Fall 11-15-2018

Analysis of Covariance with Heterogeneity of Regression and a Random Covariate

Christopher J. McLouth
University of New Mexico - Main Campus

Christopher J. McLouth
Candidate

Psychology
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Timothy E. Goldsmith, PhD, Chairperson

Harold D. Delaney, PhD

Davood Tofighi, PhD

Li Li, PhD

ANALYSIS OF COVARIANCE WITH HETEROGENEITY OF VARIANCE AND A RANDOM COVARIATE

by

CHRISTOPHER J. MCLOUTH

B.A., Psychology, University of Michigan, 2008
M.S., Psychology, University of New Mexico, 2013

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctorate of Philosophy in Psychology

The University of New Mexico
Albuquerque, New Mexico

December, 2018

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Harold Delaney, for his wisdom, thoughtful guidance, and unwavering support and encouragement throughout this process. This project would not have been possible without it. I also want to thank my committee members, Drs. Goldsmith, Tofighi, and Li for their helpful contributions. I thank my wife, Laurie, whose love and support provided the encouragement to complete this dissertation. Finally, I thank my friends and family for being understanding and supporting my academic mission.

ANALYSIS OF COVARIANCE WITH HETEROGENEITY OF REGRESSION AND A RANDOM COVARIATE

by

Christopher J. McLouth

B.A., Psychology, University of Michigan, 2008

M.S., Psychology, University of New Mexico, 2013

Ph.D., Quantitative Psychology, University of New Mexico, 2018

ABSTRACT

The analysis of covariance (ANCOVA) is a statistical technique originally developed by Fisher (1932) to increase the precision of the estimate of a treatment effect in experimental data. It is used when researchers have both qualitative and quantitative predictors of a continuous outcome. ANCOVA's most basic assumptions are similar to those of analysis of variance (ANOVA) and other related linear models, but also include an additional assumption regarding the regression line relating the outcome and the covariate. This assumption, referred to as homogeneity of regression, requires that the within-group regression slopes be the same for all groups. Typically, one also assumes that the covariate is a fixed effect, but some methodologists question this practice.

Consequences of either employing a model allowing for heterogeneity of regression (an ANCOHET model) or presuming that the covariate is random can be dealt with fairly easily if both do not occur simultaneously. However, a problem arises when one simultaneously encounters both a random covariate and heterogeneity of regression: the interaction between the random covariate and the fixed factor of treatment will affect the apparent evidence for the main effect of treatment. The current study investigated the utility of different methods of testing for the main effect of treatment in the presence of heterogeneity of regression with a random covariate. Using a Monte Carlo simulation, a 2x3x5x2x3 design manipulated the number of groups, sample size per group, extent of heterogeneity of regression, presence of a group effect, and test location to investigate this issue. For each combination of these factors, different types of tests of the group effect were conducted, as explained in the Method section. Two error terms, the mean square for the interaction and an average of the mean square error for an ANCOHET model and the interaction mean square, performed poorly across all simulations for all metrics. On the other hand, the ANCOHET and ANCOVA error terms produced Type I error rates, power, confidence interval coverage rates and widths, as well as average standard errors under low and medium levels of heterogeneity of regression that were acceptable. When heterogeneity of regression was high or extreme, the ANCOHET approach underestimated the true standard error, whereas the ANCOVA error term did not. Using an approach to increase the ANCOHET standard error based on Chen (2006) did not result in a large enough increase to make up for this underestimation. In conclusion, with the low and moderate levels of heterogeneity of regression typically reported in the literature the ANCOHET test of the group effect can be recommended

even with a random covariate. Under large or extreme levels of heterogeneity of regression, an error term from a standard ANCOVA should instead be used.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

ANCOVA – An Overview

The analysis of covariance (ANCOVA) is a statistical technique originally developed by Fisher (1932) to increase the precision of the estimate of a treatment effect in experimental data, and hence to increase the power of detecting such an effect. It can be used when researchers have both qualitative and quantitative predictors of a continuous outcome. In experimental data, where units are randomly assigned to different groups, the qualitative predictor is defined by two or more treatment conditions, with one or more groups often serving as a control for an active treatment. The quantitative predictor is often, though not necessarily, a pre-measure of the dependent variable, and is referred to as the concomitant variable or covariate.

ANCOVA Assumptions

The following is the basic ANCOVA model:

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the score of the *i*th individual in the *j*th group on the dependent variable, μ is a "grand mean" parameter (but should be conceived as an average of intercepts when the covariate is not centered at its mean), α_j is the effect associated with group *j*, β is the population within-group regression coefficient characterizing the linear relationship between *Y* and *X*, X_{ij} is the score of the *i*th individual in the *j*th group on the covariate, and ε_{ij} is the error term for the same subject. The most basic assumptions involve the

error term, ε_{ij}. That is, the errors are assumed to be independent and normally distributed with a mean of zero and a constant variance.

There is an additional assumption regarding the regression line relating the outcome and the covariate, namely, that the within-group regression slope be the same for each group. This assumption can be seen in the model where the slope, β, does not have a subscript and takes on the same value regardless of an individual's group membership. This assumption is referred to as homogeneity of regression. Finally, one typically assumes that the effects of both the treatment and the covariate are fixed (cf. Maxwell, Delaney, & Kelley, 2018, Chapter 9).

Regarding the Homogeneity of Regression Assumption

The conventional ANCOVA model as specified does not allow for a distinct slope for each group. Thus, predictions for each individual are based on a single slope estimate that is computed as a weighted average of the slope estimates for each group considered separately.

In many situations, the within-group regression slopes will be similar, and can be reasonably represented by a single slope parameter. However, as these within-group slopes become more and more disparate, using a single parameter or weighted average of the separate slopes to represent the relationship between the covariate and the outcome in these groups will be misleading.

Consequences of Violating the Homogeneity of Regression Assumption

Far from being disastrous, the presence of heterogeneity of regression in ANCOVA is fairly easily dealt with analytically. How one should deal with

heterogeneity of regression in ANCOVA can be approached in a similar fashion to how one would proceed when a significant interaction is observed in a two-way ANOVA. In a two-way ANOVA, a significant interaction between the two factors means that the difference on the dependent variable across levels of one factor is not consistent across levels of the other factor. In an ordinal interaction, the superiority of one group over the other is maintained across levels of the off factor. For a disordinal interaction, this superiority is not maintained. Whereas the presence of an interaction in factorial ANOVA can require a more nuanced interpretation of main effects, so too can the presence of unequal within-group regression slopes in ANCOVA. In other words, the presence of a significant main effect of treatment in the presence of heterogeneity of regression (which heterogeneity would be detected via a test of the interaction between the continuous covariate and the grouping variable) often cannot be adequately followed up by the comparison of adjusted, conditional means at the grand mean of the covariate, particularly in the case of a disordinal interaction. Because the within-group regression lines are not parallel, the difference in conditional means is not consistent across the values of *X*. As a result, the presence of a significant effect of treatment does not necessarily allow one to conclude that the means of the treatment conditions are significantly different at all values of *X*. Instead, it may be the case that the conditional means are significantly different at certain values of the covariate but not others. Heterogeneity of regression should signal to the researcher that a more thorough investigation of the relationship between the treatment effect and level of the covariate may be warranted.

In the presence of heterogeneity of regression, researchers may choose to use a different approach to interpret the significant effect of treatment. Two options, neither of which is recommended, would be to either ignore the covariate completely or to block on the covariate and then perform a factorial ANOVA (Glass, Peckham, & Sanders, 1972). Both options throw away important information by either ignoring the covariate all together, or by discretizing a continuous variable and decreasing the amount of error variance in the dependent variable that the otherwise continuous covariate would be able to account for (Maxwell, Delaney, & Dill, 1984).

There are two notable approaches for assessing differences in treatment conditions in the presence of heterogeneous slopes: the Johnson-Neyman technique (J-N) (D'Alonzo, 2004; Johnson & Neyman, 1936; Preacher, Curran, & Bauer, 2006), and the simple slopes technique (Aiken & West, 1991; Cohen, Cohen, West, & Aiken, 2003; Preacher, Curran, & Bauer, 2006) which Rogosa referred to as a pick-a-point procedure (Rogosa, 1980, p. 313ff.). The J-N technique allows one to determine the range of values of X at which there are significant treatment differences, whereas the simple slopes or pick-a-point technique can test for between-group differences at certain selected values of X. When using the simple slopes approach, in lieu of a priori relevant values at which to test for group differences, a common practice is to test at the grand mean of the covariate and at one standard deviation on either side of this mean (Maxwell, Delaney, & Kelley, 2018; Preacher et al., 2006; Preacher & Hayes, 2004).

However, there are also legitimate reasons to be interested in the omnibus test for the effect of treatment even in the presence of heterogeneity of regression. One potential application of research findings involving heterogeneity of regression would involve

treatment assignment based on an individual's measured level of the covariate of interest. Returning to a two-way ANOVA as an example, consider a 2x2 ANOVA comparing two possible treatment conditions for subjects classified as one of two personality types (e.g. Type A or not Type A). Although one might have hypothesized that one could achieve optimal results by assigning each personality type to the treatment matched to that type, it is possible that even if the treatment and personality type factors interacted, one might observe an ordinal interaction where the superiority of one group is maintained across the levels of the off factor. If treatment A is always better than treatment B, then there is no reason to ever assign an individual to treatment B. In ANCOVA, however, the off factor is continuous as opposed to discrete. Due to the fact that the regression lines are not parallel, unlike ANOVA, in theory there exists some point at which the lines will intersect. However, this "cross over" point, or the point at which the direction of the treatment effect changes, may not be within the range of possible values of the covariate.

Similarly, in certain real-world settings the collection of data on the covariate may be time-prohibitive and/or cost-prohibitive in practice (e.g., genetic testing, performing a full cognitive battery, etc.). In other real-world settings, assignment of individuals to different treatments may not be practical (e.g., because of the cost or staff required to administer different treatments) even if the score on the covariate were known. In either case, there are many situations where a treatment assignment based on the level of a covariate is not a realistic option. Instead, it may be sufficient to know the impact of the treatment for the "average" individual. Thus, even in the presence of heterogeneity of regression, conducting a test of between-group differences at the center of the distribution of the covariate can still be of substantive interest.

Type I Error Rates in the Presence of Heterogeneity of Regression

Whereas Type I error rates, the probability of incorrectly rejecting the null hypothesis of no group differences, are relevant in the case of homogeneity of regression, the issue is not as clear cut when the relationship between the covariate and the outcome differs across two or more groups. In lieu of theoretically relevant points along X at which to test for group differences, researchers will often begin by conducting the test at the sample grand mean of X, or $\bar{X}$, and then one standard deviation above and below $\bar{X}$, as previously mentioned. The absence of a main effect for treatment in the presence of heterogeneity of regression could be represented graphically by a plot showing that the heterogeneous regression lines for two groups intersect at the population mean of X, or μ_X. In such a case, given the test of the main effect of group would typically be conducted at the sample grand mean of X, or $\bar{X}$, it is only when $\bar{X}$ and μ_X are equal that the expected difference between the predicted means on the dependent variable at the sample grand mean of X would be zero. However, when $\bar{X}$ and μ_X are not the exact same, as will typically be the case since $\bar{X}$ is an estimate of μ_X, the rejection of the null hypothesis is no longer an incorrect rejection, since the two groups are only the same on the dependent variable at μ_X. Since $\bar{X}$ in a sample will rarely ever equal μ_X in the case where X is a random variable (more on this below), it is not necessarily correct to attribute rejection of the null hypothesis of no group main effect to a Type I error in the presence of heterogeneity of regression, because testing for group differences anywhere other than μ_X, that is, at the exact point of intersection between the two regression lines, will result in the test being conducted at a point where the true difference is non-zero.

Accuracy in Parameter Estimation

Along with the push for a decreased reliance on null hypothesis significance testing (NHST) (e.g., Cohen, 1994; Schmidt, 1996; Wasserstein & Lazar, 2016) has come the recommendation for the increased reporting of confidence intervals (e.g., Cumming & Finch, 2001; Thompson, 2007). Confidence intervals not only provide the same information as NHST (i.e., is the parameter significantly different than the null value hypothesized?), but they also provide the direction of the difference and a range of plausible values. As covered more extensively in sources devoted to the topic (see Kelley, Maxwell, & Rausch, 2003 and Lai & Kelley, 2012), just because a confidence interval excludes the value posited under the null hypothesis, does not mean that the range of possible values gives a researcher a high degree of certainty in the value they found. For instance, a study might find that the difference in means between two groups constitutes a medium effect according to benchmarks proposed by Cohen (1992). However, the confidence interval constructed around this standardized difference might include as plausible values effect sizes that range from small to large, even if the study were adequately powered based on conventional power analysis to detect a significant difference. As a result, even though the null hypothesis was rejected, the corresponding confidence interval might still be "embarrassingly large" (Cohen, 1994, p. 1002).

Thus, it is not only important that a confidence interval contains the population parameter, but that it does so in a way that the range of plausible values is sufficiently narrow (Maxwell, Kelley, & Rausch, 2008). This is what is referred to as accuracy in parameter estimation (AIPE). In line with this, the current dissertation will report information on coverage and width of confidence intervals around estimated effects.

To Fix or To Presume Random?

With the basic ANCOVA model shown above, researchers can make inferences to hypothetical replications that either involve the same *X* values observed in the original study, or involve different random samples from a population of *X* values. As Henderson observed, "It is the fixed model that is almost always intended when covariance analysis is discussed in the statistical literature. It is mixed models, however, that usually best represent the real world from which most meaningful inferences are drawn" (Henderson, 1982, p. 624).

Mixed models are models that contain a combination of both fixed and random effects. Fixed effects are variables whose specific values are of interest, and researchers are not intending to make inference beyond those values. Good examples of fixed effects are the classification or treatment variables in ANOVA or ANCOVA. When one is comparing the efficacy of two separate treatments, say Cognitive Behavioral Therapy (CBT) versus Motivational Interviewing, it is not typical to attempt to make inferences beyond the two treatments under examination, and therefore these two treatments are considered the levels of a fixed factor. Similarly, when the covariate is treated as a fixed factor, as Rogosa remarked, "Inferences from these data are restricted to subpopulations having the same values or configuration of *X* because inferences from the linear model are conditional on the observed values of *X*" (Rogosa, 1980, p. 308).

On the other hand, random effects are involved when researchers aim to draw conclusions regarding levels of a factor that were not included in the design of a study. When the levels of a discrete factor such as therapist are treated as random, this would mean that replications of a study would include different therapists than in the original

study. What would it mean for a covariate *X* to be presumed to be random? This would imply that the new sample of interest in any replication would have different individual values of the covariate and hence the groups in the new sample would also have different mean levels on the covariate than the original sample because of sampling variability. This assumption is regarded as more realistic by many methodologists. For example, Huitema remarked "Because subjects are randomly sampled from the population, it is realistic to view the *X*-values in a given experiment as a random sample of *X*-values from a population of values… One reason, then, why future samples will generally have different values of *X* is because *X* is a random variable" (Huitema, 1980, p. 188).

Consequences of Treating a Factor as Random

There are several major consequences of treating a factor as random. The first deals with the inferences that can be made. When treating a factor as fixed, researchers can make inferences to only the levels of the factor that were included in the original study. This might be fine in some situations, but it is often the case that one is interested in making inferences to a population that has levels not included in the original study. Just as one implicitly treats participants as a random sample from a population as a way of justifying inferences being made about the population from which the participants are selected, treating a factor as random allows making inferences to the population of possible levels of that factor, not just the levels included in the study.

The second consequence of presuming a factor is random has to do with how the tests of effects are to be carried out. In a two-way ANOVA where factor A is fixed and factor B is random, it can be shown that any interaction between the two factors will intrude on the expected mean square of the fixed factor (Maxwell et al., 2018, Ch. 10).

This in turn implies that an appropriate test of factor A must use as a denominator error term a different effect than mean square within, which would be used in a conventional two-way ANOVA with two fixed factors. In fact, the appropriate error term for the test of the fixed effect in the mixed two-way ANOVA can be shown to be the mean square for the A x B interaction.

Additionally, Crager (1987) presented work on the impact of a random covariate on the standard error of the difference in adjusted means for a standard ANCOVA. In particular, he asserts that the variance of the difference in predicted means would be:

$$Var(\hat{c}) = \sigma^2 \left[\frac{2}{n} + \left(\frac{2n-2}{2n-4}\right)\frac{1}{n(n-1)}\right]$$

where $\hat{c}$ is the estimated difference in adjusted means (see Appendix A for further discussion). In an update to Crager, Chen (2006) extended this work to a random covariate in the presence of heterogeneity of regression. In particular, Chen suggested that the square of the standard error in adjusted means derived under the assumption of a fixed covariate will be too small by the following term:

$$Var\left[E\left(\hat{c}|\vec{X}\right)\right] = (\beta_1 - \beta_2)^2 \frac{\sigma_X^2}{(n_1 + n_2)}$$

where β_1 and β_2 are the regression coefficients in the two groups, n_1 and n_2 are the sample sizes in the two groups, $\vec{X}$ designates the observed set of scores on the covariate, and σ_X^2 is the variance of the population of the X covariate scores.

The Issue at Hand: Heterogeneity of Regression with a Random Covariate

Consequences of individually violating either of the previous two assumptions, that is, homogeneity of regression and that all factors are fixed, are easily dealt with. However, a problem arises when one simultaneously encounters both a random covariate and heterogeneity of regression. This problem is illustrated in Figure 1.

This figure represents three replications of an experiment. In each replication, the group means of both the control and experimental groups are constant ($\bar{Y}_C$ and $\bar{Y}_E$, respectively), as are the within-group slopes ($\hat{\beta}_j$). Thus, neither of these factors can contribute to the variability in adjusted means (i.e., the difference between $\bar{Y}'_E$ and $\bar{Y}'_C$, represented by the braces "{"). The only difference between the three replications is that the group means of the covariate ($\bar{X}_j$) are varying. This variability in the covariate means produces estimates of the adjusted treatment effect that differ from replication to replication. The principle illustrated by this simple diagram of the implications of variability in the covariate means for the estimated variability in the predicted means on the dependent variable (and also for difference in such estimated means across groups) is developed more rigorously in Appendix A.

METHOD

The current study sought to investigate the appropriateness of different methods of testing for the main effect of treatment in the presence of heterogeneity of regression with a random covariate.

How to Test for Treatment Main Effects

Given the potential for the effect of a random factor to intrude upon the mean square for the treatment effect, the traditional ANOVA approach to a two-way design with a fixed factor and a random factor indicates that using the typical *MS* error as the denominator of the *F* test would not test exclusively for a consistent treatment main effect. Instead, it could be regarded as testing for the presence of a treatment main effect or the interaction between the treatment and the random factor. To test for the treatment main effect one must use as a denominator error term the interaction mean square. This suggests a potentially more adequate test of treatment in an ANCOVA with a random covariate and heterogeneous regressions might utilize in the denominator an error term that takes into account the impact of the random covariate. This study considered the following as potential error terms to use in the test of a main effect of treatment:

1. A procedure using as an error term that associated with a model allowing for heterogeneity of regression (an "ANCOHET" model).
2. Using the interaction between the fixed factor of treatment and the random covariate, i.e. $MS_{A \times X}$, as the error term.
3. An error term based on a standard ANCOVA or equivalently an average of the error terms in (1) and (2) where each is weighted by its degrees of freedom.

4. A pooled error term using an unweighted average of the $MS_{residual}$ for the ANCOHET model used in (1) and the interaction error term used in (2).

Rogosa described a variation of method (1) as a "safer ANCOVA" (1980, p. 312) test of the treatment effect, because the estimate of mean square error used in the denominator of the test does not presume equality of regression slopes across groups. However, the numerator of the Rogosa test, which was implemented in the simulations of Harwell and Serlin (1988), is based on the adjusted means from a conventional ANCOVA (see Appendix B for more detailed discussion). As such, theoretical as well as simulation results suggest that the method is inappropriately liberal. Thus, the ANCOHET method used in the current dissertation employs a numerator that takes into account the fact that different slopes are being estimated in each group.

What exact degree of departure from the nominal alpha level warrants a judgment that a test is either liberal or conservative is not entirely clear. In one early paper, Cochran (1952) declared that a difference from a nominal .05 alpha level "is regarded as unimportant…if the exact *P* lies between .04 and .06" (p. 328). Bradley (1978) suggested that a stringent criterion of robustness might be requiring the true alpha level to depart from the nominal alpha by no more than .1α, that is, in the case of a nominal .05 alpha level if the true alpha was between .045 and .055, whereas a liberal criterion might allow the true alpha to depart from the nominal alpha by .5α, that is, the true alpha could be acceptable if it were between .025 and .075. Serlin (2000) suggested one could test non-specific null hypotheses that when rejected would allow the inference that the true alpha of a procedure was within pre-specified limits, and suggested appropriate limits of up to .25α, or from .0375 to .0625 in the case of a nominal alpha of .05 (see Serlin, 2000, Table

1, p. 237). In numerous simulation studies, rejection rates are highlighted as liberal or conservative if they are outside the range specified by the sampling error given the number of simulations being run (see Appendix D, Equation D.1). For example, in the Harwell and Serlin (1988) study where 2,000 simulations were run in each condition, the range of 1.96 standard errors below or above the nominal alpha was from .040 to .060. In the current research which used 10,000 simulations in each condition, the comparable range of 1.96 standard errors around the nominal .05 level yielded limits of .046 and .054, which is approximately equal to Bradley's stringent criterion for robustness.

A further issue is that the exact pick-a-point test derived by Rogosa and others presumes that the values of the covariate would be fixed over replications, which was not the case in the current simulation study. Because of that, an additional factor that needed to be considered was the exact point on the covariate where the difference across groups was to be assessed. In most practical situations with a random covariate, the population mean will be unknown. Thus, it is of interest to evaluate the impact of testing the treatment effect at other reasonable values of the covariate at which investigators might choose to test for a treatment effect. In rare situations, the population mean on the covariate may be known and could be used as the point at which to conduct the test. Much more often, the population mean will be unknown but could be approximated by the sample grand mean on the covariate. A third reasonable alternative would be to test for the treatment effect at the point where the standard error of the difference is at its lowest value, a point known in the literature as the center of accuracy (see, e.g., Rogosa, 1980). Each of these three points will be used as the test location in the current study, with the predicted difference in means in each case being estimated by using the

ANCOHET model. However, it should be noted that in the two-group case this predicted difference using heterogeneous regressions at the center of accuracy will exactly equal the difference in adjusted means computed in a standard ANCOVA with homogeneous slopes (Maxwell et al., 2018, p. 528; Rogosa, 1980, p. 310). As a result of the test being conducted at a point other than μ_X, the rejection rate is expected to be greater than the nominal .05 rate, at least when the ANCOHET error term is used as the denominator error term in the test. On the other hand, utilizing the mean square for the interaction between the treatment and covariate as the error term, while having what would be regarded as the appropriate mean square in a traditional ANOVA approach to determining error terms in a mixed design, may well result in a lack of power for testing the treatment main effect. If the treatment factor has two levels, using the interaction mean square as the denominator of the F test will only have one degree of freedom, since $df_{denom} = (a\text{-}1)(1)$ where a is the number of levels in the treatment factor. This will result in an $F_{critical}$ of 161. Consequently, it may be possible to construct and utilize a pooled error term that will allow empirical α levels to remain close to the nominal .05 level without a substantial loss in power. Two pooled error terms, as described in the next section, were constructed using (1) and (2) above with different weighting schemes.

Additional Error Terms

Along with the mean square error for the ANCOHET model and the mean square for the interaction between the covariate and the grouping variable, two additional error terms were utilized in testing for the main effect of group:

ANCOVA. The first weighting scheme consisted of a weighted average of the mean square error from the ANCOHET model that allowed for heterogeneity of

regression and the mean square for the interaction, where the weights were the degrees of freedom associated with that mean square. This turned out to be equivalent to using the mean square error from an ANCOVA analysis, and will henceforth be referred to as the ANCOVA error term.

Equal Weights. The second weighting scheme (referred to as Unweighted or UNW henceforth) was calculated as an unweighted average of the mean square error from an ANCOHET model and the mean square for the interaction. Given the ANCOHET error term was anticipated to potentially be an underestimate of error and thus lead to overly liberal tests and the interaction error term with its small degrees of freedom was anticipated to lead to overly conservative tests, this alternative was included to allow an investigation of a potential compromise between the ANCOHET and the interaction error term approaches. The denominator degrees of freedom used in tests with this error term also was simply the average of the degrees of freedom associated with the ANCOHET error mean square and with the interaction mean square.

Simulation Design

The simulation study was conducted by manipulating the following factors displayed in Table 1. The design manipulated the number of groups (2 or 3), sample size per group (10, 30, or 100), extent of heterogeneity of regression (none, low, medium, high, extreme), presence or absence of a treatment effect, and location of test (at the population mean of the covariate μ_X, at the sample grand mean $\bar{X}$, or at the center of accuracy C_a which is defined in Appendix A). The 2x3x5x2x3 design created 180 different conditions. To decrease the complexity of the study design and interpretation, data were generated from a normal distribution where the homogeneity of variance

assumption was met (see more on this below). Differing levels of heterogeneity of regression were defined based on the extent of difference in within-group correlations of the covariate and dependent variable (see below for a discussion of how these and other levels were chosen). Regarding the different levels of effect size (i.e., difference between the means of the control and treatment), a null effect size was used for estimates of Type I errors, and non-null conditions were designed, generally following the approach of Harwell and Serlin (1988, Table 2, p. 272), in order to achieve a power of 80% for a conventional ANCOVA test of the treatment effect (see section on Non-null conditions below). For each condition, 10,000 samples were generated. Only equal-*n* cases were considered.

Data Generation

Data were generated using the data processing and generation abilities of SAS 9.4, commonly referred to as the "data step." See Appendix C for examples of the SAS syntax used to generate and analyze the simulated data, including comments highlighting the goal of each step of the program. For the two-group null condition, data were generated according to the following model:

$$Y_i^s = b_1 X_i + e_i$$

for individuals in the first group and

$$Y_i^s = b_2 X_i + e_i$$

for individuals in the second group. In each of these equations, *s* indexes the simulation or sample number (ranging from 1 to 10,000) and b_1 and b_2 are respectively the raw or

unstandardized population regression coefficients for the first and second groups. Both X_i and e_i were generated from a normal distribution with a mean of zero and a standard deviation and variance of one (i.e., X and $e \sim N(0,1)$). This data generation method results in population regression lines that intersect when $\mu_X = 0$.

Following the data generation, the data step computed and saved the values at which the test of the grouping variable would occur. Then, the GLM Procedure (i.e., PROC GLM) was used to analyze each of the 10,000 simulated datasets per condition and the relevant p values and confidence intervals for the difference between the adjusted means were computed and saved. Type I error rates were calculated as the proportion of the total number of simulations where the p values were less than the nominal .05.

Non-Null Conditions to Determine Power

After the null conditions were simulated to determine Type I error rates, an additional set of conditions was constructed to determine the power of the separate approaches. For these non-null cases, a constant value was added to specific groups during the data generation phase.

For each of the three sample size conditions, constants were added to a single group mean in the two-group condition. The constants were chosen so that the null hypothesis of no between-group difference in adjusted means would be rejected 80% of the time in a conventional ANCOVA if the slope were equal to the mean of the values in the heterogeneous slope conditions. For the three-group case, a constant was added to the simulated data points in only one of the three groups. Rather than powering according to an omnibus test of between group differences, powering based on a contrast was

employed instead. In particular, the nonzero group was compared to the average of the two groups whose means were set to be zero. In real world applications, this type of situation might be seen where researchers are comparing a single active condition to two distinct control conditions.

Table 2 contains a list of the constants used to produce a power of 80% for each combination of sample size and number of groups. Also included are two measures of effect size: Cohen's *d* and *f*. For the smallest sample size condition ($n = 10$; $N = 20$ or 30), the difference in means would constitute an effect size that is very large (i.e., exceeds the conventional cutoff for a large effect). For the middle sample size condition ($n = 30$; $N = 60$ or 90), the effect size would be considered medium to large, that is, between Cohen's cutoffs for a medium and a large effect. For the largest sample size condition ($n = 100$; $N = 200$ or 300), the effect size would be classified as small to medium.

Confidence Interval Construction and Accuracy Estimation

Given rejection rates in testing the group factor that are somewhat above .05 when there is no main effect for treatment can be misleading in the presence of heterogeneity of regression – particularly when the covariate is random – confidence interval coverage was also evaluated. To accomplish this, we calculated the actual between-group difference that would be seen in the population when evaluated at a value other than the population mean of the covariate. Since the population values of the slopes are known, this exact value can be calculated as the difference between the expected means of the groups at the X value, say X_p, used as the point to conduct the test. This would be calculated as:

$$Y_1' - Y_2' = b_1 X_p - b_2 X_p$$

where b_1 and b_2 denote the known population raw regression coefficients. Confidence interval coverage was then calculated as the proportion of individual confidence intervals that contain the actual difference in adjusted population means. In general, confidence intervals for the difference in adjusted means are calculated as:

$$estimate \pm (critical\ value)(estimated\ standard\ error)$$

In the three-group case, confidence intervals were constructed around the estimated value of a contrast, with the contrast comparing one estimated mean with the average of the two other estimated means. The standard errors of the estimated difference in means corresponding to the four different error terms used in the tests of the treatment effect in the two-group case will be given next, followed by the standard errors of the contrast estimates of interest in the three-group case.

Standard Errors for Two-group Case

ANCOHET. In the case of ANCOHET, the estimated standard error of the difference in predictions at X_p , derived as shown in Appendix A under the assumption X is fixed, is:

$$ANCOHET: \hat{\sigma}_{\hat{Y}_{p1}-\hat{Y}_{p2}} = \sqrt{\frac{E_F}{df_F}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(X_p - \overline{X_1}\right)^2}{\sum_i (X_{i1} - \overline{X_1})^2} + \frac{\left(X_p - \overline{X_2}\right)^2}{\sum_i (X_{i2} - \overline{X_2})^2}\right]}$$

This equation takes into account the sampling error of the Y group means along with the sampling error of the estimates of each group's slope and how far X_p is from the group means on X. For the ANCOHET approach, the precision of the difference in group means

is inversely related to the distance X_p is from the group means. The ratio $\frac{E_F}{df_F}$ above is the mean square error for the ANCOHET model which allows for a different slope in each group:

$$Y_{ij} = \mu + \alpha_j + \beta_j X_{ij} + \varepsilon_{ij}$$

ANCOVA. In the conditions where the ANCOVA error term is used, the standard error of the difference between group means is calculated as follows:

$$ANCOVA: \hat{\sigma}_{\hat{Y}_{p1}-\hat{Y}_{p2}} = \sqrt{\frac{E_F}{df_F}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\overline{X_1} - \overline{X_2})^2}{\sum_j \sum_i \left(X_{ij} - \bar{X}_J\right)^2}\right]}$$

In this case, $\frac{E_F}{df_F}$ refers to the mean square error for the traditional ANCOVA model assuming homogeneity of regression, i.e.

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{ij}$$

Interaction. For the interaction error term the standard error is calculated as it would be calculated for testing a contrast in an ANOVA where the mean square for the interaction is used as the error term, i.e.

$$Interaction: \hat{\sigma}_{\hat{Y}_{p1}-\hat{Y}_{p2}} = \sqrt{MS_{AxX}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Equal Weights. As a final alternative, the standard error for the equal weights or unweighted case is calculated by using in place of the mean square for interaction an unweighted average of the mean square for interaction and the mean square error from the ANCOHET model, i.e.

$$UNW: \hat{\sigma}_{\hat{Y}_{p1}-\hat{Y}_{p2}} = \sqrt{MS_{UNW}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

where $MS_{UNW} = \frac{MS_e+MS_{AxX}}{2}$

Chen's Increment to ANCOHET Standard Error. As previously mentioned, the impact of a random covariate in the presence of heterogeneity of regression may result in an estimated standard error that is too small when using the ANCOHET approach (Chen, 2006; Crager, 1987). Chen (2006) concludes that the square of the standard error derived under the assumption of a fixed covariate will be too small by the following term:

$$Var\left[E\left(\hat{c}|\vec{X}\right)\right] = (\beta_1 - \beta_2)^2 \frac{\sigma_X^2}{(n_1+n_2)}$$

For estimates of confidence interval coverage and width, Chen's increment was also evaluated.

On the other hand, Chen's increment to the ANCOHET standard error assumes that both the standardized regression slopes and covariate variability (i.e., β_j and σ_X^2, respectively) are known, population values. Due to this assumption, Chen's recommended increment in practice could also be an underestimation.

Standard Errors for Three-group Case

In the three-group case, attention was focused on a contrast between the predicted mean in one group and the average of the predicted means in the other two groups. That is, ψ was defined by the coefficients $c_1 = 1$, $c_2 = -.5$, and $c_3 = -.5$. Standard errors of the estimate of this contrast at point X_p were estimated as follows.

ANCOHET. In the case of ANCOHET, the estimated standard error of the contrast in predictions at X_p is:

$$ANCOHET: \hat{\sigma}_{\hat{\psi}_p} = \sqrt{\frac{E_F}{df_F}\left[\sum_j \frac{c_j^2}{n_j} + \sum_j \frac{c_j^2\left(X_p - \bar{X}_J\right)^2}{\sum_i\left(X_{ij} - \bar{X}_J\right)^2}\right]}$$

ANCOVA. In the conditions where the ANCOVA error term is used, the standard error of the contrast in the estimated group means is calculated as follows:

$$ANCOVA: \hat{\sigma}_{\hat{\psi}_p} = \sqrt{\frac{E_F}{df_F}\left[\sum_j \frac{c_j^2}{n_j} + \frac{\left(\sum_j c_j \bar{X}_J\right)^2}{\sum_j \sum_i\left(X_{ij} - \bar{X}_J\right)^2}\right]}$$

Interaction. For the interaction error term the standard error for the test of the contrast is calculated as:

$$Interaction: \hat{\sigma}_{\hat{\psi}_p} = \sqrt{MS_{AxX}\left[\sum_j \frac{c_j^2}{n_j}\right]}$$

Equal Weights. The standard error for the contrast in the equal weights or unweighted case is calculated as:

$$UNW: \hat{\sigma}_{\hat{\psi}_p} = \sqrt{MS_{UNW}\left[\sum_j \frac{c_j^2}{n_j}\right]}$$

Chen's Increment to ANCOHET Standard Error. The Chen increment to the square of the ANCOHET standard error of a contrast is:

$$Var\left[E\left(\hat{c}|\vec{X}\right)\right] = (c_1\beta_1 + c_2\beta_2 + c_3\beta_3)^2 \frac{\sigma_X^2}{(n_1 + n_2 + n_3)}$$

Along with sample size planning to ensure an a priori probability of rejecting the null hypothesis, it is also possible to plan a study that results in parameter estimates that are sufficiently accurate. While developing an AIPE framework in the context of ANCOHET is beyond the scope of this paper, accuracy of parameter estimates based on different error terms was incorporated. Specifically, confidence interval coverage and average confidence interval width are presented. In the case of unstandardized mean differences, which the current study dealt with, the width of the confidence interval does not depend on the mean difference (Maxwell et al., 2008). Thus, confidence interval coverage width will not be distinguished between the null and non-null conditions.

Average Standard Error compared to the True Standard Deviation

In addition to carrying out tests and constructing confidence intervals for each simulated data set, the estimated difference in predicted means across groups was retained for further analysis. Specifically, the standard deviation of the estimated differences in conditional means across the 10,000 replications for each cell in the simulation design was computed and interpreted as the "true standard deviation" of the sampling distribution of the estimated mean difference across groups which was being estimated by the various methods for computing the standard error of these differences (Muthén & Muthén, 2002). This permitted the average standard error corresponding to each of the denominator error terms to be compared to this true standard deviation, which would be the ideal value to use for testing the main effect of group as it accurately reflects the impact of both the random covariate and the heterogeneous regressions on the

distribution of the group differences across replications. Because the CI coverage rates and widths, which are also reported, are driven by the estimated standard error for a particular simulation and based on that simulation's unique distribution of covariate scores, those values can be thought of as applying to hypothetical replications where the same distribution of *X* values were obtained. On the other hand, since the current study investigated the impact of a random covariate in the presence of heterogeneity of regression, the goal is to make inferences to a broader range of *X* values. Comparing the average standard errors associated with different error terms to the true standard deviation is helpful in determining which error term comes closest to what might be considered ideal for use in testing the main effect of group.

Homogeneity of Variance Assumption

The importance of the homogeneity of variance assumption for ANOVA has been researched extensively (Glass et al., 1972; Sawilowsky & Blair, 1992; Scheffé, 1999). This same assumption also applies to ANCOVA. In the case of ANCOVA, however, it is no longer the variability of the *Y* scores that is required to be homogeneous. Instead, it is assumed that the variability of the residuals is the same across groups. When the assumption of homogeneity of regression lines between groups holds, the variability of both the residuals and the *Y* scores will be homogeneous.

When generating data according to the methods described above, it is only possible to ensure that either the variability of the outcome or the variability of the residuals will be homogeneous across groups – but it is impossible for both to be homogeneous. Preliminary work showed that when the variability of the original *Y* scores was homogeneous (and thus the variability of the residuals was heterogeneous), Type I

error rates were far greater than both the nominal .05 and the elevated rates reported in Harwell and Serlin (1988) when sample sizes were unequal between groups – as high as .19 (see Appendix D). As a result of these findings, coupled with the suggestions of Rogosa (1980, see p. 317, first full paragraph), data were generated in a way that produced homogeneous residuals, even though this resulted in heterogeneity of variance across groups on the original *Y* variable.

Justification for Levels of Simulation Factors

The decision to use the levels of the factors as described in the preceding Table 1 was based on two sources of information: previous simulation work in the area of analysis of covariance allowing for heterogeneous regressions (ANCOHET) and empirical findings involving ANCOHET, displayed in Tables 3 and 4, respectively. In order to assess the extent of empirical heterogeneity of regression, a reverse citation search was performed within several prominent psychological journals (*Developmental Psychology*, *Health Psychology* and *Journal of Personality and Social Psychology*) over the years 2015 – 2018 looking for articles that cited Aiken and West (1991). Although many of the 82 articles citing Aiken and West (1991) involved only interactions defined as the product of two continuous measures, thirteen articles, reporting on 16 different experiments involving a total of 19 analyses, were identified involving a categorical grouping variable and a covariate that presented evidence of heterogeneity of regression, and that provided enough information to allow a judgment about the magnitude of the heterogeneity of regression observed. Standardized regression coefficients for the covariate-dependent variable relationship within each group, or the standardized regression coefficient for the interaction between the covariate and the grouping variable,

were recorded for each experiment. These regression coefficients were then used to calculate Cohen's *q* (more info on this measure of effect below), which allowed the categorization of the extent of heterogeneity of regression for each analysis. These published results were combined with analyses from four other studies investigating heterogeneity of regression, for three of which complete data were available locally. This permitted the extent of heterogeneity of regression observed empirically in a total of 23 analyses to be categorized in Table 4 as follows: Small: 0 - .2; Medium: .2 - .4; Large: .4 - .6; and Extreme: > .6, with the lower limit of the interval being inclusive of the value. Table 4 also reports the number of groups and sample size for each study.

Number of Groups. Only one of the simulation studies (Klockars & Beretvas, 2001) reported in Table 3 and none of the empirical studies reported in Table 4 employed study designs utilizing more than two or three groups. As a result, the current study will also use only these two levels.

The decision of which pattern of group differences to test in the two-group conditions was straight-forward: with only two means there is only one difference that can be investigated, and the omnibus test will suffice. With three groups, on the other hand, a significant omnibus tests allows one to conclude that at least one adjusted mean is not equal to the others. As a result, it is unlikely that a researcher would be interested in only an omnibus significance test. With three groups, three pairwise comparisons and numerous complex contrasts can be tested. To mimic real-world scenarios, and to aid in the construction of a single confidence interval per simulation condition, the decision was made to test a complex comparison. As mentioned previously in the section on power,

this complex comparison and corresponding confidence interval compared the adjusted mean of a single group to the average of the remaining groups.

Sample Size per Group. Marszalek and colleagues (2011) found that the median group sample size for psychology studies published in 2006 ranged from 18 to 26 per group, depending on the area of psychological research (Abnormal $n = 26$, Applied $n = 21$, Developmental $n = 25$, Experimental $n = 18$). Sample sizes of 10 and 30 per group were clear favorites in the simulation studies reported. Given the overlap between the empirical and simulation studies, sample sizes of 10 and 30 were used in the current dissertation. However, given the "persistence of underpowered studies in psychological research" (see Cohen, 1962 and more recently Maxwell, 2004) a condition with group sizes of 100 was added to represent a sample size that would not result in a lack of power. Of the empirical articles examined, the smallest sample size examined was 32 per group, whereas the largest sample totaled over 2,500 participants.

Extent of Heterogeneity. The simulation studies reviewed employed numerous methods of depicting heterogeneity of regression in their research designs. Two studies used only one condition to represent heterogeneity of regression (Chen, 2006; Harwell & Serlin, 1988). One used a single mean correlation about which different levels of heterogeneity varied (Klockars & Beretvas, 2001). Finally, three studies used both different mean levels of correlation and amounts of heterogeneity (see Hamilton, 1977; Levy, 1980; Wu, 1984). Because varying both the mean correlation and extent of heterogeneity could quickly lead to an unnecessarily complicated simulation with potentially thousands of conditions, the current study used one level of mean correlation ($r = .3$), and heterogeneity was balanced around this value. The value of .3 was chosen

because this is roughly the average correlation between two variables in psychological research (Cohen, 1992). Based on examination of common differences between within-group correlations from simulation studies and those found in empirical studies, low, medium, and high levels of heterogeneity of regression were represented by differences in the within-group correlations of .1, .3, and .5, respectively. These values for the difference in the within-group correlations also closely align with what is considered a small, medium and large effect according to the effect size measure Cohen's q (Cohen, 1992). Cohen's q is calculated as the difference between two correlation coefficients after having performed Fisher's r-to-z-transformation as:

$$z_r = \frac{1}{2} log\left(\frac{1+r}{1-r}\right)$$

Cohen's q is then calculated as:

$$q = z_{r_1} - z_{r_2} = \frac{1}{2} log\left(\frac{1+r_1}{1-r_1}\right) - \frac{1}{2} log\left(\frac{1+r_2}{1-r_2}\right)$$

When correlations are centered around $r = .3$ and the differences in the two correlations are .1, .3 and .5, these translate to Cohen's q values of .11, .33, and .57, which are nearly equal to the cutoffs for what Cohen (1992) established as small, medium and large differences, respectively. Tables 5 and 6 provide the raw coefficients used to achieve the standardized coefficients listed for two- and three-group designs, respectively. Also included are two effect sizes for the difference between the regression coefficients: Cohen's q and f^2.

Given that so few of the empirical studies reviewed found within-group correlations with opposite signs (see observed heterogeneity in Table 4 for Blaire et al.,

2015; Lam et al., 2018; Rudolph, Davis, & Monti, 2017; Sturge-Apple et al., 2016), and often one of the opposite-signed regression coefficients was not significantly different from zero, it was decided that only one condition would include within-group correlations of opposite sign such as -.2 and +.8, and this would represent an extreme level of heterogeneity. A difference this large between two correlations would result in a Cohen's *q* or 1.30, or nearly two times greater than the cutoff for a large effect.

RESULTS

Rejection Rates for Two-Group, Null Conditions

The rejection rates for the omnibus test for the two-group designs where there was no difference between population means on the dependent variable at the grand mean on the covariate are presented in Table 7. The omnibus tests were performed at three separate locations (μ_x, $\bar{X}$, and the center of accuracy (C_a)) for three equal-*n* conditions using four separate error terms (see above for detailed description). Furthermore, these three factors were crossed with five levels of heterogeneity of regression: none, low, medium, high, and extreme. Rejection rates were considered outside the range specified by sampling error if they deviated by more than .0043 on either side of .05 (i.e., outside the interval [.0457, .0543]). Computation of the standard error of the rejection rate statistic is shown in Appendix D (Equation D.1).

In what follows, the error rates shown in Table 7 will be described separately for each error term.

Error Term: ANCOHET Error. For none, low, and medium levels of heterogeneity of regression, using the error term associated with the ANCOHET model produced rejection rates within the sampling error range of .05 regardless of the point on *X* at which the test was conducted. Moreover, when the test was being conducted at the population grand mean (μ_X), rejection rates were always within sampling error even under high and extreme levels of heterogeneity of regression. On the other hand, higher rejection rates were observed under the high and extreme heterogeneity conditions when the test was conducted at the sample mean ($\bar{X}$) and the center of accuracy (C_a). Under

extreme levels of heterogeneity of regression, rejection rates were over twice the nominal .05. However, as noted in the Introduction in the section on "Type I Error Rates in the Presence of Heterogeneity of Regression," at points other than μ_X there will be a true difference in expected conditional means across groups and so these should not be regarded as Type I errors but as detecting the small true treatment effect at those points.

Error Term: Interaction. When there was no heterogeneity of regression, using the mean square for interaction between the covariate and the grouping variable as an error term produced rejection rates within sampling error of the nominal level. This would be expected in that the mean square for interaction in such cases reflected only residual error variance. However, even beginning with low levels of heterogeneity, rejection rates began to drop significantly below .05. These rates approached zero as the level of heterogeneity of regression increased. This was also true for the effect of sample size within a single level of heterogeneity of regression: there was a negative relationship such that as sample size increased the rejection rates decreased. The same pattern of results was observed for tests conducted at all three locations.

Error Term: ANCOVA. The first alternative weighting scheme sought to combine the mean square error term from the ANCOHET model with the mean square for the interaction term – essentially a weighted average of the previous two error terms discussed. When testing the main effect of group at either $\bar{X}$ or C_a, this ANCOVA error term produced rejection rates within sampling error for all sample sizes and across all levels of heterogeneity of regression, with one exception. The one minor exception (i.e., $n = 10$, medium heterogeneity, tested at $\bar{X}$) where the rejection rate was outside of the interval around .05 was only .0002 below the cutoff of .0457. Given that this occurred in

1/30 or 3.33% of the conditions (tested at either $\bar{X}$ or C_a), it is possible that this result was due to chance. Changing the starting seed of the random number generator saw the rejection rate for this condition fall back within the range of sampling error at 5.03%, supporting the hypothesis that this aberrant finding was due to chance.

However, when the test occurred at μ_X, this error term produced rejection rates that were always significantly below .05 for the medium, high, and extreme heterogeneity of regression conditions.

Error Term: Unweighted. This error term very rarely produced rejection rates within sampling error of .05 for any of the conditions. In fact, of the 45 conditions where this error term was evaluated, only seven were neither liberal nor conservative.

Rejection Rates for Three-Group, Null Conditions

The rejection rates for the three-group conditions are presented in Table 8. As previously mentioned, rather than an omnibus test, the decision was made to test a contrast comparing the adjusted mean of one group to the average of the remaining two groups. Where heterogeneity of regression was present, the two adjusted means averaged together came from the groups with equal population regression slopes.

The pattern of results observed for the three-group simulations mirrored that of the two-group simulations almost exactly. Due to this similarity, the specifics of the findings will not be discussed in detail.

In summary, it was clear that using UNW as an error term would be ill-advised as it consistently produced rejection rates outside of the bounds of sampling error: it was

either too liberal for low and absent levels of heterogeneity of regression, or it was too conservative for moderate to extreme levels. Similarly, using the mean square for the interaction as an error term consistently produced rejection rates significantly below the nominal .05 even at the lowest levels of heterogeneity. On the other hand, the error terms based on an ANCOHET model (HET columns in Tables7 and 8) and an ANCOVA model (ANC columns in Tables 7 and 8) produced much more acceptable rejection rates across all samples sizes and when evaluated at both $\bar{X}$ and the center of accuracy. As mentioned previously, and will be discussed further below, the fact that the ANCOHET error term produced rejection rates significantly greater than .05 for tests at these points other than μ_X for high and extreme levels of heterogeneity of regression is not necessarily a problem. On the other hand, the fact that ANCOVA produced rejection rates significantly below .05 when tests were conducted at μ_X could be regarded as problematic.

Power Results for Two-Group Conditions

Power results for the two-group conditions are presented in Table 9. Deviating from the methods used in previous studies upon which the current research is based (i.e., Harwell & Serlin, 1988), power will be discussed even in cases where rejection rates were significantly greater than .05.

Error Term: ANCOHET Error. Regardless of sample size, extent of heterogeneity of regression, or the location at which the test was conducted, power based on analyses using the ANCOHET error term saw high levels of power. Despite the attempt to adjust effect size so that power would be approximately the same for each level of sample size, power increased slightly as sample size increased. Additionally, holding sample size

constant, for tests conducted at $\bar{X}$ or at C_a, power decreased slightly as heterogeneity of regression increased, with the only substantial decrease occurring in the jump from high to extreme levels. However, these decreases only saw a loss in power of approximately 4%.

Error Term: Interaction. As predicted, using the mean square from the interaction between the grouping variable and covariate produced extremely low levels of power. When there was no heterogeneity of regression, where power was the highest, none of the conditions had power levels over 20% (highest 18.21%). Interestingly, within each level of heterogeneity of regression, power decreased as sample size increased. Also, holding sample size constant, increasing heterogeneity was associated with substantial decreases in power. These results were consistent across all three locations at which the test was conducted.

Error Term: ANCOVA. When heterogeneity of regression was absent, low, medium or even high, using ANCOVA as an error term produced power levels that were high and similar to those seen in simulations using the ANCOHET error term. It was only for extreme levels of heterogeneity that power levels were lower (between 59-60%). Again, within each level of heterogeneity of regression, power increased as sample size increased.

Error Term: Unweighted. When heterogeneity of regression was absent or low, using UNW as an error term produced power that was slightly less than the ANCOHET and ANCOVA error terms though still relatively high. However, power dropped off steeply when heterogeneity was medium, high and extreme. Within each level of

heterogeneity of regression, power decreased as sample size increased. Also, holding sample size constant, increasing heterogeneity was associated with substantial decreases in power. These results were consistent between all three locations at which the test was conducted.

In summary, using the error terms from both the ANCOHET model and the ANCOVA model (ANC) produced consistently high levels of power. The only notable exception was for the ANCOVA error term under extreme levels of heterogeneity of regression where there was lower power. These findings were consistent regardless of where the test was conducted. The other two error terms produced power results that were considerably lower.

Power Results for Three-Group Conditions

As previously mentioned for the rejection rates, the pattern of results for the three-group power scenarios was nearly identical to that of the two group conditions. Those results are presented in Table 10. The only notable differences between the two- and three-group cases occurred when the mean square for the interaction was used as an error term. For these cases, power for the three-group simulations was nearly twice what it was for the two-group condition. This is likely due to the change in error degrees of freedom. In the three group case, df_{error} doubled from one to two, explaining this increase in power as a result of doubling the degrees of freedom. However, power was still low, with a maximum of just below 35% when no heterogeneity of regression was present, and quickly declined as the extent of heterogeneity increased.

Confidence Interval Coverage

Confidence interval coverage for the two- and three-group conditions is presented in Tables 11 and 12, respectively. Given the similarity in coverage rates, the results will be covered simultaneously. Across the board, the population parameter was captured within the confidence intervals a high proportion of the time (range: 88.0 - 100%), with a majority of the coverage rates falling above 93%. The location at which the test of between group differences was conducted did not impact the coverage rates. Under high and extreme levels of heterogeneity of regression, using mean square interaction or UNW as an error term produced confidence intervals that often had 100% coverage rates.

Even under high and extreme levels of heterogeneity of regression, the confidence interval coverage rates for the ANCOHET approach were between 94% and 96% for the two-group case and between 91% and 93% for the three-group case. This is in contrast to what might have been predicted based on Chen (2006)'s suggested increment to the ANCOHET standard error based on its purported under-estimation particularly at high levels of heterogeneity of regression. When this increment was implemented, as presented in Tables 13 and 14, coverage rates increased to 95% to 97% in the two-group scenario and between 91% to 95% in the three-group scenario. Thus, in the two-group case where the ANCOHET procedure was working very well, the Chen procedure inflated the coverage of 95%, and in the three-group case where the ANCOHET coverage was a little low, the only case in which the chen increment improved the coverage substantially was in the case of extreme heterogeneity of regression.

Confidence Interval Accuracy

Two Groups. Even though the confidence interval coverage rates were high across the board, it does not mean that each error term performs similarly in capturing the true population difference in adjusted means with equally narrow intervals. To evaluate this feature of the results, the average confidence interval width was also calculated, and the results are presented in Tables 15 and 16 for the two- and three-group cases, respectively. The impact of sample size within levels of heterogeneity was consistent and unsurprising. When group sizes increased, the accuracy of the confidence intervals increased. It was only under extreme levels of heterogeneity of regression and only when using mean square interaction as the error term that the impact of increasing sample size reversed: increasing sample sizes then was associated with wider confidence intervals.

Both the ANCOHET error term and ANCOVA produced similar confidence interval accuracy for all combinations of test location and of sample size, and for most levels of heterogeneity of regression. The only situation where the accuracy deviated slightly was for extreme heterogeneity, where ANCOVA produced confidence intervals that were 1.25 times as large. It makes sense that CI widths would be smaller for ANCOHET than ANCOVA at high levels of heterogeneity of regression since the mean square error from the ANCOHET approach is smaller than the mean square error based on ANCOVA as a result of allowing for heterogeneity of regression. Tables 17 and 18 present additional CI widths for the two and three group conditions, respectively, where the Chen (2006) increment to the ANCOHET standard errors was employed. The Chen adjustment produced confidence interval widths that fell in between the widths based on

the ANCOHET and ANCOVA error terms under extreme levels of heterogeneity of regression.

For conditions where heterogeneity of regression was either low or absent, confidence intervals based on the UNW error term had similar widths to both ANCOHET and ANCOVA. This changed for medium, high, and extreme levels of heterogeneity of regression where the intervals based on UNW were 1.1 to 7.7 times as large.

The largest widths were seen for confidence intervals constructed using the mean square interaction. As the level of heterogeneity of regression increased, confidence intervals became wider under this method. When heterogeneity of regression was absent, confidence intervals using the interaction resulted in widths that were 4.9 times as large as the ANCOHET and UNW methods when $n = 10$. This difference in accuracy reached its largest level under extreme heterogeneity of regression when n = 100, where the widths were 69.1 times as large.

Three Groups. The pattern of confidence interval accuracy was very similar for the two- and three-group cases. The main difference is that the three-group conditions were more accurate, due in large part to the addition of an extra group increasing the total sample size. The largest increase in accuracy was seen for confidence intervals constructed using the mean square interaction as an error term. These confidence interval widths were anywhere from 33% to 25% as large in the three-group compared to the two-group condition. However, they were still anywhere from 1.9 to 19.4 times as large as the confidence intervals constructed by either the ANCOHET or the ANCOVA methods.

In summary, both the ANCOHET and ANCOVA error terms produced confidence intervals with the highest accuracy (i.e., smallest widths). This was consistent regardless of the location at which the test was conducted, sample size, and extent of heterogeneity of regression. While the UNW confidence intervals produced similar levels of accuracy to the other two methods under low levels of heterogeneity, they quickly became less accurate as heterogeneity of regression increased. By far, using the mean square interaction as an error term produced the least accurate confidence intervals.

Average Standard Error Compared to True Standard Deviation

Two Groups. Tables 19, 20 and 21 present comparisons in the two-group scenarios of the average standard error for each denominator error term to the true standard deviation for tests conducted at $\bar{X}$, C_a, and μ_X, respectively. As mentioned previously, the true standard deviation is the standard deviation of the difference in estimated conditional means across the 10,000 simulations. Though not presented in any tables, the average of these parameter estimates was essentially zero for all combinations of heterogeneity of regression and sample size.

For tests conducted at either $\bar{X}$, or C_a, the average standard errors for both ANCOHET and ANCOVA were nearly identical to the true standard deviation when the heterogeneity of regression was medium, low, or absent. Specifically, expressing the average standard error as a percentage of the true standard error, the mean percentage across conditions for tests at $\bar{X}$ was 99.1% for ANCOHET and 99.5% for ANCOVA, and for tests at C_a the mean percentage was 98.5% for ANCOHET and 99.1% for ANCOVA. When heterogeneity was high or extreme, the average standard error from an ANCOVA

error term was a close estimate of the true standard deviation (mean percentages were 99.3% at $\bar{X}$ and 99.0% at C_a). On the other hand, the ANCOHET error term underestimated the true standard deviation by 4% to 6% when heterogeneity of regression was high, and this number increased to over 20% when heterogeneity was extreme. The Chen increment to the ANCOHET standard error was not enough of an adjustment to improve the average standard errors, as it underestimated the true standard error by 11% to 13% when heterogeneity of regression was extreme.

Alternatively, when the test was conducted at μ_X, as shown in Table 21, the average standard errors from the ANCOHET approach were close the true standard deviations, even at high levels of heterogeneity of regression. Specifically, ANCOHET standard errors averaged 97.5% of true standard deviations for no, low and medium levels of heterogeneity, and averaged 97.2% of the true standard deviations for high and extreme levels of heterogeneity. Conversely, the ANCOVA standard error was an overestimation of the true standard deviation when heterogeneity was extreme, with the ANCOVA average standard error being 118% to 124% of the true standard deviations.

Neither the average standard errors based on using the interaction as an error term or using the UNW provided good estimates of the true standard deviation, regardless of where the test was conducted. When heterogeneity of regression was absent, the true standard deviation was underestimated regardless of sample size. When heterogeneity of regression was high or extreme, the average standard errors from both of these errors term overestimated the true standard deviation.

Three Groups. Tables 22, 23 and 24 present the three-group comparisons of the average standard error for each denominator error term to the true standard deviation for tests conducted at $\bar{X}$, C_a, and μ_X, respectively. These results largely mirror what was found for the two-group scenario. That is, for tests conducted at either $\bar{X}$, or C_a, the average standard errors for both ANCOHET and ANCOVA were again nearly identical to the true standard deviation (i.e., the average standard errors were 98% to 99% of the true standard deviation) when the heterogeneity of regression was medium, low, or absent. But when the heterogeneity of regression was high or extreme, the average standard error from an ANCOVA error term was still a close estimate of the true standard deviation (98% to 99% of the true value) whereas the ANCOHET underestimated the true standard deviation particularly when heterogeneity was extreme. For tests at μ_X, as had been seen in the two-group case, the ANCOHET average standard error was again very accurate for all levels of heterogeneity, whereas the ANCOVA average standard error badly overestimated error when heterogeneity was extreme.

DISCUSSION

As mentioned in the introduction, ANCOVA was first introduced by Ronald Fisher in the early 20th century in the field of agriculture as a method of controlling for variables that were not part of the experimental design, but were nonetheless expected to be related to the outcome of interest (Eden & Fisher, 1927; Fisher, 1932). One of this method's main benefits is to increase the precision of the treatment effect estimate, also increasing power to detect such an effect, by accounting for these individual differences. In the area of psychological research, where individual differences often account substantially more of the variability in outcomes than between treatment differences, ANCOVA provides an often needed boost to power and precision and has been widely adopted.

Early on, some researchers stressed the importance of assuming that between-group regression slopes were homogenous, with some going so far as referring to it as "this key assumption" (Kirk, 1995, p. 724). Others have gone so far as to imply that ANCOVA should be abandoned when heterogeneity of regression exists (Keppel, 1973, p. 484, 499). Fortunately, approaches accommodating heterogeneous slopes were developed. The Johnson-Neyman technique (D'Alonzo, 2004; Johnson & Neyman, 1936) provided researchers with guidance on how to determine "regions of significance" specifying where on the *X* continuum there were significant differences between groups, but this procedure is computationally tedious and not widely implemented in standard statistical software. Rogosa's work in the area of ANCOHET, particularly his development of the "pick-a-point" procedure, made the problems previously inherent to heterogeneity of regression seem more approachable, particularly with more recent

guidance on where to conduct the test in lieu of a priori meaningful values (Aiken & West, 1991). Due to the increasing acknowledgement of its importance in research across disciplines, multiple online utilities have been developed to implement the Johnson-Neyman technique and the pick-a-point procedure (e.g., Hayes & Matthes, 2009; Preacher et al., 2006).

Of direct relevance to the current dissertation, Rogosa (1980) noted that previous simulation studies investigating heterogeneity of regression in the context of ANCOVA were flawed. In particular, he noted that some of the simulations violated the assumption of equal residual variances. In the context of a randomized study, where s_X^2 will be equal in the long run between groups, whenever the interaction term is nonzero, if the variability of the *Y* scores is homogeneous, then heterogeneity of regression will mean the residual variances will necessarily be unequal between groups. Additionally, he highlighted the need for an explicit definition of a treatment effect.

Whereas the work of Harwell and Serlin (1988) made improvements over prior simulation studies, their work had several drawbacks that the current study sought to ameliorate. As covered more extensively in Appendix D, they stated in their Method section that standardized regression coefficients were employed when simulating data. However, based on preliminary work for the current study, it appears as if they actually used unstandardized coefficients, because using standardized coefficients with heterogeneity of regression as extreme as they reportedly used ($\beta_1 = 0.2$, $\beta_2 = 0.9$, $\beta_3 = 0.9$) would have resulted in significantly higher rejection rates. This point is important, because using standardized regression coefficients allows other researchers to compare the heterogeneity of regression they are experiencing in a way that does not depend on

the scale of the variables they are investigating. The current study uses standardized regression coefficients in denoting the different conditions of heterogeneity of regression and makes the contribution of characterizing the magnitude of heterogeneity of regression using multiple effect sizes measure commonly used (e.g., Cohen's q, f^2; see Table 5).

Additionally, previous simulation work in the area had a limited view of Type I error rates in the presence of heterogeneity of regression. In null conditions where rejection rates were above the bounds of sampling error, Harwell and Serlin (1988) did not report power levels stating that "liberal Type I error rates render the interpretation of power values problematic" (p. 277). Chen (2006) took the same approach, as evident in his Table 1 Scenario II. However, it can be argued that these are not actually Type I error rates in that, given the presence of heterogeneity of regression, there is some true effect at every point along the X scale except where the lines cross, which in the current simulations was at μ_X. Because tests are typically conducted at a central tendency value based on the sample, and this value rarely will be exactly the population value, one would expect higher rejections rates than the nominal .05 due to the presence of a quite small but true effect. To address this issue, in addition to reporting power for non-null conditions regardless or rejection rates for the null conditions, the current study focused on confidence interval coverage rates and widths as a way of comparing error terms.

The main impetus motivating the current study is the limited understanding of the impact of a random covariate in the context of ANCOVA with heterogeneity of regression. The main idea is exemplified by Figure 1, where the difference in adjusted means in a thrice replicated study is impacted by the obtained sample mean on the

covariate. Crager (1987) addressed this in the context of a standard ANCOVA, and concluded that a random covariate will have little impact. On the other hand, other authors based on their simulations recommended different approaches than using the ANCOHET model to test for the treatment effect (Chen, 2006; Harwell & Serlin, 1988). As argued in Appendix B, theory suggests Rogosa's "safer ANCOVA" should be liberal, as confirmed by Harwell and Serlin's results, when dealing with heterogeneity of regression and a random covariate, yet it is still unclear how the standard ANCOVA and ANCOHET would fare with varying degrees of heterogeneity of regression, setting the stage for the current dissertation.

The current study makes the contribution of examining the impact of two separate factors. The first was to clarify how the decision of where to conduct the test impacts rejection rates and power. For the location of the test, the options were: μ_x, which is typically unknown; $\bar{X}$, the value most likely to be used; and C_a, where the distance between regression lines is the same for ANCOVA and ANCOHET. The second contribution involved determining the optimal error term to use. The first error term was from an ANCOHET model where heterogeneous regression slopes were allowed. The second error term was from a standard ANCOVA model that restricted the interaction to be zero. The third error term took the traditional mixed models ANOVA approach and used the interaction between the covariate and the grouping variable. Finally, the fourth error term averaged the mean square from the ANCOHET model and the interaction. This fourth error term was included because it involved a term that was thought to be too liberal (ANCOHET) with one that was thought to be too conservative (interaction).

Study Findings

The goal of this project was to examine the performance of several different error terms in assessing between-group differences in ANCOVA models where heterogeneity of regression was present. The first error term, derived from a model allowing for heterogeneity of regression, performed quite well across all levels of the factors of the simulation. The rejection rates for the null conditions were as expected for no, low, and medium levels of heterogeneity of regression, and only exceeded the nominal .05 in cases of high and extreme heterogeneity of regression when tests were conducted at $\bar{X}$ or at C_a. Using the ANCOHET error term also saw high power and confidence interval coverage rates even in the presence of extreme heterogeneity of regression. Regarding the accuracy of how well the parameter of interest was estimated (i.e., mean difference), the ANCOHET error term resulted in the narrowest confidence intervals under extreme heterogeneity of regression, and was similar to the ANCOVA error term under the other levels. When heterogeneity of regression was absent, low, or medium, the average of the ANCOHET standard errors was 99% of the true standard deviation based on the standard deviation of estimated conditional means across the 10,000 replications. When heterogeneity of regression as high or extreme, however, the ANCOHET standard error was only 87% of the true standard deviation, and this value dropped below 80% when heterogeneity was extreme, implying that the small confidence interval widths are too small. While these results validated to some extent Chen's (2006) claim that the conventional ANCOHET error term would underestimate the impact of the random covariate, Chen's suggested increment to the ANCOHET standard error made little difference under lower levels of heterogeneity of regression, and did not bring the

average standard error close enough to the true standard deviation when heterogeneity of regression was extreme. Even though the confidence interval coverage rate was high and the interval widths were narrow, comparing the average standard error to the true standard deviation suggests that the standard error based on the ANCOHET approach is an underestimation of the variability over replications in the estimated treatment effect at either $\bar{X}$, or C_a , particularly when heterogeneity of regression is extreme.

The second error term, involving the interaction between the grouping variable and the covariate saw rejection rates within sampling error of the nominal .05 only in cases of no heterogeneity of regression (the only situation where errors are true Type I errors), and became very conservative when any level of heterogeneity of regression was present and increasingly so as sample size increased making it more likely this error term would be inflated over residual error by the presence of the interaction in the population. Power was low when this error term was used, which was unsurprising given the degrees of freedom for the error term was either one or two depending on the number of groups involved in the analysis. Confidence interval coverage rates of the true population difference were high, but this was mainly due to the fact that the width of the confidence intervals was so large.

The third error term, derived from a standard ANCOVA model, performed similarly to the ANCOHET error term. Its use resulted in rejection rates within sampling error of .05 regardless of the extent of heterogeneity of regression under the conditions of no group difference when tests were conducted at $\bar{X}$ or at C_a. In these cases the overestimation of residual error was offset by the difference in predicted means not being exactly zero. However, when the test was conducted at μ_x the ANCOVA procedure

resulted in rejection rates that were significantly below the desired .05 level for all cases where there was any heterogeneity of regression, except in the case of low heterogeneity of regression combined with the smallest sample size. Using this error term resulted in high levels of power, declining only when heterogeneity of regression was extreme. This decline is to be expected inasmuch as using a single common slope will necessarily produce an overestimate of residual error when extreme heterogeneity of regression is present. Additionally, high confidence interval coverage rates were coupled with high accuracy (i.e., narrow confidence interval widths) generally, with widths noticeably exceeding those of the ANCOHET only in the case of extreme heterogeneity where ANCOHET underestimates the true standard deviation. Use of the ANCOVA error term resulted in average standard errors that were 99% of the true standard deviation even at extreme levels of heterogeneity of regression.

The final error term, computed as an average of the mean square error from the ANCOHET model and the mean square due to the interaction, performed poorly in nearly all conditions. It was either too liberal or too conservative depending on the extent of heterogeneity of regression, and it produced low power for even medium levels of heterogeneity. While confidence interval coverage rates were high, the accuracy of the parameter estimates was low

Recommendations for Dealing with Heterogeneity of Regression

Typically, when analyzing data using a two-way ANOVA, even if the interaction between factors is non-significant, it is often left in the model regardless. One potential argument for this analytic method is that just because the interaction is non-significant in the sample does not mean that it is null in the population. Additionally, power analyses

typically focus on main effects, and interaction tests are generally underpowered. The same argument could be made for analyzing ANCOVA data, regardless of whether the interaction term is significant. When testing for an interaction in the context of ANCOVA, one suggestion is to use a larger value of α than one typically uses (e.g., α= .10 or .25) in order to avoid a Type II error (Kirk, 1995). This makes sense given a typical study's propensity to be underpowered when testing such effects.

While incorporating this interaction into an analysis has the impact of removing degrees of freedom from the error term, its impact is likely to be negligible. Since a continuous covariate only accounts for one degree of freedom, its interaction with a grouping variable will only remove from the error term the total number of groups minus one (i.e., $a - 1$ where a is the number of groups). Even in the current student's smallest sample, where $a = 2$ and $n = 10$, the difference in $F_{critical}$ between the ANCOVA and ANCOHET models is 0.04. The impact of this loss of a degree of freedom on power will only diminish as a study's sample size increases.

However, the literature review of empirical studies suggested that the vast majority of studies finding heterogeneity of regression report either a small or medium level of heterogeneity of regression. In such cases, the current dissertation indicates that using an ANCOHET model that allows for heterogeneity of regression to conduct a test of the treatment effect near the center of the distribution of covariate scores could be used. At these levels of heterogeneity of regression, with nominal 95% confidence intervals, the ANCOHET approach achieves coverage of approximately 95% in the two-group case and between 91% to 93% in the three-group case. Additionally, the average standard errors are 99% of the true standard deviation of the estimated difference in

conditional means across replications. On the other hand, when heterogeneity of regression is large or extreme, using an error term based on ANCOVA would be the recommended approach. In these cases, the degree of overestimation of residual error associated with using a single slope approximately matched the additional variability induced by the random covariate and consequently resulted in average standard errors that were quite close to the true standard deviation. Although differences between the ANCOHET and ANCOVA methods for assessing treatment effects for an "average" individual might be regarded for some practical purposes as inconsequential, the ANCOHET procedure is recommended for general use based on levels of heterogeneity of regression that are most likely to be encountered in practice, that is, where the extent of heterogeneity of regression corresponds to a medium effect size or less. In such cases, ANCOHET's standard error approximates well the true standard deviation of estimated differences, and in addition achieved greater power and narrower confidence intervals than ANCOVA in general. In the rare case of high or extreme heterogeneity of regression, using ANCOVA to test for the main effect of treatment is recommended as its overestimation of residual error was in the current simulations demonstrated to be approximately the correct adjustment needed to match the increased true standard deviation in treatment effects over replications resulting from the presence of a random covariate.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Sage Publications.

Blair, C., Ursache, A., Mills-Koonce, R., Stifter, C., Voegtline, K., & Granger, D. A. (2015). Emotional reactivity and parenting sensitivity interact to predict cortisol output in toddlers. *Developmental Psychology*, *51*(9), 1271–1277. https://doi.org/10.1037/dev0000031

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Chen, X. (2006). The adjustment of random baseline measurements in treatment effect estimation. *Journal of Statistical Planning and Inference*, *136*(12), 4161–4175. https://doi.org/10.1016/j.jspi.2005.08.046

Cheval, B., Sarrazin, P., Isoard-Gautheur, S., Radel, R., & Friese, M. (2015). Reflective and impulsive processes explain (in)effectiveness of messages promoting physical activity: A randomized controlled trial. *Health Psychology*, *34*(1), 10–19. https://doi.org/10.1037/hea0000102

Cochran, W. G. (1952). The $\chi 2$ test of goodness of fit. *The Annals of Mathematical Statistics*, *23*(3), 315–345. https://doi.org/10.1214/aoms/1177729380

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.) Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Compas, B. E., Bemis, H., Gerhardt, C. A., Dunn, M. J., Rodriguez, E. M., Desjardins, L., … Vannatta, K. (2015). Mothers and fathers coping with their children's cancer: Individual and interpersonal processes. *Health Psychology*, *34*(8), 783–793. https://doi.org/10.1037/hea0000202

Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics*, *43*(4), 895–901. https://doi.org/10.2307/2531543

Crotwell, S. (2016). Incorporating alternative sources of reinforcement through online CRA goal setting and the effect on substance use in college students. *Psychology ETDs*. Retrieved from http://digitalrepository.unm.edu/psy_etds/31

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

D'Alonzo, K. T. (2004). The Johnson-Neyman procedure as an alternative to ANCOVA. *Western Journal of Nursing Research*, *26*(7), 804–812. https://doi.org/10.1177/0193945904266733

Eden, T., & Fisher, R. A. (1927). Studies in crop variation: IV. The experimental determination of the value of top dressings with cereals. *The Journal of*

Agricultural Science, *17*(4), 548–562. https://doi.org/10.1017/S0021859600018827

Fisher, R. A. S. (1932). *Statistical methods for research workers* (Fourth ed.-revised and enlarged). Edinburgh Oliver & Boyd. Retrieved from http://trove.nla.gov.au/version/50609123

Friesen, J. P., Campbell, T. H., & Kay, A. C. (2015). The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of Personality and Social Psychology*, *108*(3), 515–529. https://doi.org/10.1037/pspp0000018

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.2307/1169991

Hamilton, B. L. (1977). An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, *37*(3), 701–712. https://doi.org/10.1177/001316447703700313

Harwell, M. R., & Serlin, R. C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin*, *104*(2), 268–281. https://doi.org/10.1037/0033-2909.104.2.268

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*(3), 924–936. https://doi.org/10.3758/BRM.41.3.924

Henderson, C. R. (1982). Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. *Biometrics*, *38*(3), 623–640. https://doi.org/10.2307/2530044

Ho, A. K., Kteily, N. S., & Chen, J. M. (2017). "You're one of us": Black Americans' use of hypodescent and its association with egalitarianism. *Journal of Personality and Social Psychology*, *113*(5), 753–768. https://doi.org/10.1037/pspi0000107

Hostinar, C. E., Johnson, A. E., & Gunnar, M. R. (2015). Early social deprivation and the social buffering of cortisol stress responses in late childhood: An experimental study. *Developmental Psychology*, *51*(11), 1597–1608. https://doi.org/10.1037/dev0000029

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. John Wiley & Sons.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions*, *26*(3), 258–287. https://doi.org/10.1177/0163278703255242

Keppel, G. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, N.J.: Prentice-Hall.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole.

Klockars, A. J., & Beretvas, S. N. (2001). Analysis of covariance and randomized block design with heterogeneous slopes. *The Journal of Experimental Education*, *69*(4), 393–410. https://doi.org/10.1080/00220970109599494

Lam, P. H., Levine, C. S., Chiang, J. J., Shalowitz, M. U., Story, R. E., Hayen, R., … Chen, E. (2018). Family obligations and asthma in youth: The moderating role of socioeconomic status. *Health Psychology*, *37*(10), 968–978. https://doi.org/10.1037/hea0000655

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology*, *65*(2), 350–370. https://doi.org/10.1111/j.2044-8317.2011.02029.x

Levy, K. J. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, *40*(4), 835–840. https://doi.org/10.1177/001316448004000404

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Matos, M., Bernardes, S. F., Goubert, L., & Beyers, W. (2017). Buffer or amplifier? Longitudinal effects of social support for functional autonomy/dependence on older adults' chronic pain experiences. *Health Psychology*, *36*(12), 1195–1206. https://doi.org/10.1037/hea0000512

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin*, *95*(1), 136–147. https://doi.org/10.1037/0033-2909.95.1.136

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed). New York: Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*(4), 437–448. https://doi.org/10.3102/10769986031004437

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(4), 717–731.

Reid, A. E., Taber, J. M., Ferrer, R. A., Biesecker, B. B., Lewis, K. L., Biesecker, L. G., & Klein, W. M. P. (2018). Associations of perceived norms with intentions to learn genomic sequencing results: Roles for attitudes and ambivalence. *Health Psychology*, *37*(6), 553–561. https://doi.org/10.1037/hea0000579

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321. https://doi.org/10.1037/0033-2909.88.2.307

Rudolph, K. D., Davis, M. M., & Monti, J. D. (2017). Cognition–emotion interaction as a predictor of adolescent depressive symptoms. *Developmental Psychology*, *53*(12), 2377–2383. https://doi.org/10.1037/dev0000397

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*(2), 352–360. https://doi.org/10.1037/0033-2909.111.2.352

Scheffé, H. (1999). *The analysis of variance*. New York: Wiley-Interscience.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. https://doi.org/10.1037/1082-989X.1.2.115

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, *5*(2), 230–240.

Simons, R. L., Lei, M.-K., Beach, S. R. H., Barr, A. B., Simons, L. G., Gibbons, F. X., & Philibert, R. A. (2018). Discrimination, segregation, and chronic inflammation: Testing the weathering explanation for the poor health of Black Americans. *Developmental Psychology*, *54*(10), 1993–2006. https://doi.org/10.1037/dev0000511

Song, R., Over, H., & Carpenter, M. (2015). Children draw more affiliative pictures following priming with third-party ostracism. *Developmental Psychology*, *51*(6), 831–840. https://doi.org/10.1037/a0039176

Stock, M. L., Gibbons, F. X., Beekman, J. B., & Gerrard, M. (2015). It only takes once: The absent-exempt heuristic and reactions to comparison-based sexual risk information. *Journal of Personality and Social Psychology*, *109*(1), 35–52. https://doi.org/10.1037/a0039277

Sturge-Apple, M. L., Suor, J. H., Davies, P. T., Cicchetti, D., Skibo, M. A., & Rogosch, F. A. (2016). Vagal tone and children's delay of gratification: Differential sensitivity in resource-poor and resource-rich environments. *Psychological Science*, *27*(6), 885–893. https://doi.org/10.1177/0956797616640269

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/pits.20234

Trautwein, U., Lüdtke, O., Nagy, N., Lenski, A., Niggli, A., & Schnyder, I. (2015). Using individual interest and conscientiousness to predict academic effort: Additive, synergistic, or compensatory effects? *Journal of Personality and Social Psychology*, *109*(1), 142–162. https://doi.org/10.1037/pspp0000034

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Winkens, B., van Breukelen, G. J. P., Schouten, H. J. A., & Berger, M. P. F. (2007). Randomized clinical trials with a pre- and a post-treatment measurement: Repeated measures versus ANCOVA models. *Contemporary Clinical Trials*, *28*(6), 713–719. https://doi.org/10.1016/j.cct.2007.04.002

Wu, Y.-W. B. (1984). The effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance. *Educational and Psychological Measurement*, *44*(3), 647–663. https://doi.org/10.1177/0013164484443011

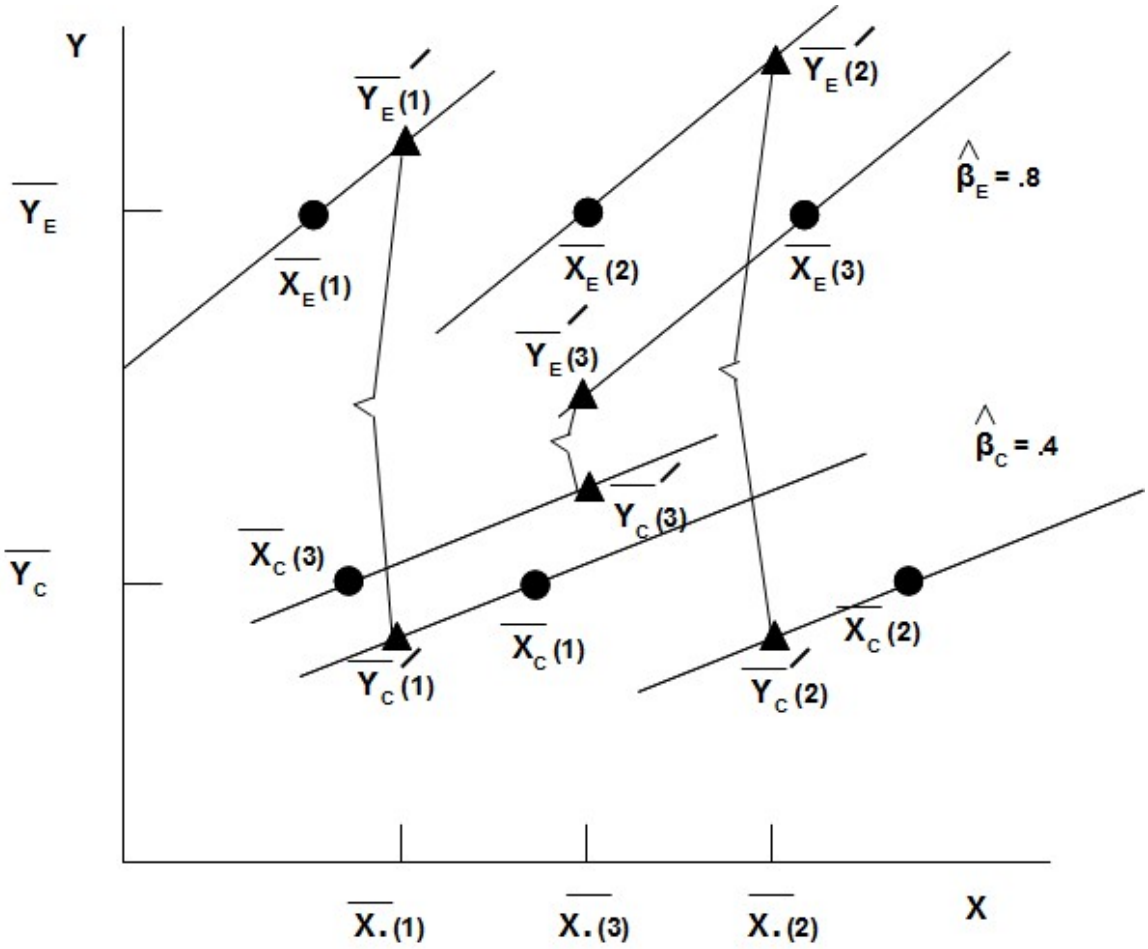

Figure 1. Reproduces Fig 3.1 from Maxwell et al. (1993). This figure illustrates the impact of having a random covariate in the presence of heterogeneity of regression across three replications (indicated by the number in parentheses). Each replication has identical slopes for the experimental condition, the control condition, and also unadjusted group means on *Y*. It is therefore the variability in the group means on *X* that results in different estimates of the adjusted treatment effect.

Table 1.
Simulation Design

Factors	Levels
Number of groups	2, 3
Sample Size Per Group	10, 30, 100
Extent of Heterogeneity of Regression	- None - Low (r = 25, .35) - Medium (r = .15, .45) - High (r = .05, .55) - Extreme (r = -.20, .80)
Effect Size	Null, Non-Null[a]
Type of Test for Group Main Effect	- At Population Grand Mean - At Sample Grand Mean - At Center of Accuracy

[a] See section regarding power for non-null simulation conditions

Table 2.
Effect Sizes and Non-Zero Means Used in Non-Null Simulation Conditions

Two-Group Case				Three-Group Case			
		Effect Size				Effect Size	
n	Mean	*f*	*d*	*n*	Mean	*f*	*d*
10	1.33473	0.66736	1.27324	10	1.12725	0.53139	1.07532
30	0.73607	0.36804	0.70216	30	0.63356	0.29866	0.60437
100	0.39816	0.19908	0.37982	100	0.34424	0.16228	0.32838

Table 3.
Simulation Studies Investigating ANOVA Models with Heterogeneity of Regression

Article	Heterogeneity	Homogeneity	Groups	Sample Size	Unequal *n*	Outcome
Harwell and Serlin (1988)	.2, .9	Yes	3	10, 30	Yes	Power, Type I Error
Levy (1980)	.3, .7; .1, .7; .0, .8; -.1, .9; -.3, .9	Yes	2	10, 20, 30	Yes	Power, Type I Error
Hamilton (1977)	Mean Slope: .3 (.2, .4; .1, .5; .0, .6; -.1, .7; -.2, .8; -.3, .9) Mean Slope: .4 (.3, .5; .2, .6; .1, .7; .0, .8; -.1, .9) Mean Slope: .5 (.4, .6; .3, .7; .2, .8; .1, .9) Mean Slope: .6 (.5, .7; .4, .8; .3, .9) Mean Slope: .7 (.6, .8; .5, .9)	Yes	2	10, 20, 30	Yes	Power, Type I Error
Klockars and Beretvas (2001)	Amount of heterogeneity manipulated so that the difference in slopes would be detected by the ANCOVA tests of slopes 20, 50 or 75% of the time. Mean slope was always 1. Ex. .04, 1, 1.96; -.65, 1, 2.65; -1.25, 1, 3.25	Yes	3, 5	12, 36	No	Power, Type I Error
Wu (1984)	.2, .6; .0, .8; .3, .7; .1, .9; .4, .6; .3, .9; .5, .9; .3, .5; .5, .7; .6, .8	Yes	2	10, 20, 30	Yes	Power, Type I Error
Chen (2006)	.2, .8	Yes	2	Equal n: 10,20, 50 Unequal n: 20, 10; 50, 25; 50, 40	Yes	Power, Type I Error

Table 4.
Empirical Findings of Heterogeneity of Regression

Study	Groups	Sample Size	Observed Heterogeneity r_j or β_j	*Interaction* β_j	Cohen's q	Heterogeneity of Regression Effect Size Category
Winkins et al. (2007)	3	210, 215, 220	0.77, 0.68, 0.51		0.36	Medium
Blaire et al. (2015)	2	963	0.21, -0.06		0.27	Medium
Cheval et al. (2015)	2	41, 41		-0.156	0.17	Small
Hostinar, Johnson and Gunnar (2015)	2	41, 41	-0.17, -0.07		0.10	Small
Friesen et al. (2015) Study 1	2	103		0.12	0.13	Small
Friesen et al. (2015) Study 2	2	179		0.24	0.27	Medium
Song, Over and Carpenter (2015)	2	32, 32	0.33, 0.09		0.25	Medium
Stock et al. (2015) Study 1	2	85, 88		-0.28; 0.26	0.31; 0.29	Medium; Medium
Stock et al. (2015) Study 2	2	111, 111		-0.21; 0.30	0.23; 0.33	Medium; Medium
Trautwein et al. (2015) Study 1	2	2,557		-0.22	0.24	Medium
Trautwein et al. (2015) Study 2	2	415		-0.20	0.22	Medium
Crotwell (2016)	2	79, 89	0.557, 0.424		0.18	Small
Ho, Kteily and Chen (2017)	2	424		0.28	0.31	Medium
Matos et al. (2017)	2	170		-0.101	0.11	Small
Rudolph, Davis and Monti (2017)	3	338	0.12, -0.02, -0.14		0.20	Medium
Lam et al. (2018)	2	172	0.26, -0.08; 0.11, -0.19		0.35; 0.30	Medium; Medium
Simulation of Sturge-Apple et al. (2016) by Maxwell et al. (2018)	2	69, 71	0.255, -0.235		0.50	Large
Rosenthal's Pygmalion data re-analyzed by Maxwell et al. (2018)	2	64, 246	0.781, 0.750		0.08	Small

Reid et al. (2018)	2	372	0.43, 0.07	0.39	Medium
Simons et al. (2018)	2	409	0.137, 0.027	0.11	Small

Note. Extent of heterogeneity of regression based on Cohen's *q* was defined as follows: Small: 0 - .2; Medium: .2 - .4; Large: .4 - .6; Extreme: > .6, with the lower limit of the intervals being inclusive of the value.
Studies with multiple groups but only one sample size did not involve randomization to treatments, but created groups based on either median splits or evaluating at the mean and/or the mean ± 1 standard deviation.
Multiple examples of heterogeneity of regression and the corresponding effect size measure within one study are separated by a semicolon. If an article included multiple experiments, each is presented on a separate row of the table.

Table 5.
Two-Group Simulation Standardized and Raw Regression Coefficients and Associated Effect Sizes

Standardized Coefficients		Raw Coefficients		q	f^2
β_1	β_2	b1	b2		
0.250	0.350	0.258	0.374	0.110	0.003
0.150	0.450	0.152	0.504	0.334	0.031
0.050	0.550	0.050	0.659	0.568	0.093
-0.200	0.800	-0.204	1.333	1.301	0.642

Table 6.
Three-Group Simulation Standardized and Raw Regression Coefficients and Associated Effect Sizes

Standardized Coefficients			Raw Coefficients			q	f^2
β_1	β_2	β_3	b1	b2	b3		
0.250	0.350	0.350	0.258	0.374	0.374	0.110	0.003
0.150	0.450	0.450	0.152	0.504	0.504	0.334	0.028
0.050	0.550	0.550	0.050	0.659	0.659	0.568	0.082
-0.200	0.800	0.800	-0.204	1.333	1.333	1.301	0.525

Table 7.
Rejection Rates for Two-Group, Null Conditions

		Location of Test											
		μ_X				$\bar{X}$				C_a			
Extent of Heterogeneity of Regression	n	Denominator Error Term				Denominator Error Term				Denominator Error Term			
		HET	Inter	ANC	UNW	HET	Inter	ANC	UNW	HET	Inter	ANC	UNW
None	10	.0483	.0502	.0471	.0542	.0475	.0520	.0478	.0589*	.0471	.0523	.0480	.0585*
	30	.0481	.0492	.0483	.0802*	.0491	.0495	.0491	.0814*	.0491	.0495	.0493	.0810*
	100	.0514	.0512	.0516	.0852*	.0510	.0511	.0509	.0850*	.0510	.0512	.0507	.0848*
Low	10	.0483	.0459	.0475	.0542	.0471	.0468	.0475	.0580*	.0465	.0469	.0477	.0583*
	30	.0481	.0452†	.0480	.0754*	.0489	.0454†	.0487	.0759*	.0494	.0454†	.0490	.0754*
	100	.0514	.0386†	.0508	.0689*	.0519	.0387†	.0516	.0697*	.0518	.0387†	.0515	.0700*
Medium	10	.0483	.0381†	.0449†	.0483	.0487	.0398†	.0455†	.0515	.0478	.0398†	.0458	.0512
	30	.0481	.0205†	.0451†	.0438†	.0518	.0211†	.0489	.0460	.0529	.0212†	.0493	.0460
	100	.0514	.0021†	.0481	.0103†	.0533	.0021†	.0510	.0111†	.0535	.0021†	.0504	.0113†
High	10	.0483	.0226†	.0395†	.0342†	.0538	.0234†	.0470	.0391†	.0544*	.0235†	.0473	.0400†
	30	.0481	.0043†	.0394†	.0126†	.0590*	.0043†	.0483	.0151†	.0594*	.0043†	.0488	.0150†
	100	.0514	.0000†	.0418†	.0006†	.0602*	.0000†	.0500	.0006†	.0603*	.0000†	.0504	.0007†
Extreme	10	.0483	.0009†	.0185†	.0048†	.1027*	.0010†	.0466	.0078†	.1094*	.0011†	.0506	.0085†
	30	.0481	.0000†	.0135†	.0000†	.1139*	.0000†	.0472	.0000†	.1153*	.0000†	.0488	.0000†
	100	.0514	.0000†	.0140†	.0000†	.1195*	.0000†	.0515	.0000†	.1195*	.0000†	.0521	.0000†

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of rejections across 10,000 simulations.
* signifies value significantly above sampling error
† signifies value significantly below sampling error

Table 8.
Rejection Rates for Three-Group, Null Conditions

Extent of Heterogeneity of Regression	n	Location of Test: μ_X Denominator Error Term HET	Inter	ANC	UNW	$\bar{X}$ Denominator Error Term HET	Inter	ANC	UNW	C_a Denominator Error Term HET	Inter	ANC	UNW
None	10	.0489	.0486	.0488	.0540	.0470	.0491	.0472	.0544	.0473	.0496	.0473	.0548*
	30	.0463	.0476	.0463	.0684*	.0465	.0484	.0468	.0688	.0466	.0484	.0470	.0689*
	100	.0479	.0483	.0476	.0716*	.0477	.0486	.0477	.0718	.0477	.0486	.0475	.0719*
Low	10	.0489	.0459	.0486	.0524	.0473	.0465	.0476	.0536	.0475	.0464	.0485	.0536
	30	.0463	.0400†	.0453†	.0630*	.0475	.0403†	.0474	.0623*	.0476	.0404†	.0474	.0624*
	100	.0479	.0289†	.0472	.0525	.0484	.0291†	.0479	.0523	.0483	.0291†	.0475	.0524
Medium	10	.0489	.0357†	.0461	.0436†	.0501	.0372†	.0476	.0460	.0510	.0373†	.0478	.0461
	30	.0463	.0186†	.0436†	.0364†	.0501	.0186†	.0475	.0387†	.0507	.0186†	.0475	.0385†
	100	.0479	.0010†	.0445†	.0072†	.0523	.0010†	.0491	.0074†	.0524	.0010†	.0491	.0074†
High	10	.0489	.0191†	.0412†	.0288†	.0564*	.0200†	.0484	.0350†	.0572*	.0201†	.0496	.0352†
	30	.0463	.0034†	.0381†	.0099†	.0567*	.0035†	.0488	.0125†	.0564*	.0035†	.0487	.0125†
	100	.0479	.0000†	.0393†	.0002†	.0589*	.0000†	.0483	.0002†	.0594*	.0000†	.0485	.0002†
Extreme	10	.0489	.0014†	.0187†	.0032†	.1016*	.0017†	.0490	.0076†	.1016*	.0018†	.0501	.0081†
	30	.0463	.0000†	.0149†	.0000†	.1078*	.0000†	.0493	.0000†	.1085*	.0000†	.0490	.0000†
	100	.0479	.0000†	.0150†	.0000†	.1083*	.0000†	.0499	.0000†	.1089*	.0000†	.0501	.0000†

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of rejections across 10,000 simulations.
* signifies value significantly above sampling error
† signifies value significantly below sampling error

Table 9.
Power for Two-Group Conditions

Extent of Heterogeneity of Regression	n	Location of Test: μ_X Denominator Error Term HET	Inter	ANC	UNW	$\bar{X}$ Denominator Error Term HET	Inter	ANC	UNW	C_a Denominator Error Term HET	Inter	ANC	UNW
None	10	.7500	.1772	.7523	.7103	.7752	.1814	.7769	.7318	.7780	.1821	.7796	.7336
	30	.7844	.1737	.7843	.7742	.7925	.1753	.7933	.7796	.7939	.1753	.7940	.7797
	100	.7994	.1725	.7992	.7952	.8023	.1733	.8026	.7977	.8025	.1733	.8027	.7973
Low	10	.7500	.1692	.7510	.7065	.7727	.1730	.7754	.7236	.7770	.1740	.7782	.7251
	30	.7844	.1573	.7831	.7474	.7952	.1580	.7943	.7530	.7958	.1581	.7948	.7525
	100	.7994	.1304	.7985	.7197	.8020	.1305	.8017	.7205	.8015	.1305	.8013	.7206
Medium	10	.7500	.1341	.7424	.6486	.7706	.1392	.7633	.6636	.7745	.1397	.7668	.6683
	30	.7844	.0785	.7746	.5749	.7917	.0789	.7829	.5794	.7910	.0790	.7834	.5801
	100	.7994	.0093	.7904	.2709	.7977	.0093	.7894	.2732	.7972	.0093	.7884	.2730
High	10	.7500	.0844	.7235	.5259	.7635	.0866	.7422	.5463	.7677	.0875	.7478	.5489
	30	.7844	.0146	.7578	.3118	.7835	.0146	.7587	.3173	.7855	.0146	.7610	.3179
	100	.7994	.0000	.7739	.0257	.7906	.0000	.7640	.0293	.7906	.0000	.7650	.0292
Extreme	10	.7500	.0043	.5781	.1076	.7266	.0047	.5881	.1338	.7315	.0049	.5895	.1376
	30	.7844	.0000	.6090	.0010	.7427	.0000	.5931	.0028	.7426	.0000	.5946	.0028
	100	.7994	.0000	.6255	.0000	.7486	.0000	.5995	.0000	.7475	.0000	.5991	.0000

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Table entries give the proportion of rejections across 10,000 simulations.

Table 10.
Power for Three-Group Conditions

Extent of Heterogeneity of Regression	n	Location of Test: μ_X Denominator Error Term HET	Inter	ANC	UNW	$\bar{X}$ Denominator Error Term HET	Inter	ANC	UNW	C_a Denominator Error Term HET	Inter	ANC	UNW
None	10	.7481	.3334	.7504	.7232	.7633	.3444	.7653	.7368	.7640	.3446	.7674	.7371
	30	.7852	.3411	.7847	.7718	.7872	.3454	.7873	.7745	.7874	.3454	.7876	.7747
	100	.7950	.3569	.7955	.7855	.7975	.3582	.7974	.7873	.7973	.3585	.7974	.7872
Low	10	.7481	.3242	.7505	.7177	.7634	.3347	.7653	.7286	.7647	.3345	.7669	.7307
	30	.7852	.3149	.7837	.7552	.7893	.3181	.7872	.7569	.7891	.3180	.7872	.7568
	100	.7950	.2506	.7941	.7270	.7975	.2509	.7964	.7273	.7974	.2511	.7965	.7273
Medium	10	.7481	.2563	.7430	.6708	.7591	.2639	.7548	.6797	.7599	.2647	.7563	.6803
	30	.7852	.1531	.7772	.6160	.7839	.1550	.7754	.6220	.7832	.1550	.7756	.6214
	100	.7950	.0159	.7862	.3397	.7931	.0161	.7851	.3430	.7929	.0161	.7848	.3431
High	10	.7481	.1528	.7259	.5723	.7553	.1601	.7351	.5824	.7561	.1611	.7360	.5832
	30	.7852	.0315	.7620	.3775	.7785	.0322	.7557	.3830	.7781	.0322	.7577	.3831
	100	.7950	.0001	.7724	.0399	.7852	.0001	.7619	.0421	.7859	.0001	.7629	.0421
Extreme	10	.7481	.0080	.5938	.1366	.7258	.0010	.5883	.1684	.7249	.0101	.5898	.1709
	30	.7852	.0000	.6312	.0033	.7399	.0000	.6068	.0069	.7404	.0000	.6092	.0069
	100	.7950	.0000	.6373	.0000	.7478	.0000	.6111	.0000	.7493	.0000	.6122	.0000

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Table entries give the proportion of rejections across 10,000 simulations.

Table 11.
Confidence Interval Coverage for Two-Group Conditions

		Location of Test											
		μ_X				$\bar{X}$				C_a			
Extent of Heterogeneity of Regression	n	Denominator Error Term				Denominator Error Term				Denominator Error Term			
		HET	Inter	ANC	UNW	HET	Inter	ANC	UNW	HET	Inter	ANC	UNW
None	10	.9478	.9477	.9402	.9233	.9525	.9462	.9519	.9345	.9505	.9497	.9502	.9337
	30	.9506	.9483	.9487	.9143	.9519	.9482	.9524	.9142	.9493	.9476	.9497	.9175
	100	.9459	.9519	.9450	.9060	.9529	.9526	.9533	.9189	.9492	.9492	.9489	.9162
Low	10	.9506	.9503	.9448	.9304	.9532	.9494	.9534	.9368	.9491	.9494	.9501	.9320
	30	.9449	.9531	.9427	.9185	.9479	.9519	.9478	.9187	.9517	.9498	.9519	.9180
	100	.9494	.9665	.9496	.9288	.9497	.9644	.9497	.9285	.9520	.9635	.9520	.9317
Medium	10	.9511	.9590	.9473	.9417	.9523	.9615	.9534	.9452	.9504	.9591	.9541	.9428
	30	.9524	.9769	.9529	.9518	.9505	.9767	.9529	.9571	.9484	.9809	.9518	.9507
	100	.9520	.9978	.9537	.9888	.9518	.9983	.9550	.9900	.9474	.9979	.9513	.9882
High	10	.9498	.9770	.9516	.9570	.9456	.9720	.9542	.9583	.9501	.9748	.9585	.9612
	30	.9485	.9937	.9570	.9836	.9536	.9950	.9625	.9868	.9501	.9953	.9579	.9872
	100	.9487	1	.9572	.9998	.9498	1	.9594	.9998	.9550	1	.9632	.9998
Extreme	10	.9522	.9986	.9796	.9936	.9498	.9985	.9807	.9958	.9511	.9985	.9825	.9974
	30	.9480	1	.9841	1	.9513	1	.9858	1	.9520	1	.9871	1
	100	.9468	1	.9838	1	.9507	1	.9858	1	.9508	1	.9858	1

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of times the confidence interval contained the population value across 10,000 simulations.

Table 12.
Confidence Interval Coverage for Three-Group Conditions

Extent of Heterogeneity of Regression	n	Location of Test: μ_X Denominator Error Term HET	Inter	ANC	UNW	$\bar{X}$ Denominator Error Term HET	Inter	ANC	UNW	C_a Denominator Error Term HET	Inter	ANC	UNW
None	10	.9213	.9481	.9081	.9011	.9193	.9514	.9130	.9101	.9169	.9490	.9105	.9035
	30	.9218	.9552	.9183	.8971	.9174	.9484	.9152	.8941	.9173	.9499	.9159	.8880
	100	.9142	.9494	.9128	.8923	.9162	.9541	.9156	.8897	.9204	.9508	.9196	.8928
Low	10	.9238	.9517	.9135	.9077	.9215	.9530	.9177	.9137	.9225	.9529	.9155	.9093
	30	.9231	.9561	.9203	.9039	.9118	.9572	.9097	.8957	.9198	.9532	.9192	.9003
	100	.9157	.9673	.9146	.9066	.9210	.9680	.9210	.9150	.9176	.9692	.9174	.9152
Medium	10	.9194	.9588	.9109	.9199	.9242	.9613	.9218	.9206	.9212	.9624	.9198	.9262
	30	.9163	.9815	.9164	.9393	.9165	.9838	.9184	.9380	.9165	.9836	.9189	.9436
	100	.9173	.9990	.9210	.9830	.9172	.9992	.9214	.9842	.9179	.9986	.9215	.9868
High	10	.9215	.9770	.9187	.9397	.9168	.9775	.9219	.9406	.9228	.9808	.9280	.9476
	30	.9229	.9973	.9305	.9807	.9191	.9973	.9289	.9780	.9184	.9974	.9270	.9807
	100	.9191	1	.9285	.9995	.9231	1	.9349	.9994	.9169	1	.9291	.9993
Extreme	10	.9243	.9986	.9600	.9923	.9255	.9989	.9648	.9925	.9210	.9996	.9634	.9931
	30	.9226	1	.9674	1	.9205	1	.9685	1	.9239	1	.9676	1
	100	.9164	1	.9671	1	.9170	1	.9684	1	.9191	1	.9675	1

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of times the confidence interval contained the population value across 10,000 simulations.

Table 13.
Confidence Interval Coverage for Two-Group Conditions, Including Chen's Increment to ANCOHET Standard Error

		Location of Test								
		μ_X			$\bar{X}$			C_a		
Extent of Heterogeneity of Regression	n	Denominator Error Term			Denominator Error Term			Denominator Error Term		
		HET	ANC	Chen	HET	ANC	Chen	HET	ANC	Chen
None	10	.9478	.9402	.9539	.9525	.9519	.9587	.9505	.9502	.9559
	30	.9506	.9487	.9524	.9519	.9524	.9536	.9493	.9497	.9518
	100	.9459	.9450	.9468	.9529	.9533	.9538	.9492	.9489	.9497
Low	10	.9506	.9448	.9571	.9532	.9534	.9591	.9491	.9501	.9550
	30	.9449	.9427	.9465	.9479	.9478	.9506	.9517	.9519	.9533
	100	.9494	.9496	.9502	.9497	.9497	.9505	.9520	.9520	.9529
Medium	10	.9511	.9473	.9579	.9523	.9534	.9590	.9504	.9541	.9589
	30	.9524	.9529	.9560	.9505	.9529	.9543	.9484	.9518	.9532
	100	.9520	.9537	.9542	.9518	.9550	.9546	.9474	.9513	.9506
High	10	.9498	.9516	.9613	.9456	.9542	.9585	.9501	.9585	.9609
	30	.9485	.9570	.9560	.9536	.9625	.9618	.9501	.9579	.9578
	100	.9487	.9572	.9554	.9498	.9594	.9567	.9550	.9632	.9612
Extreme	10	.9522	.9796	.9733	.9498	.9807	.9706	.9511	.9825	.9740
	30	.9480	.9841	.9716	.9513	.9858	.9725	.9520	.9871	.9736
	100	.9468	.9838	.9688	.9507	.9858	.9716	.9508	.9858	.9714

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of times the confidence interval contained the population value across 10,000 simulations.

Table 14.
Confidence Interval Coverage for Three-Group Conditions, Including Chen's Increment to ANCOHET Standard Error

		Location of Test								
		μ_X			$\bar{X}$			C_a		
Extent of Heterogeneity of Regression	n	Denominator Error Term			Denominator Error Term			Denominator Error Term		
		HET	ANC	Chen	HET	ANC	Chen	HET	ANC	Chen
None	10	.9213	.9081	.9259	.9193	.9130	.9251	.9169	.9105	.9221
	30	.9218	.9183	.9227	.9174	.9152	.9186	.9173	.9159	.9197
	100	.9142	.9128	.9146	.9162	.9156	.9167	.9204	.9196	.9208
Low	10	.9238	.9135	.9282	.9215	.9177	.9336	.9225	.9155	.9265
	30	.9231	.9203	.9252	.9118	.9097	.9137	.9198	.9192	.9223
	100	.9157	.9146	.9163	.9210	.9210	.9213	.9176	.9174	.9183
Medium	10	.9194	.9109	.9260	.9242	.9218	.9312	.9212	.9198	.9296
	30	.9163	.9164	.9211	.9165	.9184	.9253	.9165	.9189	.9208
	100	.9173	.9210	.9213	.9172	.9214	.9211	.9179	.9215	.9210
High	10	.9215	.9187	.9321	.9168	.9219	.9287	.9228	.9280	.9356
	30	.9229	.9305	.9309	.9191	.9289	.9292	.9184	.9270	.9266
	100	.9191	.9285	.9261	.9231	.9349	.9317	.9169	.9291	.9253
Extreme	10	.9243	.9600	.9495	.9255	.9648	.9518	.9210	.9634	.9494
	30	.9226	.9674	.9481	.9205	.9685	.9472	.9239	.9676	.9487
	100	.9164	.9671	.9448	.9170	.9684	.9452	.9191	.9675	.9451

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the proportion of times the confidence interval contained the population value across 10,000 simulations.

Table 15.
Average Confidence Interval Width for Two-Group, Null Conditions

		Location of Test											
		μ_X				$\bar{X}$				C_a			
Extent of Heterogeneity of Regression	n	Denominator Error Term				Denominator Error Term				Denominator Error Term			
		HET	Inter	ANC	UNW	HET	Inter	ANC	UNW	HET	Inter	ANC	UNW
None	10	1.99	9.01	1.92	1.95	1.93	9.06	1.92	1.94	1.92	9.05	1.91	1.94
	30	1.05	5.24	1.04	1.01	1.04	5.28	1.04	1.01	1.04	5.19	1.04	1.01
	100	0.56	2.91	0.56	0.54	0.56	2.89	0.56	0.54	0.56	2.86	0.56	0.54
Low	10	1.99	9.33	1.92	1.97	1.93	9.25	1.92	1.96	1.92	9.33	1.92	1.97
	30	1.05	5.69	1.04	1.05	1.04	5.65	1.04	1.04	1.04	5.70	1.04	1.05
	100	0.56	3.77	0.56	0.61	0.56	3.78	0.56	0.61	0.56	3.77	0.56	0.61
Medium	10	1.99	11.20	1.94	2.14	1.94	11.19	1.95	2.14	1.92	11.33	1.94	2.15
	30	1.05	9.22	1.06	1.35	1.04	9.19	1.05	0.35	1.04	9.30	1.05	1.36
	100	0.56	8.89	0.57	1.07	0.56	8.85	0.57	1.07	0.56	8.82	0.57	1.07
High	10	1.99	15.01	2.00	2.51	1.93	15.00	2.00	2.51	1.93	15.14	2.00	2.52
	30	1.05	14.97	1.09	1.90	1.04	14.90	1.08	1.89	1.04	14.95	1.08	1.89
	100	0.56	15.26	0.58	1.73	0.56	15.29	0.58	1.74	0.56	15.36	0.58	1.74
Extreme	10	1.98	34.40	2.38	4.64	1.94	34.72	2.40	4.68	1.92	34.66	2.39	4.67
	30	1.05	37.71	1.31	4.36	1.04	37.60	1.31	4.35	1.04	37.53	1.31	4.34
	100	0.56	38.67	0.70	4.29	0.56	38.63	0.70	4.28	0.56	38.64	0.70	4.29

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the average width of the confidence intervals over 10,000 simulations.

Table 16.
Average Confidence Interval Width for Three-Group, Null Conditions

Extent of Heterogeneity of Regression	n	Location of Test: μ_X Denominator Error Term HET	Inter	ANC	UNW	$\bar{X}$ Denominator Error Term HET	Inter	ANC	UNW	C_a Denominator Error Term HET	Inter	ANC	UNW
None	10	1.51	2.98	1.44	1.46	1.48	2.98	1.44	1.47	1.48	2.99	1.44	1.47
	30	0.80	1.72	0.79	0.78	0.80	1.74	0.79	0.78	0.79	1.72	0.79	0.78
	100	0.43	0.95	0.43	0.42	0.43	0.95	0.43	0.42	0.43	0.94	0.43	0.43
Low	10	1.50	3.04	1.44	1.48	1.48	3.05	1.44	1.48	1.48	3.05	1.44	1.48
	30	0.80	1.84	0.79	0.80	0.79	1.82	0.79	0.80	0.80	1.83	0.79	0.80
	100	0.43	1.14	0.43	0.46	0.43	1.14	0.43	0.46	0.43	1.14	0.43	0.46
Medium	10	1.51	3.50	1.46	1.58	1.48	3.48	1.46	1.58	0.48	3.49	1.46	1.58
	30	0.80	2.61	0.80	0.97	0.79	2.61	0.80	0.97	0.79	2.59	0.80	0.97
	100	0.43	2.29	0.43	0.72	0.43	2.30	0.43	0.72	0.43	2.29	0.43	0.72
High	10	1.51	4.35	1.49	1.80	1.48	4.34	1.49	1.78	1.48	4.36	1.50	1.79
	30	0.80	3.90	0.82	1.28	0.80	3.89	0.82	1.28	0.79	3.92	0.82	1.29
	100	0.43	3.80	0.44	1.12	0.43	3.80	0.44	1.12	0.43	3.80	0.44	1.12
Extreme	10	1.51	8.84	1.76	3.01	1.47	8.87	1.76	3.02	1.47	8.90	1.76	3.03
	30	0.80	9.29	0.98	2.76	0.80	9.30	0.98	2.77	0.80	9.26	0.97	2.75
	100	0.43	9.41	0.53	2.69	0.43	9.42	0.53	2.69	0.43	9.42	0.53	2.69

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA Error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the average width of the confidence intervals over 10,000 simulations.

Table 17.
Average Confidence Interval Width for Two-Group, Null Conditions, Including Chen's Increment to ANCOHET Standard Error

		Location of Test								
		μ_X			$\bar{X}$			C_a		
Extent of Heterogeneity of Regression	n	Denominator Error Term			Denominator Error Term			Denominator Error Term		
		HET	ANC	Chen	HET	ANC	Chen	HET	ANC	Chen
None	10	1.99	1.92	2.04	1.93	1.92	1.98	1.92	1.91	1.97
	30	1.05	1.04	1.06	1.04	1.04	1.05	1.04	1.04	1.05
	100	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Low	10	1.99	1.92	2.04	1.93	1.92	1.98	1.92	1.92	1.97
	30	1.05	1.04	1.06	1.04	1.04	1.05	1.04	1.04	1.05
	100	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Medium	10	1.99	1.94	2.05	1.94	1.95	2.00	1.92	1.94	1.98
	30	1.05	1.06	1.07	1.04	1.05	1.06	1.04	1.05	1.06
	100	0.56	0.57	0.57	0.56	0.57	0.57	0.56	0.57	0.57
High	10	1.99	2.00	2.09	1.93	2.00	2.03	1.93	2.00	2.02
	30	1.05	1.09	1.09	1.04	1.08	1.08	1.04	1.08	1.08
	100	0.56	0.58	0.58	0.56	0.58	0.58	0.56	0.58	0.58
Extreme	10	1.98	2.38	2.23	1.94	2.40	2.18	1.92	2.39	2.17
	30	1.05	1.31	1.17	1.04	1.31	1.17	1.04	1.31	1.17
	100	0.56	0.70	0.63	0.56	0.70	0.63	0.56	0.70	0.62

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the average width of the confidence intervals over 10,000 simulations.

Table 18.
Average Confidence Interval Width for Three-Group, Null Conditions, Including Chen's Increment to ANCOHET Standard Error

		Location of Test								
		μ_X			$\bar{X}$			C_a		
Extent of Heterogeneity of Regression	n	Denominator Error Term			Denominator Error Term			Denominator Error Term		
		HET	ANC	Chen	HET	ANC	Chen	HET	ANC	Chen
None	10	1.51	1.44	1.53	1.48	1.44	1.50	1.48	1.44	1.50
	30	0.80	0.79	0.80	0.80	0.79	0.80	0.79	0.79	0.80
	100	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
Low	10	1.50	1.44	1.53	1.48	1.44	1.50	1.48	1.44	1.50
	30	0.80	0.79	0.80	0.79	0.79	0.80	0.80	0.79	0.80
	100	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
Medium	10	1.51	1.46	1.55	1.48	1.46	1.51	0.48	1.46	1.51
	30	0.80	0.80	0.80	0.79	0.80	0.81	0.79	0.80	0.81
	100	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43
High	10	1.51	1.49	1.56	1.48	1.49	1.53	1.48	1.50	1.54
	30	0.80	0.82	0.82	0.80	0.82	0.82	0.79	0.82	0.82
	100	0.43	0.44	0.44	0.43	0.44	0.44	0.43	0.44	0.44
Extreme	10	1.51	1.76	1.67	1.47	1.76	1.63	1.47	1.76	1.63
	30	0.80	0.98	0.89	0.80	0.98	0.88	0.80	0.97	0.88
	100	0.43	0.53	0.47	0.43	0.53	0.47	0.43	0.53	0.47

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA Error; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error; C_a = Center of Accuracy. Refer to text for more information regarding error terms used. Numbers in table represent the average width of the confidence intervals over 10,000 simulations.

Table 19.
True Standard Deviation and Average Standard Errors for Tests Conducted at $\bar{X}$, Two-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4595	.4556	.3589	.4544	.4271	.4664
	30	.2620	.2596	.2049	.2595	.2465	.2614
	100	.1392	.1416	.1136	.1416	.1358	.1419
Low	10	.4591	.4559	.3651	.4552	.4305	.4671
	30	.2627	.2592	.2229	.2595	.2554	.2613
	100	.1416	.1415	.1475	.1418	.1530	.1420
Medium	10	.4668	.4549	.4391	.4599	.4674	.4699
	30	.2650	.2596	.3595	.2634	.3282	.2642
	100	.1437	.1415	.3479	.1437	.2692	.1434
High	10	.4844	.4562	.5937	.4745	.5517	.4790
	30	.2698	.2597	.5873	.2712	.4623	.2692
	100	.1482	.1416	.6026	.1480	.4385	.1462
Extreme	10	.5811	.4552	1.3569	.5661	1.0188	.5130
	30	.3269	.2597	1.4837	.3266	1.0657	.2913
	100	.1779	.1416	1.5202	.1784	1.0797	.1584

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Table 20.
True Standard Deviation and Average Standard Errors for Tests Conducted at the Center of Accuracy, Two-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4608	.4530	.3568	.4534	.4252	.4638
	30	.2611	.2589	.2082	.2589	.2477	.2608
	100	.1410	.1416	.1115	.1415	.1347	.1418
Low	10	.4618	.4549	.3673	.4560	.4318	.4662
	30	.2625	.2593	.2232	.2598	.2556	.2615
	100	.1419	.1414	.1484	.1417	.1537	.1419
Medium	10	.4694	.4542	.4405	.4611	.4692	.4694
	30	.2660	.2591	.3640	.2631	.3304	.2639
	100	.1451	.1415	.3508	.1437	.2711	.1434
High	10	.4838	.4541	.5961	.4744	.5527	.4775
	30	.2721	.2596	.5894	.2713	.4633	.2693
	100	.1469	.1414	.6024	.1478	.4383	.1461
Extreme	10	.5808	.4532	1.3538	.5658	1.0172	.5113
	30	.3310	.2593	1.4772	.3258	1.0610	.2906
	100	.1793	.1416	1.5223	.1786	1.0811	.1585

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Table 21.
True Standard Deviation and Average Standard Errors for Tests Conducted at μ_x, Two-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4800	.4705	.3563	.4545	.4258	.4809
	30	.2632	.2620	.2052	.2597	.2468	.2639
	100	.1432	.1419	.1129	.1415	.1353	.1422
Low	10	.4816	.4697	.3636	.4547	.4293	.4806
	30	.2674	.2621	.2284	.2602	.2583	.2643
	100	.1411	.1419	.1475	.1418	.1529	.1424
Medium	10	.4753	.4676	.4481	.4595	.4718	.4827
	30	.2621	.2626	.3596	.2631	.3280	.2662
	100	.1419	.1420	.3489	.1438	.2698	.1438
High	10	.4774	.4687	.5950	.4726	.5512	.4912
	30	.2647	.2620	.5842	.2710	.4602	.2714
	100	.1435	.1420	.6013	.1480	.4376	.1466
Extreme	10	.4761	.4686	1.3503	.5642	1.0142	.5253
	30	.2648	.2619	1.4780	.3261	1.0616	.2930
	100	.1440	.1420	1.5191	.1784	1.0789	.1588

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Table 22.
True Standard Deviation and Average Standard Errors for Tests Conducted at $\bar{X}$, Three-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4050	.4007	.3423	.3910	.3758	.4068
	30	.2278	.2257	.1994	.2244	.2183	.2269
	100	.1239	.1227	.1094	.1225	.1197	.1229
Low	10	.4063	.4004	.3504	.3914	.3796	.4070
	30	.2298	.2253	.2086	.2242	.2229	.2266
	100	.1219	.1228	.1309	.1228	.1308	.1231
Medium	10	.4123	.4008	.3996	.3964	.4051	.4104
	30	.2309	.2255	.2997	.2271	.2709	.2287
	100	.1251	.1229	.2635	.1243	.2072	.1243
High	10	.4265	.3999	.4982	.4054	.4573	.4157
	30	.2368	.2258	.4465	.2332	.3572	.2328
	100	.1265	.1228	.4357	.1275	.3205	.1263
Extreme	10	.4931	.3997	1.0170	.4784	.7740	.4429
	30	.2772	.2260	1.0665	.2769	.7711	.2501
	100	.1522	.1228	1.0801	.1513	.7687	.1358

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Table 23.
True Standard Deviation and Average Standard Errors for Tests Conducted at the Center of Accuracy, Three-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4078	.4000	.3432	.3911	.3759	.4061
	30	.2274	.2256	.1977	.2242	.2176	.2266
	100	.1217	.1228	.1076	.1226	.1190	.1230
Low	10	.4016	.4005	.3499	.3923	.3799	.4070
	30	.2267	.2256	.2099	.2246	.2237	.2270
	100	.1232	.1229	.1313	.1228	.1309	.1232
Medium	10	.4089	.4007	.4000	.3970	.4055	.4105
	30	.2281	.2252	.2977	.2269	.2697	.2285
	100	.1237	.1228	.2624	.1242	.2066	.1241
High	10	.4181	.4006	.5005	.4072	.4593	.4165
	30	.2373	.2255	.4499	.2331	.3591	.2326
	100	.1276	.1228	.4355	.1275	.3203	.1263
Extreme	10	.4903	.3991	1.0210	.4795	.7769	.4428
	30	.2761	.2258	1.0616	.2763	.7677	.2497
	100	.1518	.1227	1.0800	.1512	.7686	.1357

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Table 24.
True Standard Deviation and Average Standard Errors for Tests Conducted at μ_X, Three-Group Scenario

Extent of Heterogeneity of Regression	n	$\hat{Y}_1 - \hat{Y}_2$ Estimate Standard Deviation	Average Standard Error				
			HET	Inter	ANC	UNW	Chen
None	10	.4154	.4081	.3415	.3905	.3754	.4141
	30	.2257	.2266	.1969	.2239	.2170	.2277
	100	.1236	.1230	.1086	.1226	.1194	.1232
Low	10	.4129	.4078	.3484	.3904	.3781	.4143
	30	.2246	.2268	.2116	.2244	.2245	.2281
	100	.1237	.1230	.1312	.1228	.1309	.1233
Medium	10	.4180	.4094	.4013	.3968	.4062	.4191
	30	.2286	.2269	.2989	.2272	.2706	.2301
	100	.1236	.1230	.2623	.1243	.2066	.1244
High	10	.4164	.4081	.4987	.4059	.4576	.4237
	30	.2248	.2270	.4474	.2332	.3577	.2341
	100	.1226	.1230	.4365	.1275	.3210	.1265
Extreme	10	.4123	.4096	1.0138	.4792	.7724	.4515
	30	.2256	.2273	1.0657	.2767	.7705	.2513
	100	.1235	.1229	1.0800	.1512	.7686	.1359

Note. HET = ANCOHET; Inter = Interaction; ANC = Standard ANCOVA error; UNW = Unweighted average of $MS_{residual}$ from ANCOHET and $MS_{A \times X}$; Chen = Chen (2006)'s suggested increment to the ANCOHET standard error

Appendix A

Derivation of the Standard Error of the Difference across Groups in Predicted Scores in Analysis of Covariance: Implications of Heterogeneity of Regression and a Random Covariate

Heterogeneity of regression in analysis of covariance (ANCOVA) can be assessed by comparing a model that allows for a different slope in each of the $j = 1, 2, \ldots a$ groups with one that assumes a common within-group slope:

$$\text{Full}: \; Y_{ij} = \mu + \alpha_j + \beta_j X_{ij} + \varepsilon_{ij}$$
$$\text{Restricted}: \; Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{ij}$$

Rogosa (1980) has shown that, if there is heterogeneity of regression in the population, the typical ANCOVA test of treatment effects is not distributed appropriately. An alternative procedure in the presence of mild to moderate heterogeneity suggested by Rogosa (1980) is to compute the adjusted treatment sum of squares as in a typical ANCOVA but use as an error term the error associated with the ANCOHET model, just as would be done in ANOVA when the interaction was nonsignificant. This provides a test of the hypothesis that there are no treatment effects in the case of a covariate whose values are assumed to be fixed, and achieves an unbiased estimate of residual variance that does not assume homogeneity of regression at the cost of only $a - 1$ degrees of freedom for error (see Appendix B for further discussion of Rogosa's "safer ANCOVA").

To characterize the treatment effect more completely, it is desirable with moderate to pronounced heterogeneity to assess the treatment effect as a function of the value of the covariate. If the traditional ANCOVA model were exactly right the vertical distance

between the population regression lines would be a constant for all values of *X*. When there is reason to believe this is not the case, one would like to estimate the magnitude of the treatment effect as a function of *X* and have a way of assessing its significance. A test of the treatment effect at a given *X* value may be arrived at by developing an estimate of the treatment effect somewhat like is done in a standard ANCOVA test of the difference between adjusted means—that is, the difference between the predicted scores for different conditions at a given value of *X*—and then deriving the variability of this estimated difference. A ratio of the square of the estimated effect to its variance estimate can then be used as a statistical test.

The basic problem involves the estimation of the vertical distance between regression lines. Because this is difficult to envision, let us begin our consideration of this problem by referring to the simple regression situation involving a single group with one predictor and one dependent variable. Besides deriving estimates of the dependent variable in this case using a simple regression equation, we can also relatively easily derive estimates of the variability of our predictions. The model for this situation is typically written in standard regression texts (e.g. Neter, Wasserman, & Kutner, 1983, p. 60) as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where the intercept β_0 and slope β_1 parameters are to be estimated by least squares, the X_i values are assumed to be fixed constants, and the errors of prediction ε_i are assumed to be normally distributed with mean of 0 and variance σ^2 .

We will begin our derivation by considering a deviation form of the regression equation using the least squares estimates of the parameters

$\hat{\beta}_1 = b = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and $\hat{\beta}_0 = \bar{Y} - b\bar{X}$. Let X_p be the particular X value at which we wish to estimate Y, and let the corresponding predicted value $\hat{Y}_p$ be the estimated mean of the conditional probability distribution. Then, in this simple (i.e., two-variable) regression situation, we can write

$$\hat{Y}_p = \bar{Y} + b(X_p - \bar{X}) \tag{A.1}$$

Under the assumption that the X values are fixed and that the errors are normally distributed in the population, the variability of $\hat{Y}_p$ can be shown[1] to be decomposable into the following two components:

$$\sigma^2_{\hat{Y}_p} = \sigma^2_{\bar{Y}} + (X_p - \bar{X})^2 \sigma^2_b \tag{A.2}$$

The first component, the variability of $\bar{Y}$ is $\sigma^2_{\bar{Y}} = \sigma^2 / n$. However, we now have the magnitude of the estimate of error depending on the X value as well as the variability in Y. That is, because β is not known but is estimated by a statistic, we expect our slope estimates to vary somewhat from sample to sample. How much difference the error in b makes gets larger and larger as X_p moves farther away from $\bar{X}$.

The variance of our slope statistic itself can be derived fairly easily once we rewrite the definitional formula for the slope in a convenient form, namely

$$b = \sum k_i Y_i \tag{A.3}$$

where the k_i are simple functions[2] of the X values:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \tag{A.4}$$

Now, because the variance of a linear combination of independent random variables is simply the sum of the original variances, each weighted by the square of the original weight, we immediately have the following expression for σ_b^2, the variance of the slope estimate b:

$$\sigma_b^2 = \text{Var}\,(b) = \text{Var}\left(\sum k_i Y_i\right) = \sum k_i^2 \text{Var}(Y_i) \tag{A.5}$$

where Var is to be read as "the variance of" the expression that follows within parentheses. Making use of the fact that the variances of Y_i are constant and equal to σ^2, then substituting for k_i we obtain

$$\begin{aligned} \sigma_b^2 &= \sigma^2 \sum k_i^2 = \sigma^2 \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 \\ &= \sigma^2 \frac{\sum (X_i - \bar{X})^2}{\left[\sum (X_i - \bar{X})^2\right]^2} = \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2} \end{aligned} \tag{A.6}$$

We are now ready to substitute our results into Equation A.2 to obtain the final form of the variability of our estimated conditional mean $\hat{Y}_p$:

$$\begin{aligned} \sigma_{\hat{Y}_p}^2 &= \frac{\sigma^2}{n} + (X_p - \bar{X})^2 \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \end{aligned} \tag{A.7}$$

Thus, we have now derived the variance of the estimated mean Y score for a particular X score X_p in simple regression, and we have shown that it is more variable than the sample mean Y score, and increasingly so as X_p departs more from $\bar{X}$.

Neter, Wasserman, and Kutner (1983, pp. 83-84) assert that, if X is random, estimation and testing can proceed in simple regression just as if X were fixed, as long as the following two conditions are met:

1. The conditional distributions of Y given X_i are normal.

2. The X_i are independent random variables whose distribution does not depend on the intercept or slope parameters, or on the variance of the errors, σ^2.

Assuming these conditions obtain, we can write the expected variance of a prediction shown in Equation A.7 in terms of the population variance of the X scores, σ_X^2. Given the denominator of the term shown on the right in brackets above is the numerator of the sample variance, s_X^2, and

$$\mathscr{E}\left(s_X^2\right) = \mathscr{E}\left(\frac{\sum\left(X_i - \bar{X}\right)^2}{n-1}\right) = \sigma_X^2$$

In the case of random X, our model would have two sources of random variability, X and error, so we now explicitly denote the sigma in Equation A.7 as referring to the error variability, i.e. σ_ε^2. Thus, in the case of random X we could write the variance of the predicted scores as:

$$\sigma^2_{\hat{Y}_p} = \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{(X_p - \bar{X})^2}{(n-1)\sigma^2_X} \right] \tag{A.8}$$

A similar but somewhat different result obtains in ANCOVA. The similarity concerns the variance of the estimated mean Y score for a particular X score in a particular group. For $X = X_p$ and group j, with the assumption of homogeneity of regression, the slope would be estimated by the pooled within-group slope, $\hat{\beta} = b_{\mathrm{W}}$, and the intercepts would be estimated as $\hat{\mu} + \hat{\alpha}_j = \bar{Y}_j - b_{\mathrm{W}}\bar{X}_j$. Hence, the predicted scores at X_p could be written:

$$\hat{Y}_p = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}X_p = \bar{Y}_j - b_{\mathrm{W}}\bar{X}_j + b_{\mathrm{W}}X_p = \bar{Y}_j + b_{\mathrm{W}}(X_p - \bar{X}_j)$$

(A.9)

Thus, as in the simple-regression situation, the variance of our estimated conditional mean Y score increases as X_p departs from $\bar{X}_j$:

$$\begin{aligned} \sigma^2_{\hat{Y}_{p_j}} &= \mathrm{Var}(\bar{Y}_j) + \mathrm{Var}[b_{\mathrm{W}}(X_p - \bar{X}_j)] = \frac{\sigma^2}{n_j} + (X_p - \bar{X}_j)^2 \mathrm{Var}(b_{\mathrm{W}}) \\ &= \sigma^2 \left[\frac{1}{n_j} + \frac{(X_p - \bar{X}_j)^2}{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2} \right] \end{aligned} \tag{A.10}$$

(The intermediate steps of the derivation follow along the same lines as those for Equation A.7.) However, in ANCOVA, interest centers on the predicted scores at the grand mean on X (i.e., the adjusted Y means) and in the vertical distance between them.

Letting $X_p = \bar{X}$ in Equation A.9 results in the standard equation for the adjusted mean in group j:

$$\hat{Y}_j = \bar{Y}_j + b_W(\bar{X} - \bar{X}_j) = \bar{Y}_j - b_W(\bar{X}_j - \bar{X})$$

Thus, the square of the standard error of this adjusted mean, following Equation A.10, is

$$\sigma^2_{\hat{Y}_j} = \sigma^2 \left[\frac{1}{n_j} + \frac{(\bar{X}_j - \bar{X})^2}{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2} \right] \tag{A.11}$$

In one-way designs, the contrasts that are most often of interest are pairwise comparisons between groups. Because interpretation of a treatment effect is considerably more complicated in the case of heterogeneous regressions, where the magnitude of the difference between groups changes continuously as a function of the covariate, it is even more likely that contrasts would focus on only two groups at a time. Thus, for these reasons and for simplicity of development in what immediately follows, we consider only the two-group case. In the two-group case, under the assumption of homogeneous slopes, we would be most interested in the difference between the two adjusted means:

$$\begin{aligned} \hat{Y}_1 - \hat{Y}_2 &= \bar{Y}_1 - b_W(\bar{X}_1 - \bar{X}) - [\bar{Y}_2 - b_W(\bar{X}_2 - \bar{X})] \\ &= (\bar{Y}_1 - \bar{Y}_2) - b_W(\bar{X}_1 - \bar{X}_2) \end{aligned} \tag{A.12}$$

Notice that, although the comparison is a comparison of the estimated Y means at $\bar{X}$, $\bar{X}$ does not appear in the final form of Equation A.12. Furthermore, this would be true regardless of the particular value X_p at which we might compute the difference between our estimates of the conditional Y means. Thus, it perhaps should not be surprising that,

although it is unlike the simple regression situation, the standard error of this estimated treatment effect does not depend on the value of X at which we estimate it. That is, when homogeneous slopes are assumed, the precision of our estimate of the treatment effect is "maintained for all values of X" (Rogosa, 1980, p. 311), with the variance of our estimate in Equation A.12 being

$$\sigma^2_{\hat{Y}_1-\hat{Y}_2} = \sigma^2\left[\frac{1}{n_1}+\frac{1}{n_2}+\frac{(\bar{X}_1-\bar{X}_2)^2}{\sum_j\sum_i (X_{ij}-\bar{X}_j)^2}\right] \tag{A.13}$$

This variance expression is like those for the conditional mean (in Equation A.7) and for the adjusted mean (in Equation A.11) in that there is a component for the variability of the mean estimates and another component for the variability of the slope estimate. For the component reflecting the mean estimates, because we now are concerned with the difference between two independent group means (see Equation A.12), the variance of their difference is the sum of the variances of each mean separately. For the component reflecting the slope estimate, the variance of the slope is simply multiplied, as in Equations A.7 and A11, by the square of the relevant coefficient, which here is $\left(\bar{X}_1-\bar{X}_2\right)$ as shown in Equation A.12. We can estimate the variance of the difference in adjusted means by replacing σ^2 in Equation A.13 by the mean square error associated with the traditional ANCOVA full model. Denote this mean square error s^2. Thus, $(N-3)s^2$ would be equal to the residual sum of squares associated with the model using a common, pooled estimate of the slope in this two-group case.

In the case where the covariate is considered to be a random variable, and the distribution of the dependent variable and the covariate is bivariate normal, the consistent conclusion of statisticians (e.g., Scheffé, 1959, pp. 195-197) and behavioral science methodologists (e.g. Huitema, 1980, p. 111) is that, with the conventional assumption of homogeneity of regression, tests of the treatment effect in ANCOVA and the regression of the dependent variable on the covariate can be conducted in exactly the same fashion as when the covariate was considered to be fixed. As Winer, Brown and Michels (1991) affirmed, "the analysis of covariance need not be restricted to the case in which X is a fixed variable" (1991, p. 770). Crager (1987) reached a similar conclusion asserting that with a random covariate, the usual "ANCOVA estimates of the slope parameter and treatment effect contrasts are unbiased" and "the usual ANCOVA treatment effect contrast t-tests are valid significance tests for treatment effects" (1987, p. 895). In his derivations, Crager (1987, p. 901) suggests relating one part of the formula for the variance of the difference in adjusted means shown in Equation A.13 to Student's t distribution. In the case of a two-group, equal-n design, Crager states that the variance of this difference could be expressed, in the case of a random covariate, as follows:

$$\sigma^2_{\hat{Y}_1-\hat{Y}_2} = \sigma^2\left[\frac{2}{n}\right] + \sigma^2\mathscr{C}\left[\frac{(\bar{X}_1-\bar{X}_2)^2}{\sum_i (X_{i1}-\bar{X}_1)^2 + \sum_i (X_{i2}-\bar{X}_2)^2}\right] \qquad \text{(A.14)}$$

Note that the last term within brackets above can be seen to be directly related to a conventional two-group t test comparing the means on the X variable, which in the case of equal-n could be written as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_i (X_{i2} - \bar{X}_2)^2}{2(n-1)}\left[\frac{2}{n}\right]}} \tag{A.15}$$

Squaring this t statistic and re-arranging terms, we could write this as

$$t^2 = n(n-1)\frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_i (X_{i2} - \bar{X}_2)^2} \tag{A.16}$$

Thus, we see that $n(n-1)$ times the term within brackets on the right in Equation A.14 is distributed as a t^2 variable with $2(n-1)$ degrees of freedom, or equivalently as an F with 1 and $2(n-1)$ degrees of freedom. Given the expected value of an F random variable with df_{denom} denominator degrees of freedom is $df_{\text{denom}} / (df_{\text{denom}} - 2)$, we can write the expected value of the term within brackets on the right in Equation A.14 as

$$\mathscr{E}\left[\frac{(\bar{X}_1 - \bar{X}_2)^2}{\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_i (X_{i2} - \bar{X}_2)^2}\right] = \left(\frac{2n-2}{2n-4}\right)\left[\frac{1}{n(n-1)}\right] = \left[\frac{2(n-1)}{2(n-2)}\right]\left[\frac{1}{n(n-1)}\right] = \frac{1}{n(n-2)}$$

Thus, the variance of the difference in adjusted means in ANCOVA with a random covariate shown in Equation A.14 could be written as

$$\sigma^2_{\hat{Y}_1 - \hat{Y}_2} = \sigma^2\left[\frac{2}{n} + \frac{1}{n(n-2)}\right] \tag{A.17}$$

To understand the differing impacts of the various contributions to this variance, it is worth noting that even with a random covariate, the variability contributed by the two group means indicated by the $2/n$ term in Equation A.17 will be $2(n - 2)$ times as large as the variability due to the sampling error in the slope estimate and the sampling error in the group means of the random covariate. For example, with $n = 10$, $2/n$ would be 16 times as large as $1/n(n\text{-}2)$, and with $n = 50$, $2/n$ would be 96 times as large as $1/n(n\text{-}2)$.

We are now finally ready to return to the problem of estimating the vertical distance between two nonparallel regression lines and determining the variability of that estimate. We begin by returning to the case of X being fixed where the results are well understood. These results build on those we have presented previously for the simple regression situation and for ANCOVA with homogeneous slopes. The prediction equation for the ANCOHET model can be written:

$$\hat{Y}_{ij} = a_j + b_j X_{ij} \tag{A.18}$$

Thus, if we substitute for X_{ij} some particular value of the covariate—for example, X_p—the difference in estimated conditional means for the two groups would be

$$\hat{Y}_{p1} - \hat{Y}_{p2} = a_1 + b_1 X_p - \left(a_2 + b_2 X_p\right) = a_1 - a_2 + \left(b_1 - b_2\right) X_p \tag{A.19}$$

An alternative way of writing this estimated difference, in which we substitute the expressions for our estimated values of the intercepts, makes it easier to understand the

variance estimate. That is, we can write the vertical distance between the two regression lines:

$$\hat{Y}_{p1} - \hat{Y}_{p2} = (\bar{Y}_1 - b_1 \bar{X}_1) - (\bar{Y}_2 - b_2 \bar{X}_2) + (b_1 - b_2)X_p$$

$$= \bar{Y}_1 - \bar{Y}_2 + b_1(X_p - \bar{X}_1) - b_2(X_p - \bar{X}_2)$$

(A.20)

To determine the variability of this estimate, we must consider not only the sampling error of the Y group means, but also both the variance of our estimate of b_1, which equals $\sigma^2/\Sigma_i(X_{i1} - \bar{X}_1)^2$, and the variance of our estimate of b_2, $\sigma^2/\Sigma_i(X_{i2} - \bar{X}_2)^2$. Thus, similar to Equation A.13, but now allowing for heterogeneous slopes, the variability of our estimate of the vertical distance between the lines with X being regarded as fixed can be written:

$$\sigma^2_{\hat{Y}_{p1}-\hat{Y}_{p2}} = \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(X_p - \bar{X}_1)^2}{\sum_i (X_{i1} - \bar{X}_1)^2} + \frac{(X_p - \bar{X}_2)^2}{\sum_i (X_{i2} - \bar{X}_2)^2} \right] \quad \text{(A.21)}$$

A comparison with the variance of the estimate of a single mean in regression (Equation A.7) or ANCOVA (Equation A.10) shows that the variance of the distance between two regression lines is simply the sum of the variances of conditional means estimated by each. We can estimate this variance, and thereby move toward carrying out a test of the significance of the difference between the regression lines at any arbitrary value of X, by simply replacing σ^2 in Equation A.21 by the mean square error associated with the model allowing for heterogeneous slopes, which we denote s^2_{het}. In the two-group situation in which we estimate a slope and an intercept for each group, our model would have $N - 4$

degrees of freedom. Thus, a test of the significance of the difference between the two lines—that is, of the treatment effect at an X value X_p—would be carried out as a simple t test with $N - 4$ degrees of freedom. That is,

$$t = \frac{\hat{Y}_{p1} - \hat{Y}_{p2} - 0}{\hat{\sigma}_{\hat{Y}_{p1} - \hat{Y}_{p2}}} \tag{A.22}$$

where the denominator is

$$\hat{\sigma}_{\hat{Y}_{p1} - \hat{Y}_{p2}} = s_{\text{het}} \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(X_p - \bar{X}_1)^2}{\sum_i (X_{i1} - \bar{X}_1)^2} + \frac{(X_p - \bar{X}_2)^2}{\sum_i (X_{i2} - \bar{X}_2)^2} \right]^{1/2} \tag{A.23}$$

with s_{het} being the square root of s^2_{het}, which, as we suggested previously, is the sum of squares error E_F divided by $N - 4$ for the ANCOHET Full model:

$$\text{Full: } Y_{ij} = \mu + \alpha_j + \beta_j X_{ij} + \varepsilon_{ij} \tag{A.24}$$

As can be seen in this expression for the estimated standard error (Equation A.23), the precision of our estimate of the treatment effect decreases the farther the particular point X_p at which we are evaluating it is from the group means of the covariate. This is similar to what we saw in the simple regression situation (Equation A.7). Thus, if X_p is chosen near the center of the distribution of X scores, the accuracy of our estimation of the treatment effect increases. In fact, it turns out that the accuracy is greatest at a point corresponding to a weighted average of the group means on the covariate (with the

weight for each mean being the sum of squares on the covariate in the other group). This point is referred to in the literature as the *center of accuracy,* denoted C_a, and so we have:

$$C_a = \frac{\sum_i \left(X_{i2} - \bar{X}_2\right)^2 \bar{X}_1 + \sum_i \left(X_{i1} - \bar{X}_1\right)^2 \bar{X}_2}{\sum_{j=1}^{2} \sum_i \left(X_{ij} - \bar{X}_j\right)^2} \tag{A.25}$$

Surprisingly, the vertical distance between the two nonparallel regression lines at the center of accuracy corresponds exactly to the estimate of the difference between adjusted means in a typical ANCOVA assuming a common slope. Thus, one can interpret the difference between adjusted means in ANCOVA as the treatment effect for an "average" individual—that is, an individual whose X score is roughly at the center of the distribution of X scores—regardless of whether the regressions are parallel. The difference between the ANCOHET and the ANCOVA tests of this difference is in the error term. The ANCOVA test is perfectly valid only if the assumption of parallelism is exactly met. The ANCOHET test is actually more like the tests commonly used in factorial ANOVA in that it is valid regardless of whether there is an interaction in the population (nonparallelism). The form of the error term for the ANCOHET test of the treatment effect at the center of accuracy reduces to

$$\hat{\sigma}_{\hat{Y}_{Ca1} - \hat{Y}_{Ca2}} = s_{\text{het}} \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2} \right]^{1/2} \tag{A.26}$$

If X is a random variable *and* the slopes are heterogeneous, the implications for the variability of the difference in predicted means in the two groups are not well understood. Chen (2006) considered this situation and argued when tests are conducted, as is often the case, at the observed grand mean on the covariate, that is, letting $X_p = \bar{X}$, that the standard error shown in Equation A.23 would be too small because it ignores the sampling error that causes $\bar{X}$ to depart from μ_X. He suggested that an expression for the variance of a contrast in means like that shown in Equation A.21, which was derived under the assumption of a fixed covariate, would need to be increased when the covariate was random by an amount dependent on the sampling variability in $\bar{X}$ and on the difference in the slopes in the two groups; Chen specifically indicated (2006, p. 4163) the needed increment was

$$\left(\beta_1 - \beta_2\right)^2 \mathrm{Var}\left(\bar{X}\right) = \left(\beta_1 - \beta_2\right)^2 \frac{\sigma_X^2}{n_1 + n_2} \tag{A.27}$$

Although Chen simply said about this result "it can be shown" rather than presenting a proof, it appeared he was presuming the population slopes could be treated as known, fixed constants rather than themselves being subject to sampling variability. Thus, it seems plausible that the needed increment might be even greater if the sampling variability in the slopes was also considered.

This conjecture was leant some indirect support by recent work on a different but related issue. Instead of the situation considered by Chen (2006) that is the focus of the current dissertation where the treatment factor is a fixed factor and only the covariate is

treated as random, a recent article by Liu, West, Levy, and Aiken (2017) considered the situation where one is interested in two random predictors, *X* and *Z*, and their interaction. What is analogous to the difference in adjusted means in that situation is what Liu et al. (2017), following Cohen, Cohen, West and Aiken (2003), term the simple slope of the outcome *Y* on the predictor *X* at the sample mean of Z. Their mathematical derivations indicate that the variability in the estimate of this simple slope is extremely complex, involving the sum of more than 10 different terms, with the final result depending not only on the sampling variability in the estimated *Z* mean and sampling variability in the slope estimates but also the covariances of various terms such as the intercept and slope, the intercept and mean, and the slope and mean. Because of this complexity, Liu et al. (2017) proposed comparing an estimated effect in their situation to a distribution generated for each empirical data set collected by use of bootstrapping methods applied to that data set.

It is hoped that in the simpler situation of a fixed treatment factor considered in the current dissertation that the variability in the estimate of the treatment effect in the case of heterogeneous regressions on a random covariate might be adequately approximated by some simpler method. Toward that end, multiple estimates of denominator error terms are considered and evaluated in terms of the tests and confidence intervals that result from using such error terms.

[1] The proof makes use of the fact that both $\bar{Y}$ and *b* can be expressed as linear combinations of the Y_i and that the covariance of $\bar{Y}$ and *b* can be shown to be zero.
[2] This is a legitimate rewriting of the definitional formula for the slope because $\sum\left(X_i-\bar{X}\right)\left(Y_i+\bar{Y}\right)=\sum\left(X_i-\bar{X}\right)Y_i$. This in turn is true because $\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=\sum\left(X_i-\bar{X}\right)Y_i-\sum\left(X_i-\bar{X}\right)\bar{Y}$, but $\sum\left(X_i-\bar{X}\right)\bar{Y}=\bar{Y}\sum\left(X_i-\bar{X}\right)=0$ because the sum of the deviations from the mean must equal zero. Thus, we have

$$b = \frac{\sum (X_i - \bar{X})\ Y_i}{\sum (X_i - \bar{X})^2}$$

which may be rewritten

$$b = \sum \frac{\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} Y_i$$

References

Chen, X. (2006). The adjustment of random baseline measurements in treatment effect estimation. *Journal of Statistical Planning and Inference, 136,* 4161-4175.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral science* (3rd ed.). New York: Routledge.

Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics, 43,* 891-901.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: John Wiley & Sons.

Liu, Y., West, S. G., Levy, R., & Aiken, L. S. (2017). Tests of simple slopes in multiple regression models with an interaction: Comparison of four approaches. *Multivariate Behavioral Research, 52*, 445-464.

Neter, J., Wasserman, W., & Kutner, M. H. (1983). *Applied linear regression models.* Homewood, IL: Irwin.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307-321.

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

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* (3rd ed.). New York: McGraw-Hill.

Appendix B

On the Rogosa "Safer" Test of the Group Effect

As noted by Maxwell, Delaney, and Kelley (2018, p. 528), one of the more surprising results of ANCOVA analyses with heterogeneous slopes (or "ANCOHET") is that in the two-group case, as shown by Rogosa (1980, Equation 7), the difference in predicted values in the two groups at the center of accuracy is exactly equal to the difference in adjusted means in a conventional ANCOVA. The distribution of the error term used in a conventional ANCOVA, as also noted by Rogosa (1980, p. 311), is exactly correct only if the within-group slopes are perfectly homogeneous. Because of this Rogosa proposed what he termed a "safer ANCOVA" (1980, p. 312), where the numerator is like that in a conventional ANCOVA but the denominator is computed using a model allowing for heterogeneous regressions. It was this procedure that was, reasonably enough, used by Harwell and Serlin (1988) in conducting what they denoted as a Rogosa *F* test (e.g., in their Table 8, p. 275).

Thus, it is important to understand how the Rogosa "safer" *F* test of the group effect compares to the test of the group effect at a given value of the covariate (e.g., as developed in Appendix A). One challenge in relating Rogosa's (1980) formulas to a traditional ANOVA or ANCOVA formulation of an *F* test of a group effect is that he describes tests only for the two-group situation and approaches these as one might if one were doing a two-group *t* test in which the numerator only has the difference between adjusted means, rather than being a mean square for an effect that is to be compared to a mean square error as in a conventional ANCOVA *F* test.

To think of how the numerator of the ANCOVA *F* compares to the numerator of an ANCOHET *F*, it is helpful to think of the multiplier of the error estimate in the denominator of the Rogosa approach as a term that could be shifted to the numerator when one wants to approximate the numerator used in a conventional ANCOVA. The basic idea is seen clearly if one reverts back to a simple two-group *t* test and relates this to the *F* that would be used if an ANOVA were performed instead.

A conventional two-group, independent *t* test might be written:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{B.1}$$

Rogosa's tests are expressed in a fashion analogous to the square of such a form of the *t* test, but keeping the multiplier of the variance estimate in the denominator, i.e.

$$t^2 = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{s^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \tag{B.2}$$

If one were doing an ANOVA of these data, the term involving sample sizes would in essence be shifted to the numerator of the test statistic, which if one first expressed the reciprocals of the sample sizes using a common denominator before moving to the numerator would result in the following form of the test statistic:

$$F = t^2 = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{s^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} = \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{s^2\left[\frac{n_2 + n_1}{n_1 n_2}\right]} = \frac{\left[\frac{n_1 n_2}{n_1 + n_2}\right]\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{s^2} \tag{B.3}$$

Now in the equal-*n* case, the multiplier of the difference in group means would be just half of sample size in each group. This results in a simple form that is clearly equal to the conventional ANOVA *F* given the difference in group means would be just twice the absolute value of the difference between either group mean and the grand mean:

$$F = \frac{\left[\frac{n}{2}\right]\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{s^2} = \frac{\left[\frac{n}{2}\right]\left\{2\left(\bar{Y}_1 - \bar{Y}\right)\right\}^2}{s^2} = \frac{2n(\bar{Y}_1 - \bar{Y})^2}{s^2} = \frac{\sum_{j=1}^{2} n\left(\bar{Y}_j - \bar{Y}\right)^2}{s^2} = \frac{MS_{\text{Betw}}}{MS_{\text{With}}} \tag{B.4}$$

Rogosa (1980) writes the *F* statistic used in a conventional ANCOVA in the first part of his Equation 11 in a form analogous to the square of a two-group *t* where the multiplier indicating the sample sizes is in the denominator rather than the numerator:

$$F = \frac{\left(\bar{Y}_1' - \bar{Y}_2'\right)^2}{s^2_{\text{ANCOVA}}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{SSX_1 + SSX_2}\right]} \tag{B.5}$$

Now a conventional ANCOVA *F* would compute a numerator of *MS* for the group effect by shifting the multiplier to the numerator. In the equal *n* case this could be expressed as follows:

$$F = \frac{\left[\frac{1}{\frac{2}{n} + \frac{\left(\bar{X}_1 - \bar{X}_2\right)^2}{SSX_1 + SSX_2}}\right]\left(\bar{Y}_1' - \bar{Y}_2'\right)^2}{s^2_{\text{ANCOVA}}} \tag{B.6}$$

The multiplier of the squared difference in adjusted means will of course be less than $\frac{n}{2}$ to the extent that the other term (i.e., the squared difference between group means on the covariate over that sum of squares within on the covariate) is nonzero. The reason that using the ANCOVA numerator, as Harwell and Serlin apparently did in their Rogosa *F*, produces more Type I errors than the pick-a-point tests that one might carry out instead is that the multiplier shown in the numerator above will be larger than the one that would be used in a pick-a-point test. That test could be expressed as (see Equations 19 and 20 of Appendix A):

$$F = \frac{\left(\hat{Y}_{p1} - \hat{Y}_{p2}\right)^2}{s^2_{\text{het}}\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(X_p - \bar{X}_1)^2}{\sum_i (X_{i1} - \bar{X}_1)^2} + \frac{(X_p - \bar{X}_2)^2}{\sum_i (X_{i2} - \bar{X}_2)^2}\right]} \qquad \text{(B.7)}$$

The sum of the last two terms in the multiplier in the denominator will be larger than the last term in the ANCOVA denominator shown in Equation B.5 whenever the test is conducted at a point other than the true center of accuracy, as almost certainly would be the case in realistic situations with a random covariate. Because the difference in adjusted mean in ANCOVA would be expected to be the same as that in the ANCOHET test, the difference in the adjusted numerator mean squares in the ANCOVA as opposed to the ANCOHET test will boil down to the difference in the multipliers. One way of expressing this intuitively is that the estimation of a single pooled slope reduces the effective sample size per group less than does the estimation of two separate slopes in the

two groups. Thus the numerator of the ANCOVA would be expected to be somewhat larger than that in the ANCOHET test, and hence may be positively biased.

Appendix C

SAS Syntax Used to Simulate Data, Perform Data Analysis and Analyze Results

Presented in this appendix is an example of the SAS code used to generate and analyze the data for the simulations. The syntax for the two-group case is presented first followed by the three-group case. Most of the syntax is to prepare for a macro program that can be called to generate the data for a specific cell of the design. This macro program is a general framework that incorporates macro variables, whose values change based on user specification, to simulate data and generate the desired output. While the macro program is presented below, the following macro variables are used within the program. The only difference between the two-group and three-group macro variables is that the three-group program contains the additional macro variables 'slope3' and 'sampsize3' as a result of having an additional group.

The nine macro variables, in order of appearance, are:

1. numsamples – number of simulation samples, always 10,000 in this project, but could be specified as something else
2. slope1 – unstandardized regression slope for the first group
3. slope2 – unstandardized regression slope for the second group
4. sampsize1 – sample size for the first group
5. sampsize2 – sample size for the second group
6. mean – constant added to produce difference in adjusted means (takes on a value of zero for the null conditions)

7. location – specifies the location at which the test of between group differences occurs. Can take on the following values:
    a. x
    b. xbar
    c. ca
8. output – specifies the output to be generated. Can take on the following values:
    a. TypeI
    b. Power
    c. ci
    d. ci_width
    e. trueSD
9. seed – establishes the starting seed for the random number generator, which is important for replicating results exactly

Macro Program for Two-Group Conditions

```
%macro ODSOff;      /*Macro to turn off output for certain SAS procedures*/
ods graphics off;   /*This prevents output from 10,000 ANCOHETs from being*/
ods exclude all;    /*Displayed*/
ods noresults;
%mend;
%macro ODSOn;       /*Macro to turn output back on*/
ods graphics on;
ods exclude none;
ods results;
%mend;
```

```
%macro two_group_simulation(numsamples, slope1, slope2, sampsize1, sampsize2,
mean, location, output, seed);
```

```
data CI_2grp;
        call streaminit(&seed);
        do sampleid=1 to &numsamples;       *Loops over total number of simulations;
        do i=1 to (&sampsize1 + &sampsize2);        *Loops over total sample size;
                x = rand("Normal", 0, 1);                   *Randomly generates x
values from Normal Distribution with mean = 0 and standard deviation = 1;
                e = rand("Normal", 0, 1);                   *Randomly generates error
values from Normal Distribution with mean = 0 and standard deviation = 1;
                        if i LE (&sampsize1) then do;
                                group=0;
                                y = &mean + &slope1*x + e; *Creates Y values based on
macro values of mean and slope and randomly generated values of x and e for first group;
                        end;
                        if i GT (&sampsize1) then do;
                                group = 1;
                                y = &slope2*x + e;     *Creates Y values for second group;
                        end;
                output;
        end;
end;
run;

proc sort data=work.ci_2grp;by sampleid group;run;

/*Calculates Correlations and Saves to Output Dataset*/
%odsoff;
proc corr data=work.ci_2grp outp=corr;
by sampleid group;
var x;
with y;run;
%odson;

data work.corr_grp0;set work.corr (where=(_type_="CORR" and group=0));drop
_name_ _type_ group;rename x=corr_grp0;run;
data work.corr_grp1;set work.corr (where=(_type_="CORR" and group=1));drop
_name_ _type_ group;rename x=corr_grp1;run;
data work.corr_both;merge work.corr_grp0 work.corr_grp1;by sampleid;run;

/*Calculates Mean of X for Each Simulation*/
%odsoff;
proc means data= CI_2grp mean stddev;
by sampleid;
var x;
output out=simmean mean=xbar var=var;run;
proc means data=CI_2grp mean stddev;
by sampleid group;
```

```
var x;
output out=simmean_grp mean=xbar var=var n=n;run;
%odson


proc transpose data=simmean_grp out=simmean_trans_mean
(rename=(col1=xmean_grp0 col2=xmean_grp1));
var xbar;
by sampleid;
run;
proc transpose data=simmean_grp out=simmean_trans_var (rename=(col1=xvar_grp0
col2=xvar_grp1));
var var;
by sampleid;
run;
proc transpose data=simmean_grp out=simmean_trans_n (rename=(col1=n_grp0
col2=n_grp1));
var n;
by sampleid;
run;

/*Merge Group and Overall Means for X*/
data simmean_bygroup;merge simmean_trans_mean simmean_trans_var
simmean_trans_n;drop _name_;run;

/*Merge mean of X for Each Simulation with Original Data*/
data CI_2grp_merge; merge CI_2grp simmean simmean_trans_mean simmean_trans_var
simmean_trans_n corr_both end=lastobs;
        by sampleid;
SSx_grp0=(n_grp0-1)*xvar_grp0;              *Calculate Sums of Squares for Each Group;
SSx_grp1=(n_grp1-1)*xvar_grp1;
Ca=((SSx_grp1*xmean_grp0)+(SSx_grp0*xmean_grp1))/(SSx_grp1+SSx_grp0);
/*Calculate Center of Accuracy*/
Xdev_Xbar= x-xbar;                    /*Centers X around X_bar*/
Xdev_Ca = x-ca;                       /*Centers X around the Center of Accuracy*/
drop _type_ _freq_ _name_;
run;

/*Chen Calculation Dataset*/
data chen_calc;set work.ci_2grp_merge;by sampleid; if first.sampleid;drop i x e y
xdev_xbar xdev_ca;run;

data work.chen_calc;set work.chen_calc;
chen_incr= ((corr_grp0-corr_grp1)*(corr_grp0-corr_grp1))*(var/(n_grp0+n_grp1));
partial_chen_incr=2.3*chen_incr;
run;
```

```
/*Perform ANCOHET for Each Simulation Iteration based on Test Location and Output
Results to ANCOHET_Stat Dataset*/
%if &location= xbar %then %do;
%odsoff
proc glm data=CI_2grp_merge outstat=ANCOHET_stat;
        by sampleid;
        class group;
        model y = xdev_xbar group group*xdev_xbar/solution;
        lsmeans group/ pdiff=all cl at xdev_xbar = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

%if &location=ca %then %do;
%odsoff
proc glm data=CI_2grp_merge outstat=ANCOHET_stat;
        by sampleid;
        class group;
        model y = xdev_ca group group*xdev_ca/solution;
        lsmeans group/ pdiff=all cl at xdev_ca = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

%if &location= x %then %do;
%odsoff
proc glm data=CI_2grp_merge outstat=ANCOHET_stat;
        by sampleid;
        class group;
```

```
        model y = x group group*x/solution;
        lsmeans group/ pdiff=all cl at x = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

/*Save Sums of Squares and Degrees of Freedom from Individual ANCOHET Models
and then Merge*/
data ANCOHET_stat;set ANCOHET_stat (where=(_type_ ne "SS1"));run;
proc transpose data=ANCOHET_stat out=ancohetSS_wide prefix=SS;
        by sampleid;
        id _source_;
        var SS;
run;
proc transpose data=ANCOHET_stat out=ancohetDF_wide prefix=df;
        by sampleid;
        id _source_;
        var df;
run;

data ANCOHETstat_wide; merge ancohetss_wide ancohetdf_wide;by sampleid;drop
_name_;
run;

/*Create Confidence Interval Data Set*/
data ci_info; set ci_2grp_merge;by sampleid; if first.sampleid;drop i x e group y
xdev_xbar xdev_ca;run;
data ci_info;merge work.ci_info lsmeans_y_trans ancohetstat_wide chen_calc;by
sampleid;run;

/*Calculate Quantities for Type I Error Rates, Power, and CI's*/

data work.ci_info;set work.ci_info;

%if &location ne x %then %do;
MSE_ANCOHET = (sserror/dferror);
MSE_ANCOVA =
(sserror+SSXdev_&location._group)/(dferror+dfXdev_&location._group);
```

```
MSE_interaction = (SSXdev_&location._group/dfXdev_&location._group);
MSE_weight1 =
(sserror+SSXdev_&location._group)/(dferror+dfXdev_&location._group);
MSE_weight2 = (MSE_ANCOHET + MSE_Interaction)/2;
%end;

%if &location= x %then %do;
Xp_diff_Xbar_grp0 = (0 - xmean_grp0)*(0 - xmean_grp0);
Xp_diff_Xbar_grp1 = (0 - xmean_grp1)*(0 - xmean_grp1);

MSE_ANCOHET = (sserror/dferror);
MSE_ANCOVA = (sserror+SS&location._group)/(dferror+df&location._group);
MSE_interaction = (SS&location._group/df&location._group);
MSE_weight1 =  (sserror+SS&location._group)/(dferror+df&location._group);
MSE_weight2 = (MSE_ANCOHET + MSE_Interaction)/2;
%end;

%if &location= xbar %then %do;
Xp_diff_Xbar_grp0 = (xbar - xmean_grp0)*(xbar - xmean_grp0);
Xp_diff_Xbar_grp1 = (xbar - xmean_grp1)*(xbar - xmean_grp1);
%end;

%if &location= ca %then %do;
Xp_diff_Xbar_grp0 = (ca - xmean_grp0)*(ca - xmean_grp0);
Xp_diff_Xbar_grp1 = (ca - xmean_grp1)*(ca - xmean_grp1);
%end;

ratio_grp0 = Xp_diff_Xbar_grp0/ssx_grp0;
ratio_grp1 = Xp_diff_Xbar_grp1/ssx_grp1;
ratio_01 = ((xmean_grp0 - xmean_grp1)*(xmean_grp0 -
xmean_grp1))/(ssx_grp0+ssx_grp1);

samp_inv = (1/n_grp0) + (1/n_grp1);

Stderr_ANCOHET = sqrt(MSE_ANCOHET*(samp_inv + ratio_grp0 + ratio_grp1));
ANCOHET_variance=(stderr_ancohet)*(stderr_ancohet);
stderr_ancohet2 = sqrt(mse_ancohet)*sqrt((samp_inv + ratio_grp0 + ratio_grp1));
Stderr_ANCOVA = sqrt(MSE_ANCOVA*(samp_inv + ratio_01));
Stderr_Interaction = sqrt(MSE_interaction*samp_inv);
Stderr_Weight1 = sqrt(MSE_weight1*(samp_inv + ratio_01));
Stderr_Weight2 = sqrt(MSE_weight2*samp_inv);

stderr_ancohet_chen=sqrt(ANCOHET_variance+chen_incr);
stderr_partial_chen=sqrt(ANCOHET_variance+partial_chen_incr);

%if &location ne x %then %do;
```

```
CV_ancohet=sqrt(finv(.95, dfgroup, dferror));
CV_ANCOVA=sqrt(finv(.95, dfgroup, (dferror+1)));
CV_Interaction=sqrt(finv(.95, dfgroup, dfXdev_&location._group));
CV_Weight1=sqrt(finv(.95, dfgroup, (dferror+dfXdev_&location._group)));
CV_Weight2=sqrt(finv(.95, dfgroup, ((dferror+dfXdev_&location._group)/2)));
%end;

%if &location = x %then %do;
CV_ancohet=sqrt(finv(.95, dfgroup, dferror));
CV_ANCOVA=sqrt(finv(.95, dfgroup, (dferror+1)));
CV_Interaction=sqrt(finv(.95, dfgroup, df&location._group));
CV_Weight1=sqrt(finv(.95, dfgroup, (dferror+df&location._group)));
CV_Weight2=sqrt(finv(.95, dfgroup, ((dferror+df&location._group)/2)));
%end;

lsmeans_y_diff = y_lsmean_grp0 - y_lsmean_grp1;

ci_halfwidth_ancohet=CV_ancohet*Stderr_ANCOHET;
ci_lower_ancohet= lsmeans_y_diff-ci_halfwidth_ancohet;
ci_upper_ancohet=lsmeans_y_diff+ci_halfwidth_ancohet;
ci_width_ancohet=2*CV_ancohet*Stderr_ANCOHET;

ci_halfwidth_chen=CV_ancohet*stderr_ancohet_chen;
ci_lower_chen= lsmeans_y_diff-ci_halfwidth_chen;
ci_upper_chen=lsmeans_y_diff+ci_halfwidth_chen;
ci_width_chen=2*CV_ancohet*stderr_ancohet_chen;

ci_halfwidth_chen_partial=CV_ancohet*stderr_partial_chen;
ci_lower_chen_partial= lsmeans_y_diff-ci_halfwidth_chen_partial;
ci_upper_chen_partial=lsmeans_y_diff+ci_halfwidth_chen_partial;
ci_width_chen_partial=2*CV_ancohet*stderr_partial_chen;

ci_halfwidth_ANCOVA=CV_ANCOVA*Stderr_ANCOVA;
ci_width_ANCOVA=2*CV_ANCOVA*Stderr_ANCOVA;
ci_lower_ANCOVA= lsmeans_y_diff-ci_halfwidth_ANCOVA;
ci_upper_ANCOVA=lsmeans_y_diff+ci_halfwidth_ANCOVA;

ci_halfwidth_interaction=CV_Interaction*Stderr_interaction;
ci_width_interaction=2*CV_Interaction*Stderr_interaction;
ci_lower_interaction= lsmeans_y_diff-ci_halfwidth_interaction;
ci_upper_interaction=lsmeans_y_diff+ci_halfwidth_interaction;

ci_halfwidth_weight1=CV_Weight1*Stderr_weight1;
ci_width_weight1=2*CV_Weight1*Stderr_weight1;
ci_lower_weight1= lsmeans_y_diff-ci_halfwidth_weight1;
ci_upper_weight1=lsmeans_y_diff+ci_halfwidth_weight1;
```

```
ci_halfwidth_weight2=CV_Weight2*Stderr_weight2;
ci_width_weight2=2*CV_Weight2*Stderr_weight2;
ci_lower_weight2= lsmeans_y_diff-ci_halfwidth_weight2;
ci_upper_weight2=lsmeans_y_diff+ci_halfwidth_weight2;

%if &location= x %then %do;
MU_adj_g0=&mean + (&slope1*0);
Mu_adj_g1=&slope2*0;
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location = xbar %then %do;
MU_adj_g0=&mean + (&slope1*xbar);
Mu_adj_g1=&slope2*xbar;
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location = ca %then %do;
MU_adj_g0=&mean + (&slope1*ca);
Mu_adj_g1=&slope2*ca;
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location ne x %then %do;
F_ANCOHET = (ssgroup/dfgroup)/(sserror/dferror);
F_ANCOVA =
(ssgroup/dfgroup)/((sserror+SSXdev_&location._group)/(dferror+dfXdev_&location._gr
oup));
F_interaction=(ssgroup/dfgroup)/(ssXdev_&location._group/dfXdev_&location._group);
F_weight1=(ssgroup/dfgroup)/((SSXdev_&location._group +
sserror)/(dfXdev_&location._group + dferror));
F_weight2 = (ssgroup/dfgroup)/(MSE_weight2);
p_ANCOHET = 1-probf(F_ANCOHET,dfgroup,dferror);
p_ANCOVA = 1-probf(F_ANCOVA,dfgroup,(dferror+dfXdev_&location._group));
p_interaction=1-probf(f_interaction,dfgroup,dfXdev_&location._group);
p_weight1=1-probf(F_weight1, dfgroup, (dferror+dfXdev_&location._group));
p_weight2=1-probf(F_weight2, dfgroup, ((dferror+dfXdev_&location._group)/2));
%end;

%if &location=x %then %do;
F_ANCOHET = (ssgroup/dfgroup)/(sserror/dferror);
F_ANCOVA =
(ssgroup/dfgroup)/((sserror+SS&location._group)/(dferror+df&location._group));
F_interaction=(ssgroup/dfgroup)/(ss&location._group/df&location._group);
```

```
F_weight1=(ssgroup/dfgroup)/((SS&location._group + sserror)/(df&location._group +
dferror));
F_weight2 = (ssgroup/dfgroup)/(MSE_weight2);
p_ANCOHET = 1-probf(F_ANCOHET,dfgroup,dferror);
p_ANCOVA = 1-probf(F_ANCOVA,dfgroup,(dferror+df&location._group));
p_interaction=1-probf(f_interaction,dfgroup,df&location._group);
p_weight1=1-probf(F_weight1, dfgroup, (dferror+df&location._group));
p_weight2=1-probf(F_weight2, dfgroup, ((dferror+df&location._group)/2));
%end;

RejectHO_ANCOHET = (p_ancohet<=.05);
RejectHO_ANCOVA = (p_ANCOVA<=.05);
RejectHO_Interaction = (p_interaction<=.05);
RejectHO_Weight1 = (p_weight1<=.05);
RejectHO_Weight2 = (p_weight2<=.05);
Pop_ParamInCI_ancohet = (ci_lower_ancohet<difference_population &
ci_upper_ancohet>difference_population);
Pop_ParamInCI_ancohet_chen = (ci_lower_chen<difference_population &
ci_upper_chen>difference_population);
Pop_ParamInCI_ANCOVA = (ci_lower_ANCOVA<difference_population &
ci_upper_ANCOVA>difference_population);
Pop_ParamInCI_interaction = (ci_lower_interaction<difference_population &
ci_upper_interaction>difference_population);
Pop_ParamInCI_weight1 = (ci_lower_weight1<difference_population &
ci_upper_weight1>difference_population);
Pop_ParamInCI_weight2 = (ci_lower_weight2<difference_population &
ci_upper_weight2>difference_population);
run;

%if &output= TypeI %then %do;
proc freq data=ci_info;
title1 "Type I Error Rates for";
title2 "b0=&slope1, b1=&slope2";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2";
title5 "at &location";
table RejectHO_ANCOHET RejectHO_ANCOVA RejectHO_Interaction
RejectHO_Weight1 RejectHO_Weight2/nocum;
run;
%end;

%if &output= Power %then %do;
proc freq data=ci_info;
title1 "Power Rates for";
title2 "b0=&slope1, b1=&slope2";
title3 "and";
```

```
title4 "n0=&sampsize1, n1=&sampsize2";
title5 "at &location";
table RejectHO_ANCOHET RejectHO_ANCOVA RejectHO_Interaction
RejectHO_Weight1 RejectHO_Weight2/nocum;
run;
%end;

%if &output= ci %then %do;
proc freq data=ci_info;
title1 "Confidence Interval Coverage Rates for";
title2 "b0=&slope1, b1=&slope2";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2";
title5 "at &location";
table pop_paraminCI_ancohet pop_paraminCI_ancova pop_paraminCI_interaction
pop_paraminCI_weight1 pop_paraminCI_weight2
Pop_ParamInCI_ancohet_chen/nocum;run;
%end;

%if &output=ci_width %then %do;
proc means data=ci_info n mean stddev min max;
title1 "Average Confidence Interval Width for";
title2 "b0=&slope1, b1=&slope2";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2";
title5 "at &location";
var ci_width_ancohet ci_width_ancova ci_width_interaction ci_width_weight1
ci_width_weight2 ci_width_chen ci_width_chen_partial;
run;
%end;

%if &output= trueSD %then %do;
proc means data=ci_info n mean stddev min max;
title1 "True SD and Empirical SD";
title2"b0=&slope1, b1=&slope2";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2";
title5 "at &location";
var lsmeans_y_diff Stderr_ANCOHET Stderr_ANCOVA Stderr_Interaction
Stderr_Weight1 Stderr_Weight2 stderr_ancohet_chen;
run;
%end;

%mend two_group_simulation;
```

Macro Program for Three-Group Conditions

```
%macro three_group_simulation(numsamples, slope1, slope2, slope3, sampsize1,
sampsize2, sampsize3, mean, location, output, seed);

data CI_3grp;
call streaminit(&seed);
do sampleid=1 to &numsamples;
        do i=1 to (&sampsize1 + &sampsize2 + &sampsize3);
                x=rand("Normal", 0, 1);
                e=rand("Normal", 0, 1);
                        if i LE (&sampsize1) then do;
                                group = 0;
                                group_contrast=0;
                                y = &mean + &slope1*x + e;
                        end;
                if i GT (&sampsize1) and i LE (&sampsize2+&sampsize1) then do;
                                group = 1;
                                group_contrast=1;
                                y = &slope2*x + e;
                        end;
                        if i GT (&sampsize1+&sampsize2) then do;
                                group=2;
                                group_contrast=1;
                                y = &slope3*x + e;
                        end;
                output;
        end;
end;
run;

data ci_3grp;set ci_3grp;
if group=0 then group_2grp=0;
if group in (1,2) then group_2grp=1;
run;

/*Calculates Correlations and Saves to Output Dataset*/
%odsoff;
proc corr data=work.ci_3grp outp=corr;
by sampleid group;
var x;
with y;run;
%odson;

/*Calculates Correlations Grouping 1 and 2 Together*/
%odsoff;
```

```
proc corr data=work.ci_3grp outp=corr_2grp;
by sampleid group_2grp;
var x;
with y;run;
%odson;

/*Assembles Correlation Data Sets */
data work.corr_grp0;set work.corr (where=(_type_="CORR" and group=0));drop
_name_ _type_ group;rename x=corr_grp0;run;
data work.corr_grp1;set work.corr (where=(_type_="CORR" and group=1));drop
_name_ _type_ group;rename x=corr_grp1;run;
data work.corr_grp2;set work.corr (where=(_type_="CORR" and group=2));drop
_name_ _type_ group;rename x=corr_grp2;run;
data work.corr_both;merge work.corr_grp0 work.corr_grp1 work.corr_grp2;by
sampleid;
avg_corr_1_2=mean(corr_grp1,corr_grp2);run;

data work.corr_grp0_2grp;set work.corr_2grp (where=(_type_="CORR" and
group_2grp=0));drop _name_ _type_ group_2grp;rename x=corr_grp0_other;run;
data work.corr_grp1_2grp;set work.corr_2grp (where=(_type_="CORR" and
group_2grp=1));drop _name_ _type_ group_2grp;rename x=corr_grp1_other;run;

data work.corr_both; merge work.corr_both work.corr_grp0_2grp
work.corr_grp1_2grp;by sampleid;
corr_diff= avg_corr_1_2-corr_grp1_other;run;

/*Calculates Mean of X for Each Simulation*/
%odsoff;
proc means data= CI_3grp mean stddev;
by sampleid;
var x;
output out=simmean mean=xbar var=var;run;
proc means data=CI_3grp mean stddev;
by sampleid group;
var x;
output out=simmean_grp mean=xbar var=var n=n;run;

proc means data=ci_3grp mean stddev;
by sampleid group_contrast;
var x;
output out=simmean_grp_contrast mean=xbar var=var n=n;run;
%odson

proc transpose data=simmean_grp out=simmean_trans_mean
(rename=(col1=xmean_grp0 col2=xmean_grp1 col3=xmean_grp2));
var xbar;
```

```
by sampleid;
run;
proc transpose data=simmean_grp out=simmean_trans_var (rename=(col1=xvar_grp0
col2=xvar_grp1 col3=xvar_grp2));
var var;
by sampleid;
run;
proc transpose data=simmean_grp out=simmean_trans_n (rename=(col1=n_grp0
col2=n_grp1 col3=n_grp2));
var n;
by sampleid;
run;

proc transpose data=simmean_grp_contrast out=simmean_trans_mean_contrast
(rename=(col1=xmean_grp0_contrast col2=xmean_grp1_contrast));
var xbar;
by sampleid;
run;
proc transpose data=simmean_grp_contrast out=simmean_trans_var_contrast
(rename=(col1=xvar_grp0_contrast col2=xvar_grp1_contrast));
var var;
by sampleid;
run;
proc transpose data=simmean_grp_contrast out=simmean_trans_n_contrast
(rename=(col1=n_grp0_contrast col2=n_grp1_contrast));
var n;
by sampleid;
run;

data simmean;set simmean;drop _type_ _freq_;run;

/*Merge Group and Overall Means for X*/
data simmean_bygroup;merge simmean_trans_mean simmean_trans_var
simmean_trans_n simmean_trans_mean_contrast simmean_trans_var_contrast
simmean_trans_n_contrast simmean;drop _name_;run;

/*Merge mean of X for Each Simulation with Original Data*/
data CI_3grp_merge; merge CI_3grp simmean_bygroup corr_both end=lastobs;
        by sampleid;
        xbar=(xmean_grp0+xmean_grp1+xmean_grp2)/3;
SSx_grp0=(n_grp0-1)*xvar_grp0;              *Calculate Sums of Squares for Each Group;
SSx_grp1=(n_grp1-1)*xvar_grp1;
SSx_grp2=(n_grp2-1)*xvar_grp2;
SSx_grp1_contrast = (n_grp1_contrast-1)*xvar_grp1_contrast;
grp0_weight=(ssx_grp1+ssx_grp2);
grp1_weight=(ssx_grp0+ssx_grp2);
```

```
grp2_weight=(ssx_grp1+ssx_grp0);
ca=((grp0_weight*xmean_grp0)+(grp1_weight*xmean_grp1)+(grp2_weight*xmean_grp
2))/(grp0_weight+grp1_weight+grp2_weight);
Xdev_Xbar= x-xbar;
                        /*Centers X around X_bar*/
Xdev_Ca = x-ca;
                        /*Centers X around the Center of Accuracy*/
run;

/*Chen Calculation Dataset*/
data chen_calc;set work.ci_3grp_merge;by sampleid;if first.sampleid; drop i x e y
xdev_xbar xdev_ca;run;

data work.chen_calc;set work.chen_calc;
chen_incr = ((corr_grp0-avg_corr_1_2)*(corr_grp0-
avg_corr_1_2))*(var/(n_grp0+n_grp1+n_grp2));
partial_chen_incr=2.3*chen_incr;
run;

%if &location= xbar %then %do;
/*Perform ANCOHET for Each Simulation Iteration and Output Results to
ANCOHET_Stat Dataset*/
%odsoff
proc glm data=CI_3grp_merge outstat=ANCOHET_stat;
        by sampleid;
        class group;
        model y = xdev_xbar group group*xdev_xbar/solution;      *&location is a
MACRO variable that specifies the location of the test for between group differences;
Contrast "Contrast" group 1 -.5 -.5;
lsmeans group/ pdiff=all cl at xdev_xbar = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1 col3=y_lsmean_grp2));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

%if &location=ca %then %do;
%odsoff
proc glm data=CI_3grp_merge outstat=ANCOHET_stat;
        by sampleid;
```

```
        class group;
        model y = xdev_ca group group*xdev_ca/solution;  *&location is a MACRO
variable that specifies the location of the test for between group differences;
        Contrast "Contrast" group 1 -.5 -.5;
        lsmeans group/ pdiff=all cl at xdev_ca = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1 col3=y_lsmean_grp2));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

%if &location=x %then %do;
%odsoff
proc glm data=CI_3grp_merge outstat=ANCOHET_stat;
        by sampleid;
        class group;
        model y = x group group*x/solution; *&location is a MACRO variable that
specifies the location of the test for between group differences;
        Contrast "Contrast" group 1 -.5 -.5;
        lsmeans group/ pdiff=all cl at x = 0 ;
        ods output lsmeans=lsmeans_y;
run;quit;
%odson

proc transpose data=work.lsmeans_y out=lsmeans_y_trans
(rename=(col1=y_lsmean_grp0 col2=y_lsmean_grp1 col3=y_lsmean_grp2));
var lsmean;
by sampleid;
run;
data work.lsmeans_y_trans;set work.lsmeans_y_trans;drop _name_ _label_;run;
%end;

/*Save Sums of Squares and Degrees of Freedom from Individual ANCOHET Models
and then Merge*/
data ANCOHET_stat;set ANCOHET_stat (where=(_type_ ne "SS1"));run;
proc transpose data=ANCOHET_stat out=ancohetSS_wide prefix=SS;
        by sampleid;
        id _source_;
        var SS;
run;
```

```
proc transpose data=ANCOHET_stat out=ancohetDF_wide prefix=df;
        by sampleid;
        id _source_;
        var df;
run;

data ANCOHETstat_wide; merge ancohetss_wide ancohetdf_wide;by sampleid;drop
_name_;
run;

/*Create Confidence Interval Data Set*/

data ci_info; set ci_3grp_merge;by sampleid; if first.sampleid;drop i x e group
group_contrast y xdev_xbar xdev_ca;run;
data ci_info;merge work.ci_info lsmeans_y_trans simmean_bygroup ancohetstat_wide
chen_calc;by sampleid;run;

data work.ci_info;set work.ci_info;

%if &location ne x %then %do;
MSE_ANCOHET = (sserror/dferror);
MSE_ANCOVA =
(sserror+SSXdev_&location._group)/(dferror+dfXdev_&location._group);
MSE_interaction = (SSXdev_&location._group/dfXdev_&location._group);
MSE_weight1 =
(sserror+SSXdev_&location._group)/(dferror+dfXdev_&location._group);
MSE_weight2 = (MSE_ANCOHET + MSE_Interaction)/2;
%end;

%if &location = x %then %do;
Xp_diff_Xbar_grp0 = (0 - xmean_grp0)*(0 - xmean_grp0);
Xp_diff_Xbar_grp1 = (0 - xmean_grp1)*(0 - xmean_grp1);
Xp_diff_Xbar_grp2 = (0 - xmean_grp2)*(0 - xmean_grp2);

MSE_ANCOHET = (sserror/dferror);
MSE_ANCOVA = (sserror+SS&location._group)/(dferror+df&location._group);
MSE_interaction = (SS&location._group/df&location._group);
MSE_weight1 =  (sserror+SS&location._group)/(dferror+df&location._group);
MSE_weight2 = (MSE_ANCOHET + MSE_Interaction)/2;
%end;

%if &location = xbar %then %do;
Xp_diff_Xbar_grp0 = (xbar - xmean_grp0)*(xbar - xmean_grp0);
Xp_diff_Xbar_grp1 = (xbar - xmean_grp1)*(xbar - xmean_grp1);
Xp_diff_Xbar_grp2 = (xbar - xmean_grp2)*(xbar - xmean_grp2);
%end;
```

```
%if &location = ca %then %do;
Xp_diff_Xbar_grp0 = (ca - xmean_grp0)*(ca - xmean_grp0);
Xp_diff_Xbar_grp1 = (ca - xmean_grp1)*(ca - xmean_grp1);
Xp_diff_Xbar_grp2 = (ca - xmean_grp2)*(ca - xmean_grp2);
%end;

ratio_grp0 = (1*Xp_diff_Xbar_grp0)/ssx_grp0;
ratio_grp1 = (.25*Xp_diff_Xbar_grp1)/ssx_grp1;
ratio_grp2 = (.25*Xp_diff_Xbar_grp2)/ssx_grp2;
ratio_01 = (((1*xmean_grp0) + (-.5*xmean_grp1) + (-
.5*xmean_grp2))**2)/(ssx_grp0+ssx_grp1+ssx_grp2);

samp_inv_other = (1/n_grp0) + (1/n_grp1_contrast);
samp_inv = (1/n_grp0) + (.25/n_grp1) + (.25/n_grp2);

Stderr_ANCOHET = sqrt(MSE_ANCOHET*(samp_inv + ratio_grp0 + ratio_grp1 +
ratio_grp2));
ANCOHET_variance=(stderr_ancohet)*(stderr_ancohet);
stderr_ancohet2 = sqrt(mse_ancohet)*sqrt((samp_inv + ratio_grp0 + ratio_grp1 +
ratio_grp2));
Stderr_ANCOVA = sqrt(MSE_ANCOVA*(samp_inv + ratio_01));
Stderr_Interaction = sqrt(MSE_interaction*samp_inv);
Stderr_Weight1 = sqrt(MSE_weight1*samp_inv + ratio_01);
Stderr_Weight2 = sqrt(MSE_weight2*samp_inv);

stderr_ancohet_chen=sqrt(ANCOHET_variance+chen_incr);
stderr_partial_chen=sqrt(ANCOHET_variance+partial_chen_incr);

%if &location ne x %then %do;
CV_ancohet=sqrt(finv(.95, dfgroup, dferror));
CV_ANCOVA=sqrt(finv(.95, dfgroup, (dferror+1)));
CV_Interaction=sqrt(finv(.95, dfgroup, dfXdev_&location._group));
CV_Weight1=sqrt(finv(.95, dfgroup, (dferror+dfXdev_&location._group)));
CV_Weight2=sqrt(finv(.95, dfgroup, ((dferror+dfXdev_&location._group)/2)));
%end;

%if &location = x %then %do;
CV_ancohet=sqrt(finv(.95, dfgroup, dferror));
CV_ANCOVA=sqrt(finv(.95, dfgroup, (dferror+1)));
CV_Interaction=sqrt(finv(.95, dfgroup, df&location._group));
CV_Weight1=sqrt(finv(.95, dfgroup, (dferror+df&location._group)));
CV_Weight2=sqrt(finv(.95, dfgroup, ((dferror+df&location._group)/2)));
%end;

y_lsmean_grp1_contrast=(y_lsmean_grp1 + y_lsmean_grp2)/2;
```

```
lsmeans_y_diff = y_lsmean_grp0 - y_lsmean_grp1_contrast;

ci_halfwidth_ancohet=CV_ancohet*Stderr_ANCOHET;
ci_lower_ancohet= lsmeans_y_diff-ci_halfwidth_ancohet;
ci_upper_ancohet=lsmeans_y_diff+ci_halfwidth_ancohet;
ci_width_ancohet=2*CV_ancohet*Stderr_ANCOHET;

ci_halfwidth_chen=CV_ancohet*stderr_ancohet_chen;
ci_lower_chen= lsmeans_y_diff-ci_halfwidth_chen;
ci_upper_chen=lsmeans_y_diff+ci_halfwidth_chen;
ci_width_chen=2*CV_ancohet*stderr_ancohet_chen;

ci_halfwidth_chen_partial=CV_ancohet*stderr_partial_chen;
ci_lower_chen_partial= lsmeans_y_diff-ci_halfwidth_chen_partial;
ci_upper_chen_partial=lsmeans_y_diff+ci_halfwidth_chen_partial;
ci_width_chen_partial=2*CV_ancohet*stderr_partial_chen;

ci_halfwidth_ANCOVA=CV_ANCOVA*Stderr_ANCOVA;
ci_width_ANCOVA=2*CV_ANCOVA*Stderr_ANCOVA;
ci_lower_ANCOVA= lsmeans_y_diff-ci_halfwidth_ANCOVA;
ci_upper_ANCOVA=lsmeans_y_diff+ci_halfwidth_ANCOVA;

ci_halfwidth_interaction=CV_Interaction*Stderr_interaction;
ci_width_interaction=2*CV_Interaction*Stderr_interaction;
ci_lower_interaction= lsmeans_y_diff-ci_halfwidth_interaction;
ci_upper_interaction=lsmeans_y_diff+ci_halfwidth_interaction;

ci_halfwidth_weight1=CV_Weight1*Stderr_weight1;
ci_width_weight1=2*CV_Weight1*Stderr_weight1;
ci_lower_weight1= lsmeans_y_diff-ci_halfwidth_weight1;
ci_upper_weight1=lsmeans_y_diff+ci_halfwidth_weight1;

ci_halfwidth_weight2=CV_Weight2*Stderr_weight2;
ci_width_weight2=2*CV_Weight2*Stderr_weight2;
ci_lower_weight2= lsmeans_y_diff-ci_halfwidth_weight2;
ci_upper_weight2=lsmeans_y_diff+ci_halfwidth_weight2;

%if &location = x %then %do;
MU_adj_g0=&mean + (&slope1*0);
Mu_adj_g1=&slope2*0;
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location = xbar %then %do;
MU_adj_g0=&mean + (&slope1*xbar);
Mu_adj_g1=&slope2*xbar;
```

```
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location = ca %then %do;
MU_adj_g0=&mean + (&slope1*ca);
Mu_adj_g1=&slope2*ca;
difference_population = mu_adj_g0 - mu_adj_g1;
%end;

%if &location ne x %then %do;
F_ANCOHET = (sscontrast/dfcontrast)/(sserror/dferror);
F_ANCOVA =
(sscontrast/dfcontrast)/((sserror+SSXdev_&location._group)/(dferror+dfXdev_&location
._group));
F_interaction=(sscontrast/dfcontrast)/(ssXdev_&location._group/dfXdev_&location._gro
up);
F_weight1=(sscontrast/dfcontrast)/((SSXdev_&location._group +
sserror)/(dfXdev_&location._group + dferror));
F_weight2 = (sscontrast/dfcontrast)/(MSE_weight2);
p_ANCOHET = 1-probf(F_ANCOHET,dfcontrast,dferror);
p_ANCOVA = 1-probf(F_ANCOVA,dfcontrast,(dferror+dfXdev_&location._group));
p_interaction=1-probf(f_interaction,dfcontrast,dfXdev_&location._group);
p_weight1=1-probf(F_weight1, dfcontrast, (dferror+dfXdev_&location._group));
p_weight2=1-probf(F_weight2, dfcontrast, ((dferror+dfXdev_&location._group)/2));
%end;

%if &location=x %then %do;
F_ANCOHET = (sscontrast/dfcontrast)/(sserror/dferror);
F_ANCOVA =
(sscontrast/dfcontrast)/((sserror+SS&location._group)/(dferror+df&location._group));
F_interaction=(sscontrast/dfcontrast)/(ss&location._group/df&location._group);
F_weight1=(sscontrast/dfcontrast)/((SS&location._group + sserror)/(df&location._group
+ dferror));
F_weight2 = (sscontrast/dfcontrast)/(MSE_weight2);
p_ANCOHET = 1-probf(F_ANCOHET,dfcontrast,dferror);
p_ANCOVA = 1-probf(F_ANCOVA,dfcontrast,(dferror+df&location._group));
p_interaction=1-probf(f_interaction,dfcontrast,df&location._group);
p_weight1=1-probf(F_weight1, dfcontrast, (dferror+df&location._group));
p_weight2=1-probf(F_weight2, dfcontrast, ((dferror+df&location._group)/2));
%end;

RejectHO_ANCOHET = (p_ancohet<=.05);
RejectHO_ANCOVA = (p_ANCOVA<=.05);
RejectHO_Interaction = (p_interaction<=.05);
RejectHO_Weight1 = (p_weight1<=.05);
RejectHO_Weight2 = (p_weight2<=.05);
```

```
Pop_ParamInCI_ancohet = (ci_lower_ancohet<difference_population &
ci_upper_ancohet>difference_population);
Pop_ParamInCI_ancohet_chen = (ci_lower_chen<difference_population &
ci_upper_chen>difference_population);
Pop_ParamInCI_ancohet_chenpart = (ci_lower_chen_partial<difference_population &
ci_upper_chen_partial>difference_population);
Pop_ParamInCI_ANCOVA = (ci_lower_ANCOVA<difference_population &
ci_upper_ANCOVA>difference_population);
Pop_ParamInCI_interaction = (ci_lower_interaction<difference_population &
ci_upper_interaction>difference_population);
Pop_ParamInCI_weight1 = (ci_lower_weight1<difference_population &
ci_upper_weight1>difference_population);
Pop_ParamInCI_weight2 = (ci_lower_weight2<difference_population &
ci_upper_weight2>difference_population);
run;

%if &output= TypeI %then %do;
proc freq data=ci_info;
title1 "Type I Error Rates for";
title2 "b0=&slope1, b1=&slope2, b2=&slope3";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2, n2=&sampsize3";
title5 "at &location";
table RejectHO_ANCOHET RejectHO_ANCOVA RejectHO_Interaction
RejectHO_Weight1 RejectHO_Weight2/nocum;
run;
%end;

%if &output= Power %then %do;
proc freq data=ci_info;
title1 "Power Rates for";
title2 "b0=&slope1, b1=&slope2, b2=&slope3";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2, n2=&sampsize3";
title5 "at &location";
table RejectHO_ANCOHET RejectHO_ANCOVA RejectHO_Interaction
RejectHO_Weight1 RejectHO_Weight2/nocum;
run;
%end;

%if &output=ci %then %do;
proc freq data=ci_info;
title1 "Confidence Interval Coverage Rates for";
title2 "b0=&slope1, b1=&slope2, b2=&slope3";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2, n2=&sampsize3";
```

```
title5 "at &location";
table pop_paraminCI_ancohet pop_paraminCI_interaction pop_paraminCI_ancova
pop_paraminCI_weight1 pop_paraminCI_weight2 Pop_ParamInCI_ancohet_chen
Pop_ParamInCI_ancohet_chenpart/nocum;run;
%end;

%if &output=ci_width %then %do;
proc means data=ci_info n mean stddev min max;
title1 "Average Confidence Interval Width for";
title2 "b0=&slope1, b1=&slope2, b2=&slope3";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2, n2=&sampsize3";
title5 "at &location";
var ci_width_ancohet ci_width_interaction ci_width_ancova  ci_width_weight1
ci_width_weight2 ci_width_chen ci_width_chen_partial;
run;
%end;

%if &output=trueSD %then %do;
proc means data=ci_info n mean stddev min max maxdec=4;
title1 "True SD and Average Standard Errors for";
title2 "b0=&slope1, b1=&slope2, b2=&slope3";
title3 "and";
title4 "n0=&sampsize1, n1=&sampsize2, n2=&sampsize3";
title5 "at &location";
var lsmeans_y_diff Stderr_ANCOHET Stderr_Interaction Stderr_ANCOVA
Stderr_Weight1 Stderr_Weight2 stderr_ancohet_chen;
run;
%end;


%mend three_group_simulation;
```

Running the previous code does not generate output, but instead establishes a macro program that will run when SAS receives the proper input. The following line of code will generate and output the results for the true standard deviation in the two-group, low heterogeneity of regression condition, where $n = 10$ and the test is being conducted at $\bar{X}$.

```
%two_group_simulation(10000, .258, .374, 10, 10, 0, xbar, trueSD, 600);
```

Appendix D

An Attempted Replication of Harwell and Serlin (1988)

The current research project was partially motivated by the findings of Harwell and Serlin (1988). Harwell and Serlin investigated the Type I error rates and power of several methods – both parametric and nonparametric - of analyzing data that include a quantitative and qualitative predictor of a continuous outcome. While their paper focused on multiple distributions for the dependent variable, the current paper focused solely on the results where the outcome was normally distributed. They found that the Type I error rates of the Rogosa approach were significantly greater than the nominal .05 for different combinations of sample size (both equal and unequal) and levels of heterogeneity. In fact, all eight combinations produced Type I error rates significantly greater than what would be expected due to sampling error when $\alpha = .05$ (see Table 8 on p. 275).

Because the current project was based on the work of Harwell and Serlin, the first step was to replicate their results. The following sections contains these findings, and is followed by a section regarding Harwell and Serlin's purported use of standardized regression coefficients.

Harwell and Serlin Method Overview

Along with the normal distribution, Harwell and Serlin generated data under exponential, double exponential and Cauchy distributions. They used combinations of two group sizes (n = 10, 30) and two sets of regression coefficients to model between group heterogeneity of regression for three groups: .2, .2, .9 and .9, .9, .2. There were eight combinations of slopes and sample sizes in total. They performed 2,000 simulations

for each combination, and Type I error rates were calculated as the total number of simulations out of 2,000 where the *p* value would have resulted in the rejection of the null hypothesis given the specified nominal α (they looked at α = .01, .05 and .10). The standard error of a proportion was used to calculate sampling error for each α used the following equation:

$$SE_p = \sqrt{\frac{pq}{N}} \quad \text{(D.1)}$$

where p is the nominal α, $q = 1 - p$, and N is the number of replications of the simulation. Any Type I error rate greater than 2 standard errors above or below α was considered to be outside of the range specified by sampling error.

For combinations of regression heterogeneity and sample sizes where the Type I error rate was below or within the sampling error around nominal α, they also looked at the power of the test. Though, as mentioned previously, for the Rogosa procedure all eight of the combinations had Type I error rates significantly greater than what would be expected when $\alpha = .05$. As a result, power was not calculated for the Rogosa procedure for any of the combinations when $\alpha = .05$. In fact, for the data where the outcome followed the normal distribution, only one condition out of 18 had a Type I error rate within sampling error of the nominal α.

Attempted Replication of Harwell and Serlin

The current paper attempted an exact replication of the Harwell and Serlin findings. Initially, it was discovered that regardless of the heterogeneity of regression level used (i.e., .2, .2, .9 vs. .9, .9, .2), the Type I error rates were equal for each

combination of sample size when the same starting seed value was used for the random number generator. This was not the case for Harwell and Serlin, who likely used different seed values for each of the conditions. As a result, the current paper presents results only from one of the two levels of heterogeneity of regression. Consequently, the Type I error rates from Harwell and Serlin were averaged for each combination of sample size in order to making comparisons with the current simulation.

Table D1 contains the Type I error rates from the original Harwell and Serlin article, combined with the error rates from the current project. There are also upper and lower limits around each set of Type I error rates. The Harwell and Serlin upper and lower limits were based on 4,000 simulations whereas the current study's limits were based on 10,000 simulations. The Type I error rates from each study were compared in two ways: first, it was determined if the Harwell and Serlin average fell within the confidence interval for the current study. Subsequently, the overlap between the confidence intervals from each of the studies was evaluated. As seen in the table, all confidence intervals from the current student did not contain $\alpha = .05$. Additionally, the confidence intervals from each study overlapped, but the Harwell and Serlin Type I error rate average fell within only two of the four confidence intervals for the current study. In conclusion, while Harwell and Serlin's Type I error rates may have been slightly higher than what the current study found in the case of equal-*n*, the overlap of the confidence intervals from both studies suggests that the Rogosa procedure is indeed a liberal test.

Use of Standardized Regression Coefficients

Until this point, it was assumed that Harwell and Serlin's coefficients of .2 and .9 were the raw regression coefficients used to simulate the data. With a more careful reading, however, Harwell and Serlin claim that they used standardized regression coefficients of .2 and .9 to produce their simulated data (see simulation factor (b) in the first full paragraph on pg. 272). However, this seems highly unlikely given results from preliminary work undertaken for the current study. In order to calculate the raw regression coefficients to produce standardized regression weights of these values, one method involves using the standard deviation of *Y* computed separately in each group. It is necessary to use the standard deviation of *Y* separately in each group because, as mentioned previously in the body of this paper, the standard deviation of *Y* will not be the same across groups in order to meet the assumption of heterogeneity of residual variances.

In general, the formula for a standardized regression coefficient is:

$$\beta_k = b_k \times \frac{s_{x_k}}{s_y} \tag{D.2}$$

Where β_k and b_k are the *k*th standardized and unstandardized regression coefficients, respectively, and s_{x_k} and s_y are the standard deviations of the *X* and *Y* variables. Given $\sigma_\varepsilon = 1$ and $\sigma_x = 1$ (based on the simulation set-up), one can solve for the raw slopes to achieve particular standardized slopes as follows:

$$Known: \rho_{yx} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{D.3}$$

$$Known: \sigma_y^2 = b_1^2\sigma_x^2 + \sigma_\epsilon^2 = b_1^2(1) + 1 = 1 + b_1^2 \tag{D.4}$$

$$Known: cov(x, y) = b_1\sigma_x^2 = b_1(1) = b_1 \quad \text{(D.5)}$$

Combining D.2, D.3 and D.4 after squaring the terms in D.2 produces

$$\rho_{yx}^2 = \frac{b_1^2}{1^2(1+b_1^2)} = \frac{b_1^2}{1+b_1^2} \quad \text{(D.6)}$$

Using equation D.5, if ρ_{yx} = .9, then solving for b_1 results in a raw regression coefficient of b_1 = 2.06474. Similarly, the same equation can be used to solve for the necessary raw regression coefficient to produce ρ_{yx} = .2, resulting in b_2 = 0.204124.

Table D2 contains the Type I error rates using raw regression coefficients to produce standardized coefficients of .2, .2, and .9. As can be seen, these Type I error rates are significantly greater than the rates Harwell and Serlin found, making it unlikely that their study used standardized regression coefficients.

Conclusions

The current study partially replicated the findings of Harwell and Serlin (1988) regarding Type I error rates for testing the main effect of a qualitative grouping variable in the presence of heterogeneity of regression. However, despite the claims that their simulations were conducted using standardized regression coefficients of .2 and .9, the current study found this claim to be highly unlikely. Instead, standardized regression coefficients would have produced Type I error rates significantly higher than what Harwell and Serlin originally found.

Table D1

Comparing Type I Error Rates for Harwell and Serlin (1988) and the Current Replication Attempt

		Harwell and Serlin			**Using .2, .2, .9 as unstandardized coefficients to generate data, as suggested by Harwell & Serlin (1988, Equation 10, p. 273)**					
					McLouth					
B1, B2, B3	N	Average based on 4,000 Simulations	Lower Limit	Upper Limit	10,000 Simulations	Lower Limit	Upper Limit	Does Interval Contain .05?	Does Interval Contain H&S Average?	Do Intervals Overlap?
.2, .2, .9	10,10,10	0.072	0.06525	0.07875	0.0641	0.05983	0.06837	NO	NO	YES
	10,10,30	0.071	0.06425	0.07775	0.0728	0.06853	0.07707	NO	YES	YES
	30,30,10	0.079	0.07225	0.08575	0.0767	0.07243	0.08097	NO	YES	YES
	30,30,30	0.068	0.06075	0.07425	0.0619	0.05763	0.06617	NO	NO	YES

Table D2
Type I Error Rates Using Raw Regression Coefficients to Produce Standardized Coefficients of .2, .2, .9

B1, B2, B3	*n*	Type I Error Rate
.204, .204, 2.0647	10,10,10	0.180
	10,10,30	0.238
	30,30,10	0.232
	30,30,30	0.169