UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE ACHILLES HEEL OF PSYCHOLOGY:

HOW CONVENIENCE SAMPLING AFFECTS PARAMETER ESTIMATES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

DUSTIN ALAN FIFE
Norman, Oklahoma
2013

THE ACHILLES HEEL OF PSYCHOLOGY:
HOW CONVENIENCE SAMPLING AFFECTS PARAMETER ESTIMATES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Jorge Mendoza, Chair

Dr. Robert Terry

Dr. Scott Gronlund

Dr. Rick Thomas

Dr. Kevin Grasse

Acknowledgements

First, I wish to express my gratitude to my research advisor, Professor Jorge Mendoza, for teaching me a great deal of statistics, having patience with me as I grew in understanding, guiding and supporting me during all these years of graduate school, and for being a excellent mentor. I am grateful for having had the opportunity to work with him.

I also wish to thank my committee members, Robert Terry; Scott Gronlund; Rick Thomas; and Kevin Grasse, for their enthusiasm and feedback during the early stages of my project. With their guidance and feedback, the impact and clarity of this paper has been greatly enhanced.

Finally, I wish to thank my wife, Amber Fife, who never complained about the long hours of work and graciously tended to the needs of our children during my frequent absences.

Contents

List of Figures

List of Tables

Abstract

Many have commented on potential problems associated with using undergraduate psychology students as research participants (e.g. Arnett, 2008; Henrich, Heine, & Norenzayan, 2010; Highhouse & Gillespie, 2009; Rosenthal & Rosnow, 1969). However, little research has been directed at demonstrating the extent of bias that may result from such practices and how to address this bias. In this dissertation, I investigate how the F statistic and treatment effects are affected when researchers use a convenience sample. I show that without measuring and modeling the selection variable, these parameter estimates are biased. I also show that covariate adjustments can mitigate bias when interactions do not occur between the treatment effect(s) and the selection variable. When interactions do exist, however, it is impossible to eliminate bias, particularly for the F statistic.

Introduction

Many statistical procedures rely on several assumptions, including independence, normality, and homoskedasticity. In addition, there are two critical assumptions that are sometimes overlooked in statistical textbooks. These assumptions are random sampling and random assignment to treatment conditions (see Rubin, 1974; West & Sagarin, 2000). Unfortunately, both of these last two assumptions are seldom met (Rubin, 1974).

The problem of non-random assignment to treatment conditions has been thoroughly investigated, and an entire body of literature is devoted to overcoming this problem (e.g., Cook & Campbell, 1979; Rubin, 2005, 1974; Shadish, Cook, & Campbell, 2002). However, the problem of non-random selection has not received much research attention, despite the fact that it is so common. It has been estimated that between 67% (Arnett, 2008) and 92.7% (Kulich, Seldon, Richardson, & Servies, 1978) of published experiments in psychology are performed on undergraduate psychology students. Although many acknowledge this as a limitation when conducting research (Highhouse & Gillespie, 2009), the majority do not. Furthermore, few researchers test whether their sample can be considered as a random sample from the referent population (Arnett, 2008).

The problem of non-random selection becomes problematic when researchers wish to generalize beyond the convenience sample. It has been suggested that findings found within convenience samples may not generalize to the referent population (e.g., Rosenthal & Rosnow, 1969). For example, Henrich et al. (2010) list several psychological findings that fail to generalize across cultures. We show in this paper that parameter estimates obtained from convenience samples may overestimate or underestimate population parameters. Furthermore, we also show

that some effects may be detected in a convenience sample that do not exist in the population (and vice versa).

In the following section, we begin with a working definition of generalizability. We also illustrate three ways in which results may not generalize to referent populations. We then review several approaches other researchers have taken to address problems with generalizability. Finally, we will introduce our approach, then investigate its performance using Monte Carlo simulations. We show that many parameter estimates can be recovered even when the researcher is working within a selected sample.

Generalizability

Throughout this paper, we define generalizability as the ability of obtaining unbiased estimates of population parameters when working with a subset of the population. Put differently, generalizability fails when statistics computed within a selected sample are biased estimates of their referent population. We address three different conditions within the ANOVA paradigm under which parameter estimates may be biased: a one-way ANOVA, a one-way ANOVA when the selection variable interacts with the treatment effect, and a *two*-way ANOVA when the selection variable interacts with the treatment effects.

One-way ANOVA

To illustrate a generalizability issue for the one-way ANOVA, we will use an example. Suppose a researcher is interested in determining whether study skills training helps improve classroom performance in the general population. The researcher recruits undergraduate psychology students to participant in an experiment where students are randomly assigned to treatment or control conditions.

Naturally, the students in this convenience sample are likely more intelligent than the average population. Consequently, the students represent a biased sample of IQ scores. For simplicity, we will assume that IQ is the sole characteristic on which these students differ from a random sample. The researcher wishes to generalize findings from this sample to the population.

Figure 1 shows several boxplots of the hypothetical distribution of class performance scores. The shaded boxes represent the treatment condition. The white ones represent the control conditions. The plots on the left are from a random sample and the plots on the right are from a convenience sample (i.e., a sample of only those who are highly intelligent).

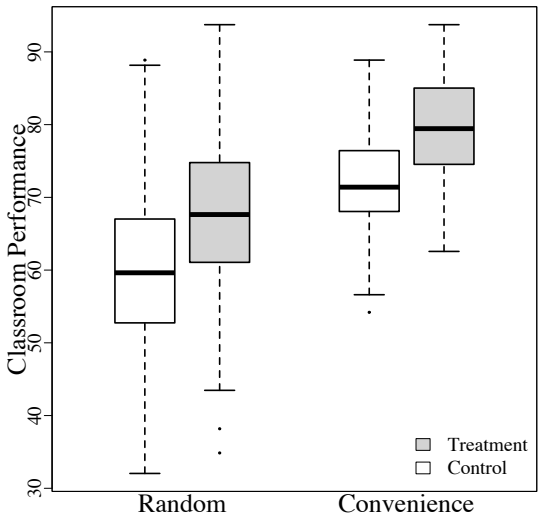

Figure 1. This figure plots a hypothetical scenario where individuals have been selected on IQ, randomly assigned to treatment conditions, then measured on the DV (Classroom Performance). The shaded boxes represent the treatment condition. The white ones represent the control conditions. Also, the plots on the left are from the random sample and the plots on the right are from the convenience sample.

There are several things worth mentioning about Figure 1. First, assuming random assignment has occurred, estimated treatment effects (i.e., α, or

$\bar{X}_{Treatment} - \bar{X}_{Control}$) will be unbiased. Notice in the plot that the difference between medians in the control versus treatment is nearly identical for the random and convenience samples. This is one of the reasons randomization is so important: it tends to balance the treatment and control on all unconsidered covariates. The end result of randomization is that the numerator of the F test[1] (MS_B) is unbiased. However, the denominator of the F test (MS_W) is *not* unbiased; selection on a variable correlated with the DV shrinks the within group variability, thus inflating the F statistic. This can be seen from the length of the boxplots, which is proportional to MS_W. We see that the length of the boxplots for the random sample are much larger than the boxplots for the convenience sample. In other words, when subject pools are used, and the selection variable(s) are correlated with the dependent variable, Type I error rates are inflated relative to a random sample. Under this condition, the F statistic computed in the selected sample will always overestimate the population F. (See Appendix for a mathematical explanation of this).

Using the terminology of Cook and Campbell (1979), Statistical Conclusions Validity is threatened when the sample is non-random. Relative to a random sample, parameters estimated from a convenience sample misestimate the significance of the treatment effect. It is important to note that this problem will not be fixed by replacing estimates of statistical significance with effect sizes. Since effect sizes often require an estimate of the variance (e.g., Cohen's d), they too will be affected.

[1] Although a t test would typically be used in this situation, we focus on the F statistic for consistency throughout the paper.

One-way ANOVA, Selection Variable Interacts

Now suppose the selection variable (in this case, IQ) interacts with the treatment effect such that the size of the treatment effect varies as a function of the selection variable. The researcher may or may not be aware of this interaction. Figure 2 illustrates this situation. The vertical gray line represents the cut-off point for IQ. In other words, no students were sampled with an IQ lower than approximately 112. The solid line represents the treatment condition, while the dashed line represent the control. Notice that for intelligent students, the treatment is beneficial. However, for students who are lower in intelligence, the treatment is not and may in fact harm their performance. This interaction between the selection variable and the treatment effect distorts power and the interpretation of the main effect. In other words, the numerator of the F statistic is biased.

Under this situation, estimates of the treatment effect (i.e., the "marginal effect"; Cramer & Applebaum, 1980) will be biased estimates of the population effect. Note that even with random assignment, these estimates will be biased because they are only estimated within the convenience sample.

Two-way ANOVA, Selection Variable Interacts

For our final example, let us suppose that in addition to study-skills training, the researcher also manipulates whether students receive memory training. These two variables are crossed so that interactions can be detected. Let us also suppose that the selection variable (IQ) interacts with these two variables. Is it possible to find a significant two-way interaction in the sample that does not exist in the population?

Figure 2 shows one example of this situation. The solid lines represent the

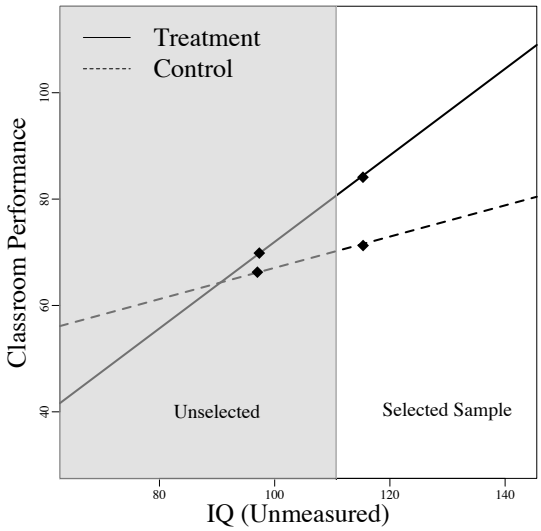

Figure 2. This figure plots a hypothetical relationship between IQ and Classroom Performance, conditional on whether subjects receive the treatment (a study skills training). If the researcher is working with a selected sample (i.e., the individuals to the right of the vertical gray line), estimates of the treatment effect will be misleading if there is an interaction between the selection variable (IQ in this case) and the treatment effect.

means of those who received memory training, while the dashed lines represent those who did not. Also, the left dots represent the means of those in the study skills training while the right dots represent the mean of those who did not. The Y-axis is the score on the outcome variable. Notice how the nature of the interaction is reversed from the selected (top half of the plot) to the unselected sample (bottom half of the plot). For example, in the selected group performance is best when both or neither memory/study skills training are administered. On the other hand, the selected group performs best when only one training or the other is performed.[2]

Although interactions were found in either the selected or unselected group

[2]We offer no theoretical reason why this may happen in empirical data. We only offer this example for illustrative purposes.

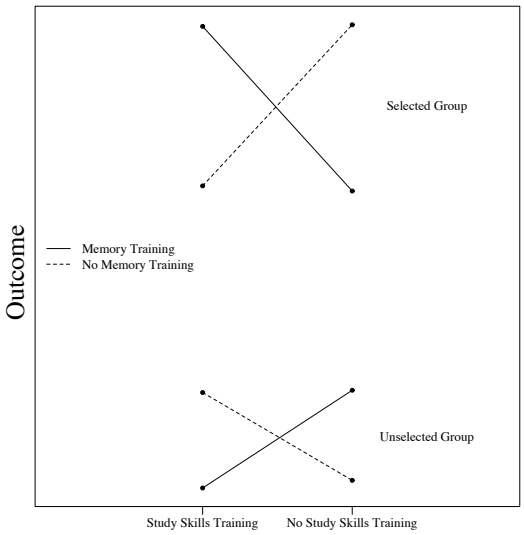

Figure 3. This figure plots a hypothetical relationship between IQ and the two treatment effects (memory training and study skills training). The top half of the plot shows the interaction in the selected group between the two treatments and the bottom half of the plot shows the interaction in the unselected group. Notice that the nature of the interaction reverses from the selected to the unselected group.

alone, at the population level the two-way interaction does not exist. Put differently, after averaging the two-way effects across the selection variable, the net effect is zero. Recall that a three-way interaction is present when the nature of the two-way interaction changes as a function of a third variable. This example has a significant three-way interaction, but the two-way interaction is non-existent. If the F statistic of the A by B interaction were estimated in the selected sample, it would be quite biased.

Summary

We have shown that generalizing from selected samples to referent populations presents several difficulties. If the selection variable is correlated with the DV, non-random sampling leads to biased F statistics, as well as misleading treatment

and interaction effects. In fact, sometimes estimated effects may reverse directions in the unselected sample. Given the fact that the majority of experimental research is performed on non-random samples, we think these results cannot be ignored. In the following section we will review several approaches that have been proposed to obtain unbiased estimates. We then introduce our approach to handling bias.

Previous Approaches to Non-random Sampling

Alternative Convenience Samples

Several authors have noted that convenience sampling is not a "lazy-dodge" on the part of the researcher, but an intelligent choice given the cost of random sampling (Farber, 1952, p. 102). Consequently, it is understandable that many researchers might be reluctant to abandon convenience sampling.

Some have suggested drawing from other "convenient" samples besides undergraduate students. Murray, Rugeley, Mitchell, and Mondak (2013), for example, commented on the practice of sampling from jury pools. Because jury pools are randomly sampled within communities, they will likely be more heterogeneous and thus better reflect population characteristics.

Another alternative convenience sample that has been suggested is campus staff (Kam, Wilking, & Zechmeister, 2007). Kam et al. compared a sample of local residents[3] to a sample of campus staff. They found few significant differences in terms of demographics between the two.

While both sampling methods will certainly increase the heterogeneity of participants, neither is ideal for two reasons. First, although more convenient than a

[3]The authors did not randomly select local residents. Instead, they drew a random sample of 1,500 individuals to invite to the study. 11.9% of that random sample chose to participate. Thus their comparison sample was self-selected.

truly random sample, both methods are still not as convenient as undergraduates. Consequently, it is unlikely such practices would be adopted in mass. Second, even with these samples there still may be substantial bias in parameter estimation, depending on the referent population. Neither sampling procedure escapes the problem of regional effects. For example, a random sample of individuals from Omaha Nebraska likely will not generalize to New York City or Tokyo.

Potential Outcomes (Counterfactual) Approach

The second approach to address non-random sampling is called the counterfactual or potential outcomes approach. This method of causal modeling considers two scenarios: Y(1) is the potential outcome had the treatment been received. Y(0) is the potential outcome had the treatment *not* been received. The difference for a particular individual between Y(1) and Y(0) is defined as the causal effect of the treatment. When averaged across individuals, it is called the average causal effect. However, only one of the two will be observed; either the subject will receive treatment or he/she will receive a control. Assuming random assignment, the potential outcome score for the treatment condition not assigned is considered missing completely at random (MCAR). For example, if Subject A had been assigned the treatment, their potential outcome for the control is missing, or counterfactual. (For a review, see Rubin, 1974, 2004; Shadish, 2010).

This potential outcomes model is often called Rubin's Causal Model (Holland, 1986), although a similar framework was also proposed by Neyman (1923). When proposed by Rubin (1974), the potential outcomes approach was a stepping stone towards generalizing to a well-defined population of interest; one first generalized to the potential outcome not received, then generalized beyond the sample. The second generalization requires either random sampling or "subjective random

sampling," where the researcher has reason to believe that the individuals in the study can be considered a random sample of the population (Rubin, 1974, p. 698; see also Fisher, 1955).

Although Rubin originally conceptualized the potential outcomes approach as a stepping-stone to generalizing to the population, other researchers have advocated methods that seek to generalize only within the sample. For example, Reichardt and Gollob (1999) introduced an alternative equation for the t test that enhances power. However, the increase of power comes at the cost of generalizability; it assumes the potential outcomes model, and thus can only be "transported" (Pearl & Bareinboim, 2011) to the potential outcomes observed within the sample. Because most researchers would rather think their results have application beyond the sample, we do not recommend this procedure.

Bayesian Networks

Bayesian Networks (Pearl, 1985) are an approach to causal inferences that grew out of computer science. The methodology was developed as an efficient approach to machine learning, but has broad implications for causal inferences. The details of Bayesian Networks (or Probabilistic Directed Acyclic Graphical Models) is beyond the scope of this paper. Interested readers are invited to read Pearl (2009).

Recent papers (e.g. Bareinboim & Pearl, 2012; Cooper, 1995; Didelez, Kreiner, & Keiding, 2010; Geneletti, Richardson, & Best, 2009; Didelez et al., 2010) have used Bayesian Networks to address the problem of non-random sampling and have developed a set of theorems to test whether results from a sample can be transported back to the population. The basic approach is similar to the one we introduce, namely using covariates to adjust treatment effects. However, our

approach will be couched within familiar ANOVA/ANCOVA terminology. Interested readers are invited to read Bareinboim and Pearl (2012) for an excellent review.

ANCOVA Approach

The approach we advocate adjusts treatment effects using carefully selected covariates. This approach to adjusting sample-based estimates is not new. Rubin (1974) suggested it in passing many years ago. However, we know of no other publications in the social sciences that have fully explored the strengths and limitations of this method. Additionally, in our literature search, no experimental studies attempted to adjust treatment effects for non-random selection of subjects.

We begin by introducing several definitions and assumptions, after which we will explain the rationale behind covariate adjusted treatment effects and why they should yield unbiased estimates of some population parameters. We also note similarities between our approach and common approaches to handling non-random assignment (e.g., propensity score matching and covariance adjustments). We will then investigate the performance of this method using Monte Carlo simulations.

Definitions

Throughout our simulations, we make use of four variables: A, B, Y, and Z. We define A and B as the treatments, and Y will be the outcome measure. Z is a variable that is correlated with Y on which selection has taken place. To simulate a non-random sample, we sorted the dataset according to Z, then selecting the top 50% of observations. Z can be considered a single variable or a collection of

variables. However, for simplicity of explication, we treat Z as if it is a single variable.

We also make a distinction between missing scores and missing cases. Missing scores occur when data is missing on one, but not all variables. The left image in Figure 4 illustrates missing scores. The solid boxes represent information that is available while the dashed lines represent information that is unavailable. Labels that are subscripted with OBS represent information that is observed, while MZ represents missing information due to selection on Z. The left image illustrates missing scores, where half of the Y scores are missing because of Z, while the right image illustrates missing cases. Missing cases occur when data are missing for all variables. Notice all information for those who scored poorly on Z is missing. Though we have graphed this figure such that half the scores are missing, in reality the number missing may be unknown.

Returning to our previous example, A is the study skills training, B is the memory training, and Y is classroom performance. However, the sample of undergraduate psychology volunteers represent a non-random sample. The collection of variables that differentiates them from a random sample is Z, which may be IQ, SES, conscientiousness, etc. An unknown quantity of certain types of individuals (e.g., a 95-year-old retired man) have almost a zero probability of being selected, and thus those people would be in the dashed boxed area.

In this paper, we attempt to tackle estimation under missing cases. Readers interested in estimating under missing scores are invited to read the selection literature (see Sackett & Yang, 2000; Thorndike, 1949, for a review), where corrections are fairly straightforward.

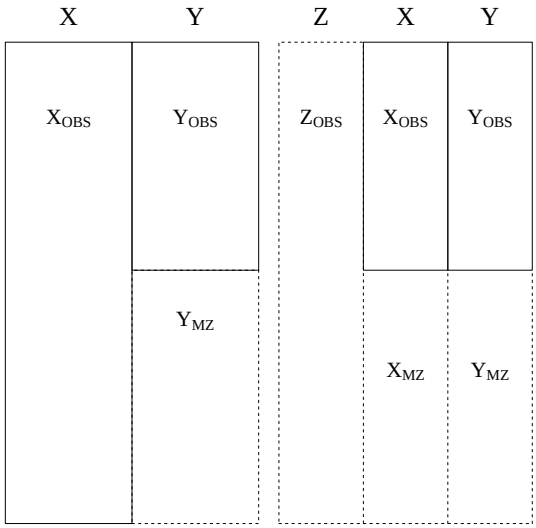

Figure 4. This image illustrates the difference between missing cases and missing scores. Solid lines represent data that is available (as indicated by the subscript *OBS*), while dashed lines represent information that is unavailable (i.e., it is missing because of selection on Z) as indicated by the subscript *MZ*. Missing scores have missing information on some, but not all variables (the left figure) while missing cases have information missing on all variables (right figure).

Assumptions

Since our procedure relies on the Analysis of Variance, we make the same assumptions made in all linear models. Namely, we assume homoscedasticity, linearity, independence, and normality. In addition, we assume that subjects have been randomly assigned to treatment conditions and that top-down selection has occurred. This last assumption is not critical; if selection takes place from the bottom-up, the results will often be opposite of that presented.

ANCOVA with Missing Subjects

Recall that an analysis of covariance (ANCOVA) conditionalizes a treatment effect on the value of one or more covariates. In order to understand how this fact helps with the missing cases problem, consider Figure 5. Suppose this sliced

ellipsoid represents the shape of a three-dimensional scatterplot. However, notice that values below a particular Z are missing (i.e., they are missing cases). Recall that a partial correlation measures the relationship between two variables, conditional on a third variable. Geometrically, the partial can be thought of as the correlation in the light colored area: it is the relationship between X and Y, at a particular level of Z. Notice that it does not matter that we have limited information on Z; under standard statistical assumptions, the partial is approximately the same at every level of Z (see Fife & Mendoza, 2013). In other words, we could cut Z at a different level, and the shape of the light ellipse will be the same.

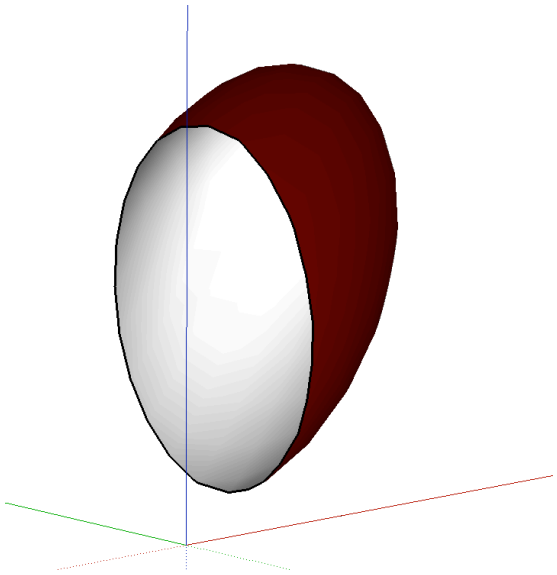

Figure 5. This image shows a three-dimensional ellipsoid where selection has occurred on Z. Note that the partial between X and Y is unaffected by selection—-It does not matter whether the researcher has full or partial information, the shape of the two-dimensional ellipse between X and Y does not change.

Like a partial correlation, an ANCOVA conditionalizes on the values of the covariates. Theoretically, we can obtain unbiased estimates of the Type III Sums

of Squares (partial) of the F statistic, whether we have full information or not on the selection variables, provided that they are included in the ANCOVA model. The procedure we propose for handling missing cases is to simply covary the variable(s) that cause selection. Note that this method is very similar to how one might handle non-random assignment. Under non-random assignment, the researcher may identify the variables that distinguish the two groups then covary them out. (Although propensity-score matching is another attractive alternative). Likewise, we suggest carefully identifying the variable (or set of variables) that cause selection, then covary those variables out, similar to how propensity score matching is done. In the following section, we introduce the Monte Carlo method we used to investigate the ability of the ANCOVA to recover population parameters from a selected sample.

Method

Earlier we illustrated how generalizability is affected under three conditions: a one-way ANOVA, a one-way ANOVA when an interaction exists between the treatment effect and the selection variable, and a two-way ANOVA when an interaction exists between the interaction effect and the selection variable. In the Monte Carlo, we sought to determine under which of these three conditions population parameters could be recovered by covarying the selection variable. To do this, we did the following.

1. *Generate 100 scores for Z.* We first created a normally distributed random variable which we called Z. It had a mean of zero and a standard deviation of 1.

2. *Generate 100 scores for Y, conditional on the treatment(s)/covariate.* Using the equations shown in Table 1, we generated 100 dependent variable (Y) scores. The coefficients for these models were chosen somewhat arbitrarily. How-

ever, the resulting data had the problems illustrated in Figures 2 and 3.[4]

3. *Generate a "convenience sample" based on values for Z.* To simulate a convenience sample, the dataset was sorted according the values on Z, then the top 50% of scores were selected for subsequent analysis. This sample we will call S_C, where C denotes its a convenience sample.

4. *Generate a random sample for comparison.* For comparison purposes, a random sample of 50 participants was selected. We will call this sample S_R.

5. *Compute parameter estimates.* For each ANOVA/ANCOVA condition, the F statistics and the treatment/interaction effects were estimated for all of the variables included in the model. This was done using both an ANOVA (ignoring Z) as well as the ANCOVA (which included Z).

6. *Estimate percent bias.* Each of the parameter estimates for the convenience sample (S_C) were compared to the estimates from the random sample (S_R). To compute percent bias, we used the equation

$$\% \text{ Bias} = \frac{E_C - E_R}{E_R} \times 100 \tag{1}$$

where E_C refers to the estimate in the convenience sample, and E_R refers to the estimate in the random sample. Percent bias was computed using both an ANOVA and an ANCOVA, where the selection variable was covaried out.

7. *Repeat 10,000 times.* Each of these steps were repeated 10,000 times in order to simulate a sampling distribution and to compute standard errors.

[4]The problems we refer to are as follows: Model 2 generated data where the estimate of the main effect differed in the random versus convenience sample. Model 3 generated data where the nature of the two-way interaction (between A and B) reversed from the convenience to the unselected sample and the net effect of the two-way interaction is zero.

Table 1: Summary of the Three Simulated Conditions. The Column Labeled "Data Generating Model" was Used to Simulate the Value for the Response Variable. These Data were then fit with the Traditional ANOVA Model (Third Column) as well as the ANCOVA Model (Which was the Generating Model). The Parameter Values used to Generate the data are Listed on the Right-Most Column.

Condition	Data Generating (ANCOVA) Model	ANOVA Model	Parameter Values
One-way	$Y = \mu + A + .6Z + \epsilon$	$Y = A + \epsilon$	$\mu = 0$, A=1
One-way, interaction present	$Y = \mu + A + .8AZ + \epsilon$	$Y = A + \epsilon$	$\mu = 0$, A=-1
Two-way, interaction present	$Y = \mu + A + B + -10Z + 9AZ + 9BZ + 0AB - 6ABZ + \epsilon$	$Y = A + B + AB + \epsilon$	$\mu = 1, A = 0, B = 0$

Results

One-Way Anova, No Interaction

Table 2: Median Percent Bias in Parameter Estimates for the One-Way ANOVA

	ANOVA	ANCOVA
F	42	-1
A	-1	-1

Table 2 shows the results of the first simulation, where no interaction exists between the selection variable and the independent variable. Recall that the data were generated using the equation $Y = A + .6Z + \epsilon$. The left column of the table (labeled ANOVA) shows the median degree of bias when the main effect (Type I SS) of A is estimated using the model $Y = A$ for both the selected and unselected sample. The right column of the table (labeled ANCOVA) shows the results of estimating the main effect (Type III SS) of A using the data generation model ($Y = A+Z$), for both the selected and unselected sample. The median percentage difference between the selected and the unselected F is shown. As mentioned previously, when the selection variable (Z) is not included in the model, the F statistic is positively biased (42%). On the other hand, the estimates for A are unbiased with or without including Z in the model in Table 2. Furthermore, including Z using an ANCOVA makes the Type III estimate of the F unbiased (see right column of Table 2).

Figure 6 plots the distribution of $100 \times (F_C - F_R)/F_R$, or the percentage difference between estimates in the random and convenience samples for the first Monte Carlo. As indicated in Table 2, the ANCOVA distribution centers around zero, while the ANOVA distribution does not. Furthermore, the variability is much greater in the ANOVA distribution.

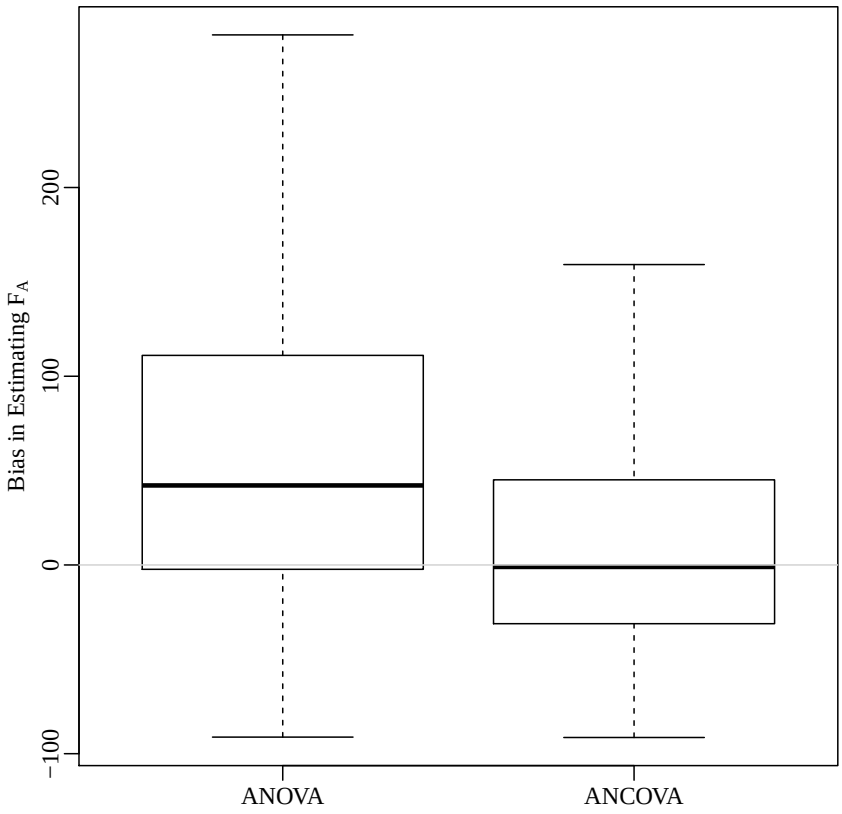

Figure 6. Distribution of the bias in estimating the F statistic for the ANOVA and the ANCOVA for the A effect. For these results, the interaction variable does not interact with the treatment effect.

One-Way Anova, With Interaction

The results for the second simulation are shown in Table 3. Recall that the data were generated using the equation $Y = \mu + A + .8AZ + \epsilon$, where μ was zero and A was -1. This resulted in a model where there was an interaction between the treatment effect and the selection variable such that the treatment improved performance for those selected, but hurt performance for those not selected (see Figure 2 for a graphical depiction). As before, the left column tries to estimate the model $Y = A$ using the convenience and random samples, while

Table 3: Median Percent Bias in Parameter Estimates for the One-Way ANOVA when an Interaction Exists Between the Selection Variable (Z) and the Treatment Effect (A).

	ANOVA	ANCOVA
F_A	-81	-87
F_Z		-64
F_{AZ}		-65
A	-63	-0
β_Z		-51
β_{AZ}		-1

the right model uses the model that actually generated the data. The results, as before, are presented as the median percentage difference between the random and convenience sample. Notice that the F statistic was underestimated in all cases, whether Z was included in the model or not. For example, the F statistic for the A effect was biased in both the ANOVA (-81%) as well as the ANCOVA (-87%). However, including the selection variable mostly removes bias in estimating the β parameters.[5]

Figure 7 shows the distribution of bias for both the ANOVA and the ANCOVA when the selection variable interacts with the treatment effect. Note that the estimate of the F is consistently underestimated using a selected sample when an ANOVA is used. However, when the selection variable is covaried out of the model, the estimate of the main effect of A is unbiased even with a selected sample.

Two-Way Anova, With Interaction

Our final table (Table 4) shows what happens when an interaction existed between the selection variable (Z) and the two treatment effects. Recall that the

[5]The raw value for β_Z was very near zero. Because the percent bias computation divided by a value near zero, it tended to make the percentage bias look quite extreme. However, the average raw difference between the two was quite small (in the third decimal place).

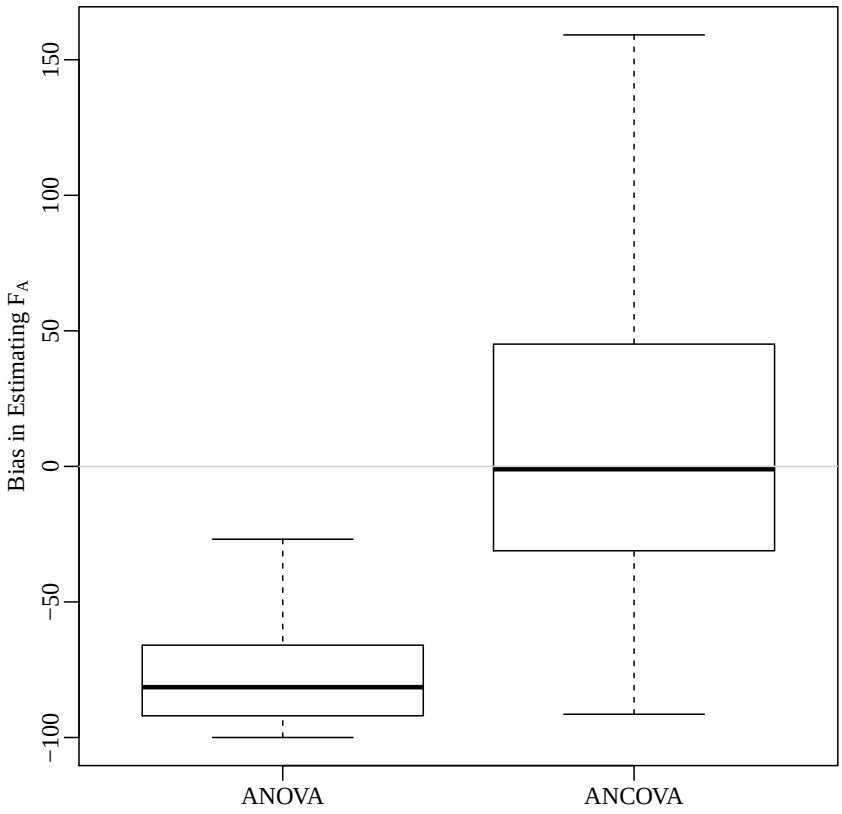

Figure 7. Distribution of bias in estimating the F statistic for the ANOVA and the ANCOVA for the A effect. For these results, the interaction variable does interact with the treatment effect, such that the treatment effect depends on the level of the selection variable.

data were simulated in such a way that the two-way interaction was non-existent in the population, but existed in the selected sample. As mentioned previously the data generating model was $Y = \mu + -10Z + 9AZ + 9BZ - 6ABZ + \epsilon$, where $\mu = 1$. The ANOVA model fitted was $Y = A + B + AB$. Note that nearly all of the estimates (both F and β) are quite biased. The degree of bias for the main effects of A and B in the ANOVA were 7% and 18%, while the maximum for the main effects in the ANCOVA were -52% and -16%. However, the β parameters that involve Z are unbiased, never exceeding 1%.

Table 4: Percent Bias in Parameter Estimates for the Two-Way ANOVA when an Interaction Exists Between the Selection Variable (Z) and the Treatment Effects (A and B).

	ANOVA	ANCOVA
F_A	7	-52
F_B	18	-16
F_{AB}	2857	3402
F_Z		-64
F_{AZ}		6
F_{ZB}		-46
F_{AZB}		-66
A	-116	-49
B	-49	-48
AB	-155	-51
β_Z		-0
β_{AZ}		-1
β_{ZB}		0
β_{AZB}		-0

Discussion

The majority of psychological research violates two key assumptions: first, that subjects have been randomly assigned to treatment conditions. Second, they have been randomly selected from a well-defined population (Rubin, 1974). Our paper has focused on the second violation. We have suggested that covarying out the selection variable may reduce or eliminate bias in estimating population parameters under certain conditions. Our first Monte Carlo demonstrated that an ANCOVA was sufficient to eliminate bias in the F and treatment effect estimates in a non-random sample, assuming the selection variable has been included in the model.

However, when the selection variable interacts with the treatment effect(s), unbiased estimates of the F statistic are elusive at best. In each of our simulations where there was an interaction between the two, all estimates of treatment

effect(s) were biased, even if all the selection variable was correctly identified and the correct model was estimated. The problem in this case is not a misspecified model, but rather the problem is non-random sampling.

Although the F seems to be biased whenever there is an interaction present, the treatment effects were not as troublesome. With the one-way ANOVA, nearly all β coefficients and treatment effects could be reproduced. In other words, although the F statistic could not be recovered, unbiased linear regression equations could. Using this information, perhaps future researchers could devise a correction for the F based off of the unbiased regression function by correcting the error term.

Unfortunately, estimating population parameters becomes increasingly complicated when the selection variable interacts with two treatment effects. We showed that it is possible to detect a two-way interaction in a selected sample that does not exist in the population. Efforts to recover parameters from a selected sample fail, even if all the correct variables are included in the model. Table 4 shows that all parameters related to the main effects and two-way interactions between A and B are biased. Although, estimates related to Z itself are unbiased, typically these estimates are not of interest.

It may be tempting to suggest that estimating effect sizes instead of the F will solve the problem of convenience sampling. Indeed, many suggested that effect size estimates could solve the "problem" of Null Hypothesis Significance testing (see Rodgers, 2010, for a review). However, the problems we have demonstrated will not be resolved by resorting to effect sizes. Recall that the F statistic can be expressed as a function of an effect size. For example, on effect size measure (R^2) is related to the F as follows

$$F = \frac{R^2}{1 - R^2} \times \frac{df_2}{df_1}$$

In other words, the F is a product of two functions, one related to degrees of freedom and the other related to the effect size. Since the degrees of freedom are unaffected by convenience sampling, the effect size will also be biased.

One may also consider another effect size estimate, Cohen's d. Recall that it is computed by dividing the mean difference between groups by the standard deviation. We have seen that the standard deviation is affected by selection, which will also yield biased estimates of Cohen's d.

In summary, covarying the selection variable only works when the selection variable does not interact with one or more treatment effects. Consequently, we recommend future researchers carefully consider the variables that make their sample non-random. If there is reason to believe any of these variables may interact with the treatment effect(s), then we recommend researchers use other sampling methods in order to obtain a more representative sample.

Is all this necessary?

Discussion centered around convenience sampling have been a hot topic in psychological journals for decades. Many have already argued that convenience sampling threatens the validity of psychological findings (see Arnett, 2008; Gordon, Slade, & Schmitt, 1986; Henrich et al., 2010; McNemar, 1946; Rosenthal, 1965; Rosenthal & Rosnow, 1969). Despite this fact, convenience sampling is very much alive and shows no signs of yielding to random sampling.

Perhaps part of the reason for this is the mistaken belief that testing and developing psychological theories does not require random sampling. Highhouse and Gillespie (2009), for example, argued in behalf of the practice of undergrad-

uate sampling. Citing several meta-analyses from the organizational literature (see Anderson, Lindsay, & Bushman, n.d.; Eagly, Karau, & Makhijani, 1995; Kluger & DeNisi, 1996; Kubeck, Delp, Haslett, & McDaniel, 1996; Sagie, 1994) they concluded that effect sizes from samples rarely differ significantly from effect sizes collected from organizations. They also argue that random samples are not required in order to generalize theories. Rather, they are only required when one wishes to describe the population of interest (see also Farber, 1952). To illustrate, they offered the following example.

> [I]magine a group of researchers with a theory about why some shows are more popular than others. For example, what is it about Wheel of Fortune that makes so many people want to watch it? One theory might be that people like to solve puzzles. Another might be that people enjoy seeing others compete for prizes. The researchers might test these theories by surveying a sample of television viewers using measures of attitudes toward puzzles and prizes. Another approach would be to randomly assign shows that differ in degree of emphasis on puzzles and prizes to a sample of television viewers. It is not necessary that these samples are representative of the population of all television viewers. (p. 257)

Let us further suppose the researcher couches his or her predictions using sophisticated psychological theories and terminology such as "need for cognition." If one theory is supported on an undergraduate sample, does that mean it is a sound theoretical development?

We argue that it is not. The problem with this assertion is that it assumes the theoretical effect (in this case, choosing puzzle-focused shows versus prize-based

shows) does not depend on characteristics unique to the sample. However, it is not hard to think of situations where sample-specific effects distort experimental findings. For example, suppose the researcher finds through experimentation that subjects tend to prefer puzzle-focused shows, supporting a need for cognition theory. Unfortunately, this characteristic (enjoying puzzles) may be unique to undergraduate psychology students. Suppose the same experiment was performed on a sample of homemakers and the findings were in reverse—homemakers prefer shows where hefty prizes are won or lost. If such were the case, any theory developed within a sample of undergraduate students would be misguided.

Highhouse and Gillespie (2009) did recognize this as a limitation: "It is only necessary that the sample does not systematically differ from the population in a way that would plausibly interact with the constructs of interest" (p. 257). They then recommend the researcher use theory to determine whether such an interaction might exist. For example, Birnbaum and Martin's (2003) theory predicted that students in a particular context, given a choice between two slot machines, would choose the one that gave fewer payoffs. This was indeed the case. Due to concerns that these findings may not generalize to more sophisticated decision-makers, they subsequently sampled decision-making scholars and the results were the same.

We too recommend the researchers carefully consider characteristics that might distort experimental findings. However, if a particular theory was developed in the lab it may not provide ample understanding of the unselected population to make such determinations. Furthermore, it is quite possible that the selection variable(s) interact with the treatment effect in ways that are difficult to anticipate. Consequently, although careful consideration may guide researchers in better understanding the limitations of their theory, it is no substitute for

better sampling.

Implications

We have demonstrated that population parameters such as the F statistic are poorly estimated from convenience samples. In some cases, the F is overestimated, while it is underestimated in other cases. Because the majority of experimental research in psychology is performed on undergraduate psychology students, we have reason to suspect that many psychological findings have been overstated, understated, or unfairly lost to the null hypothesis. In other words, psychological journals may be rife with both Type I and Type II errors.

This is particular problematic for cross-cultural studies. As mentioned earlier, many psychological findings fail to generalize across cultures (Henrich et al., 2010). Perhaps covariate adjustments may help mitigate this problem and help researchers understand how these findings differ across cultures.

Future Research

We have noted that covariate adjustments require information about the selection variable. In reality, it may be difficult to determine what variable or set of variables make subject pools non-random. Future research may compare undergraduate subject pools to a random sample on many potential variables to help determine where the significant differences lie. Using propensity score analysis, perhaps researchers could discover a relatively small collection of variables where the two samples differ. This may then inform future researchers on what variables ought to be collected from experimental subjects.

In this paper we have assumed that the researcher perfectly measured the selection variable(s). In reality, this would seldom occur. At best, researchers

will have a variable or collection of variables that are highly correlated with the selection variable(s). Future research may be directed at understanding how the results presented in this study would be affected by covarying a proxy variable, rather than the selection variable itself. We suspect that the results presented would be even less promising and that bias would increase.

Summary

In summary, we have demonstrated that convenience sampling may have unanticipated statistical problems that threaten the validity of experimental research. The best approach to mitigating bias is to carefully consider what variables make the selected sample non-random, then covary these out through an ANCOVA. Unfortunately, when an interaction exists between the selection variable(s) and the treatment effect(s), it is presently impossible to generalize the findings beyond the convenience sample. Although covarying the selection variable may mitigate or eliminate bias, there is no substitute for better sampling.

References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (n.d.). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3-9.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602 - 614.

Bareinboim, E., & Pearl, J. (2012). *Controlling selection bias in causal inference.* La Palma, Canary Islands.

Birnbaum, M. H., & Martin, T. (2003). Generalization across people, procedures, and predictions: Violations of stochastic dominance and coalescing. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research.* Cambridge, England.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Cooper, G. (1995). Causal discovery from data in the presence of selection bias. *Artificial Intelligence and Statistics*, 140-150.

Cramer, E. M., & Applebaum, M. I. (1980). Nonorthogonal analysis of variance–once again. *Psychological Bulletin*, *87*, 51-57.

Didelez, V., Kreiner, S., & Keiding, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statistical Science*, *25*(3), 368-387.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin*.

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

Fife, D., & Mendoza, J. L. (2013). The estimation of incremental validity in the presence of missing data.
(Submitted for publication)

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, *17*, 69-78.

Geneletti, S., Richardson, S., & Best, N. (2009). Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*, *10*(1).

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2/3), 61 - 135.

Highhouse, S., & Gillespie, J. Z. (2009). Do samples really matter that much? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in organizational and social sciences.* Taylor and Francis.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Kam, C., Wilking, J., & Zechmeister, E. (2007). Beyond the "narrow data base": Another convenience sample for experimental research. *Political Behavior*, *29*(4), 415 - 440.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: Historical review, meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Kubeck, J. E., Delp, N. D., Haslett, T. K., & McDaniel, M. A. (1996). Does job-related training performance decline with age? *Psychology and Aging*.

Kulich, R., Seldon, J. W., Richardson, K., & Servies, S. (1978, May). *Frequency of employing undergraduate samples in psychological research and subject reaction to forced participation.* Chicago, IL. (Paper presented at the meeting of the Midwest Psychological Association)

Levin, J. (1972). The occurrence of an increase in correlation by restriction of range. *Psychometrika*, *37*, 93-97. Available from http://dx.doi.org/10.1007/BF02291414 (10.1007/BF02291414)

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289-374.

Murray, G. R., Rugeley, C. R., Mitchell, D. G., & Mondak, J. L. (2013). Convenient yet not a convenience sample: Jury pools as experimental subject pools. *Social Science Research*, *42*(1).

Neyman, J. (1923). *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes.* Unpublished master's thesis, University of Kharkov.

Pearl, J. (1985, August 7-11, 2011). *Bayesian networks: A model of self-activated memory for evidential reasoning.* University of California, Irvine, CA..

Pearl, J. (2009). *Causality: Models, reasoning, and inference.* New York, NY: Cambridge University Press.

Pearl, J., & Bareinboim, E. (2011, August). *External validity and transportability: A formal approach.* San Francisco, CA.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. on the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *200*, 1-66.

Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*(1), 117-128.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1 - 12.

Rosenthal, R. (1965). The volunteer subject. *Human Relations*, *18*, 389-406.

Rosenthal, R., & Rosnow, R. (1969). The volunteer subject. In Rosenthal & Rosnow (Eds.), *Artifact in behavior research.* Academic Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29*, 343-367.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322-331.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112-118.

Sagie, A. (1994). Participative decision making and performance: A moderator analysis. *Journal of Applied Behavioral Science*.

Shadish, W. R. (2010). Campbell and rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, *15*(1), 3 - 17.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston, MA: Houghton Mifflin.

Thorndike, R. L. (1949). *Personnel selection: test and measurement techniques.* Oxford England: Wiley.

West, S. G., & Sagarin, B. J. (2000). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Research design: Donald campbell's legacy* (Vol. 2, p. 117-154). Thousand Oaks, CA: Sage Publications.

Appendix

Effects of Selection on the ANOVA

We begin by making several assumptions with regard to the ANOVA

1. The sample size of the selected sample ($\tilde{N}$) is the same as the sample size of the random sample (N). We make this assumption in order to make the estimates comparable.[6]

2. The number of treatment levels (a) is the same for both the random and selected sample. Otherwise, the estimates will be incomparable.

3. Subjects have been randomly assigned to treatment conditions

4. Selection results in a *reduction* in variance on the selected variable rather than an enhancement. Although selection can cause enhancement (Levin, 1972), this sort of selection is rare. However, if selection does actually increase variance, then the results presented will be opposite of that shown (e.g., the F test will be underestimated rather than overestimated).

Suppose we have three variables: Z (the selection variable), X (the treatment assignments), and Y (the outcome of interest). Under selection, it is known that

$$\sigma^2_{y.z} = \tilde{\sigma}^2_{y.z} \tag{A.1}$$

or the conditional variance of Y, given Z is unaffected by direct selection on Z (Pearson, 1903). However, suppose we are interested in estimating $\sigma^2_{y.x}$ when selection occurs on Z. It is generally *not* the case that $\sigma^2_{y.x} = \tilde{\sigma}^2_{y.x}$. In this case, Y has been restricted *indirectly* via Z. (X, on the other hand, has not been

[6]Of course selection will reduce the net sample size. However, when this occurs, it is difficult to determine whether differences in the F and/or standard errors are due to the differences in sample size or differences in estimation. Consequently, we will assume the same N for both estimates.

indirectly selected as long as random assignment has occurred, simply because the correlation between X and Z is zero.)

It is well known that the expected value of MSE is

$$E(MSE) = \sigma^2_{y.x} \tag{A.2}$$

However, under selection, the restricted expected value of MSE ($\tilde{\sigma}^2_{y.x}$) tends to underestimate the population value of $\sigma^2_{y.x}$ (because the poor performers within a group are removed, making the scores more homogenous). In other words,

$$\sigma^2_{y.x} > \tilde{\sigma}^2_{y.x} \implies \tilde{\sigma}^2_{y.x} = \sigma^2_{y.x} - c \tag{A.3}$$

where c is some positive constant that indicates the degree of bias of $\tilde{\sigma}^2_{y.x}$ in estimating $\sigma^2_{y.x}$.

The expected value of MSB (mean squares between) is

$$E(MSB) = \sigma^2_{y.x} + \sum_j n_j \alpha_j^2/(a-1) \tag{A.4}$$

where n_j is the sample size of treatment group j, α_j is the treatment effect of group j, and a is the number of treatments. Under selection, the following hold

$$\begin{aligned} &\tilde{\alpha}_1^2 \neq \alpha_1^2 \\ &\tilde{\alpha}_j^2 = \alpha_j^2 \text{ where } j \neq 1, \\ &\quad \implies \tilde{\alpha}_j^2 = \alpha_j^2 + d \end{aligned} \tag{A.5}$$

where group $j = 1$ is the control group, and d is a positive value indicating bias in estimating α_1^2 from $\tilde{\alpha}_1^2$ (and by implication indicates bias in estimating α_j^2 from $\tilde{\alpha}_j^2$). In other words, the effect of the treatment is unaffected by selection because of random assignment. However, the "effect" of the control group is not unaffected because the untreated sample is made up of individuals with higher scores. Consequently, the quantity $\sum_j \tilde{n}_j \tilde{\alpha}_j^2$ will overestimate $\sum_j n_j \alpha_j^2$ by the degree to which $\tilde{\alpha}_1^2 > \alpha_1^2$, which is indicated by d.

Recall that the F statistic is computed as follows

$$\begin{aligned} F &= \frac{MSR}{MSE} \\ \tilde{F} &= \frac{\tilde{MSR}}{\tilde{MSE}} \end{aligned} \tag{A.6}$$

Under selection $\tilde{F}$ will overestimate F. In order to demonstrate this fact, we will use Equations A.2 and A.4, and using the inequalities expressed in Equations A.3 and A.5

$$\begin{gathered} \tilde{F} - F > 0 \\ \frac{(\sigma_{y.x}^2 - c) + \frac{1}{a-1}\sum_j n_j \alpha_j^2 + d}{\sigma_{y.x}^2 - c} - \frac{(\sigma_{y.x}^2) + \frac{1}{a-1}\sum_j n_j \alpha_j^2}{\sigma_{y.x}^2} > 0 \end{gathered} \tag{A.7}$$

After some simplification, we get

$$\frac{d\sigma^2_{y.x} + \frac{c}{a-1}\sum_j n_j\alpha_j^2}{\sigma^2_{y.x}\tilde{\sigma}^2_{y.x}} > 0 \tag{A.8}$$

Notice that all terms in the numerator are positive (c and d are positive be definition, n_j will be positive since it is the number of people in a treatment group, $a-1$ will always be positive since it is the number of treatment groups, α^2 will always be positive because it is a squared term, and the variance $[\sigma^2_{x.y}]$ will always be positive barring heywood cases). Likewise, the denominator will always be positive since both terms are variances. Therefore, under selection the $\tilde{F}$ will always overestimate F.

Recall that the F statistic can be expressed as the product of an effect size (SSB/SSE) and some function of the degrees of freedom (df_E/df_B). Because degrees of freedom are unaffected by selection (according to Assumptions 1 and 2), effect sizes will also be overestimated.