Estimating Individual-Level Interaction Effects in Multilevel Models: A Monte Carlo Simulation Study with Application

Julie A. Lorah

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Elizabeth A. Sanders, Chair

Elizabeth Litzler

Deborah McCutchen

Robert Abbott

Program Authorized to Offer Degree:

Educational Psychology

University of Washington

Abstract

Estimating Individual-Level Interaction Effects in Multilevel Models: A Monte Carlo Simulation Study with Application

Julie A. Lorah

Chair of the Supervisory Committee:
Assistant Professor Elizabeth A. Sanders
Measurement and Statistics, Educational Psychology

Moderated multiple regression (MMR) provides a useful framework for understanding moderator variables. When the relationship between an independent variable (predictor) and a dependent variable (criterion) varies as a function of a third variable, that third variable is considered a moderator of the predictor-criterion relationship. Moderated relationships can also be examined within datasets that include nesting, for example individuals might be nested within groups. However, the literature is not clear on the best way to assess data for significant moderating effects, particularly within a multilevel modeling framework. This study explores potential ways to test moderation at the individual level (level 1) within a 2-level multilevel modeling framework, with varying effect sizes, cluster sizes, and numbers of clusters that represent realistic conditions in applied educational research. The study examines five potential methods for testing interaction effects: the Wald test, *F*-test, likelihood ratio test, BIC, and AIC. For each method, the simulation study examines how Type I error rates vary as a function of number of clusters and cluster size and how power varies as a function of number of clusters, cluster size, and interaction effect size. Following the simulation study, an applied study uses real data to assess interaction effects using the same five methods. Results indicate that the Wald test, *F*-test, and likelihood ratio test all perform similarly in terms of Type I error rates and power. Type I error rates for the AIC are more liberal, and for the BIC typically more conservative. A four-step procedure for applied researchers interested in examining interaction effects in multi-level models is provided.

Table of Contents

List of Figures

List of Tables

Chapter 1: Review of Literature

Researchers may want to test hypotheses about how a given relationship varies among different groups. For example, Litzler, Samuelson, and Lorah (2014) examined the relationship between various covariates, such as student perception of professors and student confidence. It would be interesting to test whether this relationship differs by group membership. For example, perhaps the relationship is stronger for women than for men. This would indicate that the relationship between perception of professors and confidence is moderated by gender. In other words, an interaction between perception of professors and gender is present. Further, the data for this study (Litzler et al., 2014) included students nested within multiple schools, indicating that a multilevel model would be appropriate, given the non-independence of individual student observations. Moderator hypotheses such as these can be examined with moderated multiple regression models and this can be done within a multilevel framework for data that includes nesting. The results of these types of models can represent important and relevant contributions to the literature. This study will investigate methods for testing multilevel moderated regression models with interaction effects at the individual level only (or more generally, at level-one only). Within this study, this model includes a continuous dependent variable, a continuous independent variable, and a binary moderator variable. Methods for testing this interaction effect will be evaluated on Type I error rates and power based on a simulation study and an applied study will follow to illustrate the technique. An introduction to moderation, multilevel models, and testing for interaction effects follows in this chapter.

Moderated Multiple Regression (MMR)

In order to assess the relationship between one dependent variable and one or more independent variables, a researcher can use regression analyses (e.g. Tabachnick & Fidell, 2007, p. 117). Some researchers may further be interested in if and how the relationship between a given independent variable (IV) and a dependent variable (DV) varies as a function of a third variable. This third variable is referred to as a moderator variable (MV) because it moderates the relationship between the IV and the DV and these questions can be addressed with moderated multiple regression (MMR) models (Aiken & West,

1991; Jose, 2013). The MMR model includes a product term, the IV multiplied with the MV, as an additional predictor making these models more complex than simple linear regression models.

Although interactions between two discrete variables are commonly covered in the ANOVA literature, there is less literature examining continuous IV and/or MV variables (Aiken & West, 1991) although it does exist (see Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990; Jose, 2013). Complex hypotheses that cannot be tested with simple linear models are common across many disciplines (Aiken & West, 1991). However, researchers often incorrectly test these hypotheses with simple additive regression models, omitting the correct higher-order terms, which does not test the moderation hypothesis (Aiken & West, 1991). Other times, researchers may choose to dichotomize the IV and MV in order to test the interaction within an ANOVA framework, which results in a substantial loss of power (Aiken & West, 1991). Researchers have found that it is rare for substantive research to use the correct test for hypotheses regarding moderation rather than one of the incorrect methods just described (Aiken & West, 1991).

Typically, to test for a significant interaction effect, an *F*-test that assesses the change in R^2 value for two regression models is used (Aiken & West, 1991; Jaccard et al., 1990) which essentially tests whether the more complicated model (including the product term) explains significantly more variance in the outcome compared with the simpler model (not including the product term). If a significant interaction effect is found, the next step will be interpretation. The coefficient for an interaction effect is not as easily interpretable as those for main effects, and so researchers should typically graph the interaction in order to interpret it appropriately (Jose, 2013).

There are many topics and fields that make substantial use of moderation including research examining aptitude-by-treatment interactions (DeShon & Alexander, 1996; Dretzke, Levin, & Serlin, 1982), gender difference studies (DeShon & Alexander, 1994), differential prediction studies (Culpepper & Davenport, 2009; DeShon & Alexander, 1996), differential validity studies (Linn, 1987), industrial/organizational psychology (Champoux & Peters, 1987), and the job design literature (Champoux & Peters, 1987). It may be easier for researchers in fields with more mature bodies of

literature to hypothesize about complex relationships such as moderation. For example, a study examining the moderating effects of maladaptive perfectionism on the relationship between adult attachment and depressive mood commented on the recent trend within their field of moving past the "examination of simple bivariate linear relationships" linking insecurity and distress to these more complex moderated relationships (Wei, Mallinckrodt, Russell, & Abraham, 2004).

Another example of a study using MMR examined the relationship between the mean group midterm scores (aptitude, the IV) and group article critique scores (outcome, the DV) for several groups of graduate students (Onwuegbuzie, Collins, & Elbedour, 2003). They were interested in whether the treatment (MV), which was the amount of heterogeneity (variance) in the group's midterm scores moderated the relationship between aptitude and the critique score. They found evidence of an aptitude-by-treatment interaction where the group heterogeneity didn't impact the article critique scores for high aptitude groups, but low aptitude groups that were heterogeneous scored higher on the article critique than the groups that were not heterogeneous. In other words, the relationship between aptitude (IV) and article critique score (DV) depends on whether the group is heterogeneous or not (MV).

MMR methods are also used in examining differential validity and differential prediction (Culpepper & Davenport, 2009; Linn, 1978). Differential validity assesses whether the magnitude of association between IVs and DVs (for example, between SAT score and college GPA) differs by subgroup and is typically assessed by looking for significant differences in correlation coefficients by group. Differential prediction examines whether a single regression equation is appropriate for all subgroups (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008). For example, one differential prediction study (Culpepper & Davenport, 2009) used MMR to find that the relationship between high school achievement measures (high school GPA and SAT scores) and college GPA differed by race/ethnicity. Specifically, given the same level of high school achievement, African American students are predicted to have lower average college achievement compared with White students. In other words, high school achievement (IV) interacts with race/ethnicity (MV) in predicting college GPA (DV) meaning that one prediction equation for all racial/ethnic groups would not be appropriate. Understanding these prediction

equations is critical since postsecondary institutions use them for high-stakes judgments, such as admission decisions (Culpepper & Davenport, 2009).

In addition to examining substantive research questions, MMR can also be used to assess whether the assumption of homogeneity of regression slopes is tenable before proceeding with an analysis of covariance (ANCOVA; DeShon & Alexander, 1994; Dretzke et al., 1982; Lomax, 2007; Tabachnick & Fidell, 2007). This assumption states that the slope of the regression line for all groups is equal (Lomax, 2007) meaning that the IV doesn't interact with the covariate (Tabachnick & Fidell, 2007).

For example, Barrett & Fish (2011) investigated whether participation in a 30-week chess program had a relationship with middle school special education students' math achievement. They conducted an experiment where classrooms were randomly assigned to either a treatment group which participated in the chess program or a control group which did not participate in the chess program. To analyze this data, the researchers conducted an analysis of covariance (ANCOVA) with math achievement as the DV and experimental condition as the IV. They also wanted to control for previous math achievement, so they included a math pretest variable as a covariate in the model. In order to assess the homogeneity of slopes assumption for ANCOVA, they would need to test whether the regression slope for pretest (IV) on posttest scores (DV) was the same for both the experimental and control group (MV). This could be investigated with a MMR where the desired outcome would be a non-significant interaction between the IV and MV. Although within a hypothesis testing framework a non-significant finding could indicate either absence of a true interaction or low power, other methods such as information criteria methods have a less ambiguous interpretation for a non-significant finding.

It is interesting to note that when the question of parallel slopes within a MMR model is of substantive interest, the goal of the researcher is typically to reject the null hypothesis of equal slopes, whereas when the goal is to assess the tenability of the ANCOVA assumption of homogeneity of regression slopes, the goal is to find evidence suggesting the slopes are not different. Since the hypothesis testing framework is not an ideal method for providing evidence for a null hypothesis (i.e.

equal slopes), in the case of assumption tenability, it appears an information criteria approach, which is designed to provide evidence either for or against a null hypothesis, may be preferable.

The method that researchers refer to as moderated multiple regression (MMR; Morris, Sherman, & Mansfield, 1986) has been referred to by various other names in the literature, including interaction effects in multiple regression (Jaccard et al., 1990) and moderating effects (Dunlap & Kemery, 1987) when the MV is continuous or discrete. When referring to MMR specifically with a discrete MV, researchers have called this regression slope differences (Alexander & DeShon, 1994), homogeneity of two straight lines (Bofinger, 1999), regression slope equality (DeShon & Alexander, 1994), regression slope homogeneity (DeShon & Alexander, 1996), regression homogeneity, or parallelism of K independent regression lines (Dretzke et al., 1982), and comparing two straight-line regression equations (Kleinbaum & Kupper, 1978). This study will continue to refer to this method as MMR and to the effects themselves as interaction effects.

Assumptions and Difficulties with MMR

In order to provide evidence for significant interaction effects, an *F*-test on the change in R^2 values may be conducted. There are multiple assumptions associated with this test, including independence of errors, independence of errors with the IV, a linear relationship between the DV and the IVs, normality of the conditional distribution of Y within each level of the MV, and equality of error variances within each level of the MV (DeShon & Alexander, 1996). Researchers have investigated the effects of violation of some of these assumptions, including violation of error variance homogeneity (Alexander & DeShon, 1994; DeShon & Alexander, 1994; DeShon & Alexander, 1996; Dretzke et al., 1982; James, 1951). Researchers studying MMR have investigated topics including multicollinearity (Dunlap & Kemery, 1987; Jaccard et al., 1990); issues of low power (MacCallum & Mar, 1995; Morris et al., 1986), and the effects of measurement error on MMR models (Aiken & West, 1991; Baron & Kenny, 1986; Holmbeck, 1997).

Researchers have suggested various modified procedures for the *F*-test in the presence of heterogeneity of error variance (DeShon & Alexander, 1994; Dretzke et al., 1982). Another modified

procedure was suggested to address the low power inherent in testing interactions (Morris et al., 1986) although this particular procedure involving the use of principal components regression was subsequently shown to be faulty (Cronbach, 1987; Dunlap & Kemery, 1987).

Researchers have discussed the effects of a possible large correlation coefficient between the product term (i.e. interaction term) and its component variables which are the IV and the MV (Aiken & West, 1991; Holmbeck, 1997). This high correlation can be introduced within the MMR model when the product term is created from uncentered IV and MV terms (Aiken & West, 1991; Holmbeck, 1997) and has caused some researchers to point to multicollinearity between the product term and one or both of its component variables as a problem with moderation models (Morris, Sherman, & Mansfield, 1986). Centering IV and MV variables will reduce the correlation between the product term and its component variables which will help to avoid technical difficulties in estimating an MMR model (Aiken & West, 1991) although this may be less of an issue since computing power and the sophistication of estimation algorithms may have increased in the intervening years. Although some researchers have claimed that multicollinearity contributes to the low power typically observed in detecting significant interactions (Morris et al., 1986), further research has indicated that multicollinearity is not related to power since additive transformation (i.e. centering) of the MV and IV predictors has been shown to reduce the correlation between the product term and its component variables while having no effect on the interaction coefficient itself (Aiken & West, 1991; Dunlap & Kemery, 1987; Holmbeck, 1997; Jaccard et al., 1990). In addition to possible benefits in terms of ease of estimation, centering of predictors is a good idea for ease of interpretation of main effects (Aiken & West, 1991).

The *F*-test for interaction effects has been shown to generally have low power (Aiken & West, 1991; Alexander & DeShon, 1994; Holmbeck, 1997; Jose, 2013; MacCallum & Mar, 1995) with continuous as well as categorical variables (Aiken & West, 1991). A reduction in power also occurs in the presence of measurement error (Aiken & West, 1991; Holmbeck, 1997; Jose, 2013). In fact the decrease in power for interaction effects due to measurement error is larger than the decrease for main effects (Aiken & West, 1991). Some researchers have attempted to address the issue of measurement error with

latent variable models that utilize multiple indicators of a given construct (Aiken & West, 1991; Jaccard et al., 1990). Low power is further exacerbated in the presence of violation of the assumption of error variance homogeneity (Alexander & DeShon, 1994).

Moderation versus Mediation: Quick Clarification

Although moderation and mediation are conceptually and statistically distinct, they are often confused in practice (Baron & Kenny, 1986; Holmbeck, 1997; Jose, 2013). This study will focus on moderation; a brief discussion of mediation is provided to clarify the distinction. A moderator is defined as "a qualitative (e.g., sex, race, class) or quantitative (e.g. level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (Baron & Kenny, 1986, p. 1174). A mediator "accounts for the relation between the predictor and criterion" (Baron & Kenny, 1986, p. 1176) which implies that mediators may be thought of in causal terms, whereas moderators are not (Baron & Kenny, 1986; Jose, 2013). In other words different levels of a moderator variable imply different relationships between the IV and DV whereas with a mediator, there is a causal path from the IV to the mediator, and then from the mediator to the DV. Mediation and moderation are also historically distinct: mediation derives from the path modeling framework, whereas moderation was introduced within the ANOVA framework (Jose, 2013).

In addition, it is possible to combine the two techniques. Researches can use one variable as both a mediator and moderator within separate models (Holmbeck, 1997; Wei et al., 2004), such as including interactions within path models which could be considered mediated moderation (Jose, 2013), or seeing if a mediational model functions differently for different groups which could be considered moderated mediation (Holmbeck, 1997; Jose, 2013).

Introduction to Multilevel Models

As this study examines MMR in the context of multilevel data, an introduction to multilevel models is provided. Many datasets have a multilevel structure, for example, students nested within schools; repeated measurements nested within subjects; or animals nested within litters (Snijders & Bosker, 1999). This nesting of individuals within these clusters implies that observations are not

independent of one another, violating the independence assumption of many statistical procedures, including linear regression procedures. Ignoring the nesting while conducting analyses is not recommended as it underestimates standard errors and leads to errors in inference (O'Connell & McCoach, 2008). Although this nesting may be simply an added difficulty for data analysis, the nesting of observations within clusters can also be substantively interesting. Multilevel modeling provides one option for addressing non-independence of observations and for further investigation of effects at multiple levels (O'Connell & McCoach, 2008; Snijders & Bosker, 1999). Within this study, the individual measurements will be referred to as level-one or individual-level (for example, students) and the clusters will be referred to as level-two (for example, schools).

For example, a researcher might be interested in the relationship between Catholic school attendance and math achievement after controlling for socio-economic status (SES). Data from the students within this dataset would be clustered within schools and this research question could be investigated with a multilevel model (Raudenbush & Bryk, 2002). To estimate the parameters for these models, maximum likelihood estimation methods are typically used which are asymptotic, meaning they are unbiased for large sample sizes (Hox, 2010). Researchers typically begin a study such as this by estimating the empty model (model with no covariates), which is equivalent to a random effects analysis of variance and serves to partition the variance between student-level variance and school-level variance (Snijders & Bosker, 1999). With equal cluster sizes, the ANOVA variance partition estimates are equivalent to the residual maximum likelihood (REML) variance partition estimates; however these two estimates may differ from the maximum likelihood (ML) estimate and from each other when cluster sizes are not equal (Snijders & Bosker, 1999). The estimates for variance components may be more accurate for larger sample sizes when using these asymptotic estimators (Hox, 2010) which indicates that the ANOVA partition of variance may be preferred for small sample sizes. Researchers have shown that with REML estimation, approximately 6-12 level-two clusters are needed, and with ML estimation, approximately 48 level-two clusters are needed (Hox, 2010).

From the empty model results, the researcher can compute the intraclass correlation coefficient (ICC) which represents the degree of similarity between level-one units among a given group and can be interpreted as the proportion of variance in the outcome accounted for by cluster membership (Snijders & Bosker, 1999). The formula to compute the ICC is given by Snijders and Bosker (1999) as:

$$\mathrm{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2} \quad (1)$$

Where τ^2 is the between-cluster variance and σ^2 is the within-cluster variance (Snijders & Bosker, 1999). After estimation of the empty model, additional models with explanatory variables at level-one and/or level-two can be estimated (Snijders & Bosker, 1999).

Within a multilevel model, an interaction effect could be examined at level-one, at level-two, or as a cross-level interaction. A cross-level interaction is an interaction between a level-one and a level-two variable and can occur as a result of including a level-two predictor of a random slope coefficient (Snijders & Bosker, 1999). Cross-level interactions may be included on theoretical grounds or as a way to explain variance after identifying significant randomly varying slopes (Snijders & Bosker, 1999).

Prevalence of Methods

In order to provide some sense of the prevalence of the use of moderated multiple regression (MMR) and multilevel modeling in education research, all articles published in the *American Educational Research Journal (AERJ)* during 2014 were examined by the author. This journal was specifically selected for examination due to its wide readership across subfields of education as well as its rigorous standards for both quantitative and qualitative methodologies. The six issues published in 2014 each contained six articles, for a total of36 articles. Of these, 25 (69%) used quantitative research methods for all or part of the study. Of the 25 quantitative articles, five used MMR (20%), seven used multilevel modeling (28%), and two used both multilevel modeling and moderation (8%).

The first of the two studies that used MMR within a multilevel modeling framework (Cooper, 2014) focused on predicting learning engagement for students who were nested within classrooms. For this study, three interaction terms between predictors at the individual level were included in the model.

Significance of the interactions was assessed based on coefficient *t*-test *p*-values less than .05 (referred to as a Wald test within the present study). The second study (Rimm-Kaufman et al., 2014) uses a moderated mediation model which is estimated within a structural equation modeling framework. This study specifically evaluated the impact of a teaching approach intervention on student achievement through a randomized controlled trial. Students in this study were nested within 24 elementary schools. Although there are many possible conditions for testing moderation within a multilevel modeling framework, the present study will focus on providing baseline data on methods for evaluating moderators in multilevel modeling by focusing on interactions specified at the individual level only.

Formally Testing an Interaction Moderator Effect

Although model selection should be guided by theory, observed data is also used in the model selection process (McCoach & Black, 2008). The question of whether or not to include an interaction term in a given model can be viewed from a model selection perspective. In order to test the significance of any fixed effect, including an interaction effect, various procedures can be used. Significance of the slope coefficient of the interaction term itself may be directly assessed with a Wald test. Analogously, comparison of the model fit of two models can be used. The first model includes only the main effects and the second includes the addition of the product term. These models will be referred to as the main effects only model (equation 2) and the full model (equation 3), respectively, and are specified as follows:

$$Y = B_0 + B_1*X + B_2*M + e \quad (2)$$

$$Y = B_0 + B_1*X + B_2*M + B_3*X*M + e. \quad (3)$$

Where Y is the dependent variable; X is the independent variable; X*M is the product term which is created by multiplying X times M; and e is the residual term. The product term represents the interaction effect in a model that also includes both main effects. The model comparison of these two models (equations 2 and 3) tests the significance of the interaction effect and may be conducted with an *F*-test, likelihood ratio test, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC) model fit indices. These five methods will be examined in this study, as these methods represent the most

common model selection methods including both hypothesis testing and information criteria approaches (McCoach & Black, 2008).

In all cases, any model that includes product terms should also include the main effects as well. This is necessary so that the results will actually represent the moderation hypotheses being tested (Aiken & West, 1991). Without the main effects, the product term would no longer represent an interaction effect.

It is unclear which of these five methods for testing the interaction effect works best for controlling Type I error or for maximizing power within a multilevel model. In order to understand more about these specific tests, the following sections will describe each of the five tests separately.

Wald test. To assess the significance of any fixed effect, a *t*-test can be conducted where the null hypothesis is that the given parameter is equal to zero. The test statistic for this test is computed by dividing the parameter estimate by its standard error and it is distributed approximately as a *t*-distribution (Snijders & Bosker, 1999) although specifying the degrees of freedom for this test is more complicated within a multilevel framework due to the presence of data at multiple levels (Snijders & Bosker, 1999). This test is also referred to as a Wald test (Hox, 2010; Snijders & Bosker, 1999). In order to use the Wald test to test for a significant interaction, the researcher can assess significance of B_3, which is the slope coefficient for the product term in the full model (Aiken & West, 1991; Baron & Kenny, 1986; Jose, 2013). This method is suggested by some researchers specifically addressing moderation (Aiken & West, 1991; Baron & Kenny, 1986; Jose, 2013) and has been used in substantive research for assessing interaction effects within multilevel models (Culpepper & Davenport, 2009).

For hypotheses involving multiple fixed-effect parameters, a multivariate Wald test can be used. This test statistic uses the standard errors of the parameter estimates as well as their covariances and is distributed approximately as a chi-square distribution (Snijders & Bosker, 1999). The Wald test can be used for fixed but not random parameter estimates (Snijders & Bosker, 1999) and the standard errors used in this test are asymptotic, meaning they are valid for large samples (Hox, 2010).

The presence of data at multiple levels complicates the process of defining the number of degrees of freedom for the *t*-test (Snijders & Bosker, 1999), but the Wald test for level-one fixed effects can be approximated with a *z*-test rather than a *t*-test if the degrees of freedom (level-one sample size minus number of explanatory variables minus one) is larger than about 40 (Snijders & Bosker, 1999). Therefore, for larger sample sizes computation of degrees of freedom is not an issue, since the *z*-test does not use degrees of freedom. For the smaller sample sizes within a multilevel model there a few different methods for computing the degrees of freedom that have been suggested, including using the sample size of level-two units or using the Sattertwaithe approximation (Hox, 2010). An alternative method for computing degrees of freedom is used by the HLM software program, which is the number of level-two units minus the number of parameters minus one (Hox, 2010). Some researchers argue that the Sattertwaithe approximation, which uses the values of the residual variances, is preferable to the HLM software's approach and works best when the sample size is less than 30 (Hox, 2010).

***F*-test.** Another method to assess the significance of the interaction term in the MMR model is an *F*-test that assesses the change in R^2 value for the main-effects-only model and the full model (Aiken & West, 1991; DeShon & Alexander, 1996; Jaccard et al., 1990) which essentially tests whether the more complicated model explains significantly more variance in the outcome. Jaccard et al., (1990, p. 18) provide the formula for *F*:

$$F = \frac{(R_2^2 - R_1^2) \,/\, (k_2 - k_1)}{\frac{(1 - R_2^2)}{N - k_2 - 1}} \tag{4}$$

Where R_2^2 indicates the R^2 for the full model (equation 3); $R_1^2$is the R^2 for the main effects only model (equation 2); k_2 and k_1 are the number of parameters in the full model and main effects only models respectively; and *N* is the total sample size.

If the *F*-test is significant, it suggests that the relationship between the IV and DV depends on the MV; in other words the MV moderates the IV-DV relationship. If the *F*-test is not significant, it suggests

that the IV-DV relationship is independent of the MV, given sufficient power. For this interaction effect test, the test of significance of B_3 (Wald test) is equivalent to the *F*-test on the change in R^2 value (DeShon & Alexander, 1996; Paunonen & Jackson, 1988). The *F* value is the square to the Student *t* statistic (Champoux & Peters, 1987; Jaccard et al., 1990) which is used in the Wald test specified previously.

The formula for *F* requires computation of R_1^2 and R_2^2. Model adequacy may be assessed with the R^2 statistic, which represents the explained proportion of variance in a multiple regression (Snijders & Bosker, 1999). However, since multilevel models include variance at two or more levels, the concept of R^2 becomes problematic (McCoach & Black, 2008; Snijders & Bosker, 1999). The *F*-test is not typically cited as a model comparison procedure in the multilevel modeling literature. Another complication arises because sample size computation is unclear due to having data at multiple levels (Snijders & Bosker, 1999), and so it is not clear how to compute *N* for this *F*-test as well.

Some researchers suggest this *F*-test as an alternative to the Wald test for testing global hypotheses (assessing more than one term simultaneously) (Aiken & West, 1991). Although the test is equivalent to the Wald test for the model specified above (DeShon & Alexander, 1996; Paunonen & Jackson, 1988), it is not clear how the *F*-test would be computed for a multilevel model. Both R^2 and *N* become ambiguous in a multilevel model. Snijders and Bosker (1999) have suggested one way to compute R^2 with multilevel models based on the estimated level-one and level-two variances. However, these are estimates of variance, rather than a simple partitioning of variance. Although random effects at the lowest level (individual-level) of the model have been shown to be fairly accurate (Hox, 2010), estimates for level-two variance can be downwardly biased with smaller samples (Hox, 2010). For example, simulation research has shown that more than 100 groups would be needed for accurate level-2 variance estimates (Hox, 2010). So although it seems that *F* computed as a function of the sum of squares (based on the ANOVA/ANCOVA models) would be analogous to the square of the *t* statistic, computing *F* based on the R^2 formulas suggested by Snijders and Bosker (1999) might lead to a different value for *F*.

Snijders and Bosker (1999) provide a formula to compute R^2 for a multilevel model (equation 5). This quantity represents the proportional reduction in prediction error at the individual level and can be computed as:

$$R^2 = 1 - \frac{\sigma_F^2 + \tau_F^2}{\sigma_E^2 + \tau_E^2} \quad (5)$$

Where σ_F^2 represents the level-one variance for the full model; τ_F^2 represents the level-two variance for the full model; σ_E^2 represents the level-one variance for the empty model; and τ_E^2 represents the level-two variance for the empty model. It is unclear whether using this approximation to R^2 will appropriately approximate the *F*-test of interaction effects within multilevel modeling. As mentioned previously, the variance components estimated from multilevel models can be biased with small sample sizes (Hox, 2010), meaning that the R^2 value could also be biased. Since the variance components from ML typically don't correspond exactly with the ANOVA partition of variance (Snijders & Bosker, 1999), it is unlikely that the *F* value computed with this version of R^2 would be exactly equal to the square of *t*.

Likelihood ratio test (LRT). The likelihood ratio test, also known as the chi-square difference test (Snijders & Bosker, 1999), is a hypothesis testing procedure that allows for the comparison of two nested models (Hox, 2010; McCoach & Black, 2008). The deviance from each of the two models is compared. The deviance can be computed by taking the natural logarithm of the likelihood function value at convergence, and multiplying this value by -2 (Hox, 2010). The difference in deviance values between two competing models is distributed according to the chi-square distribution with degrees of freedom equal to the difference in number of parameters between the two models (McCoach & Black, 2008). The likelihood ratio test can be used for fixed or random parameters and is often used for multiparameter tests (Snijders & Bosker, 1999). For tests of random components, restricted maximum likelihood (REML) or full maximum likelihood (ML) estimations methods may be used, but for tests of the fixed components, ML must be used (Snijders & Bosker, 1999). For larger sample sizes, the likelihood ratio test will provide very similar results to the Wald test since they are asymptotically equivalent (Hox, 2010; Snijders & Bosker, 1999). The likelihood ratio test is preferable compared with

the Wald test for testing random components (Hox, 2010). However, for fixed effects, the Wald test is more convenient and more often used (Hox, 2010) although in cases where these two tests lead to different conclusions, the likelihood ratio test is preferable as it is invariant to re-parameterization (Hox, 2010).

BIC and AIC. Hypothesis testing procedures have been criticized as being strongly influenced by sample size (McCoach & Black, 2008; Raftery, 1995; Weakliem, 2004) and for providing a framework that can reject a null hypothesis but never actually provide evidence in favor of a null hypothesis (McCoach & Black, 2008; Raftery, 1995). Further, hypothesis testing procedures fail to give information about the degree to which models may differ and they are only valid for nested models (McCoach & Black, 2008). Hypothesis testing procedures using a conventional α (Type I error rate) of 0.05 or 0.01 tend to work better with smaller samples (these procedures were developed with sample sizes of about 30 to 200) and although researchers have suggested choosing a smaller α value for analyses with larger samples, it is unclear precisely how to do this (Raftery, 1995). Additionally, hypothesis testing does not always indicate a single best model and makes use of significance which is chosen arbitrarily (Weakliem, 2004).

Given the drawbacks of hypothesis testing procedures for model comparison purposes, researchers have developed additional methods for comparing models, including information criteria methods. The AIC and BIC are two of these measures. Both use the deviance and can be computed as:

$$\text{AIC} = \text{deviance} + 2*k \tag{6}$$

$$\text{BIC} = \text{deviance} + k*\ln(N) \tag{7}$$

Where k is the number of parameters estimated in the model and N is the sample size (Hox, 2010). Deviance was given previously as -2 times the log-likelihood at convergence (Hox, 2010). Both AIC and BIC place a penalty on more complex models, and for larger sample sizes the BIC in particular shows a preference for simpler models compared to both the AIC (Hox, 2010) and the likelihood ratio test (McCoach & Black, 2008) although all three will favor more complex models at larger sample sizes to some degree (Weakliem, 2004). In practice the AIC and likelihood ratio test will often provide similar

results, but when the number of additional parameters is small, the likelihood ratio test tends to favor the simpler model (McCoach & Black, 2008).

The BIC is based on Bayesian principles and the AIC is based on finding an approximate model rather than a true model (Weakliem, 2004). The BIC and hypothesis testing criteria can at times lead to substantially different results (Weakliem, 2004). The BIC is particularly popular in sociology while the AIC is often used in economics (Weakliem, 2004).

The AIC and BIC offer several advantages over hypothesis testing. They are valid for non-nested models (McCoach & Black, 2008), and they provide a ranking of all models and the degree to which they differ (Weakliem, 2004) so they can essentially provide evidence in favor of a simpler model (Weakliem, 2004). In either case, the information criterion for competing models can be computed for each model and compared. Although researchers have debated whether the AIC or BIC is preferable, others advocate for their combined use in practice (McCoach & Black, 2008). Within a multilevel modeling framework, one problem with the BIC is that defining the sample size is ambiguous (Hox, 2010; Raftery, 1995) and although it is clear that the value of *N* should be reduced somehow (Raftery, 1995), in practice different software may compute this quantity differently (Hox, 2010) with some software using the number of level-one units and others using the sample size of level-two units (McCoach & Black, 2008). This has led some to recommend AIC over BIC for multilevel models (Hox, 2010). In order to compare fixed effects with AIC or BIC, ML estimation methods (as opposed to REML) should be used (Hox, 2010).

Using information criteria methods, the interaction effect can be tested by estimating the main-effects-only model and the full model (equations 2 and 3) and then computing and comparing BIC and/or AIC values for the two models. A lower value indicates a better fit and the model with the lowest AIC or BIC value is considered best (Hox, 2010). A difference of 0-2 between the two BIC values indicates weak evidence and a probability of 50-75% in favor of the more complex model; a difference of 2-6 indicates positive evidence with a probability of 75-95% in favor of the more complex model; a difference of 6-10 indicates strong evidence with a probability of 95-99% in favor of the more complex model; and a difference of greater than 10 indicates very strong evidence with a probability of greater

than 99% in favor of the more complex model (Raftery, 1995). Further, Raftery (1995, p. 140) provides a table indicating the approximate minimum *t* values for various strengths of evidence (difference between BIC scores) and various sample sizes, indicating that even to achieve weak evidence of a difference, the approximate *t* values for most conditions is greater than 1.96. As the sample size increases, the approximate *t* value increases as well. For sample sizes of about 30-50 the *t* values are roughly in agreement with conventional cut-offs (for example $p < 0.05$); but for larger samples, the *t* values are much larger (analogously the *p* values are much smaller) than those used in conventional hypothesis testing.

This indicates that although with conventional hypothesis testing, a larger sample size makes it "easier" to reject the null, the BIC is to some degree correcting for this. Further, this indicates that the rate of falsely rejecting the null (Type I error) should be usually less than 0.05 and this number would decrease as the sample size increases.

Choice of model selection method. The literature is not clear on how best to test moderation effects. The Wald test, *F*-test, and likelihood ratio test are hypothesis testing procedures, while the information criteria approaches do not use the hypothesis testing framework. One advantage of the information criteria approach is the ability to provide evidence for or against interaction effects as compared with the hypothesis testing procedures which are only able to provide evidence for interaction effects (rejecting the null).

Considering the hypothesis testing procedures, the Wald test and the *F*-test are identical for one-level models with *F* being equal to the *t* value squared (Champoux & Peters, 1987; Jaccard et al., 1990). Within a multilevel modeling framework, however, the value of *F* could be computed from R^2 that are based on estimated variance components rather than the ANOVA partition of variance. This can lead to a different value for R^2 and therefore a different value for *F* as well. It is not clear whether this specification for *F* will result in Type I error equal to the nominal value and/or if it will result in high power.

For fixed effects, such as an interaction effect, the Wald test is typically the most convenient and used most often in practice (Hox, 2010). Although the Wald test and likelihood ratio test are

asymptotically equivalent, they do not always produce the same result, and in general the likelihood ratio test is preferable because of its invariance to re-parameterization (Hox, 2010). Further, when the likelihood ratio test diverges from the AIC and/or BIC, researchers suggest using discretion and substantive knowledge to make a final model selection decision (McCoach & Black, 2008).

Johnson-Neyman regions of significance. If a significant interaction effect is found, the researcher may want to assess where the levels of the moderator variable differ significantly on the outcome. In order to do this, the Johnson-Neyman technique can be used to find the values of the independent variable where levels of the moderator differ significantly (Pedhazur, 1982). Researchers have extended this procedure to a multilevel modeling setting and provide an analogous procedure for examining regions of significance with cross-level interaction effects (Miyazaki & Maier, 2005).

Current Study Parameter Specification

This study will compare all five of the model selection methods described, as these methods represent the most common model selection methods including both hypothesis testing and information criteria approaches (McCoach & Black, 2008). The following sections will review considerations in selecting the parameters for the simulated data that will be used to examine these five model selection methods.

Effect size. Effect size represents the magnitude of an association and shows the proportion of variance in a dependent variable that is associated with a given independent variable (Tabachnick & Fidell, 2007). Researchers have suggested and make use of various ways to measure effect size including primarily standardized effect sizes (Snijders & Bosker, 1999; Spybrook, 2008) and measures of variance explained (Jaccard et al., 1999; Tabachnick & Fidell, 2007).

A standardized effect size implies that the effect size is not dependent on the original units of measurement (Spybrook, 2008). Standardized coefficients represent one measure of standardized effect size; they are the estimated coefficients from a model where the variables have been standardized with standard deviation equal to one (Snijders & Bosker, 1999). These standardized coefficients can also be obtained by multiplying the coefficient by the standard deviation of X (the independent variable) and

dividing by the standard deviation of Y (the outcome variable). The standardized coefficient indicates how many standard deviations Y would be expected to change for one standard deviation increase in X (Snijders & Bosker, 1999). Standardized effect sizes can also be converted to eta squared, which is a measure of explained variance (Tabachnick & Fidell, 2007).

For a standardized effect size, researchers may consider above 0.80 a large effect, between 0.50 and 0.80 a medium effect, and between 0.20 and 0.50 a small effect, although researchers may also examine the literature in their own field to assess what is considered small, medium, and large (Spybrook, 2008). Within education, small effect sizes are particularly common (Spybrook, 2008). Further, interaction terms have been shown to have particularly low power (Aiken & West, 1991; Alexander & DeShon, 1994; Holmbeck, 1997; Jose, 2013; MacCallum & Mar, 1995) and often fail to manifest themselves (Jaccard et al., 1990) which may indicate that interactions could be expected to have particularly small effect sizes.

Eta squared and the difference in squared multiple correlations are two measures that both represent the proportion of variance explained in the outcome by the given predictor (Jaccard et al., 1990). Measures that represent proportion of explained variance can range in value from zero to one (Tabachnick & Fidell, 2007). Eta squared (η^2) for the interaction effect can be found by dividing the sum of squares for the interaction term by the total sum of squares. This proportion can be multiplied by 100 to indicate the percentage of variance in the outcome that is accounted for by the interaction effect in the sample data (Jaccard et al., 1990). For eta squared, researchers suggest guidelines for the magnitude of various values: a small eta squared is about 0.01; a medium eta squared is about 0.09 and a large eta squared is about 0.25 (Tabachnick & Fidell, 2007). Eta-squared has been critiqued for the fact that its value for any given covariate depends on the number and significance of the other independent variables in the given model (Tabachnick & Fidell, 2007).

The difference in variance explained (squared multiple correlation) between two models can be evaluated as $R_2^2 - R_1^2$. This represents the strength of the interaction in the sample data and can be multiplied by 100 to indicate the percent of variance that the interaction accounts for in the outcome

variable. This difference can be assessed for statistical significance with the *F*-test described above (Jaccard et al., 1990). Although the formulas for eta squared and difference in squared multiple correlation differ, they are both based on the ANOVA partition of variance and their values should be identical.

Although some researchers consider the eta squared and difference in squared multiple correlation measures to be measures of effect size (Jaccard et al., 1990), others argue that they do not represent effect size and in fact are not linearly related to the correct measure of effect size for an interaction effect (Aiken & West, 1991) which can be evaluated as:

$$f^2 = \frac{R_2^2 - R_1^2}{1 - R_2^2} \tag{8}$$

This quantity represents the variance accounted for by the interaction effect relative to the unexplained variance in the outcome (Aiken & West, 1991). Values for f^2 around 0.02 are considered small, around 0.15 are considered medium and around 0.35 are considered large (Aiken & West, 1991).

Effect sizes for interaction effects tend to be very small in practice (Aiken & West, 1991). For example, in a broad review of psychology literature, observed interaction effects accounted for about 1% of the variance in outcomes and in the job design literature, about 3% (Aiken & West, 1991). For effect sizes such as these, very large sample sizes are necessary (Aiken & West, 1991). Previous simulation research may have found high power for interactions by analyzing unrealistically large effect sizes (Aiken & West, 1991).

For this study, effect size for the level-one interaction term will be set at 0 (to assess Type I error rates), and at f^2 values of approximately 0.02 and 0.05 (to assess power). Other interaction effects, such as cross-level (IV and MV measured at different units of analysis) or level-two (IV and MV both measured at level-two) are beyond the scope of this study.

Sample size. Guidance related to the sample size for empirical studies reflects concerns about both power of statistical tests as well as the validity and reliability of statistical methods for smaller sample sizes, particularly those that use estimators that are asymptotically unbiased (Hox, 2010). These

concerns inform the selection of sample sizes for this study in addition to reflecting the approximate sizes of samples that might be found in substantive research.

In ordinary (not multilevel) regression, researchers have suggested a minimum sample size of 104 plus the number of parameters when the interest is in interpreting slope coefficients (Hox, 2010). For reliability of prediction, one rule of thumb suggests the sample should have 15 subjects for each predictor (Stevens, 2002).

Jaccard et al. (1990, pg. 37) provide a table with the approximate sample sizes needed to achieve a power of 0.80 given various effect sizes in a model testing one interaction term between two continuous variables. For example, if the R^2 value for equation 1 is 0.10 and for equation 2 is 0.15 the necessary sample size is 135; if for equation 2 it is 0.25, the sample size should be at least 41. The table indicates that as the difference in R^2 values becomes larger, the required sample size becomes smaller. Although this table provides useful information, most of the effects sizes examined within it are larger than what would be typically found in practice (Aiken & West, 1991).

Raftery (1995) provides tables with common sample sizes within various types of studies in sociology: 30 represents the number of industrialized countries; 50 represents the number of U.S. states; 100 represents the number of U.S. SMSAs (standard metropolitan statistical areas); and 1000, 10,000, and 100,000 represent a small, medium, and large survey respectively.

Within multilevel models, researchers suggest at least 100 groups when the interest is in estimating level-2 fixed effects or variance components, but fewer clusters would be needed for individual-level fixed effects (Hox, 2010). One rule of thumb, called the '30/30 rule' suggests a minimum of 30 clusters and 30 individuals per cluster in order to accurately estimate fixed effects and their standard errors (Hox, 2010). Other researchers have suggested that when cluster variability needs to be modeled and there are more than about 10 clusters, it makes sense to model this variability as a random effect rather than fixed effects (Snijders & Bosker, 1999).

In one multilevel simulation study, researchers chose to vary the number of clusters with three conditions (30, 50, and 100) and the cluster size with three conditions (5, 30, and 50) (Maas & Hox,

2004). They chose these conditions based on practical experience and past literature. They found that the literature specified 30 clusters as a minimum while 100 clusters was specified as sufficient, and 50 was added as a common size in actual research. Cluster size of 5 is typical for family research; cluster size of 30 is typical for educational research; and cluster size of 50 is chosen based on the literature (Maas & Hox, 2004). Based on additional simulations, Hox & Maas (2004) conclude that for examination of fixed effects, 10 clusters may be sufficient; for contextual effects, 30 clusters are needed; and for correct estimates of standard errors, 50 clusters are needed.

This study will use four conditions for the number of clusters: 10, 30, 50, and 100 and three cluster size conditions: 5, 30, and 50 similarly to previous multilevel simulation studies (Maas & Hox, 2004) and these two sets will be fully crossed for a total of 12 separate sample size conditions. These conditions are similar to Maas and Hox (2004) but with the addition of 10 clusters which may more closely reflect the difficulties of obtaining several groups in applied educational research. Further, this condition will be interesting to explore as it represents a sample that is at or just below the recommended minimum sample size for multilevel modeling and testing interaction effects. These conditions will allow for assessment of a wide range of sample size conditions, as may be seen in substantive research. In addition, sample size may be particularly relevant due to the low power associated with tests of moderation, meaning that a wide variety of sample size conditions could be particularly useful.

ICC. The intraclass correlation coefficient (ICC) is a measure of the proportion of variance accounted for at level 2 (Snijders & Bosker, 1999). Since the ICC is a ratio of the variability between level 2 units to total variability (Spybrook, 2008), values for ICC can range only between zero and one. The average ICC for K-12 academic achievement has been shown to be about 0.22 for nationally representative schools and about 0.09 for low-achieving schools (Hedges & Hedberg, 2007). Similarly, other studies have shown that the ICC for school achievement in the United States ranges between 0.10 and 0.20 (Spybrook, 2008). Simulation research examining multilevel models has used ICC values of 0.10, 0.20, and 0.30 to represent the levels customarily found in substantive multilevel research (Maas & Hox, 2004).

As the ICC increases, for any given sample size, the effective amount of information decreases (Snijders & Bosker, 1999). The effective *N* can be computed as a function of the ICC value, the total sample size, and the cluster size (Snijders & Bosker, 1999), indicating that the effect of varying the ICC value should be captured in the various cluster size and number of cluster conditions specified above. Therefore, only a value of approximately 0.10 will be used for ICC in this study, which is typical within the range of ICC values for achievement outcomes.

Significance level. The significance level, or Type I error rate (α), represents the rate at which the null hypothesis is falsely rejected and is typically set at 0.05 by researchers (Cohen, 1992; Hox, 2010).

For hypothesis testing procedures, this study will maintain a nominal significance level of 0.05 and deviations from this value based on simulated data will be noted. An observed significance level greater than 0.05 indicates a liberal test, which is considered problematic, and an observed significance level of less than 0.05 indicates a conservative test, which is also problematic, but presumably less so, depending on the substantive implications of a Type I error. Note that the BIC and AIC are information criteria and as such, do not make use of the hypothesis testing framework and therefore cannot be used with a specified significance level. Raftery (1995, p. 141) provides the expected Type I error rates for the BIC based on various strengths of evidence and sample sizes.

Researchers have used simulation procedures to confirm that the observed Type I error rate for the moderated multiple regression procedure (*F*-test) is the same as the expected nominal Type I error rate, which they had set to 0.05 (Paunonen & Jackson, 1988). The nominal Type I error rate for the hypothesis testing procedures used in this study is set at 0.05.

Power. Power cannot be specified for this simulation experiment, but it will be measured as an outcome of interest. Power is the probability of correctly rejecting the null hypothesis (Spybrook, 2008) and typically should be approximately 0.80 or larger (Cohen, 1992; Spybrook, 2008).

Distribution of variables and error variance. The continuous variables (independent variable and dependent variable) in this study will all be approximately normally distributed as will the distribution of error variance. This choice is based on the fact that distribution shape is not a particular

area of focus for this study, along with the fact that ANCOVA models have been shown to be relatively robust to the violations of the normality assumption (Glass, Peckham, & Sanders, 1972). Further, previous simulation research examining moderation has simulated independent variables according to the normal distribution despite the fact that empirical data may not always be normally distributed; the authors argue that their sample size of $n = 200$ is sufficient to invoke the central limit theorem and allow for valid statistical test results (Champoux & Peters, 1987). Continuous variables in the present study may not be exactly normally distributed due to the nature of the simulation process. Further details regarding these deviations can be found in the methods section. The moderator variable in this study will be binary with equal numbers in both groups.

Chapter II: Research Questions and Hypotheses

Research Questions

The proposed research will address the following three research questions:

Question 1. What is the Type I error rate when testing a level-one interaction effect in a multilevel model using a Wald test, an *F*-test, a likelihood ratio test, the BIC, and the AIC and how does it vary as a function of cluster size and number of clusters?

Question 2. How does each of these same five methods perform in terms of the power to detect level-one interaction effects and how does it vary as a function of interaction effect size, cluster size, and number of clusters?

Question 3. How does each of these methods compare when used to analyze data from an engineering education dataset?

Hypotheses

The following hypotheses are suggested for each research question respectively.

Question 1. Although choice of degrees of freedom for the Wald test can be an issue in multilevel modeling for small sample sizes (Snijders & Bosker, 1999), all conditions in this study (smallest $N = 150$) will exceed the threshold of 40 degrees of freedom suggested by Snijders and Bosker (1999) meaning that the Z test rather than *t*-test will be used for all Wald tests in this study. Given this, an observed Type I error rate close to the nominal rate of 0.05 is expected for the Wald test for all conditions.

Although the *F*-test based on the ANOVA partition of variance should be identical to the Wald test (DeShon & Alexander, 1996; Paunonen & Jackson, 1988), conducting the *F*-test based on the ML estimate of variance components may differ. With smaller sample size, the estimates of variance components may be too small (Hox, 2010) indicating that both R^2 and *F* may be biased. Since R^2 is computed for this study based on the level-one and level-two variance components for both the empty model and the full model being tested (Snijders & Bosker, 1999), it is unclear whether R^2 will be biased and if so, whether it will be an under-estimate or an over-estimate. Therefore, it is also unclear whether *F*

will be biased or not and whether or not the Type I error rate will approximate the nominal rate of 0.05. Any observed bias should, however, decrease as the sample size increases, as the ML estimate of the variance components is expected to be asymptotically unbiased (Hox, 2010).

The likelihood ratio test represents another hypothesis testing method which indicates that the observed Type I error rate should approximate the nominal rate of 0.05. Previous research has shown that the likelihood ratio test controls Type I error rates well for multilevel models (Hox, 2010).

The BIC and AIC model comparison procedures are not hypothesis testing procedures which means that the researcher does not choose a Type I error rate for these tests. So although the Type I error rate can be measured and compared with other model selection procedures, there is no nominal Type I error rate associated with these tests. Raftery (1995) has provided a table that specifies exactly how Type I error rates associated with use of the BIC vary as a function of sample size and strength of evidence for a given model. This table indicates that for most conditions the Type I error rate is less than the conventional cut-off of 0.05. Based on this information, Type I error rates for BIC (weak evidence; difference of 0-2 points) are expected to be generally smaller than 0.05 and to decrease as sample size increases. Examining BIC in terms of very strong evidence (greater than 10 point difference), Type I error rates should be much smaller, perhaps even approaching zero. As AIC and the likelihood ratio test have been shown to give similar results in practice (McCoach & Black, 2008), and as AIC value does not depend on sample size (Hox, 2010), it is expected that the Type I error rates for AIC may stay close to 0.05 for all sample size conditions. It should also be noted that the BIC and AIC are the only methods of those examined that have the ability to provide evidence for the null condition.

Question 2. As effect size increases, power will increase as well (Hox, 2010) and this finding is expected across all model selection methods. Also, as sample size increases, power will increase (Spybrook, 2008) which should also be consistent across all sample size conditions. Research has shown that the number of clusters may have a bigger impact on power than cluster size (Spybrook, 2008), which indicates that power may increase more as the number of clusters increases in this study, compared with

the increases in cluster size. It is unclear how each of these five model selection methods will compare with each other in their overall power to detect interaction effects.

Question 3. The results from question 1 and 2 should inform the analysis of an engineering education dataset with respect to a substantive moderation question.

Chapter III: Methods

Monte Carlo Simulation Study Methodology

Monte Carlo simulation was used to generate data for research questions 1 and 2. Monte Carlo simulation can be used to understand statistical processes and provides a good alternative to analytical mathematics for more complicated statistics (Mooney, 1997). Data for this study was simulated and analyzed using R (R Core Team, 2014). Models were estimated using the *lmer* function within the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Full maximum likelihood estimation (ML), as opposed to restricted maximum likelihood estimation (REML), was used for all models in this study, as this is required for the likelihood ratio test (Snijders & Bosker, 1999) and the BIC and AIC model comparison methods (Hox, 2010) to be valid.

A total of 36 conditions were examined. Research shows that Monte Carlo studies using at least 1000 trials are typical, and even as large as 10,000 to 25,000 trials may be common (Mooney, 1997); the decision on number of trials is a balance between precision of the estimates (the greater the number of simulations, the better) and complexity of the model estimated (the greater the complexity, the longer the time in processing the data). Past simulation research examining multilevel data has used 1000 simulated data sets for each condition examined (Maas & Hox, 2004). For this study, given the increases in computing power in the intervening years, the number of simulated datasets was increased to 5000 for each of the 36 conditions. The number of simulations was not increased past 5000 due to the computing time for generating the data of approximately one hour per condition. Data was generated according to the following multilevel linear regression model:

$$Y_{ij} = B_0 + B_1*X_{ij} + B_2*M_{ij} + B_3*X_{ij}*M_{ij} + u_j + e_{ij} \quad (9)$$

where Y_{ij} is the outcome for individual i within cluster j. X is the independent variable and is normally distributed with mean equal to zero and standard deviation equal to one. X was simulated using the *rnorm* function in R. M is a binary moderator variable (p=0.5) and was constructed using effect coding (values of -1 or 1). M represents membership in either group 1 or group 2. The random error at level 1 (e_{ij}) is normally distributed with mean equal to zero and standard deviation equal to two and was also

simulated using the *rnorm* function in R. The random error at level 2 (u_j) is normally distributed with mean equal to zero and standard deviation equal to 0.78 also simulated with *rnorm*. This value was chosen to create an intraclass correlation coefficient of approximately 0.10 (which can be seen given the variance of Y which can be approximately computed as the variance of a mixture distribution and is a function of the variance of e_{ij}, the variance of u_j, and the intercept and slope for each level of M).

Y was created as the mixture distribution of two normal distributions (one for each level of M) with adjustments for cluster membership added based on a random draw from u_j for each cluster. To create Y, two variables were created and then vertically concatenated (Mooney, 1997). The first variable was computed as alpha for group one, plus beta for group one times the first half of X values plus the first half of the error term values (note that this specification describes the standard linear regression equation). The second variable was computed analogously as alpha for group two, plus beta for group two times the second half of X values plus the second half of the error term values. These values were vertically concatenated into one variable, and then modified to induce a non-zero ICC value (meaning clusters differ). To create the non-zero ICC, an adjustment for each level 2 cluster was selected (randomly drawn from the distribution of u_j values) and then added to each cluster member's Y value. Note that u_j is centered at zero so this adjustment doesn't affect the expectation of Y. A cluster ID variable was created that listed the cluster affiliation for each member of the sample. Logically, if the beta (slope) values for each group of the moderator variable are different, there should be a significant interaction effect. Conversely, if the beta (slope) values for both groups of the moderator are the same, there should be no interaction effect (interaction effect size = 0). The interaction effect is exclusively modeled at level-one, as level-two and/or cross-level interactions are beyond the scope of this study. This data simulation process resulted in a dataset with the following properties given in Table 1.

Table 1

Simulated Variables and Their Properties

Variable	Distribution	Mean	Variance
Y	mixture	3.5	varies
X	normal	0	1
M	discrete (binary)	0	1
MX	normal	0	1

The mean of Y can be computed based on the mean of a mixture distribution which is essentially a weighted average of alpha for group one and group two. Given that the sample sizes for both groups are equal, the weights are therefore also equal. So in this case, the mean of Y is the average of 3 and 4, so it is 3.5. The variance of Y can be approximated based on adding the level 1 variance (using the formula for the variance of a mixture distribution) and the level 2 variance which is simulated as 0.78 squared. Since the slope for each group varies as a function of effect size (a larger effect size indicates a larger difference in the slope coefficients for each of the two moderator groups), the variance of Y will also vary as a function of effect size. The slope for group two will be set at 1.5. The slope for group one will be set at 1.5, 1.0, and 0.5 for the three effect size conditions of no effect, small effect, and medium-small effect size respectively.

The mean of X was set at zero and the variance at one when this variable was generated. No further adjustments were made. M is effect coded with half the sample coded -1 and the other half coded 1. The mean for a discrete random variable can be computed according to the formula for the expectation of a discrete variable which is the sum of the possible values of the variable each multiplied by the proportion of that value within the sample (Devore, 2004). This indicates the mean for M is zero. In order to compute the variance of M, this expectation may be squared and then subtracted from the expectation of (X^2) according to the formula for the variance of discrete random variables (Devore, 2004). Doing this indicates that the variance of M is one. Note that M differs from a Bernoulli random variable, as the later can only take values of zero or one (Devore, 2004) whereas M takes values of one and negative one. MX is computed by multiplying M and X. In other words, half of the values of X are multiplied by one and

the other half by negative one. Therefore, each half of X will still be distributed normally with mean of zero and variance of one. Again using the formulas for the mean and variance of a mixture distribution indicates that MX will have a mean of zero and a variance of one.

In addition to the full multilevel model according to which the data was generated (equation 9), an empty multilevel model (no predictors; equation 10) was estimated in order to assess ICC, and the main-effects-only multilevel model (equation 11), which does not include the product term, was estimated for model comparison purposes. Relevant quantities were saved into a csv data file.

Summary of Conditions

The simulated dataset will vary will respect to number of clusters, cluster size, and effect size and each of these conditions will be fully crossed. For each effect size condition (f^2 = 0, 0.20, 0.50), Table 2 shows the 12 sample size conditions that were simulated.

Table 2

Sample Size Conditions for Simulated Datasets

Number of Clusters	Cluster Size	Total *N*
10	5	50
10	30	300
10	50	500
30	5	150
30	30	900
30	50	1500
50	5	250
50	30	1500
50	50	2500
100	5	500
100	30	3000
100	50	5000

With 4 values for number of clusters, 3 values for cluster size, and 3 values for effect size, this results in 4 x 3 x 3 sets of conditions for a total of 36 unique conditions. For each of these conditions, N = 5,000 replications were run.

Analysis

The following three multilevel models were estimated for each simulated data set:

$$M0: Y_{ij} = B_0 + u_j + e_{ij} \quad (10)$$

$$M1: Y_{ij} = B_0 + B_1*X_{ij} + B_2*M_{ij} + u_j + e_{ij} \quad (11)$$

$$M2: Y_{ij} = B_0 + B_1*X_{ij} + B_2*M_{ij} + B_3*X_{ij}*M_{ij} + u_j + e_{ij} \quad (12)$$

M0 (equation 10) is the empty multilevel model. This model simply partitions the variance into σ^2 (level-one variance) and τ^2 (level-two variance) and allows for the computation of ICC (equation 1). These variance components are also used for the computation of R^2 (equation 5) which is used to compute *F* (equation 4). Note that because these models are estimated using ML methods, the resulting partition of variance can differ from the ANOVA partition of variance. Next, M1 (equation 11) was estimated, which is the multilevel version of the main-effects only model. This model was used to find the change in R^2 (variance explained) value to compute *F*, the deviance value to compute the LRT, and the BIC and AIC values to perform model comparison. Finally, the full model, M2 (equation 12), was estimated. The test of significance for B_3 was performed; this is the Wald test for interaction effects.

Power and Type I error were both assessed based on the proportion of simulations (out of 5000 for each condition) with a significant interaction effect. Type I error can be computed for the conditions with no interaction effect and power can be computed for the two conditions with a non-zero interaction effect size. Significance was determined separately based on each of the five methods examined in this study.

The Wald test was computed for each dataset based on the *t* value for B_3 from M2 (equation 12). A *t* value of greater than 1.96 or less than -1.96 was recorded as being a significant interaction effect. This test statistic value was used because all sample sizes (*N*=50 or greater) were considered approximately large enough to use the *Z* rather than *t*-test.

The *F*-test (see equation 4) was computed which used R^2 values (see equation 5) based on the maximum likelihood partition of variance, as well as the observed value of *N*. Note that this partition of variance (estimated values of σ^2 and τ^2) should differ from the ANOVA partition of variance which is based on the sum of squares. So the ANOVA partition of variance would result in an *F* value that is exactly t^2, meaning the *F*-test based on this value would result in identical results to the Wald test.

However, the maximum likelihood partition of variance can result in an *F* value that differs slightly from t^2. An *F* value of greater than 3.84 implies a significant interaction effect. The critical value of 3.84 represents the value of the *F*-distribution with degrees of freedom equal to one and *N*-6 for all *N* at the 95% level.

The likelihood ratio test was performed by subtracting the deviance value of M2 (full model) from the deviance value of M1 (main effects only model). The test was considered significant if this value was greater than the critical value of the chi-square distribution with one degree of freedom which is 3.84 at the 95% level.

The BIC and AIC are given by the *BIC* and *AIC* functions provided within the lmer package in R. BIC is computed according to equation 7 and AIC is computed according to equation 6. The value of *N* used for the computation of BIC is based on the total *N*, which could potentially be misleading as the effective information conveyed with multilevel data is representative of a sample size less than the total *N* (Snijders & Bosker, 1999). The number of parameters (*k*) for both BIC and AIC is four for M1 (two slope coefficients plus two variance components) and five for M2 (three slope coefficients plus two variance components). The test for an interaction effect is performed by first subtracting the BIC (or AIC) value for M2 from the BIC (or AIC) value of M1. With the BIC, if this difference is positive, weak evidence for the interaction effect is provided; if this difference is greater than 10, very strong evidence for the interaction effect is provided (Raftery, 1995). Within this study, these two criteria are referred to as BIC (weak) and BIC (strong), respectively. With the AIC, if the difference is positive, evidence for the interaction effect is provided.

In order to assess Type I error rates, the percentage of datasets with a significant interaction was computed for the conditions with effect size equal to zero. This was done separately for each sample size condition and each of the five methods. To assess power, the percentage of datasets with a significant interaction was computed for the conditions with a non-zero effect size. This was also done separately for each sample size condition and each of the five methods.

Applied Analysis Methodology

This section provides a brief overview of the methodology used in the applied analysis. More methodological details, including a more detailed description of the data, measures, and analysis, are provided along with the results in chapter 5.

In order to demonstrate and further understand each of the five methods for detecting interaction effects within a substantive example (research question 3), data from the Project to Assess Climate in Engineering (PACE) was used. The data is from a large-scale survey conducted in 2012 within 15 engineering schools among a sample of undergraduate engineering students. The applied analysis examines a student confidence outcome (DV). The continuous IV is a measure of work-life balance and the binary MV is gender. The empty model (M0), main effects only model (M1), and full model (M2) were estimated and each of the five methods were used to assess for significant interaction effects between work-life balance perceptions and gender.

This analysis provides an example of how these methods can be used with real data and explores potential benefits and weaknesses associated with each of the methods. Although these methodological demonstrations and insights represent the main motivation for the applied analysis, the substantive aspect of this analysis provides the opportunity for a more concrete example. To this end, the research question implicit in this analysis is stated explicitly: Does gender moderate the relationship between perceptions of work-life balance and student confidence among a sample of undergraduate engineering students? The process of estimation and analysis for the applied data follows exactly the same procedure as that for the simulated data.

Chapter IV: Monte Carlo Simulation Study Results

Research Question 1

Research question 1 asks what is the Type I error rate when testing a level-one interaction effect in a multilevel model using a Wald test, an *F*-test, a likelihood ratio test, the BIC, and the AIC and how does it vary as a function of cluster size and number of clusters? To address this question, the Type I error rate for each sample size condition was computed empirically as the proportion of samples where the given test detected significant interaction effects divided by the total number of samples (5000) for each of the 12 conditions where the data was generated with an interaction effect size of zero. For each type I error rate, the standard error was computed as the square root of $[p*(1-p)]/N$ where p represents the proportion of successes where in this case the p = Type I error rate and N represents the number of simulations run (i.e. 5000). To determine whether the observed Type I error rate was significantly different from 0.05 (which represents the nominal rate for the Wald test, *F*-test, and LRT), 0.05 was subtracted from the observed Type I error rate and the quantity was divided by the relevant standard error. The Type I error rates displayed in bold font in Table 3 were significantly different from 0.05 at the 95% level. All estimated models converged. Table 3 and Figure 1 display the results.

To further understand the Type I error rate results (table 3), a series of three one-way ANOVAs was run on the factors considered in this analysis: number of clusters, cluster size, and test used. Three one-way ANOVAs were used rather than one factorial ANOVA due to limitations in sample size. If all three factors were considered simultaneously there would be one unique observation per cell which would not meet the assumptions of ANOVA. Additionally, results from the one-way ANOVA models examining Type I error and power should be interpreted with caution as the outcome does not meet the normality assumption for ANOVA (Devore, 2004) since it is a proportion that is bounded by zero and one.

Results of these one-way ANOVA models indicate that no significant differences were found for number of clusters ($p = 0.95$) or cluster size ($p = 0.94$) on Type I error rates. The method used did show a significant effect on Type I error rates ($p < 0.01$) with follow-up pairwise comparison *t*-tests using pooled

standard deviation and the Bonferroni correction showing all methods differed with each other method, except the Wald test, *F*-test, and LRT at the 95% level. The AIC seems to be somewhat of an outlier, displaying Type I error rates consistently around or above 0.15. Use of the AIC does not require the researcher to choose a significance level, and therefore there is no nominal Type I error rate to compare with the observed values. However, most researchers would undoubtedly consider a Type I error rate of 0.15 too high for general use.

Table 3

Type I Error Rates for Each Sample Size Condition

				Type I Error Rates					
Cond	No. Clusters	Cluster Size	N	Wald test	*F*-test	LRT	BIC (weak)	BIC (strong)	AIC
1	10	5	50	**0.068**	0.049	**0.063**	**0.060**	**0.000**	**0.180**
2	10	30	300	0.053	0.046	0.052	**0.016**	**0.000**	**0.167**
3	10	50	500	0.050	0.047	0.050	**0.012**	**0.000**	**0.164**
4	30	5	150	0.056	0.045	0.053	**0.026**	**0.000**	**0.161**
5	30	30	900	**0.057**	0.047	**0.057**	**0.010**	**0.000**	**0.159**
6	30	50	1500	0.051	0.044	0.051	**0.006**	**0.000**	**0.160**
7	50	5	250	0.051	**0.043**	0.051	**0.017**	**0.000**	**0.158**
8	50	30	1500	0.055	0.051	0.055	**0.008**	**0.000**	**0.163**
9	50	50	2500	0.049	0.047	0.049	**0.004**	**0.000**	**0.156**
10	100	5	500	0.045	**0.041**	0.044	**0.011**	**0.000**	**0.148**
11	100	30	3000	0.046	0.048	0.046	**0.005**	**0.000**	**0.159**
12	100	50	5000	0.048	0.045	0.048	**0.005**	**0.000**	**0.156**
Mean				0.052	0.046	0.052	0.015	0.000	0.161

Note. Observed Type I error rates based on $N = 5000$ simulations per cell. The nominal Type I error rate for the hypothesis testing procedures (Wald test, *F*-test, and LRT) was set at $\alpha = 0.05$. LRT = likelihood ratio test. Weak evidence for BIC indicates that the BIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model. Strong evidence for the BIC indicates that the BIC value for the more complex (interaction) model was at least 10 points higher than that for the simpler (main effects only) model. Using the AIC, evidence for a significant interaction effect was provided when the AIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model. Bold indicates significantly different from the value of 0.05.

The Wald test and likelihood ratio test (LRT) have been shown to be asymptotically equivalent (Hox, 2010) and these results align nicely with this finding as we can see that the Type I error rates for these two tests differ very little and basically converge as the sample size gets larger. In terms of Type I error rates, it seems that researchers can simply use the Wald test for larger sample sizes, but might want to additionally perform the LRT for smaller sample sizes, although the results seem unlikely to differ.

The *F*-test based on the ML estimate of the variance components maintains a Type I error rate very close to the nominal rate of 0.05. The difference between the Wald test and *F*-test in these results is likely due to the bias of ML variance component estimates for small samples. As the sample size increases, the *F*-test results seem to converge with the Wald test and LRT. Given this, it seems that the Wald test would be preferable given its ease of use. Further, the bias of ML with small sample sizes has been shown for random effects (Hox, 2010) but should not be as large for fixed effects.

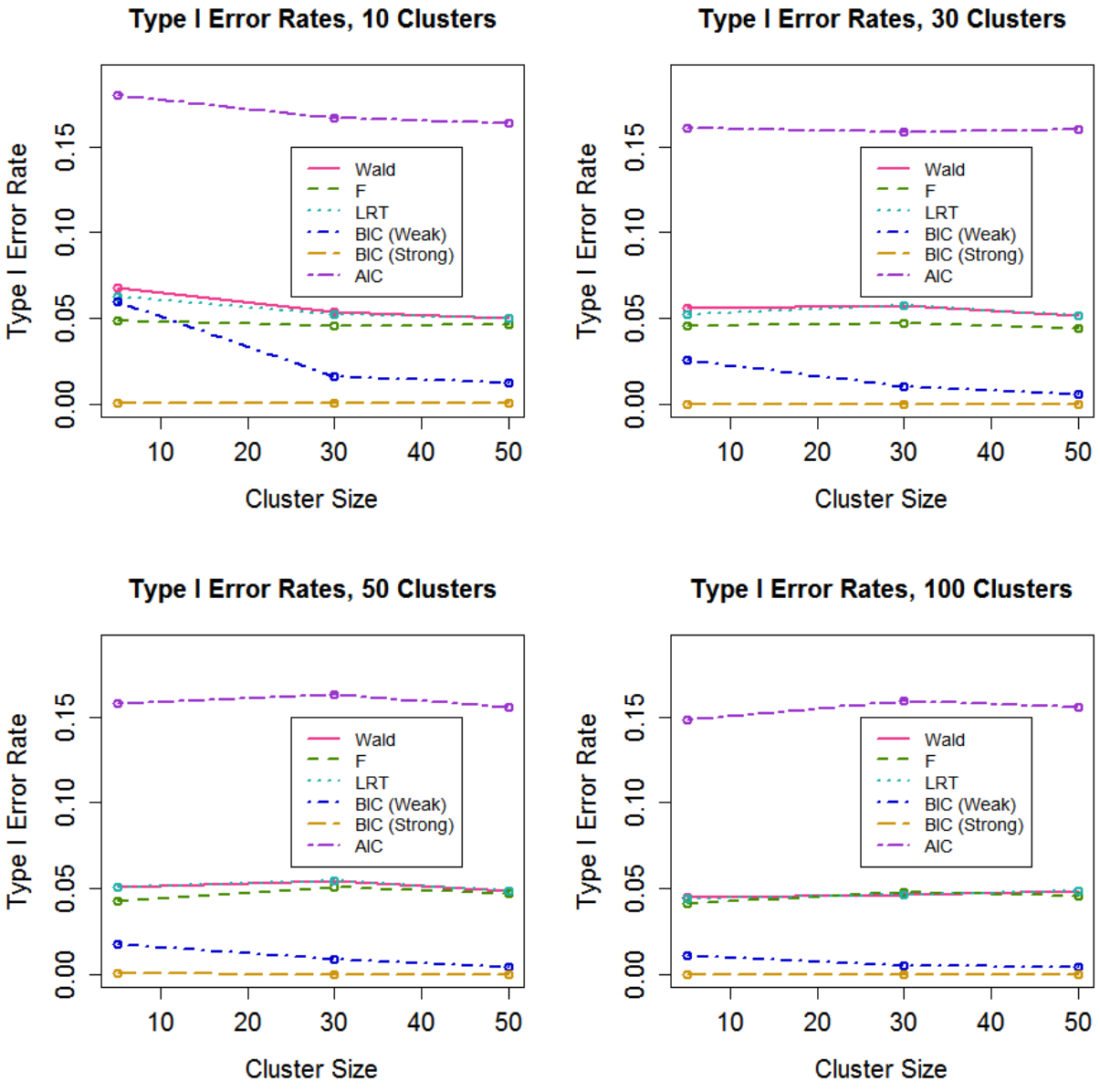

Figure 1. Type I error rates for each sample size condition.

The Wald test, LRT, and *F*-test all represent hypothesis testing procedures meaning the observed Type I error rates can be compared to the nominal rate of 0.05. It is good to see that for all three tests the observed rates are typically very close to 0.05. There are a few cases where these methods differ significantly from the nominal rate of 0.05 (see bold entries in Table 3). In these cases, the Wald test and LRT show somewhat liberal (higher than 0.05) rates at smaller sample sizes and the *F*-test shows somewhat conservative (lower than 0.05) rates in a few cases. Given that the Type I error rates were shown to not significantly vary as a function of number of clusters or cluster size, it may be that the two conditions where the *F*-test shows a Type I error rate significantly less than 0.05 are due to random variation rather than a function of those specific sample size conditions. Despite the ANOVA findings of no effect for number of clusters and cluster size, it seems likely that the inflated Type I error rates for the Wald test and LRT are a function of particularly small sample size considering the findings in the literature that state that the standard errors for these tests are asymptotic (Hox, 2010).

Using the weak evidence criteria for BIC produced a Type I error rate significantly lower than for the hypothesis testing procedures except for the condition of 10 clusters each of size 5. In addition, the test based on BIC seems to be the only one that is affected by sample size, with the Type I error rate decreasing as the sample size increases. This was expected since the BIC formula includes the *N* value and has been shown to perform more conservatively as the sample size increases (Raftery, 1995). The relatively conservative nature of this test indicates that it might be appropriate for researchers to use an adjusted *N*, rather than the total *N*, for multilevel data, so that the BIC value is not penalized for large sample sizes as much as it would be with non-nested data. Using the BIC very strong criteria (difference of 10 or more) resulted in no Type I errors out of each sample of 5000 replications. This indicates that using the strong criteria for BIC is unlikely to result in a Type I error.

Research Question 2

Research question 2 asks how does each of these five methods perform in terms of the power to detect level-one interaction effects and how does it vary as a function of interaction effect size, cluster size, and number of clusters? To address this question, the power of each sample size condition was

computed empirically as the proportion of samples where the given test detected significant interaction effects divided by the total number of samples (5000) for the 24 conditions where the data was generated with a non-zero interaction effect size. All models converged. Results for the small effect size conditions ($f^2 = 0.02$) are provided in Table 4 and for the medium-small effect size conditions ($f^2 = 0.05$) in Table 5.

Table 4

Power for Each Sample Size Condition with Small Effect Size

				Power, Small Effect Size (0.02)					
Cond	No. Clusters	Cluster Size	*N*	Wald test	*F*-test	LRT	BIC (weak)	BIC (strong)	AIC
1	10	5	50	0.163	0.124	0.151	0.148	0.003	0.312
2	10	30	300	0.561	0.509	0.559	0.389	0.032	0.754
3	10	50	500	0.776	0.733	0.774	0.576	0.089	0.901
4	30	5	150	0.324	0.284	0.319	0.226	0.011	0.535
5	30	30	900	0.956	0.932	0.955	0.858	0.358	0.988
6	30	50	1500	0.997	0.994	0.997	0.980	0.719	0.999
7	50	5	250	0.496	0.451	0.492	0.344	0.022	0.706
8	50	30	1500	0.998	0.992	0.998	0.981	0.746	1.000
9	50	50	2500	1.000	1.000	1.000	1.000	0.974	1.000
10	100	5	500	0.792	0.739	0.790	0.600	0.102	0.920
11	100	30	3000	1.000	1.000	1.000	1.000	0.994	1.000
12	100	50	5000	1.000	1.000	1.000	1.000	1.000	1.000
Mean				0.755	0.730	0.753	0.675	0.421	0.843

Note. Observed power rates based on $N = 5000$ simulations per cell. LRT = likelihood ratio test. Weak evidence for BIC indicates that the BIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model. Strong evidence for the BIC indicates that the BIC value for the more complex (interaction) model was at least 10 points higher than that for the simpler (main effects only) model. Using the AIC, evidence for a significant interaction effect was provided when the AIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model.

To further understand the power results (Tables 4 and 5, and Figures 2 and 3), a series of four one-way ANOVAs was run on the factors considered in this analysis: number of clusters, cluster size, test used, and effect size. Results of these four one-way ANOVA models indicate that all factors are significant predictors of power rate, including number of clusters ($p < 0.01$), cluster size ($p < 0.01$), test used ($p < 0.01$), and effect size ($p < 0.01$). Follow-up pairwise comparison *t*-tests using pooled standard deviation and the Bonferroni correction showed the following significant differences for cluster size: 10 and 30, 10 and 50, 10 and 100; for number of clusters: 5 and 30, 5 and 50; and for test used: all methods

and BIC (strong) except the BIC (weak) conditions which did not vary significantly from BIC (strong). Only two conditions (0.02 and 0.05) were evaluated for the effect size factor.

Table 5

Power for Each Sample Size Condition with Medium-Small Effect Size

				Power, Medium-Small Effect Size (0.05)					
Cond	No. Clusters	Cluster Size	*N*	Wald test	*F*-test	LRT	BIC (weak)	BIC (strong)	AIC
1	10	5	50	0.423	0.342	0.402	0.394	0.026	0.601
2	10	30	300	0.985	0.977	0.985	0.956	0.561	0.995
3	10	50	500	1.000	1.000	1.000	0.999	0.910	1.000
4	30	5	150	0.845	0.789	0.841	0.756	0.185	0.937
5	30	30	900	1.000	1.000	1.000	1.000	0.999	1.000
6	30	50	1500	1.000	1.000	1.000	1.000	1.000	1.000
7	50	5	250	0.973	0.954	0.971	0.937	0.476	0.992
8	50	30	1500	1.000	1.000	1.000	1.000	1.000	1.000
9	50	50	2500	1.000	1.000	1.000	1.000	1.000	1.000
10	100	5	500	1.000	0.999	1.000	0.999	0.925	1.000
11	100	30	3000	1.000	1.000	1.000	1.000	1.000	1.000
12	100	50	5000	1.000	1.000	1.000	1.000	1.000	1.000
Mean				0.935	0.922	0.933	0.920	0.757	0.960

Note. Observed power rates based on $N = 5000$ simulations per cell. LRT = likelihood ratio test. Weak evidence for BIC indicates that the BIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model. Strong evidence for the BIC indicates that the BIC value for the more complex (interaction) model was at least 10 points higher than that for the simpler (main effects only) model. Using the AIC, evidence for a significant interaction effect was provided when the AIC value for the more complex (interaction) model was higher than that for the simpler (main effects only) model.

For all conditions assessing power, the AIC shows the highest power, although overall the AIC power rates were not significantly different from most other methods. Of course, high power for AIC is not too surprising considering the high rates of Type I error that the AIC also displayed. As the sample size increases, however, the power of other tests begins to converge at 1. The Wald test and LRT display virtually identical power for all conditions meaning that for these conditions the tests are mostly interchangeable. Again, as they have been shown to be asymptotically equivalent, it is not surprising that the results from both tests correspond closely. This may indicate that performing the LRT is an unnecessary extra step when it is simpler and easier to just use the results from the Wald test for interaction effects. The *F*-test shows power that is just slightly less than that for the Wald/LRT tests. Overall it seems that the *F*-test based on the ML variance components corresponds very closely with the

two other tests using a hypothesis testing framework (Wald and LRT) and again, there is probably no need to take the extra step to compute this test when the researcher could more easily use the Wald test. Also, it is likely that the *F*-test is asymptotically equivalent to the Wald test and LRT since the variance components on which it is based have been shown to be biased for small samples, but that bias should reduce as the sample size becomes large. Using the weak evidence for BIC displays comparatively good power, just slightly less than that for the three hypothesis testing procedures. It is interesting that this test still displays good power, considering the Type I error rates were comparatively conservative and indicates this may be a useful method for practice and for further investigation. Using the strong criteria for BIC displays much lower power, which makes sense given the tremendously conservative nature of this test that was observed for the Type I error rates.

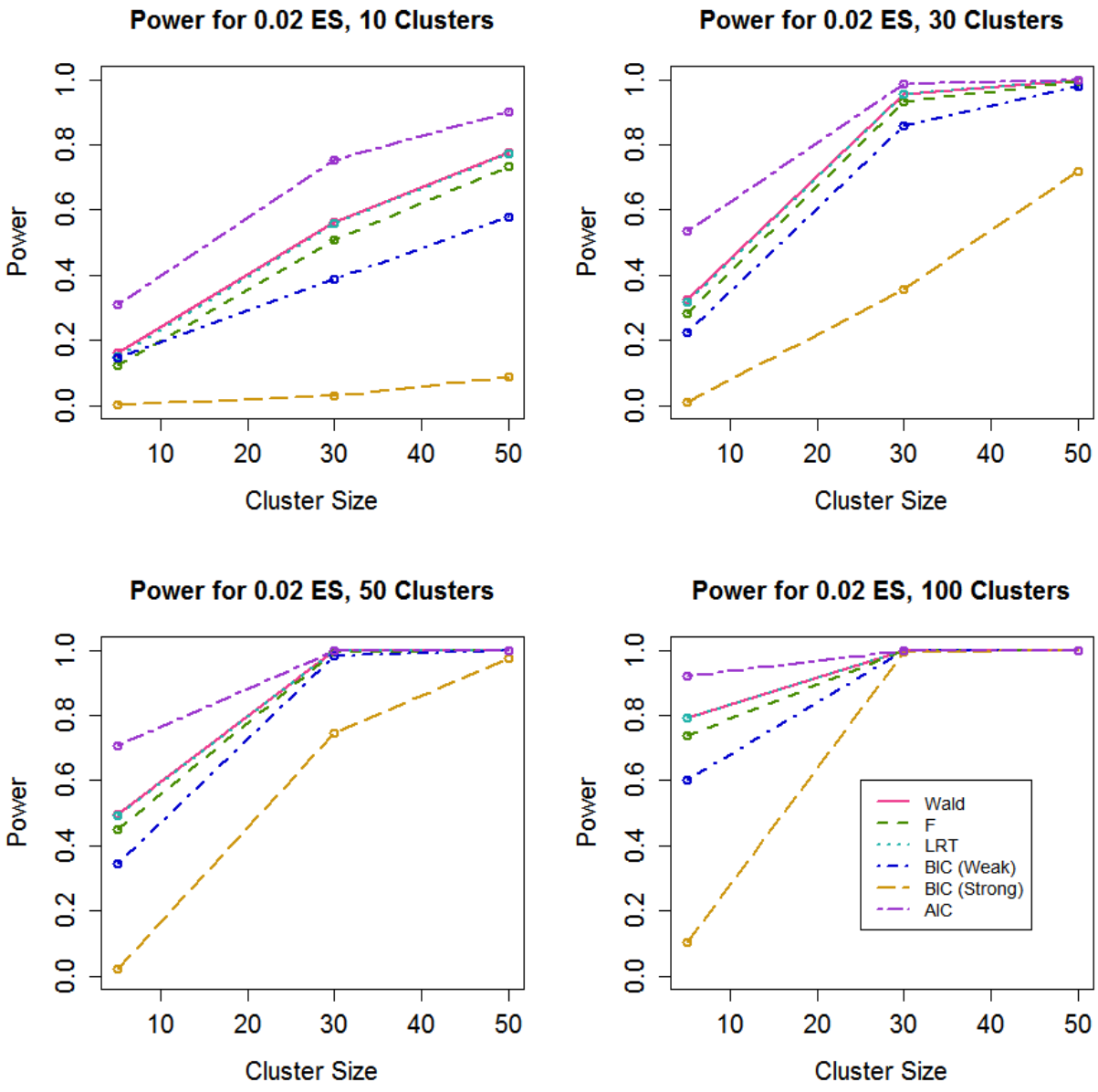

Figure 2. Power for each sample size condition with a small effect size.

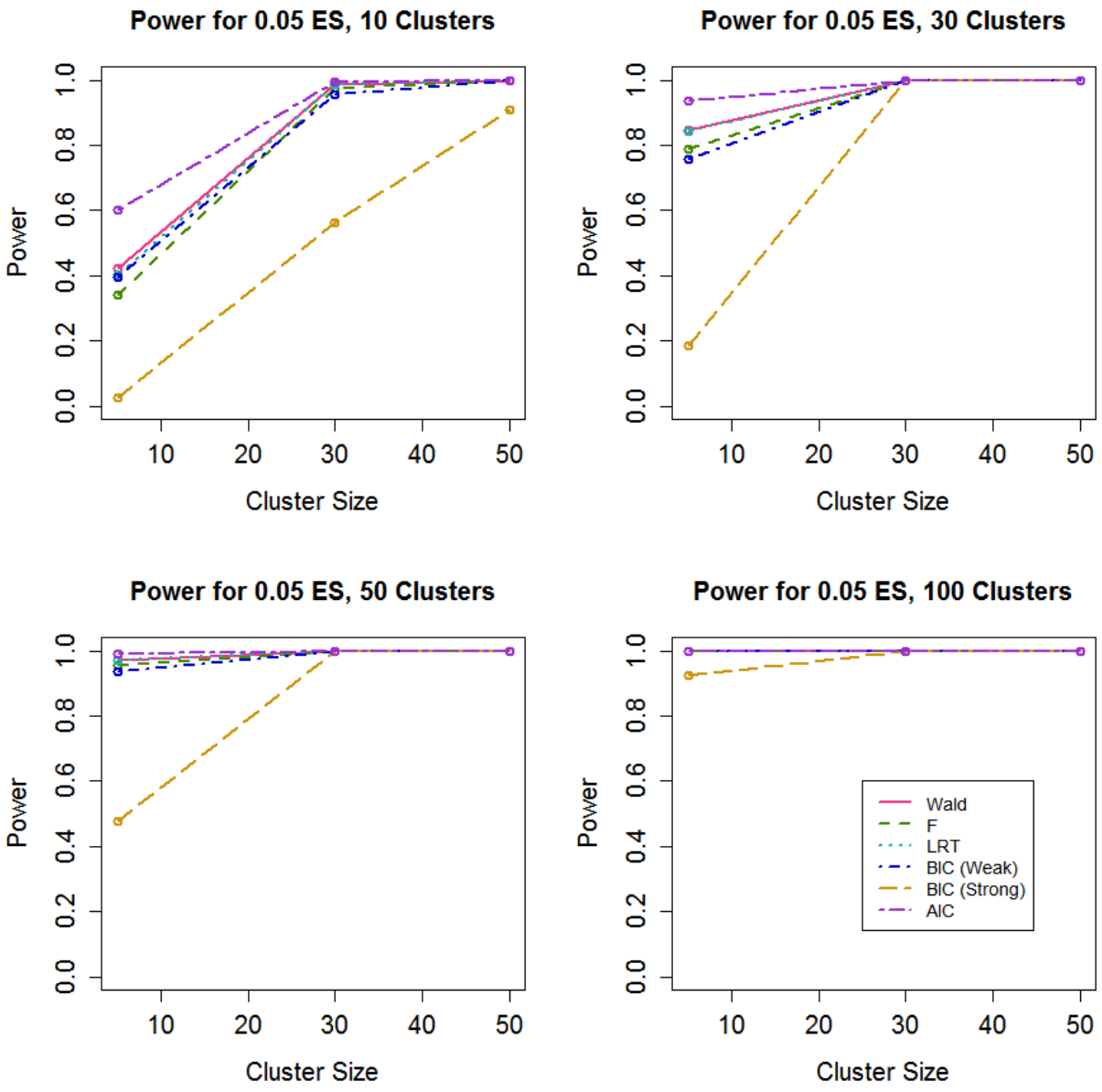

Figure 3. Power for each sample size condition with a medium-small effect size.

Across all conditions, the power significantly increases as cluster size increases. Further, the power significantly increases as number of clusters increases. These results were expected, as a larger dataset provides more information to detect effects. When there is a larger number of clusters, the increase in cluster size has less impact. Analogously, when there is a larger cluster size, the number of clusters has less impact. This is partially because the effect size is large enough that the tests are essentially hitting a ceiling effect (all 5000 samples find a significant effect). Perhaps if the effect size

was smaller, the increases in cluster size and number of clusters would have a larger impact on the power of each test. This is true for the small effect size condition, but particularly true for the medium-small effect size condition, where many of the sample size conditions displayed a power of one.

Additional Analyses

Although not explicitly performed to answer research questions 1 and 2, some additional analyses were performed based on the simulated data to gain further insight and help contextualize and interpret the results presented thus far. First, b_3 (the interaction term slope coefficient) will be examined for bias to ensure that this is not contributing to observed Type I error rates and power. Next, agreement among methods will be examined to understand whether the observed Type I (and Type II) errors are a function of specific outlying datasets, or more a function of the differences between the methods themselves. Finally, significance of number of clusters and cluster size is examined from a different modeling perspective (logistic regression) to provide additional evidence as to whether Type I error and power vary as a function of these factors. Additional analyses regarding this topic were thus performed.

Observed b_3 values are unbiased. To assess for bias in b_3 (the interaction term slope coefficient), the observed values for this coefficient were averaged. In this study, a difference of zero was specified between group 1 and group 2 slopes for the no effect size conditions; a difference of 0.50 for the small effect size conditions; and a difference of 1.00 for the medium-small effect size conditions. Since the moderator variable, M, was effect coded (-1 and 1), the slope coefficient for the interaction term variable (b_3) should be 0.00, 0.25, and 0.50 respectively. The observed b_3 values averaged across the simulated datasets come very close to these expected values as shown in Table 6.

Table 6

Observed versus Expected Values for b_3 (Interaction Term Slope Coefficient)

	b_3			
Effect Size Condition	Expectation	Observed Mean	Observed SD	No. Sims
0.00	0.0000	-0.00043	0.12546	60000
0.02	0.2500	0.24984	0.12529	60000
0.05	0.5000	0.49969	0.12681	60000

Since the observed distribution of values of a given parameter approximates its sampling distribution (Mooney, 1997), these results indicate that the ML estimate of b_3 is not biased. Further, I computed the observed mean for each sample size condition separately (rather than aggregated as in Table 6) and similarly found no bias in b_3. This was expected since ML estimates for fixed effects have been shown to be unbiased (Hox, 2010). Since the Wald test is based on both the parameter estimate and its standard error (Hox, 2010), any deviations from the nominal Type I error rate of 0.05 are likely due to bias in the standard error estimate. Given that previous research has found a downward bias for the standard errors of fixed effects with less than about 50 groups (Hox, 2010), it is expected that the standard errors for the Wald test would be downwardly biased, inflating the Type I error rates for small samples. With 10 clusters of size 5 each, the Type I error rate for the Wald test is 0.07 in this study but reduces to the nominal rate of 0.05 with 10 clusters of size 30 each. With larger numbers of clusters, the Type I error rate is generally acceptable as well.

Agreement among methods. Although it is clear that the methods have different Type I and Type II error rates, it is not clear from these overall proportions whether the errors are committed on the same datasets, or whether the methods are committing errors in different ways. One could imagine that there would exist some simulated datasets that are essentially outliers and these would result in Type I or Type II errors for all methods. Alternatively, it is possible that the methods commit errors with different types of datasets, in which case there would be less consistency in the decisions suggested by various methods for a given dataset. To further the understanding of why certain methods might suggest results that represent a Type I or Type II error, the number of samples observed for each "response pattern" was compiled for both the no effect conditions (Table 7) and the positive effect size conditions (Table 8).

According to the specification of the simulation, a value of zero indicates the correct decision. In Table 7, "No. Sims" shows the observed number of simulated datasets with the given pattern and the percent shows the percentage of observed simulations with the given pattern.

Based on these results, around 84% of the simulated datasets were found to have no significant interaction by all the methods, which is the correct result. The AIC, which was shown to be the most

liberal test, is the only test showing significant interaction in about another 10% of the datasets. After this, there are some patterns that are more common than others. The 1.3% of datasets that were incorrectly found to have a significant interaction by all 5 methods probably represent datasets that are "outliers" in some way. Also, the 2.3% of datasets that were found to have a significant interaction with all methods except the BIC are also probably partial "outliers" as the BIC was shown to be the most conservative method. In between those extremes, the patterns are more variable. Although the rest of the patterns don't represent a large percentage of datasets, they do indicate that the methods may be assessing different information since the more liberal or conservative methods are not always the most/least likely to find the significant interaction. Therefore, it can be informative for researchers to consider evidence from multiple test procedures, as they are each evaluating on somewhat different information, rather than just a more liberal or conservative version of another test.

Table 7

Patterns of Significance Decisions for No Effect Size Conditions

Wald test	LRT	*F*-test	BIC (weak)	AIC	No. Sims	Percent
0	0	0	0	0	50330	83.88%
0	0	0	0	1	5931	9.89%
0	0	1	0	0	18	0.03%
0	0	1	0	1	572	0.95%
1	0	0	0	1	39	0.07%
1	0	1	0	1	12	0.02%
1	1	0	0	1	825	1.38%
1	1	0	1	1	112	0.19%
1	1	1	0	1	1377	2.30%
1	1	1	1	1	784	1.31%

Note. For each test, 0 indicates "not significant" and 1 indicates "significant" for the interaction effect.

Table 8 shows the patterns of significance for the two effect size conditions that are positive. According to this specification, a value of one indicates the correct decision. No. Sims shows the observed number of simulated datasets with the given pattern and the percent shows the percentage of observed simulations with the given pattern.

Table 8

Pattern of Significance Decisions for Positive Effect Size Conditions

Wald test	LRT	*F*-test	BIC (weak)	AIC	No. Sims	Percent
0	0	0	0	0	11793	9.83%
0	0	0	0	1	6135	5.11%
0	0	1	0	0	6	0.01%
0	0	1	0	1	621	0.52%
1	0	0	0	1	219	0.18%
1	0	1	0	1	54	0.05%
1	1	0	0	1	1789	1.49%
1	1	0	1	1	980	0.82%
1	1	1	0	1	3660	3.05%
1	1	1	1	1	94743	78.95%

Note. For each test, 0 indicates "not significant" and 1 indicates "significant" for the interaction effect.

The highest frequency pattern for conditions with small or medium-small effect size was detection of a significant interaction by all methods. This correct decision by all methods represents almost 80% of the datasets evaluated. Among the incorrect or partially incorrect patterns, the most common pattern was that all methods missed detecting an interaction and this represents about 10% of the generated datasets, probably concentrated among the smaller sample sizes, considering the results showing increased power for higher sample sizes (Tables 4 and 5). Another common pattern was for all methods to fail to detect an interaction except the AIC, which occurred among about 5% of the datasets. Again, these different patterns show that the methods are not simply more liberal or conservative versions of each other, and instead evaluate somewhat unique information, meaning a researcher can probably gain information from conducting more than one model selection procedure. More specific suggestions for applied procedures are provided in the discussion section.

Significance of cluster size and number of clusters. This study provides the Type I error rates and power rates for each of the different sample size conditions, and this variation is further tested for significance in this section, building on the ANOVA models previously examined. With such a large number of simulations (large sample size), differences may be likely to be significant even if they are small, but it still may be illuminating to examine these differences from a statistical significance framework. To do this, a logistic regression for binary outcomes was performed for each of the five

methods and each of the three effect size conditions (15 models total) with the binary significant/not significant outcome. For each model, the number of clusters and cluster size were used as the two independent variables (Table 9).

Table 9

Significance of Number of Clusters and Cluster Size for Binary Interaction Significance Outcome

	Zero Effect		Small Effect		Medium-Small Effect	
	No. Clusters	Cluster Size	No. Clusters	Cluster Size	No. Clusters	Cluster Size
Wald test	negative	ns	positive	positive	positive	positive
LRT	negative	ns	positive	positive	positive	positive
F-test	ns	ns	positive	positive	positive	positive
BIC (weak)	negative	negative	positive	positive	positive	positive
AIC	negative	ns	positive	positive	positive	positive

Note. ns = not significant; negative = significant, negative relationship; positive = significant, positive relationship.

Results indicate that as the number of clusters increases and the cluster size increases, power increases for all methods for both the small and medium-small effect size conditions. For the no effect size conditions, as the number of clusters increased, the likelihood of finding a significant interaction effect decreased for all methods examined except the *F*-test. This is probably due to the downwardly biased standard errors that were observed with small sample sizes. The cluster size was not related to Type I error rates for all methods except the BIC. Since the BIC explicitly includes sample size within the computation, it makes sense that the cluster size is related to the likelihood of detecting a significant interaction with the BIC. As the cluster size increases, the likelihood of detecting a significant interaction decreases when using the BIC.

Chapter V: Applied Analysis Results

In order to demonstrate use of moderated regression within a multilevel framework, an applied analysis was conducted. Although the results of this analysis may have substantive significance, the main focus of this chapter is the methodological process. For demonstration purposes, the data and measures used, along with analysis results are presented. Following that, the five methods examined in the simulation study are used to assess interaction effects for this substantive example. Results from these five methods are interpreted in light of the simulation study findings. The substantive question examined asks whether gender moderates the relationship between perceptions of work-life balance and student confidence among a sample of undergraduate engineering students.

Data

The analysis used quantitative survey data from the Project to Assess Climate in Engineering (PACE) funded by the Alfred P. Sloan Foundation. Data from this project has been analyzed previously to examine patterns of attrition and covariates to attrition among undergraduate engineering students (Litzler & Young, 2012) and to examine engineering student confidence and how it varies as a function of gender and race/ethnicity as well as various self-reported measures of climate and student perception (Litzler et al., 2014). More details regarding the participating schools and survey methodology can be found in previously published work (see Litzler & Lorah, 2013; Litzler et al., 2014; Litzler & Young, 2012). The survey used for this analysis was fielded in 2012 within 15 engineering schools.

Measures

The outcome (DV) variable for this analysis was student confidence which was computed as an average of 6 individual survey items. The predictor (IV) was a measure of work-life balance computed as an average of 4 individual survey items. Nine of the ten survey items were measured based on a 5-point rating scale with the following 5 options: strongly disagree; somewhat disagree; neutral; somewhat agree; strongly agree. The exception to this set of response options was for the last confidence item (item 6) which offered the following five response options: far below average; below average; average; above

average; far above average. The binary moderator (MV) was gender. The individual confidence and work-life balance items as well as Cronbach's alpha for both variables are provided in Table 10.

Table 10

Individual Items Averaged for Confidence and Work-Life Balance Variables

Confidence ($\alpha = 0.83$)
1. I am confident in my ability to succeed in my college engineering courses
2. I am confident in my ability to succeed in my college science courses
3. I am confident in my ability to succeed in my college math courses
4. I am confident in my overall academic ability
5. I am confident that someone like me can succeed in an engineering career
6. Compared to other students in my classes, I think my abilities in my engineering classes are:
Work-Life Balance ($\alpha = 0.67$)
1. Engineers can leave and come back to their careers more easily than can people in other professions
2. Engineering is a field that supports people who want to have children and continue working
3. Engineers can design their own work schedules
4. Engineering is a field that supports a balance between work and family life

In addition to the three substantive variables (DV, IV, and MV), a variable indicating school membership was used to indicate level-two cluster membership for the multilevel analysis.

Analysis

The empty model (M0; equation 10), the main-effects only model (M1; equation 11), and the full model including the interaction term (M2; equation 12) were estimated with the *lmer* function in R using ML estimation. All descriptive statistics were computed and plots were created using R.

Results

Results for this particular substantive analysis will be presented and interpreted, followed by evaluation using the five model comparison methods examined within the simulation study. Implications will be discussed. The total sample size was 7482 students and 15 schools indicating an average cluster size of approximately 500. The actual cluster sizes ranged from 251 to 789. Descriptive statistics for the three variables are provided in Table 11.

Table 11

Descriptive Statistics for Applied Dataset

	Descriptive Statistics				Zero-Order Correlations		
Variable	*M*	*SD*	*Min*	*Max*	1.	2.	3.
1. Confidence	4.16	0.63	1	5			
2. Work-Life Balance	3.19	0.68	1	5	0.17*		
3. Female	0.43	0.50	0	1	-0.12*	-0.05*	--

Note. N = 7482 students. * $p < 0.05$. Dummy coding (yes = 1; no = 0) was used for *Female* descriptive statistics (i.e. 43% of the sample is female) but effect coding (female = 1; male = -1) was used for analysis.

The descriptive statistics indicate that average student confidence falls slightly above the somewhat agree category, while the average perception of engineering work-life balance falls slightly above the neutral category. Measures for both confidence and work-life balance perceptions ranged from one to five. Visual examination of the confidence measure's histogram indicated left-skew (higher concentration of students at higher levels of confidence), and computation indicated skew of approximately -1. The work-life balance measure showed very little skew with a value close to zero (symmetric).

The sample is somewhat evenly split between males (n = 4118) and females (n = 3105) with a slightly higher number of males. Sample correlations indicate that confidence and work-life balance perception are positively correlated. Confidence is negatively associated with being female and work-life balance perception is also negatively associated with being female, but both associations are weak.

Model results. The fixed-effect results from all three multilevel models are provided in Table 12.

Table 12

Fixed Effects for Three Applied Multi-level Models

	Coeff	*SE*	*t*	*p*
Empty Model (M0)				
Intercept	4.17	0.02	198.20	<0.01
Main Effects Only Model (M1)				
Intercept	3.68	0.04	92.04	<0.01
Work-Life Balance	0.15	0.01	14.07	<0.01
Female	-0.07	0.01	-9.65	<0.01

Full Model (M2)				
Intercept	3.69	0.04	91.91	<0.01
Work-Life Balance	0.15	0.01	13.77	<0.01
Female	0.01	0.03	0.42	0.67
Work*Female	-0.03	0.01	-2.54	0.01

Results from the empty model (M0) indicate that the average value of student confidence was about 4.17 after accounting for school membership. This corresponds closely with the observed confidence mean of 4.16. Results from the main-effects-only model (M1) indicate that Work-Life Balance is positively associated with Confidence with an increase of one point (out of the 5-point Likert scale) in engineering work-life balance perceptions associated with an average, expected increase of 0.15 points in student confidence after controlling for gender. Also, being female rather than male is associated with an average, expected confidence of approximately 0.14 points lower (the estimate is doubled because Female is effect coded) after controlling for work-life balance perceptions.

The results from M2 indicate there is a significant interaction between gender and work-life balance perceptions, based on the Wald test (t = -2.54, $p < 0.05$). Main effects within a regression including product terms must be interpreted differently from the main effects in a model with no product terms. These main effects are now called "conditional effects" (Aiken & West, 1991, p. 37) meaning the slope estimates are accurate, conditional on a value for another variable. In this case, the estimate for gender is valid only at the point where work-life balance perceptions equal zero. In other words, for students with work-life balance perceptions of zero, there is no difference in confidence between males and females (t = 0.42, $p > 0.05$). Since work-life balance perceptions ranges from one to five in this dataset, the finding conditional at zero is not very meaningful. This finding does not, however, indicate whether there is a gender difference in confidence for students at higher values of work-life balance perceptions and the subsequent plot for the interaction effects, along with the results from M1, implies that there probably is a gender different in confidence at higher levels of work-life balance perceptions

(see Figure 4). This example highlights the caution that must be taken when interpreting these conditional main effects in an interaction model.

Interaction interpretation. In order to interpret significant interaction effects, plotting should be used (Aiken & West, 1991). This is accomplished by identifying the regression line for separate levels of the MV (Aiken & West, 1991). First, consider the overall prediction equation which is simply the model specification (M2) with the estimated intercept and slope coefficients (Table 12) substituted for B_0, B_1, B_2, and B_3. This is shown below as equation 13. From there, the prediction equations for females and males can be computed separately by substituting in the relevant value for the *Female* variable. Since this variable is coded as +1 for females, simply substituting +1 for *Female* into the equation 13 yields a regression equation for female students (equation 14a). Analogously, substituting -1 for *Female* yields a regression equation for male students (equation 14b).

Overall Confidence = 3.69 + 0.15*Work-Life + 0.01*Female + -0.03*Work*Female (13)

Female Confidence = 3.70 + 0.12*Work-Life (14a)

Male Confidence = 3.68 + 0.18*Work-Life (14b)

The equations for males and for females (equations 14a and 14b) are plotted in Figure 4.

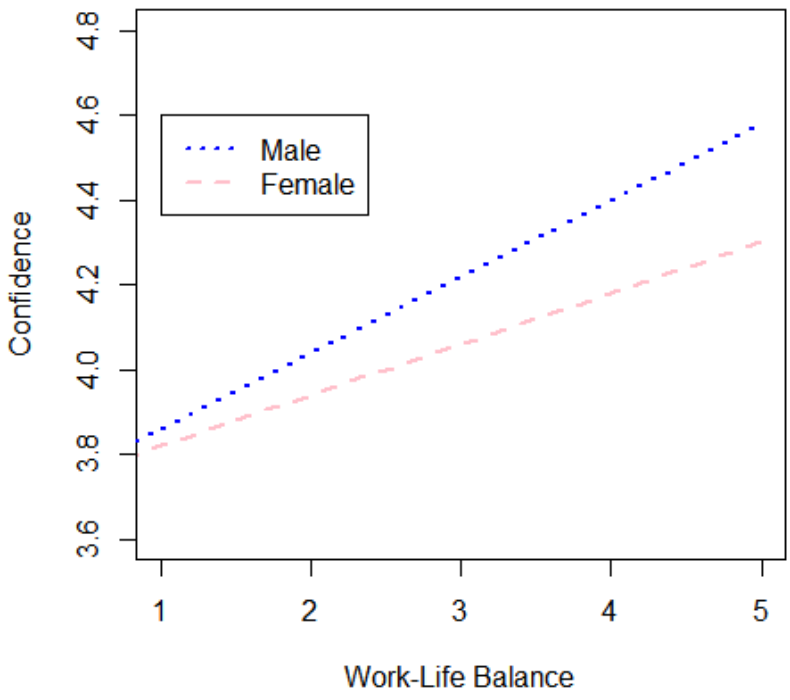

Figure 4. Full model results illustrating interaction effect

This plot demonstrates that the relationship between work-life balance and confidence is different for males and females. The significance of the interaction term indicates that this is a statistically significant difference. As positive perceptions of engineering work-life balance increase, a larger increase in confidence is observed for males than for females. At low perceptions of work-life balance, the confidence of males and females is expected, on average, to be approximately the same; however, at high perceptions of work-life balance, a confidence gap of a bit less than half a point (out of a five-point rating scale) emerges with males indicating higher expected confidence compared with females. Although it is beyond the scope of this paper to speculate as to why this difference might exist, these results provide some evidence that the mechanisms related to building confidence for female engineering students may differ from those for male engineering students.

Method Comparison. The previous analysis proceeded by assessing the interaction term for significance with a Wald test. In practice, the Wald test is used most often for evaluating fixed effects, due to its ease of use and convenience (Hox, 2010). Due to this common pattern, it is possible that

researchers might proceed with the Wald test without even considering alternative methods for evaluation of interaction effects. However, the simulation study presented in Chapter 4 suggests that further insight might be gained by conducting additional tests, and that perhaps alternative tests are preferable. These additional tests will be performed and evaluated in the following section and include all five model comparison methods examined thus far. Table 13 shows model estimates for relevant quantities.

Table 13

Random Effects and Model Fit Indices for Three Applied Multilevel Models

	M0	M1	M2
School Variance	0.01	0.01	0.01
Student Variance	0.39	0.38	0.38
No. Clusters	15	15	15
Overall *N*	7481	7219	7219
BIC	14294.1	13516.7	13519.1
AIC	14273.3	13482.3	13477.8
Deviance	14267.3	13472.3	13465.8
ICC	0.01	0.02	0.02

Note. M0 = empty model; M1 = main effects only model; M2 = full model; ICC = intraclass correlation coefficient.

The school and student variance estimates represent the ML partition of variance and are given as the random effects. The number of clusters is the number of schools in this dataset and the overall *N* represents the number of students. The BIC, AIC, and deviance are all given by the lmer summary function, as they were in the simulation study. The ICC was computed according to equation 1 and is based on the school and student variance estimates. These quantities, along with the *t* value for the interaction effect used already in the Wald test, were used in the various model selection methods used in this study.

Table 14

Multiple Procedures for Testing Work-Life Balance/Gender Interaction

Procedure	Test Statistic Value	Comparison Value	Evidence for Interaction
Wald test	-2.54	+1.96/-1.96	yes
F-test	5.43	3.84	yes

LRT	6.50	3.84	yes
BIC (weak)	-2.40	0	no
BIC (strong)	-2.40	10	no
AIC	4.50	0	yes

Note. Test statistic value must exceed comparison value in order to provide evidence for significant interaction effect. Results based on full model (M2) results.

Table 14 provides the evidence and outcome of each of the model selection methods used to assess the interaction between gender and work-life balance for significance. Results indicate that the relevant outcome varied as a function of which test was used.

The first test examined, the Wald test, was examined previously and the test statistic is the *t* value associated with the interaction term in M2. As this is a two-sided test, the test statistic value must either be larger than +1.96 or smaller than -1.96 to indicate significance at the 95% level. The Wald test is the only two-sided test examined. As demonstrated previously, the Wald test provides positive evidence for an interaction effect. The test statistic for the next test, the *F*-test, was computed according to equation 4. This computation uses R^2 values (equation 5) which are based on the ML partition of variance. This *F*-test also provides positive evidence for an interaction effect since 5.43 is larger than the critical value of the *F*-distribution (df = 1, 7213) which equals 3.84. The LRT test statistic value is computed as the difference in deviances between two models. This test statistic value is compared to the critical value for the chi-square distribution (df = 1) which equals 3.84. The likelihood ratio test also provides positive evidence for an interaction effect. The test statistic value for the BIC (or AIC) is computed as the difference between BIC (or AIC) values for two models. For BIC and AIC, a negative test statistic value indicates a better fit for the simpler (main effects only) model while a positive value indicates a better fit for the more complex (full) model. For this model, BIC indicates that the simpler (main-effects-only) model provides a better fit when using either the strong criteria (difference must exceed 10) or the weak criteria (difference must exceed zero). Conversely, the AIC test statistic exceeds zero indicating that the more complex (interaction term) model provides a better fit. Of all methods examined, only the BIC fails to provide evidence for a significant interaction effect. Rather, the BIC provides evidence for the main

effects only model that does not include the interaction effect. The proceeding sections further elaborate possibilities for making a substantive conclusion based on this conflicting evidence.

Effect size. Further insight can be gained by computing the effect size for the interaction term. Based on the ANOVA partition of variance, η^2 can be computed as the sum of squares due to the interaction term divided by the total sum of squares and a value of about 1% is considered small (Jaccard et al., 1990). In this study, η^2 for the interaction term is 0.0009 or about 0.09% which is much smaller than even the "small" effect size guideline. Another measure, f^2, was computed based on the ML partition of variance and the associated R^2 values according to equation 8. For f^2, a value of about 0.02 or 2% is considered small (Aiken & West, 1991). For this study, f^2 was 0.0008 or about 0.08% again indicating a very small effect size. This information may be of interest to the applied research when making substantive conclusions and interpreting model results, especially in the case of conflicting evidence of significance.

Implications

For this analysis, the Wald test, *F*-test, LRT, and AIC indicated a significant interaction effect, meaning that these methods could have committed a Type I error. Conversely, both the strong and weak criteria for the BIC indicated no interaction effect, meaning that this method could have committed a Type II error. The simulation study conducted in chapter 4 can inform our understanding of these possibilities. The Type I error rates did not show much change based on sample size, either in terms of number of clusters, or cluster size for any method, except at the very smallest sample sizes. For this analysis, the sample size of approximately 7000 slightly exceeds the conditions examined in the simulation, but considering the Type I error rates seem to be fairly constant, slight extrapolation seems appropriate. This extrapolation indicates that for the AIC, the Type I error rate should be around 0.15 and for the three hypothesis testing methods, the Type I error rate should be around 0.05. Generally, rejection of the null hypothesis becomes easier as the sample size increases (Tabachnick & Fidell, 2007), indicating that the large sample size may also have been a factor in the applied analysis.

The BIC was the only method that did not find the interaction to be significant. BIC is the only method that systematically makes it more difficult to reject the simpler model as the sample size increases (see equation 7). Without the presence of a significant effect, Type I error is not a possibility. Rather, it is possible that a decision based on the BIC represents a Type II error, which is the probability of failing to detect a significant interaction effect. It should be noted at this point that as the BIC is not a hypothesis testing method, the concept of Type II error does not precisely apply. However, for purposes of continuity and demonstration, this analysis will proceed by assuming that the researcher may use the BIC to make a "significance" decision, meaning the implications of this decision can still be explored. The probability of a Type II error is simply 1 minus power (Tabachnick & Fidell, 2007, p. 36). Therefore, although not explicitly considered in the simulation study, Type II error may easily be computed based on the power results displayed.

Results from the simulation study indicated that power increases as effect size increases. Although the effect sizes used in the simulation study were quite small (0.02 and 0.05), the effect size in the applied study was quite a bit smaller yet (0.0008). So although the simulation study found a ceiling effect for power as sample size increased, it is likely that power was still significantly lower at this small effect size. Although it is unclear precisely what power might have been for this analysis with a particularly small effect size, the simulation study indicates that BIC shows the lowest power, and that all methods show increasing power with increasing sample size. Since BIC consistently shows the lowest power, it makes sense that only BIC failed to detect significant interaction effects compared with the other methods. Analogously, it is also unclear exactly what Type II error rate would be expected at this very small effect size.

Another difference between the simulated data and the applied analysis is ICC. Although ICC was not systematically varied as part of the simulation study, it was set at approximately 0.10. For the applied data, the observed ICC is approximately 0.01. Research indicates that as the ICC increases, for any given sample size, the effective amount of information decreases (Snijders & Bosker, 1999). This effect then is similar to the effect we might see by varying the sample size. The implication for the

applied study is that by having a lower ICC, power might be expected to be slightly higher than with a higher value of ICC. As sample size did not seem to impact type I error rates, the type I error rates from the applied study should not differ from those shown in the simulation study as a function of ICC.

Although the results from these five methods seem to differ in their substantive conclusions, it is possible that understanding subtle differences in the stated goals of each test would indicate that they are actually consistent and that the difference in outcome depends on the desired substantive conclusion rather than the presence of a Type I or Type II error as stated above. In a hypothesis testing framework, researchers typically hope to reject the null hypothesis (Tabachnick & Fidell, 2007, p. 36) whereas the BIC and AIC can be used to provide evidence for or against the simpler model (Weakliem, 2004), where the simpler model is analogous to the null hypothesis in this case. Results from a null hypothesis test indicate whether the result was likely obtained by chance (Tabachnick & Fidell, 2007, p. 37). In comparison, the BIC is designed to "maximize the probability of picking the correct model" (Kieseppa, 2003, p. 1274) and the AIC is designed to maximize the "predictive accuracy" of a model (Kieseppa, 2003, p. 1275). So for this applied example, it is possible that making a decision in favor of the interaction effect correctly indicates that the result was not obtained by chance and that the model provides the best predictive accuracy; while making a decision in favor of removing the interaction term represents the correct model.

Hypothesis testing procedures have also been critiqued for being particularly sensitive to sample size and researchers have suggested that the practical importance of a finding, represented by the effect size, also be considered (Tabachnick & Fidell, 2007). However, this distinction between statistical and practical (or substantive) significance has been critiqued, and it has been suggested that the distinction is invoked when the real issue is the failure to attenuate *p*-values as sample sizes increase (Raftery, 1995). To this end the computation of BIC (see equation 7) explicitly includes *N* which provides the function of attenuating *p*-values as sample size increases. Overall, given the small effect size evident for this interaction, a conclusion of significance based on hypothesis testing procedures may indicate that we have "statistical" significance but not "practical significance" (Tabachnick & Fidell, 2007), or it may just

indicate that the hypothesis testing procedures are simply flawed (Raftery, 1995). Either way, more closely examining the effect size as well as considering the weaknesses of hypothesis testing indicates that the work-life balance/gender interaction may not be a particularly relevant research finding after all.

Another distinction between the simulation study and applied analysis is the limitations evident in real data, such as measurement error and failure to meet distributional assumptions. According to Tabachnick & Fidell (2007) "Regression analysis assumes that IVs are measures without error, a clear impossibility in most social and behavioral science research. The best we can do is choose the most reliable IVs possible" (p. 122). In this analysis, the IV was perceptions of work-life balance and this measure has an internal consistency of 0.67 which indicates a fair amount of measurement error in the variable and which may have attenuated power in this case. Gender, which was the MV in this analysis, may have been measured without much measurement error. In contrast to the data for the applied analysis, the data created for the simulation study was assumed to be measured without error. Generally, measurement error has been shown to be a particular problem with moderator analyses (Aiken & West, 1991; Holmbeck, 1997; Jose, 2013) so it is likely that measurement error played a significant role in the applied analysis within this study.

Another assumption of linear regression methods is the assumption of normality of errors of predictions, which can be assessed with a plot of residuals (Tabachnick & Fidell, 2007). Visual examination of the residuals plot for M2 (full model) in the applied analysis indicates that this assumption may not be met. Prediction errors are slightly more likely to be negative than positive, across the range of predicted confidence scores, probably due to the left-skew of the confidence variable, which essentially implies a ceiling effect in the data. This result calls into question the validity of model results.

In addition, the small effect size for the interaction term also aligns with the common consensus that interaction effect sizes are typically very small (Aiken & West, 1991; Alexander & DeShon, 1994; Holmbeck, 1997; Jose, 2013; MacCallum & Mar, 1995).

Chapter VI: Discussion and Conclusions

This chapter discusses findings for each of the five methods examined in this study. In addition, a suggested procedure for researchers investigating interactions with multilevel models is provided, followed by suggestions for future research and conclusions.

Wald Test

Type I error rates for the Wald test typically were not significantly different from the nominal rate of 0.05 except for a few small sample size conditions. The standard error used for this test has been shown to be downwardly biased for small samples (Hox, 2010) which will result in inflated Type I error rates. Further, the test is asymptotically unbiased (Hox, 2010) meaning it should function correctly for large samples. Since this analysis showed no bias in the slope coefficient itself (b_3), it seems very likely that the inflated Type I error with the Wald test at small sample sizes evident in this analysis is due to downwardly biased standard error estimates. This analysis showed an inflated Type I error rate significantly different from 0.05 for two sample size conditions: 10 clusters of size 5 each and 30 clusters of size 30 each. The observed Type I error rate at the smaller of the two sample size conditions (10 clusters of size 5 each) was more severely inflated at 0.068 while at the second condition the observed Type I error rate was a less-concerning 0.057. Therefore, in terms of practical guidance, it seems that researchers should interpret results from the Wald test with caution at sample sizes of 50 or less. It is not clear exactly how large the sample should be in order to maintain a Type I error rate of 0.05 with the Wald test, so future research should examine this question more precisely.

Previous research has found that at least 50 groups is needed to avoid inflated Type I error rates for the Wald test for fixed effects (Hox, 2010). This analysis indicates that around 30 groups is probably sufficient to maintain reasonable Type I error rates. Although it is not clear exactly why these results differ, it is possible that for higher values of ICC, the number of groups needed will increase, since the effective information available decreases. Further, results from this study indicated that aside from a very small sample size (10 clusters of size 5 each), Type I error rates with the Wald test did not vary as a function of sample size.

The Wald test showed increasing power as sample size increased, which was expected. For a small effect size ($f^2 = 0.02$), the Wald test achieves power of around 80% with 10 clusters of size 50 each, 30 clusters of size 30 each, 50 clusters of size 30 each, or 100 clusters of size 5 each, assuming no measurement error in the measures. For a medium-small effect size ($f^2 = 0.05$), 10 clusters of size 30 each, or 30 or more clusters of size 5 each should result in approximately 80% power or higher.

The Wald test is commonly used for fixed effects due to its ease of use. Further, for testing interaction effects in multilevel models, the Wald test provides good control of Type I error at all but the smallest sample sizes, and reasonable power as well. Based on these results, it is recommended that the Wald test is used as a first step for researchers examining interaction effects.

***F*-test**

The *F*-test performed very similarly to the Wald test and LRT. The Type I error rates were slightly lower than the Wald test and LRT at smaller sample sizes, but seemed to converge as the sample size increased and typically maintained rates very close to the nominal rate of 0.05. Although the *F*-test did not display any instances of inflated Type I error, there were a few conditions where the rates were overly conservative. This may have occurred because the estimated variance components on which the computation of *F* is based have been shown to be downwardly biased for smaller sample sizes (Hox, 2010).

Similarly to the Wald test and LRT, Type I error rates did not vary as a function of number of clusters or cluster size. One advantage to the *F*-test as compared with the Wald test and LRT, is that it did not display inflated Type I error rates at small sample size conditions. Therefore, for hypothesis testing procedures, this test might be advantageous particularly for smaller sample sizes.

Power for the *F*-test followed patterns very similar to the Wald test and LRT. At smaller sample sizes, power slightly lowers, but these power rates converged as the sample sizes increased.

The procedure for computing this test involves a few extra steps, so for larger sample sizes where the results converge with the results of the Wald test and LRT, it is probably not worth the extra step. For small sample sizes however, this test could be particularly useful for controlling Type I error rates.

LRT

The LRT virtually approximated the Wald test results under all conditions, which was not surprising since they are asymptotically equivalent (Hox, 2010). Due to the relative difficulty in performing this test as compared with the Wald test, it is not recommended to take this extra step in evaluating interaction effects in multilevel models.

BIC

The BIC represents an information criterion approach to model selection and is not a hypothesis testing procedure. Although the concepts of Type I error and power primarily refer to hypothesis testing, these concepts can also usefully be applied to the BIC results with the understanding that things like a nominal Type I error rate do not apply in this situation. Given that, the Type I error rates for BIC generally were less than 0.05 and for all but the smallest sample sizes, were significantly less than 0.05. Additionally, the Type I error rates with the BIC were the lowest among all the methods, for all but the smallest sample size condition. Raftery (1995, p. 141) provides approximate *p*-values corresponding with different sample sizes. These expected values can be roughly compared with the values observed in this study. For example, for $n = 50$, weak evidence with the BIC should result in a *p*-value of approximately 0.053; in this study, $n = 50$ represents a slightly smaller effective sample size because of nesting in the data, and we might expect the *p*-value to be slightly higher. In fact, this is the case with $p = 0.06$. Although this is larger than 0.05, for this sample size condition, it is not actually as liberal as the Wald test at $p = 0.068$. As another example, Raftery (1995, p. 141) indicates that for $n = 100$, weak evidence with the BIC results in a *p*-value of 0.032 and for $n = 1000$, p should be about 0.009. This study finds that for $n = 150$ (30 clusters of size 5 each), $p = 0.026$ and for $n = 1500$ (30 clusters of size 50 each and 50 clusters of size 30 each), $p = 0.006$ and 0.008, respectively. These values roughly align with the expected values. The very strong criteria for the BIC maintained Type I error rates around zero, indicating it is unlikely to make a Type I error. The weak criteria for BIC was the only method that showed Type I error rates that varied as a function of sample size, with rates decreasing as sample size increased.

This tendency to attenuate power as sample size increases makes this method particularly well-suited for studies with large sample sizes, as it provides a concrete way to differentiate between "practical" and "statistical" significance when sample size is large.

Although this indicates a conservative test in terms of Type I error rates, when using the weak criteria BIC still displays power almost as high as the other methods examined, but not higher. For a small effect size, power of approximately 0.80 can be attained with at least 50 clusters of size 30 each, and for a medium-small effect size, at least 10 clusters of size 30 should be used when assessing effects with the weak criteria. The strong criteria display significantly lower power than the weak criteria across most sample size conditions. Given these results, weak evidence with the BIC would be an appropriate test for determining the presence of an interaction effect, whereas the very strong criteria could potentially add strength to any significance claims.

An additional feature of the BIC is that it can be used to provide evidence for either model, in contrast to hypothesis testing methods which cannot provide evidence for the null model. Providing evidence for the null model, in this case the main-effects-only model, could be particularly useful in the case when a researcher wants to test the hypothesis of no interaction effects. Such a situation could arise, for example, when evaluating the assumption of homogeneity of regression slopes with an ANCOVA (DeShon & Alexander, 1994; Dretzke et al., 1982; Lomax, 2007; Tabachnick & Fidell, 2007). Note that the "power" for such a test would increase as sample size increases. This is evident from the decreasing Type I error rates and would be a desirable property for any test of significance. BIC is the only test that shows this property.

AIC

The AIC also represents an information criterion approach and is not a hypothesis testing procedure. This analysis showed Type I error rates consistently around 0.15 for all conditions. Although there is no nominal Type I error rate associated with this test, most researchers would consider this too liberal of a test for widespread use. Although some advocate for the use of AIC over BIC in multilevel

modeling to avoid the issue of ambiguous *N* (Hox, 2010), it seems that regardless of the way *N* is computed, the BIC still maintains more desirable Type I error rates as compared with the AIC.

Although using a threshold of zero for the difference in AIC values seems to result in unacceptably high Type I error rates, it is possible that a different threshold would yield more reasonable estimates. This question was briefly examined and results indicate that across all sample size conditions, a threshold of one (i.e. the difference between the AIC of the complex model and the AIC of the simple model is equal to one) results in a Type I error rate of 0.09; a threshold of two results in a Type I error rate of 0.05; and a threshold of three results in a Type I error rate of 0.03. Although further examination of this topic is beyond the scope of this study, it appears that this could be an interesting area for future research.

Suggested Procedure for Multilevel Moderation Research

The following procedure is suggested for evaluating interaction effects in multilevel models.

Step one: Perform the Wald test. This test is simple to use and would probably be a natural first step for most researchers anyway. It maintains good control over Type I error rates for all but the smallest sample sizes and displays reasonable power.

Step two: Perform the *F*-test if *N* is small. For particularly small samples, the Wald test displays inflated Type I error rates; in contrast, the F-test maintains rates not significantly different from 0.05 within the small samples sizes examined in this study. For a sample size of 50 or less, this procedure is recommended, but it is unclear exactly where a reasonable cut-off should be set. If results from the Wald test and *F*-test differ, the *F*-test should be preferred.

Step three: Evaluate the BIC if *N* is large or if evidence for the null model is desired. Since the BIC systematically reduces Type I error rates as sample size increases, it provides a convenient way to differentiate between "practical" significance and "statistical" significance. Alternatively, in the case where a researcher is hoping to find evidence against an interaction effect, the BIC can be used. Since BIC is the only method that shows systematic decline in Type I error rates as sample size increases, it

follows that it is the only method that would show increased "power" to detect the null hypothesis as sample size increases.

There is no exact threshold for what constitutes "large" N, but based on the simulation results, Type I error rates start to diverge more significantly around the condition of 10 clusters of size 30 each. Based on this result, a sample size of approximately 300 or larger could be considered "large."

If results from the Wald test and BIC model comparison test differ, it is not clear that either of these tests is "better" and so the researcher should use theory and substantive knowledge to make a judgement call. In addition, the researcher may want to collect additional information before making a decision, such as effect size, or an additional model comparison test, such as the *F*-test or the AIC model comparison procedure.

Step four: Compute effect size and create plot for significant interactions. After evaluating the interaction effect based on the previous three steps, if the effect is found to be significant, the effect size should be computed in order to contextualize the scope of the given effect, and a plot should be created to interpret that effect. More details regarding how to plot interaction effects can be found in Aiken & West (1991).

Future Research and Limitations

Further research should examine the choice of *N* for computing the BIC with multilevel models. Previous researchers have pointed to the fact that defining *N* with multilevel data is ambiguous (Hox, 2010; Raftery, 1995) and that the value should be reduced somehow (Raftery, 1995). However, in this study, the default BIC value produced from R was used, and this value uses observed sample size for *N*. Given the advantages of BIC demonstrated in this study, it is worth further investigating the method and better understanding the implications of various ways to adjust the value of *N* used in computation of the BIC. The effective sample size in a dataset with clustering is less than the actual sample size. Accounting for this by using a downwardly adjusted value for *N* when computing the BIC would result in a higher effective *p*-value. Given the conservative nature of the BIC test anyway, this seems entirely appropriate and would potentially result in increased power. This downward adjustment for *N* could be computed by

using the effective *N* which can be computed as a function of the ICC and cluster size (Snijders & Bosker, 1999). Further research could compare the Type I error rates and power when using BIC to detect fixed effects with the actual *N*, the effective *N*, and any other methods that appear plausible upon further investigation.

Future research should also examine Type I error rates and power for variables at different levels of measurement. For example, this study only examines a continuous outcome investigated with a linear regression model. Future research could examine interaction effects for a model with a binary or ordinal outcome through the use of logistic regression methods. In addition, different types of MV should be considered. This study examined a binary MV, but it is also possible that the MV could be continuous or discrete with three or more levels. It is not clear which methods would be best to detect interaction effects in these scenarios, and what might be a suitable procedure for doing so. It is possible that a continuous MV offers more power to detect interaction effects as they are inherently measured at a more granular level.

Another useful extension of this research could be to investigate cross-level and level-two interactions, rather than focusing solely on interactions at level-one. With educational data for example, analyzing data at the school level (level-two) can help researchers identify school-level variables that correlate with quality and equity at a school, knowledge which is critical to educational reform (Ma, Ma, & Bradley, 2008). Further, although few researchers actually examine interaction effects at the school level, such an analysis could be particularly useful, as it recognizes the complexity of school effects and their relationship with student outcomes (Ma et al., 2008). Researchers have suggested examining the interaction between school context and school climate which could help policy makers understand the context in which certain climate reforms might or might not be effective (Ma et al., 2008). Such an analysis would require an interaction term at level-two (school level). For these level-two and cross-level interactions, it is unclear what sample sizes are needed to maintain sufficient power, how many clusters should be used, and how Type I error rates vary between different methods, meaning it is unclear which method should be preferred.

This study examined only one value for ICC. It is possible that differing values of ICC affect the Type I error rates and power of different methods for examining interaction effects. As the ICC gets larger, the effective sample size and effective information available decreases. For that reason, it is likely that varying the ICC would produce similar results to varying the sample size. It would be interesting to understand whether computing effective sample size (Snijders & Bosker, 1999) would allow for sufficient understanding of variations in Type I error rate and power, or whether knowing the specific number of groups, group size, and ICC values for any given effective sample size would provide any additional information.

Another area that may require further investigation is the degrees of freedom used for the Wald test. Although research has shown that when degrees of freedom is approximately 30 (Hox, 2010) to 40 (Snijders & Bosker, 1999), the *Z* approximation may be adequate for the Wald test, this study showed inflated Type I error rates at the sample sizes slightly larger than 30-40. It would be interesting to understand if using a different approximation for degrees of freedom, such as the Sattertwaithe approximation would correct for this bias.

One limitation of this study is that the effects of measurement error are not explicitly considered within the simulation study. Measurement error has been shown to attenuate power particularly for detecting interaction effects (Aiken & West, 1991), and so the high power displayed in the simulation study for the larger sample size conditions may be attenuated by measurement error in practice. In the applied analysis, the small effect size paired with a low IV reliability seemed to indicate that measurement error was attenuating the interaction effect size, although it is not clear if in reality this was the case or if the interaction effect was simply a Type I error.

In addition, the results from the simulation study may be limited compared with real data since real data may not always meet model assumptions. For example, in the applied analysis, the distribution of residuals did not indicate normality which may have attenuated power and/or the interaction effect size compared with what might be expected from the simulation study.

As with any simulation study, the results are not immediately generalizable outside the conditions examined here and are therefore limited to specific values for number of clusters, cluster size, interaction effect size, ICC, and five specific methods.

Contributions

Researchers have a choice of methods when making a model selection decision. Given this, it is important to understand different methods and under what conditions each method might be appropriate. This study has examined this question for five popular model selection methods. Although this study looked at model selection for a model with and without an interaction term, it is likely that these results would apply to decisions about any fixed effect, rather than just a product term.

Procedures for evaluating interaction effects in single-level models have been provided previously (Aiken & West, 1991; Jose, 2013) but this study extends those findings by systematically comparing multiple methods for evaluating interaction effects. Further, this study extends the results to a generalization of the single-level model, which is the multilevel model. A step-by-step procedure is provided for applied researchers examining interaction effects in multilevel models, which was not previously available.

Although Type I error rates for the Wald test, *F*-test, and LRT can be compared to the nominal value set by the researcher, and a table of approximate Type I error rates has been previously provided for the BIC (Raftery, 1995), this study extends this understanding to the AIC by providing expected Type I error rates within the simulation study. No previous study was found that has provided any indication of expected Type I error rates for AIC.

This study further extends the literature by examining power. Approximate sample sizes necessary for power of 0.80 in detecting interaction effects has been provided by single-level models (Jaccard et al., 1990, p.37), but this was not known for interaction effects in multilevel models. Having a rough estimate for sample size can be particularly useful for researchers in the early stages of planning and designing research studies.

References

Aiken, L. S., & West, S. G. (1991). Multiple Regression: Testing and interpreting interactions. Newbury Park, CA: Sage.

Alexander, R. A., & DeShon, R. P. (1994). The effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 115*, 308-314.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Barrett, D. C. & Fish, W. W. (2011). Our move: Using chess to improve math achievement for students who receive special education services. *International Journal of Special Education, 26*(3), 181-193

Bofinger, E. (1999). Homogeneity of two straight lines: Equivalence approach using fitted regressions. *Australia and New Zealand Journal of Statistics, 41*(4), 481-491.

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

Cooper, K. S. (2014) Eliciting engagement in the high school classroom: A mixed-methods examination of teaching practices. *American Educational Research Journal, 51*, 363-402.

Champoux, J. E. & Peters, W. S. (1987). Form, effect size, and power in moderated regression analysis. *Journal of Occupational Psychology, 60*, 243-255.

Culpepper, S. A., & Davenport, E. C. (2009). Assessing differential prediction of college grades by race/ethnicity with a multilevel model. *Journal of Educational Measurement, 46*(2), 220-242.

DeShon, R. P., & Alexander, R. A. (1994). A generalization of James' second-order approximation to the test for regression slope equality. *Educational and Psychological Measurement, 54,* 328-335.

DeShon, R. P. & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*, 261-277.

Devore, J. L. (2004). *Probability and statistics for engineering and the sciences.* Belmont, CA: Thomson, Brooks/Cole.

Dretzke, B. J., Levin, J. R., & Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychological Bulletin, 91,* 376-383.

Dunlap, W. P., & Kemery, E. R. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin, 102,* 418-420.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

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.

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures. *Journal of Consulting and Clinical Psychology, 65,* 599-610.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression.* Newbury Park: Sage Publications.

Jose, P. E. (2013). *Doing statistical mediation & moderation*. New York: The Guilford Press.

Kieseppa, I. A. (2003). AIC and large samples. *Philosophy of Science, 70,* 1265-1276.

Kleinbaum, D. G., & Kupper, L. L. (1978). *Applied regression analysis and other multivariable methods.* Boston: Duxbury.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63,* 507-512.

Litzler, E. & Lorah, J. A. (2013, May). *The Intersection of Gender and Race/Ethnicity with Educational Aspirations of Undergraduate Engineering Students*. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Litzler, E., Samuelson, C. C., & Lorah, J. A. (2014). Breaking it down: Engineering student STEM confidence at the intersection of race/ethnicity and gender. *Research in Higher Education*. DOI 10.1007/s11162-014-9333-z

Litzler, E. & Young, J. (2012). Understanding the risk of attrition in undergraduate engineering: Results from the project to assess climate in engineering. *Journal of Engineering Education, 101*(2), 319-345.

Lomax, R. G. (2007). *Statistical concepts: A second course*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Ma, X., Ma, L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell, & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 59-110). Charlotte, NC: Information Age Publishing, Inc.

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127-137.

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin, 118*, 405-421.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (Research Report No. 2008-4). New York: The College Board.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell, & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245-272). Charlotte, NC: Information Age Publishing, Inc.

Miyazaki, Y. & Maier, K. S. (2005). Johnson-Neyman type technique in hierarchical linear models. *Journal of Educational and Behavioral Statistics, 30*(3), 233-259.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage Publications.

Morris, J. H., Sherman, J. D., & Mansfield, E. R. (1986). Failures to detect moderating effects with ordinary least squares-moderated multiple regression: Some reasons and a remedy. *Psychological Bulletin, 99,* 282-288.

O'Connell, A. A. & McCoach, D. B. (2008). Introduction: Pedagogy and context for multilevel models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 3-10). Information Age Publishing Inc.

Onwuegbuzie, A. J., Collins, K. M. T., & Elbedour, S. (2003). Aptitude by treatment interactions and Matthew effects in graduate-level cooperative-learning groups. *The Journal of Educational Research, 96*(4), 217-230.

Paunonen, S. V., & Jackson, D. N. (1988). Type I error rates for moderated multiple regression analysis. *Journal of Applied Psychology, 73,* 569-573.

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

R Core Team (2014). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks: Sage Publications.

Rimm-Kaufman, S. E., Larsen, R. A. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G., Abry, T., & DeCoster, J. (2014). Efficacy of the *Responsive Classroom* approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal, 51,* 567-603.

Snijders, T. & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273-311). Information Age Publishing Inc.

Stevens, J. P. (2002*). Applied multivariate statistics for the social sciences*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson.

Weaklim, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research, 33,* 167-187.

Wei, M., Mallinckrodt, B., Russell, D. W., & Abraham, W. T. (2004). Maladaptive perfectionism as a mediator and moderator between adult attachment and depressive mood. *Journal of Counseling Psychology, 51,* 201–212.

Appendix: Example R Code

```
#R code generating data for simulation, conditions with the effect size = 0

#load libraries
library(lme4)

########################
#set parameters and save in "conditions" file

ncond<-9
conditions<-as.data.frame(matrix(1:ncond,ncol=1,nrow=ncond))
colnames(conditions)<-"Cond"
nsim<-5000      #number of simulations per condition
conditions$nsim<-rep(nsim,times=ncond)
conditions$NumGroups<-c(rep(30,times=3),rep(50,times=3),rep(100,times=3))
conditions$GroupSize<-rep(c(5,30,50),times=3)
conditions$N<-conditions$NumGroups*conditions$GroupSize
conditions$HalfN<-conditions$N/2
conditions$HalfNP<-conditions$HalfN+1

beta1a<-1.5      #slope for X when M=-1
conditions$beta1a<-rep(beta1a,times=ncond)

beta1b<-1.5      #slope for X when M=1
conditions$beta1b<-rep(beta1b,times=ncond)

alpha1a<-3       #intercept for X when M=-1
conditions$alpha1a<-rep(alpha1a,times=ncond)

alpha1b<-4       #intercept for X when M=1
conditions$alpha1b<-rep(alpha1b,times=ncond)

w1<-2            #weight for adjusting sd of e1 (correlation between X and Y)
conditions$w1<-rep(w1,times=ncond)

L2sd<-.78        #Standard deviation at level 2
conditions$L2sd<-rep(L2sd,times=ncond)

MeanY<-.5*alpha1a+.5*alpha1b
conditions$MeanY<-rep(MeanY,times=ncond)

VarY1<-beta1a^2+w1^2
VarY2<-beta1b^2+w1^2
VarY<-.5*((alpha1a-MeanY)^2+VarY1)+.5*((alpha1b-MeanY)^2+VarY2)+L2sd^2
conditions$VarY<-rep(VarY,times=ncond)

write.csv(conditions,"C:/Conditions_ES0.csv")
######################
```

```
######################
#run ncond*nsim simulations

result<-as.data.frame(matrix(1:(ncond*nsim),ncol=1,nrow=ncond*nsim))
colnames(result)<-"Iterate"

result$condition<-c(rep(1,nsim),rep(2,nsim),rep(3,nsim),rep(4,nsim),rep(5,nsim),
        rep(6,nsim),rep(7,nsim),rep(8,nsim),rep(9,nsim))
result$NumGroups<-c(rep(30,times=3*nsim),rep(50,times=3*nsim),rep(100,times=3*nsim))
result$GroupSize<-rep(c(rep(5,nsim),rep(30,nsim),rep(50,nsim)),times=3)
result$N<-result$NumGroups*result$GroupSize
result$HalfN<-result$N/2
result$HalfNP<-result$HalfN+1

for(i in 1:(ncond*nsim)){

#create error term
e1<-rnorm(result$N[i])

#create X (indepdendent variable)
X<-rnorm(result$N[i])

#create Y (dependent variable)
y1<-alpha1a+beta1a*X[1:result$HalfN[i]]+e1[1:result$HalfN[i]]*w1
y2<-alpha1b+beta1b*X[result$HalfNP[i]:result$N[i]]+e1[result$HalfNP[i]:result$N[i]]*w1
Y<-c(y1,y2)

#create M (moderator variable)
M<-c(rep(-1,times=result$HalfN[i]),rep(1,times=result$HalfN[i]))

#create interaction term
MX<-M*X

#create group ID
ID<-rep(seq(from=1,to=result$GroupSize[i],by=1),times=result$NumGroups[i])

#create set of terms to add/subtract for macro-unit membership
Adjust<-rep(rnorm(n=result$GroupSize[i],mean=0,sd=L2sd),times=result$NumGroups[i])

#add adjustment for nesting in dataset
Y<-Y+Adjust

#Concatenate variables into one dataset
mydata<-data.frame(Y,X,M,MX,ID)

#run three models (empty model, and with and without interaction term)
M0<-lmer(Y~(1|ID),data=mydata, REML=F)
M1<-lmer(Y~X+M+(1|ID),data=mydata, REML=F)
M2<-lmer(Y~X+M+MX+(1|ID),data=mydata, REML=F)
```

```
result$Ymean[i]<-mean(Y)
result$Yvar[i]<-var(Y)
result$M0Dev[i]<-deviance(M0)
result$M1Dev[i]<-deviance(M1)
result$M2Dev[i]<-deviance(M2)
result$M0AIC[i]<-AIC(M0)
result$M1AIC[i]<-AIC(M1)
result$M2AIC[i]<-AIC(M2)
result$M0BIC[i]<-BIC(M0)
result$M1BIC[i]<-BIC(M1)
result$M2BIC[i]<-BIC(M2)
result$M0ResSD[i]<-summary(M0)$sigma
result$M1ResSD[i]<-summary(M1)$sigma
result$M2ResSD[i]<-summary(M2)$sigma
result$M0TotSD[i]<-sum((as.data.frame(summary(M0)$varcor))$sdcor)
result$M1TotSD[i]<-sum((as.data.frame(summary(M1)$varcor))$sdcor)
result$M2TotSD[i]<-sum((as.data.frame(summary(M2)$varcor))$sdcor)
result$M0Int[i]<-coef(summary(M0))[1,1]
result$M0SE[i]<-coef(summary(M0))[1,2]
result$M0T[i]<-coef(summary(M0))[1,3]
result$M1Int[i]<-coef(summary(M1))[1,1]
result$M1SE[i]<-coef(summary(M1))[1,2]
result$M1T[i]<-coef(summary(M1))[1,3]
result$M1Xsl[i]<-coef(summary(M1))[2,1]
result$M1XSE[i]<-coef(summary(M1))[2,2]
result$M1XT[i]<-coef(summary(M1))[2,3]
result$M1Msl[i]<-coef(summary(M1))[3,1]
result$M1MSE[i]<-coef(summary(M1))[3,2]
result$M1MT[i]<-coef(summary(M1))[3,3]
result$M2Int[i]<-coef(summary(M2))[1,1]
result$M2SE[i]<-coef(summary(M2))[1,2]
result$M2T[i]<-coef(summary(M2))[1,3]
result$M2Xsl[i]<-coef(summary(M2))[2,1]
result$M2XSE[i]<-coef(summary(M2))[2,2]
result$M2XT[i]<-coef(summary(M2))[2,3]
result$M2Msl[i]<-coef(summary(M2))[3,1]
result$M2MSE[i]<-coef(summary(M2))[3,2]
result$M2MT[i]<-coef(summary(M2))[3,3]
result$M2MXsl[i]<-coef(summary(M2))[4,1]
result$M2MXSE[i]<-coef(summary(M2))[4,2]
result$M2MXT[i]<-coef(summary(M2))[4,3]
result$Fmx[i]<-anova(M2)[3,4]
result$MSmx[i]<-anova(M2)[3,3]
result$SSx[i]<-anova(M2)[1,2]
result$SSm[i]<-anova(M2)[2,2]
result$SSmx[i]<-anova(M2)[3,2]
}

write.csv(result,"C:/Result_ES0.csv")
```