

Figure 2A. o(x,) with a=1 and b=.3

By carefully specifying *a* and *b*, we can describe a variety of alternate selection processes.

Next, assume that the selection process *s*(*x*,*a*,*b*) is applied to the standard normal distribution. The density function for the selected (conditional) distribution in this case is given by

$$f'(x) = (1/k)s(x,a,b)n(x,0,1) \tag{2A}$$

where *k* is the constant of integration, given by

$$a = \int_{-\infty}^{\infty} s(x,a,b)n(x,0,1)dx\,. \tag{3A}$$

Now, consider the two selection function previously described in Figures 1A and 2A, $s(x,0,.3)$ and $s(x,1,.3)$, and applied them to the standard normal distribution. The first density function

$$f'_0(x) = (1/.5)s(x,0,.3)n(x,0,1) \tag{4A}$$

describes the distribution in the selected population-- the subpopulation that one would had obtained had one applied the selection process to the entire population. This selected distribution would look like

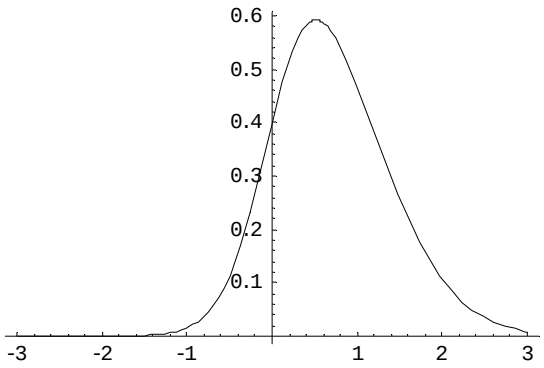

Figure 3A. Selected Distribution with $s(x, 0, 3)$

It is possible to see that the selected distribution is no longer symmetric and the shape of the distribution will vary with the selection function (with highly skewed distributions possible under extreme selection). Next consider the other the selected distribution

$$f'_1(x) = (1/.188637)s(x,1,.3)n(x,0,1)\,; \tag{5A}$$

that is,

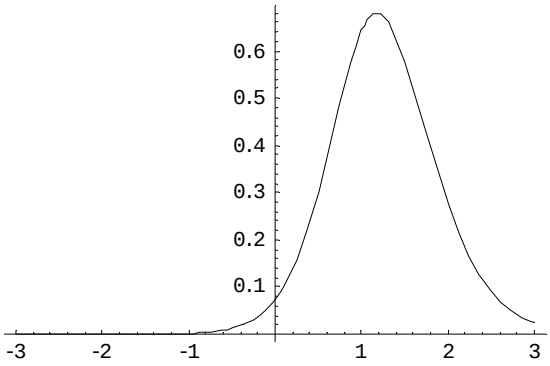

Figure 4A. Selected Distribution with with $s(x, 1, 3)$

We can see that in this case the mode of the distribution shifted right to account for the more restrictive selection process. The selected distribution is more skewed. To illustrate, we plot the two selected distributions against the normal,

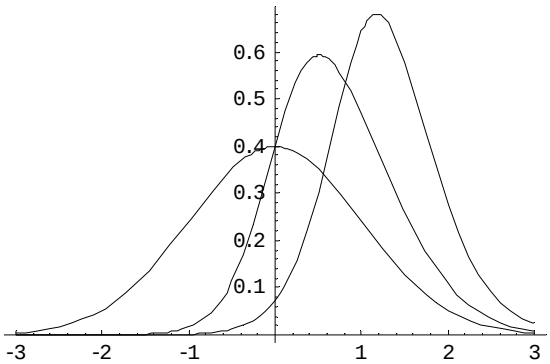

Figure 5A. Overlay of functions

Given the more extreme selection function on the right, skewness has only increased to 1.14, which is moderate, but not as large as one might expect.

Appendix B

Regardless of the selection process assumed, once a selection process $o(x)$ is applied to the general population the density function for the selected (subpopulation) distribution is

$$f'(x, y) = (1/k)o(x)n(x, y) \tag{1B}$$

where k is the constant of integration,

$$k = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} o(x)n(x, y)\partial x \partial y \, . \tag{2B}$$

Assuming that the general population follows the standard bivariate normal distribution with mean zero, variance one, and covariance equal to .6 with $o(x, a=0, b=.3)$, we numerically integrate Equation 1B to find the mean and variance of the subpopulation

$$\begin{pmatrix} \mu_x^x \\ \mu_y^x \end{pmatrix} = \begin{pmatrix} .705 \\ .423 \end{pmatrix} \text{ and}$$

$$\begin{pmatrix} \sigma_x^{x^2} & \sigma_{xy}^x \\ & \sigma_y^{x^2} \end{pmatrix} = \begin{pmatrix} .501 & .501 \\ & .820 \end{pmatrix}. \tag{3B}$$

Note that the means have been increased and variances have been reduced by the "selection" process $o(x)$.

The *MSE* in an ANOVA conducted on a random sample of individuals from the subpopulation would reflect a variance of .819 instead of the correct variance of 1. Clearly, the magnitude of this difference is a function of the properties of $o(x)$. Subject homogeneity is likely to lead in this case to liberal rather than to conservative results. In fact, we can show that

$$E(y^x) = E(\beta_o + \beta x) = \beta_o + \beta\mu_x^x = 0 + (.6)(.705) = .423 \qquad (4B)$$

$$\sigma_y^{x2} = \sigma_y^2 - \beta^2(\sigma_x^2 - \sigma_x^{x^2}) = 1 - (.36)(1 - .501) \approx .820. \qquad (5B)$$

Note that whether we are dealing with a restricted x or an unrestricted one, under normality the variance of y given is

$$\sigma_{y|x}^2 = \sigma_y^2 - \beta^2\sigma_x^2 = 1 - (.36) = .64. \qquad (6B)$$

The ANOVA's *MSE* would likely underestimate the *MSE* in the general population; whereas; the *MSE* in ANCOVA estimates the conditional variance whether it is performed in the general or subpopulation. In this case, the ANCOVA would be 128% more efficient than the ANOVA.

The inadequacy of the ANOVA analysis can also be shown using a missing data argument. The *Y* available data for the ANOVA are NI (not ignorable) on these situations, because

$$P(R|\, Z^{obs}, Z^{mis}\, X^{obs}, X^{mis}\, Y^{obs}, Y^{mis}) \neq P(R|\, Z^{obs}, X^{obs}, Y^{obs}). \quad (7B)$$

Selection Process and Conditional Normality

To develop a more appropriate analysis, consider again the organization function given as Equation 1A,

$$o(x) = \frac{1}{1+e^{-(x-a)/b}}, \quad (8B)$$

and apply it to the bivariate normal distribution. Then to investigate the shape of Y^x conditional on X, we find the conditional distribution of y given x^*. To this end we rewrite $n(x, y; \mu, \Sigma)$ as

$$n(x, y; \mu, \Sigma) = n(y \mid x; \mu_{y.x}, \Sigma_{y.x}) n(x; \mu_x, \sigma_x^2), \quad (9B)$$

and substitute into the selected distribution given in Equation 8B,

$$f'(y, x = x^*) = \frac{1}{k} o(x^*) n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) n(x^*; \mu_x, \sigma_x^2). \quad (10B)$$

Next note that for a fixed x^*, $o(x^*)$ and $n(x^*, \mu_x, \sigma_x^2)$ are constants; that is,

$$\begin{aligned} f'(y, x = x^*) &= \frac{1}{k} c_1 n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) c_2 \\ &= \frac{c_1 c_2}{k} n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) \end{aligned} \quad (11B)$$

The integral of $f'(y, x = x^*)$ must now be set to one to define the conditional distribution. Integrating it over *y*, we obtain

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{c_1 c_2}{k} n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) \partial y &= \frac{c_1 c_2}{k} \int_{-\infty}^{\infty} n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) \partial y \\ &= \frac{c_1 c_2}{k} 1 \\ &= \frac{c_1 c_2}{k}\end{aligned} \quad , \qquad (12B)$$

Then substituting into Equation (31), the conditional distribution of *y* given *x** is

$$\begin{aligned} n(y \mid x^*) &= \frac{k}{c_1 c_2} f'(y, x = x^*) \\ &= \frac{c_1 c_2}{k} \frac{k}{c_1 c_2} n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}), \\ &= n(y \mid x^*; \mu_{y.x}, \Sigma_{y.x}) \end{aligned} \qquad (13B)$$

which is the conditional normal distribution of *y* given *x**. We have shown that an *x*-selected function when applied to a multivariate normal does not affect the conditional distribution of *y* given *x*. Note that the definition of the selection function *o*(*x*) is not important as long as it is solely a function of *x* (MAR). As a function of *x*, *o*(*x*) will be a constant whenever *x* is fixed and the conditional normality will follow.

In terms of missing data theory, for conditional distribution *n*(*y*|*z*,*x*) the

$$P(R|\, Z^{obs}, Z^{mis}\, X^{obs}, X^{mis}\, Y^{obs}, Y^{mis}) \;=\; P(R|\, Z^{obs}, X^{obs}, Y^{obs}). \qquad (14B)$$

Under the conditional distribution the response pattern is not a function of the missing data.