Spring 5-7-2011

Factors that Influence Cross-validation of Hierarchical Linear Models

Tracy Widman
Georgia State University

ACCEPTANCE

This dissertation, FACTORS THAT INFLUENCE CROSS-VALIDATION OF HIERARCHICAL LINEAR MODELS, by TRACY WIDMAN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Phill Gagné, Ph.D.
Committee Chair

_____
Chris Domaleski, Ph.D.
Committee Member

_____
L. Juane Heflin, Ph.D.
Committee Member

_____
Frances A. McCarty, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R. W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission.

_____

Tracy Widman

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Tracy Widman
4341 Peachtree Dunwoody Road
Atlanta, GA 30342

The director of this dissertation is:

Dr. Phill Gagné
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303-3083

VITA

Tracy Widman

ADDRESS: 4341 Peachtree Dunwoody Road
Atlanta, GA 30342

EDUCATION:

Ph.D.	2011	Georgia State University Educational Policy Studies
M.S.	1994	Indiana State University College Student Personnel
B.S.	1991	Indiana State University Psychology
A.S.	1989	Vincennes University Psychology

PROFESSIONAL EXPERIENCE:

2008-Present	Visiting Instructor Georgia State University, Atlanta, GA
2006-2008	Director of Enrollment Services Agnes Scott College, Atlanta, GA
2003-2006	Director, Data and Enrollment Management The George Washington University, Washington, D.C.
2001-2003	Associate Director, MBA Programs The George Washington University, Washington D.C.

PRESENTATIONS AND PUBLICATIONS:

Davidson, J. D., & Widman, T. (2002). The effect of group size on interfaith marriage among Catholics. *Journal for the Scientific Study of Religion*, *41*(3), 397-404.

ABSTRACT

FACTORS THAT INFLUENCE CROSS-VALIDATION OF HIERARCHICAL LINEAR MODELS

by

Tracy Widman

While use of hierarchical linear modeling (HLM) to predict an outcome is reasonable and desirable, employing the model for prediction without first establishing the model's predictive validity is ill-advised. Estimating the predictive validity of a regression model by cross-validation has been thoroughly researched, but there is a dearth of research investigating the cross-validation of hierarchical linear models. One of the major obstacles in cross-validating HLM is the lack of a measure of explained variance similar to the squared multiple correlation coefficient in regression analysis.

The purpose of this Monte Carlo simulation study is to explore the impact of sample size, centering, and predictor-criterion correlation magnitudes on potential cross-validation measurements for hierarchical linear modeling. This study considered the impact of 64 simulated conditions across three explained variance approaches: Raudenbush and Bryk's (2002) proportional reduction in error variance, Snijders and Bosker's (1994) modeled variance, and a measure of explained variance proposed by Gagné and Furlow (2009).

For each of the explained variance approaches, a cross-validation measurement, shrinkage, was obtained. The results indicate that sample size, predictor-criterion correlations, and centering impact the cross-validation measurement. The degree and direction of the impact differs with the explained variance approach employed. Under some explained variance approaches, shrinkage decreased with larger level-2 sample sizes and increased in others. Likewise, in comparing group- and grand-mean centering,

with some approaches grand-mean centering resulted in higher shrinkage estimates but smaller estimates in others. Larger total sample sizes yielded smaller shrinkage estimates, as did the predictor-criterion correlation combination in which the group-level predictor had a stronger correlation. The approaches to explained variance differed substantially in their usability for cross-validation. The Snijders and Bosker approach provided relatively large shrinkage estimates, and, depending on the predictor-criterion correlation, shrinkage under both Raudenbush and Bryk approaches could be sizable to the degree that the estimate begins to lack meaning. Researchers seeking to cross-validate HLM need to be mindful of the interplay between the explained variance approach employed and the impact of sample size, centering, and predictor-criterion correlations on shrinkage estimates when making research design decisions.

FACTORS THAT INFLUENCE CROSS-VALIDATION OF HIERARCHICAL LINEAR MODELS

by

Tracy Widman

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2011

ACKNOWLEDGMENTS

I envisioned writing a dissertation to be an individual, isolated activity. The production of this dissertation was anything but isolated and individual. Without a doubt, I would not have accomplished this if not for a cadre of supportive people. First, I wish to express my most sincere appreciation to the chair of my dissertation committee, Dr. Phill Gagné. Phill has been a teacher, coach, and friend. Few others could tell me to think of it as a "re-organization" not a "re-write" and have me laughing about it at the same time. I am certain I would not have finished this endeavor without his humor, patience, and guidance. I would also like to thank the rest of my committee members: Dr. Chris Domaleski, Dr. Francis McCarty, and Dr. Juane Heflen for their time, kindness, and willingness to help. One could not ask for a more supportive and helpful group of people.

I am lucky to have so many encouraging people in my life. I owe thanks to friends who encouraged me and kept me grounded when I started to freak-out or when my motivation waned. Thanks to my wonderful parents, sisters, and nephews who understood when I could not be there. I appreciate Zoë and Ainslie Kate's understanding and patience when my goals lead to family sacrifices, delayed plans, and interrupted hikes. I especially want to acknowledge my late friend Rita Lawson whose influence continues to direct me. On many late nights while agonizing at the laptop, I heard her voice saying "you know you are going to write it eventually so stop agonizing about writing it and just write it." Eventually, I did write it, but not without *some* agonizing.

Finally, to my husband, Greg, who could not have realized what lie ahead when he encouraged me to pursue this degree, I could not and would not have done this without your support, encouragement, patience, and love. You are my love. Now, finally, let's go on vacation and leave the laptop at home.

TABLE OF CONENTS

LIST OF TABLES

ABBREVIATIONS

HLM	Hierarchical linear modeling
R^2_{GF}	Gagné and Furlow's (2009) R^2
R^2_{RB0}	Raudenbush and Bryk's (2002) intercept R^2
R^2_{RB1}	Raudenbush and Bryk's (2002) slope R^2
R^2_{SB}	Snijders and Bosker's (1994, 1999) level-2 R^2

CHAPTER 1

INTRODUCTION

A strong argument can be made that, in general, most research is conducted for one of three purposes: understanding the relationship between two or more variables, predicting the occurrence or degree of an outcome variable(s) from associated variables, or both understanding the relationships and predicting outcome(s) of interest. The results of research can be used for practical application or as the basis for further research. In either case, research results often become a base on which other activities ensue. If research results are not reliable or valid, the subsequent activities potentially are either misguided or for naught. This is especially true if the results are used in application and the purpose is for prediction of an outcome. Predictive research findings need to be assessed for validity in order to safeguard against ill-guided, future activities. Cross-validation is an approach to measuring the validity of a prediction model and is the focus of this research.

The basic idea underlying cross-validation is that if a prediction model is valid, it also should predict effectively in a second sample from the population. Therefore, in the most straightforward approach to this concept, cross-validation is conducted by estimating a prediction model from an initial sample and then applying that model to a second sample from the population. Since both datasets include the actual values for the outcome variable of interest, the actual values can be compared to the predicted values for each sample. The closer the second sample's predicted values are to the actual values,

the stronger the cross-validation. To be clear, this is not a measure of the accuracy of the prediction model but of how consistently the model predicts in samples from a population. It is possible that a prediction model cross-validates well even though the model consistently, ineffectually predicts the outcome, or criterion variable. Prediction with a high degree of error is rarely useful, so cross-validation is most important when the initial sample appears to have an acceptable amount of explained variance.

It is easy to comprehend the importance of cross-validation when the potential consequences of poor prediction models are considered. Take for instance a university that constructs an enrollment prediction model for the upcoming year from the previous year's applicant pool. The budget is set based on the projected enrollment figures, and when the enrollment is far less than predicted, the university finds itself operating at a multi-million dollar deficit. There are many explanations as to why the prediction model did not work. One possibility is that the there was something unique about the previous year's class, and that unique quality is calculated into the estimated model. The current applicant pool does not have the unique characteristic(s), so the model will not work as well for the incoming class as for the previous class. If this were the case, cross-validation could warn of the problem with the prediction model, and other paths for establishing the prediction model could be pursued. Despite the minimal amount of prediction error in the initial sample, a model that does not apply to other samples from the population is of little use, as research interests and the applications of research rarely are focused on a specific sample but on inference of results to the entire population.

Over the last 50 years, regression analysis emerged as a staple in social science research, and along with it, cross-validation developed and was researched. Therefore,

much of the research related to cross-validation sample sizes, approaches, and other factors happens within a regression context. Cross-validation, however, does not need to be restricted to one approach. Research exploring cross-validation within a structural equation modeling approach is common, albeit the focus is on model selection as opposed to predictive validity (Whittaker & Stapleton, 2006). Entirely absent in the literature is an exploration of cross-validation with hierarchical linear modeling. In reviewing the HLM and cross-validation literature, specifically searching for studies exploring "how-to" cross-validate HLM, only two studies were found that made an attempt to cross-validate. Afshartous and de Leeuw (2004) used mean squared prediction errors to cross-validate, while Astin and Denson (2009) attempted to cross-validate by comparing HLM predictions to regression predictions and regression cross-validation results of the same data. No studies were found in which researchers explored procedures for cross-validating HLM which may be one reason so few HLM studies attempt to cross-validate their models: There is no guidance on how to approach cross-validating HLM.

HLM models pose interesting problems when attempting to cross-validate. With regression, the accuracy of cross-validation is affected by sample size and predictor relationships. Presumably, they also would impact the accuracy of cross-validation of HLM, but that assumption currently has not been tested. Adding to the challenge of cross-validating HLM is the fact that, unlike regression, HLM has no agreed upon measure of model fit. The cross-validation measure with regression is based on an estimate of model fit. Additionally, researchers conducting HLM analyses often employ various forms of data centering for accurate interpretation of the coefficients. Centering can change the variance measures of the model, but it is unknown if centering also affects

the cross-validation measure. In this simulation study, the impact of these variables on HLM will be explored to address the following questions:

1. What is the impact on cross-validation measures using different forms of R^2?
2. What effect does sample size have on the cross-validity coefficients in HLM?
3. What effect does the magnitude of the predictor-criterion correlations have on the cross-validity performance with HLM?
4. What is the impact on the cross-validity measures of applying common forms of centering in HLM?

Answers to these questions would allow researchers not only to conduct cross-validation studies, but also to conduct such studies more effectively. The results also would provide consumers of HLM research with a basis for making judgments regarding the validity of the research findings.

CHAPTER 2

REVIEW OF LITERATURE

An educational researcher whose work makes inferences to a population based on a sample size of hand-picked subjects would probably not find wide acceptance for her work. Even novice researchers and educational practitioners know that random sampling and ample sample size are important for making any type of defendable inference. There is no point in conducting research if it is not going to be performed properly. Likewise, as important as cross-validation is, if it is not conducted judiciously, then the results have no value. Sample size, predictor-criterion correlations, model fit, and centering may all have an impact on cross-validation within an HLM framework, but there is no guidance about how these factors may affect cross-validating HLM models. This study intends to determine the impact of each.

As no work has been done on cross-validating HLM models, the basis for understanding cross-validation comes from existing research on cross-validating regression analyses. HLM can be thought of as a complex version of regression, so understanding cross-validation in the regression framework is pertinent. What follows in this review is an overview of cross-validation research. This will include a brief history of cross-validation and a survey of procedural costs of cross-validating along with suggested solutions. Although the literature relates to costs faced in cross-validating regression, the same costs are faced in cross-validating HLM. Information on the obstacles of model fit and centering present in cross-validating HLM will be explained.

The review will conclude with information regarding research on sample size when cross-validating regression as well as sample size considerations when conducting an HLM analysis. Cross-validation studies are not always as straightforward as they may seem, and a more complicated approach such as HLM only adds to the complexity.

Cross-validation is a method of measuring the validity of the predictive ability of a model and is generally conceptualized as a two-sample process. Data from the first sample of a population, often referred to as the training or screening sample, are used to estimate the model. The prediction model is then applied to the second sample of data, the validation sample, in order to predict the outcome variable for each member of the sample. Since the outcome variable for both sets of data is known, a comparison can be made of how effectively the model predicted the outcome variable for the second set of data in relationship to how effectively the outcome variable was predicted in the first sample. It might seem redundant to apply the model to a second set of data if the estimated model fits the data well, but the model fitting the initial data does not provide an indication of the quality of the model's fit to other samples from the population. In estimating the model, the relationship between the variables is determined. The resulting model represents the numeric relationship that best fits the data in that specific sample. To assume that the model is equally useful in other samples from the same population seems logical and would be useful, but is ill-advised in that there are possible idiosyncrasies in the data that become part of the estimated model. Application of that model to other samples could lead to poor predictions and misguided decisions if the relationships estimated from the first sample do not hold true in future samples. Cross-

validation is therefore necessary to determine if the model consistently predicts the outcome variable in additional samples of the population.

Cross-validation is most commonly associated with regression analysis for a number of reasons but primarily because was it was developed for regression models. As early as 1931, Larson warned that without a means to measure shrinkage of the squared multiple correlation coefficient, "the building of multiple regression equations might well be abandoned" (p. 45). The squared multiple correlation coefficient (R^2) will be described in more detail later in this paper, but in short, it is a measure of how well the model fits the data from which the model was estimated. Larson and theoretical statisticians at the time hypothesized that the multiple correlation coefficient was overstated as a measure of how representative the coefficient was of relationships in the population or to other samples of the population. Shrinkage then is the label for the concept that the multiple correlation coefficient, if it were measured in the population or other samples, would be smaller, or appear to shrink, in comparison to the multiple correlation coefficient resulting from model estimation in a given sample. Larson was the first to attempt an empirical measurement of shrinkage by employing a formula to account for shrinkage. In attempting to establish and estimate shrinkage, Larson substantiated that shrinkage was not just a theoretical construct, but could be found empirically, with the size of the shrinkage affected by the number of predictor variables and the size of R^2. This was the start of cross-validation.

Later authors demonstrated the need to cross-validate through the development and testing of measurement instruments. Two of the more commonly cited articles are from Kurtz (1948) and Cureton (1950). Kurtz developed an employment test that

validated well for the initial sample on which the test was developed, but failed miserably to predict successful managers when applied to a second sample in the same company. The primary conclusion pertinent to cross-validation is that validity is impacted by idiosyncrasies in the initial sample, so a second sample is needed to gauge accurately the predictive validity of, in this case, a test. Arriving at the same conclusion, but taking a different approach, Cureton conveyed the story of a test with amazing predictive validity only to show that the validity was based on nothing more than the effect of capitalizing on the peculiarities in the original sample data. As a warning, Cureton (p. 96) emphasized that any validity coefficient that is derived using the initial sample data should be heavily scrutinized because the only clear interpretation of such measures is that they are "baloney!"

From the publication of Larson's (1931) paper through 1950, interest in cross-validation seems to have understandably ballooned. This is evidenced by Mosier (1951) who provided a synopsis of the types of cross-validation procedures that emerged over the 20-year period along with a demarcation of the procedure that is recognized today as cross-validation. In describing five different approaches that were commonly being referred to as cross-validation, Mosier reserves the term cross-validation for the single population, two-sample approach. Another approach covered by Mosier is validity generalization in which the model calibrated on one sample from one population is applied to a sample from a different population. This approach would be taken if the research question is about the generalizability of the model to a different population. Validity extension, the third approach, quantifies the generalizability of the model to a different but similar criterion variable. An example of this would be if a researcher has an

acceptably accurate model that predicts SAT performance and wants to extend the model to see how valid it is for predicting ACT scores, a similar but different criterion variable. The fourth and fifth categories deal with combining regression weights from multiple samples and applying the resulting weights to other samples. If the combined samples were all drawn from the same population then the resulting weight is applied to other samples from the same population, but if the combined weight was established from samples of different populations, then the combined weight is applied to samples across populations.

Mosier's (1951) delineation of cross-validation has held the test of time and is conceptually today what he pronounced it to be 60 years ago. Through the years, various methods and aspects of cross-validation have been thoroughly researched and this research has, by and large, been from within a regression framework. One of the results of regression analysis is a measurement of model fit, the R^2 statistic, often referred to as the squared multiple correlation coefficient or the coefficient of determination. R^2 represents how well the model predicts the criterion variable. The same statistic provides a basis for measuring cross-validity.

Cross-validation with regression is built around R^2. It is a measurement of the amount of variance in the criterion variable that can be attributed to the linear combination of predictor variables. If the R^2 for a model is 0.25, then 25% of the variance in the criterion variable can be explained by the combination of predictor variables. Obviously, larger R^2-values indicate a model that fits the data better. Prediction is more accurate in a sample in which much of the variance of the criterion variable is explained by the predictors. It is important to keep in mind that R^2 is the amount of variance

explained in the sample and not the amount of variance explained by the combination of predictors in the population. The notion of R^2 shrinkage, as conceptualized by Larson (1931), is that the R^2 resulting from a regression analysis on sample data is larger than the R^2 that would be obtained in the application of the prediction equation to the population. R^2 from a sample is, therefore, a positively biased estimate of the variance explained in the population, because in the process of estimating the relationships between variables, sampling error and data idiosyncrasies of the sample are used to derive the regression equation that best fits the sample data. The sampling abnormalities would not be present in the entire population, so the relationships estimated from the sample data will logically fit the sample better than the entire population. The same holds true for R^2 shrinkage when applying a regression model to a second sample. The R^2 from the second sample is expected to be smaller than the R^2 from the first sample because the variable relationships are estimated to best fit the idiosyncrasies and sample abnormalities in the first sample.

Cross-validation involves estimating R^2 from a screening sample and comparing that R^2 to the squared cross-validity coefficient, R_{cv}^2, from a validation sample. R_{cv}^2 is a measurement of the amount of variance explained by the linear combination of predictor variables. The value of R_{cv}^2 is that it provides an indication of the usefulness of a regression model when applied outside the screening sample but within the same population (Algina & Keselman, 2000). Since R_{cv}^2 results from a model applied to a second set of data as opposed to R^2 which results from estimating the model, it is expected that R^2 will be larger than R_{cv}^2 due to the positive bias inherent in model estimation. The difference between R^2 and R_{cv}^2 is known as shrinkage, and the amount of shrinkage is the determinant as to whether a model cross-validates sufficiently.

Despite having a useful measure of model applicability, there are obstacles to cross-validating. One of the most obvious deterrents to cross-validating is the cost. Collecting data is expensive in terms of both time and money, and cross-validating requires more data, so the expense increases. Data collection for cross-validating can happen in one of three ways: two separate samples can be collected simultaneously, a second sample can be collected or acquired sometime after the initial sample is collected, or one sufficiently large sample can be collected and then split into at least two different sets as the screening and validation samples. If two samples are collected simultaneously or if one large sample is collected and divided into two or more groups, only half of the data available is used to estimate the equation which results in regression coefficients that are less stable and accurate than if all of the data were used in the estimation. To use all of the data in estimating the equation, however, leaves no way to cross-validate. Mosier (1951) discusses this double-bind situation and argues that the sacrifice of accuracy is an acceptable price to pay for an unbiased estimate of the predictive accuracy of the equation.

In order to minimize the costs of cross-validating, many researchers have explored analytical approaches to cross-validating. Analytical approaches to cross-validating are different than empirical approaches in that analytical approaches rely on one sample to estimate the regression equation and apply shrinkage formulas to the resulting equation in an attempt to determine the predictive validity of the regression equation. Various formulas were derived for the purpose of estimating how well the equation would perform in another sample and in the overall population. However, the analytical approach is not as trouble-free as it would appear. In evaluating shrinkage

formulas, Yin and Fan (2001) reviewed 15 different formulas, and in a similar study, Raju, Bilgic, Edwards, and Fleer (1999) reviewed 16 different formulas with only 11 formulas overlapping in the two studies. While it seems that research is focused primarily on the accuracy and formation of analytical methods, it also is mired in confusion. Of the 15 formulas in the Yin and Fan study addressing shrinkage of R^2, 9 of the formulas were for estimating the squared population cross-validity coefficient(ρ_{cv}^2) and 6 were for estimating the squared population multiple correlation coefficient (ρ^2). R_{cv}^2 is the R^2derived from a second sample, whereas the ρ^2 and ρ_{cv}^2 are the population parameters. In their analysis, Yin and Fan found confusing and misleading information in the literature in relationship to shrinkage, the sample statistics, and population parameters. This confusion comes from several factors, including finding multiple names for some formulas, cases of the same name on different formulas, formulas for estimating a population parameter with no distinction as to which parameter is the focus, and the many corrections to the formulas with different names associated with the corrections.

Formula confusion is something methodologists could work to resolve if analytical approaches provide an accurate estimate of shrinkage. The cost of resolving the confusion would seem to be minimal in comparison to the costs of empirical approaches. The confusion surrounding the formulas undoubtedly plays a part in the bigger problem with the analytical approach which is that currently, the question of whether or not formulas accurately estimate shrinkage has not been answered unequivocally. Many researchers have come to the conclusion that the accuracy in estimating the shrinkage of R^2 or in estimating R_{cv}^2 or ρ_{cv}^2 seems to be on par with results from empirical methods if not better (Claudy, 1978; Cotter & Raju, 1982; Kromrey & Hines, 1996; Murphy, 1984;

Raju et al., 1999). Many researcher, however, arrive at the conclusion of sufficient formula accuracy with the acknowledgment that the conclusion is couched in caveats (Cattin, 1980a, 1980b; Cotter & Raju, 1982; Mitchell & Klimoski, 1986; Murphy, 1984; Schmitt, 1982). Each formula is based on certain assumptions and the extent to which the data, sampling, and regression weight estimation procedures coincide with the assumptions determines the extent to which the formula will estimate the shrinkage of some intended R^2. In Cotter and Raju's (1982) study, they conclude that formula estimates are equivalent to the results of cross-validation empirical studies, but warn against generalizing or extending their findings to situations with conditions different from those in their study.

The number of predictor variables in the model, the sample size, and the magnitude of the R^2 all impact the accuracy of formula estimates. Cattin (1980b) found that formulas are more accurate when the ratio of observations to predictors is relatively large, but when the ratio is low, the formula estimates are substantially less accurate. Regression weights estimated through ordinary least squares (OLS) procedures and predictors that are selected a priori, as opposed to "data-snooping" (Schmitt, 1982, p. 5) procedures such as stepwise and other model selection approaches, are prerequisites for even contemplating the use of a formula approach to shrinkage estimation (Cattin, 1980a; Mitchell & Klimoski, 1986; Schmitt, 1982). Anastasi and Urbina (1997) suggest that shrinkage is minimized when the predictors are selected based on knowledge of the relationship between the variables. While formulas take into consideration sample size and number of predictors, Mitchell and Klimoski show that the rational basis for choosing predictor variables also impacts the cross-validation measurement even though

that is a dimension left out of formulas. Models that have been developed through a process like factor analysis have more stable validities than strictly data-level variables. Since shrinkage formulas are derived from data-level predictors, shrinkage is overestimated when formulas are applied to regression models utilizing analytically constructed variables.

The question of whether to cross-validate by empirical or analytical means seems to be easily answered given the cloud of uncertainty with analytical approaches. The confusion that Yin and Fan (2001) discussed undoubtedly plays a role in the mixed results regarding formula accuracy. Estimates will naturally appear less accurate if the formula is employed to estimate $R_{cv}^2$when the formula is intended to estimate ρ^2 (Kromrey & Hines, 1996). If the confusion over formulas is resolved, and the question of formula accuracy is settled, formula approaches to cross-validation are seemingly an answer to the cost problem. There is, however, an additional problem of cross-validating with formulas. An inherent problem with formula cross-validation is that formulas are derived to account for random sampling error but are insensitive to other problems and idiosyncrasies in the data (Murphy, 1983). Cross-validating by formula when non-random errors exist in the data will regularly produce strong, misleading cross-validation results which would not be found if cross-validating with some empirical procedures. Given the inability of formulas to account for idiosyncrasies in the data, the confusion surrounding the formulas in the literature (Yin & Fan, 2001), and the myriad of uncertainties in implementing a given formula, it seems the prudent option in estimating shrinkage is to conduct an empirical cross-validation study whenever possible. Advances in approaches to the problem of data sufficiency have made this a more viable option.

In discussing the empirical approach up to this point, the basic form of cross-validation has primarily been described. It is the approach that requires two random samples from the same population. This approach is expensive in time and money which has lead to the development of formula approaches, discussed above, and has also lead to data utilization solutions. Data utilization solutions for cross-validation are of three main types: data-splitting, sample reuse, and simultaneous estimation (Cooil, Winer, & Rados, 1987). A brief description of each in a cross-validation situation follows.

Data splitting, also known as hold-out validation, is a commonly used method. The initial approach to data splitting consisted of collecting a large sample of data and randomly dividing the data set in half. One half of the data is set aside as the validation data set while the other half is used as the screening sample for model estimation. This procedure does eliminate the need to collect two samples, but it is also accompanied by problems. Data sufficiency remains a problem with this procedure. To "lose" one half of the data results in the reduced accuracy and stability of the estimation of the regression weights. To help overcome the problem of data sufficiency with data splitting, various data splitting approaches have been developed. These approaches try to maximize the use of the data in estimation and validation; in many cases, however, either estimation or validation has to be prioritized.

Take as an example, double cross-validation which was used by Larson's (1931) and propounded by Mosier (1951). In double cross-validation, the full data set is divided into two sets of data. Each data set is used to estimate regression weights for the model, so each data set results in a regression equation. The equation from data set A is cross-validated with data set B to derive a R_{cv}^2 and the corresponding shrinkage. Likewise, the

equation from data set B is cross-validated with data set A to obtain the amount of shrinkage. The shrinkage figures are averaged to arrive at a measure of the predictive validity of the model. However, a criticism of this and similar approaches is that the resulting averaged cross-validation measurement lacks meaning and is difficult to interpret and therefore is seemingly of less use (Murphy, 1984). The resulting shrinkage measurement is no longer a measure of how well a regression equation predicts the criterion in a second sample because there are now two equations used to derive measures of shrinkage. If the interest is in prediction and applying the equation to future groups from the population, the question becomes which equation to use. Mosier (1951) suggests that after obtaining an averaged shrinkage estimate, the sample could be combined and regression weights for the entire sample calculated. The problem then is that there is no measure for how an equation based on the entire sample cross-validates. One is left to assume that the averaged cross-validation measurement is representative of how well an equation based on all the data would cross-validate.

The division of one data sample divided into data sets does not mean that the data have to be divided equally into two samples, nor does it mean that the data have to be randomly divided. To help improve data sufficiency, the data can be divided such that the majority of the data is used to estimate the model and the remaining data can be used for cross-validation purposes. This approach does improve estimation, while still leaving data unused for cross-validation. The extreme of this approach is the leave-one-out approach. With this method, a single randomly selected observation is removed from the data and the model is estimated on the remaining, *N-1*, observations. The one observation is used as the validation set. As is likely obvious, this has the weakness of a very small

sample for validation and the significant errors that can accompany a validity judgment regarding a regression model on the basis of a single case. In this approach, the importance is placed on the estimation of the regression weights while maintaining a modicum of consideration for the need to cross-validate. Another similar approach that places more importance on cross-validation while still prioritizing the regression weights is the leave-*p*-out approach in which the screening data set consists of *N*-*P* observations as *P* is the arbitrarily determined number of observations held-out as the validation sample.

In splitting the original sample into subsamples, the splitting can be done randomly or it can be done through a matching process based on either a mean response value for each set or a proportion of pertinent characteristics approximately equal in the subsamples. A more complex approach to matching and dividing data into validation and screening samples is based on the amount of difference in the data. DUPLEX is an algorithm used for splitting data into two groups based upon similarities and difference in terms of statistical properties and distances between the data points (Snee, 1977). The process begins by standardizing the data so that they are measured in the same units. The data are transformed so they are more spherical in shape when plotted on a grid. The points farthest apart in terms of Euclidean geometry are assigned to the screening set and the next two disparate in terms of distance are included in the validation set. The next data point assigned to the screening sample is determined based on the farthest distance from the points already in the set. This process continues alternating between assigning data to each set until all the original data are distributed.

Exhaustive data splitting techniques, often called sample reuse techniques, can be used to improve estimation while prioritizing cross-validation. Arlot and Celisse (2010) provide a description of many of these techniques in their overview of using cross-validation as a means of model selection. Leave-one-out and leave-*p*-out approaches can both be used exhaustively. The leave-one-out approach can be performed successively such that every observation in the sample is used as the validation sample once. Similarly, the leave-*p*-out approach employed exhaustively means that every set of possible *p* observations is used as the validation sample (Arlot & Celisse, 2010). The predictive accuracy is assessed by averaging the R_{cv}^2 over all the splits (Browne, 2000). This approach is computationally intensive because every possible set of *P* data points is to be used for validation with the remaining points used as a new validation set each time. As with the exhaustive leave-one-out, the resulting cross-validation coefficients are averaged over all the iterations. *K*-fold cross-validation is another similar approach. With K-fold cross-validation, the collected sample is divided into *K* subsamples. A subsample is selected as the validation sample and the remaining subsamples are used as the screening samples. The process is repeated *K* times with each of the subsamples serving as the validation sample. As in the leave-out procedures, the resulting cross-validation coefficients are averaged to produce a measure of predictive validity. Exhaustive procedures prioritize validation and allow for the development of a general explanatory model rather than the estimation of specific regression coefficients that could be used with future samples.

In terms of single sample empirical procedures, it was stated earlier that there are three approaches to cross-validation; data-splitting, sample reuse, and the third approach,

the simultaneous method. The simultaneous method estimates and cross-validates at the same time by estimating model parameters subject to a cross-validation constraint. This is important because if the purpose is to construct a model that predicts well, the criteria of predictive ability could be used in estimating parameters and in evaluating a model's performance. Originally developed by Stone (1974) and Geisser (1975), the foundation of the simultaneous method is based on the conceptualization of a loss function which involves a weighting of errors based on omitted observations and the estimation of the regression weights. The objective is to minimize the prediction errors in the determination of the regression weights. A common choice for the loss function is the squared-error loss. Mean loss is treated as a parameter that will minimize the prediction errors. This approach provides a way of developing parameter estimates "that will cross-validate well because they are selected on that basis" (Cooil et al., 1987, p. 274). Basically, this approach integrates the analytic stages of construction and evaluation. However, this approach has shortcomings in that special programs must be developed to estimate the mean loss parameter as well as much computational cost and effort.

After years of exploration on how to cross-validate, it seems that the conceptual origin of cross-validation still may be the most authentic approach. Cross-validation at the core is the idea that two samples are used to assess the predictive validity of a model. Shrinkage formulas and various approaches to data-splitting, data utilization, and the simultaneous method allow for cross-validation based on a single sample which eliminates the time and expense of cross-validating. Often overlooked are the costs associated with these single-sample approaches. As Murphy (1983) pointed out, single sample approaches only adjust for random sampling errors and do not address other data

irregularities due to sampling problems. These measurements then do not really gauge the true predictive validity of the model on a separate sample as they are crafted based on the same sample. It is at the point of sampling where sampling errors occur and such errors are not removed from the data through divisions or formula applications. “Spurious results should not, after all, re-occur in truly independent samples” (Murphy, p. 114). In essence, these single-sample approaches are subject to the same problems Kurtz (1948) and Cureton (1950) sought to warn against. Like his predecessors, Murphy shows that a large biased sample divided into subsamples cross-validates well despite the fact that it will not validate in a new random sample from the population. In some sense, Murphy’s findings can be taken as an argument about the pointlessness of investigating single-sample empirical approaches versus analytical approaches as neither protects against sampling abnormalities that cross-validation with two independent samples can detect.

As no work has yet been conducted on cross-validating HLM models, analytical formulas are in the distant future. For the time being, empirical methods, with their drawbacks, are better than nothing. Whether the analysis is within a regression or HLM context, decisions will have to be made about whether adopting a two-sample process is realistic or if another approach is needed in order to cross-validate. Once this decision has been made, the researcher cross-validating HLM has more obstacles to overcome. As mentioned previously, HLM is similar to regression; Hox (1998) explains HLM is “a hierarchical system of regression equations” (p. 148). Conceptually, HLM is a regression model in which the regression coefficients are modeled. It is ideal for analyzing hierarchical data since hierarchical data violates the assumption of independence on which standard regression analysis depends. While HLM avoids these errors, it also

complicates cross-validation. The partitioning of variance into levels and components and the challenges those bring to the concept of model-fit as well as the use of centering to derive meaningful coefficients pose interesting problems in cross-validating HLM. The extent to which model fit approaches for HLM and centering impact cross validating is yet to be determined. A discussion of the implications of HLM model fit and centering follows.

In general, the challenge in cross-validating regression models is about selecting and implementing the approach to cross-validation. In cross-validating HLM models, not only does the researcher need to answer the same question of approach, but HLM poses a much more challenging and problematic issue of how to measure model quality. Cross-validation is an attempt to measure the predictive validity of the overall model. With regression, it is a measure of how effective the combination of predictor variables is at explaining the variance in the criterion variable in a future sample. In regression, the squared multiple correlation coefficient, R^2, is the measure of model quality in the original sample, R_{cv}^2 is the cross-validated measure of model quality, and the difference between the two is shrinkage. While HLM, like regression, is used to explain and predict relationships between the criterion and predictor variables, there is not a commonly agreed upon analogue to regression's R^2, making questions about model quality, and as an extension, cross-validation, quite challenging.

As explained earlier, R^2 quantifies the proportion of variance in the criterion variable, Y, explained by the linear combination of predictor variables. The variability in Y that has been accounted for by the model is known as the SS_{reg}, the regression sum of squares. The variation that is not accounted for by the model is SS_{res}, the residual sum of

squares. By combining SS_{reg} with SS_{res}, the result is SS_y, or the total variation in Y. Since $R^2$is the proportion of variability in Y explained by the model, R^2 can be calculated for the regression model with the formula

$$R^2 = \frac{SS_{reg}}{SS_Y}. \quad (1)$$

One of the strengths of HLM is that it allows for a hierarchical analysis of the data. Level-1 data, typically data pertaining to individuals, and level-2 data, data pertaining to the group to which the individual belongs, are analyzed such that the variance in the criterion variable is partitioned between the levels and by component. The hierarchical structure and estimation procedure in HLM allow for the intercept and slope coefficients to vary between groups. Across all groups, the slope and intercept coefficients have a variance in addition to the variance that is measured at the individual level. Therefore, unlike regression in which there is one measure of variance, HLM provides two or more variance terms. Given the multiple variance measurements several methodologists advocated level and or component specific measurements of R^2 for HLM (Hox, 2002; Kreft & de Leeuw, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 1994, 1999). None of these suggestions, however, is a measure of explained variance that is identical to regression's R^2.

The component and level approaches are the two most widely accepted measures of explained variance for HLM. There is no disagreement; researchers should take advantage of the added insight HLM provides in understanding the level and component variance. It does seem, however, that there are also situations when understanding the amount of the variability of the criterion variable is explained by the overall model would be useful. Using the model for prediction and the accompanying need to cross-validate is

an ideal example of just such a situation in which a holistic approach to modeled variance might be needed. Gagné and Furlow (2009) proposed a holistic approached to modeled variance. What follows is an overview of the component, level, and holistic HLM R^2 measures that will be used in this cross-validation study.

The widely accepted component approach to explained variance with HLM is more aptly titled proportional reduction in variance (Raudenbush & Bryk, 2002). It is a calculation of the extent to which the original variance associated with a component has been reduced by the addition of predictor variables. This approach quantifies the change in variance when moving from a null model, or unconditional model, to the full model, also referred to as the conditional model. Because this is a component specific approach, the unconditional model to be used in the comparison varies dependent on the conditional model of interest. If the conditional model of interest is a random-intercept model, then the unconditional model has no predictor variables at either level. This model is symbolized in the equations

$$Y_{ij} = b_{0j} + r_{ij} \tag{2}$$

at level-1, and at level-2,

$$b_{0j} = \gamma_{00} + u_{0j}, \tag{3}$$

where group *j*'s intercept, b_{0j}, has a random component and varies across groups, u_{0j} is the deviation of group *j*'s intercept from the overall intercept, and r_{ij} is an individual's error reflecting how the individual *i* in group *j* differs from the outcome, given the estimate for the intercept. Variance terms are symbolized by $\tau_{00}$for the variance of the intercept and $\sigma^2$for the level-1 variance.

After the variances have been estimated for the unconditional model, predictors are added to the model. Keeping in mind this is a random-intercept model, predictors can be added at either level to achieve the conditional model, and residual variance terms are then estimated for the conditional model. With variance components for each level, the proportional reduction in explained variance can be calculated. The explained variance for level-1 is obtained by using the equation

$$R_1^2 = \left(\frac{\sigma_{uc}^2 - \sigma_c^2}{\sigma_{uc}^2}\right) \tag{4}$$

and for the intercept with

$$R_2^2 = \left(\frac{\tau_{00|uc} - \tau_{00|c}}{\tau_{00|uc}}\right). \tag{5}$$

It can be seen that each R^2 quantity is component specific and is the result of the difference between the variance in the unconditional model, subscript *uc*, and the conditional model, subscript *c*, as a proportion of the original, or unconditional model, variance.

Conceptually, there is no difference between calculating the R^2s for a random-intercepts model and a random-intercepts, random-slopes model. There are, however, differences in the number of R^2s calculated and the characteristics of the unconditional model. When the target model is a random-intercept, random-slopes model, the unconditional model must include the predictor variable(s) in the level-1 equation and no predictors in the level-2 equations. The level-1 predictor(s) must be included so that the slope variance (τ_{11}) can be estimated, allowing for a comparison with the slope variance after any level-2 predictors are added. The level-1 equation for the unconditional model is

$$Y_{ij} = b_{0j} + b_{1j}X_{1i} + r_{ij}, \tag{6}$$

and the level-2 equations for the intercept and slope are

$$b_{0j} = \gamma_{00} + u_{0j} \quad (7)$$

and

$$b_{1j} = \gamma_{10} + u_{1j}, \quad (8)$$

where u_{1j} represents the deviation of group j's slope from the predicted slope, γ_{10}. Once the variance terms for this unconditional model have been estimated, predictor variables can be added to level2 and new variance estimates calculated. Using the proportional reduction in variance formula, the reduction in slope variance with the addition of predictor variables is

$$R^2_{2_1} = \left(\frac{\tau_{11|uc} - \tau_{11|c}}{\tau_{11|uc}}\right). \quad (9)$$

The inclusion of the random slope does not change the formula for the proportional reduction in variance for the intercept or for the level-1 variance.

While not identical to the regression R^2, Raudenbush and Bryk's (2002) component approach of explained variance does provide an indication of how much variance is accounted for with the addition of predictor variables. Snijders and Bosker (1994, 1999) propose a level-based approach to explained variance that is a bit more similar to the regression R^2. Their approach is referred to as a calculation of the proportional reduction in prediction error, commonly referred to as modeled variance. It is a level-based approach in that there is one R^2 calculation for level 1 and one R^2 measurement for level 2, regardless of whether it is a random-intercepts model or random-intercepts, random-slopes model. In essence, Snijders and Bosker (1994) apply the concept of regression's R^2 to each level by determining how much of the variance in Y is explained by the model in comparison to the total variance of Y. This is done based

on the prediction performance for each level of the model. For level 1, the model prediction performance is based on how well the model predicts the individual outcome, Y_{ij}, and the level-2 model performance is based on the accuracy of the predicted outcome variable for each group, $\bar{Y}_{.j}$.

Like Raudenbush and Bryk's (2002) approach, Snijders and Bosker's (1994) approach also requires multiple sets of models. The difference between the two is that the modeled variance approach uses a baseline model in which all of the variables are included with the exception of the variable of interest (Hox, 2002). The variance from each level of the baseline model is then compared to the variance at each level of the target model. The target model is the model that includes all the selected predictors. This differs from the proportional reduction in variance approach which employs an unconditional model inclusive of the fewest possible predictor variables as the initial model.

In calculating level-1 modeled variance for a random-intercept model, a key concept is that level-1 variables can impact not only level-1 variance, but level-2 variance as well. Therefore, the prediction error for the level-1 model is $\sigma^2 + \tau_{00}$. Given R^2 is the proportional reduction in prediction error, the formula for level-1 seems logical,

$$R_1^2 = 1 - \frac{(\sigma^2+\tau_{00})_t}{(\sigma^2+\tau_{00})_b}, \quad (10)$$

1 minus the quantity of the total variance of the target model divided by the total variance of the baseline model (Snijders & Bosker, 1994).

At level-2 the prediction error in the random-intercept model refers to the accuracy of the equation to predict the group mean for the outcome variable. The mean square prediction error for level-2 is then derived from the variance of the mean value for

each group, $\bar{Y}_{.j}$, minus the linear equation involving a group-mean centered $\bar{X}_{.j}$. Snijders and Bosker (1994) equate the resulting variance of that quantity to $\sigma^2/n_j + \tau_{00}$. Similar to the level-1 formula, the level-2 model for the intercept is 1 minus the ratio of prediction error in the target model to the baseline model,

$$R_2^2 = 1 - \frac{\left(\frac{\sigma^2}{n_j} + \tau_{00}\right)_t}{\left(\frac{\sigma^2}{n_j} + \tau_{00}\right)_b}. \quad (11)$$

This modeled variance approach to $R^2$is substantially more complicated when random-slopes are added to the model. To understand the complication, it is first important to realize that as conceptualized by Snijders and Bosker (1994, 1999), an R^2 does not exist for the slope variance because for prediction error to be measured, the interest is placed on mean values. Specifically at level-2, the interest is on the prediction error of $\hat{\bar{Y}}_{.j}$ to $\bar{Y}_{ij}$. The slope quantifies the relationship between variables and plays a role in how well the model predicts, but has nothing to do with group means. According to Snijders and Bosker (1994), the inclusion of random-slopes will not make drastic changes to the R^2 values obtained under a fixed slope model unless there is a model misspecification. Random-slopes models do, however, necessitate the use of different formulas to quantify R_1^2 and R_2^2.

When random-slopes are included in the model, the formulas get quite complex with the formulas increasing in complexity as more coefficients are added to the model. The reason for the increased complexity is that in predicting Y_{ij} and $\bar{Y}_{.j}$ with random coefficients, the error is conditioned on the values of *X* and the covariates of *X* (Snijders & Bosker, 1994). Therefore, the formulas for R_1^2 and R_2^2 grow to involve the estimated covariance, the mean of the predictor variable, decomposed level-1 variance terms

quantifying, in separate terms, the between- and within-group variance comprising the predictor variable, as well as the intercept and level-1 variance terms. To calculate R_1^2 and R_2^2, the ratio of the level specific error term for the target model to the variance of the baseline model is subtracted from one. The target model error terms for level-1 and level-2 are, respectively,

$$\tau_{00} + 2\bar{X}_1\tau_{01} + \tau_{11}(\bar{X}_1 + \sigma_1^B + \sigma_1^W) + \sigma^2 \quad (12)$$

and

$$\tau_{00} + 2\bar{X}_1\tau_{01} + \tau_{11}\left(\bar{X}_1 + \sigma_1^B + \frac{1}{n_j}\sigma_1^W\right) + \frac{1}{n_j}\sigma^2. \quad (13)$$

In applying HLM measures of R^2 to cross-validation, the challenges are likely now evident. In cross-validating, the R^2 measures can still be used and shrinkage would still be expected in comparing the proportion of variance explained or reduction in prediction error between the screening sample and the validation sample. In calculating the shrinkage, the Snijders and Bosker (1994, 1999) approach would require two shrinkage calculations from four R^2 measures. The Raudenbush and Bryk (2002) approach will require three shrinkage calculations from six R^2 measures. In neither case is the R_{cv}^2 estimate providing an estimate of how well the model cross-validates; the resulting R_{cv}^2 are estimates of how well the level and/or the component cross-validates.

Cross-validating based on a component or level R^2 measure has value, but an R^2 and R_{cv}^2 that provide a regression-like measure of explained variance for the entire HLM model would be useful. Recently, there have been two works that attempt to establish a holistic measure of model quality for HLM models (Gagné& Furlow, 2009; Roberts & Monaco, 2006). Roberts and Monaco (2006) present three approaches to overall model quality. Each approach, as presented, is a creative attempt to solve the holistic model fit

problem with HLM. Gagné and Furlow (2009) present a single R^2 measure for HLM that is a straightforward extension of the regression R^2 and does not involve conceptualization that might impact the interpretation of resulting R^2_{cv} in a cross-validation context.

Roberts and Monaco (2006) propose three approaches to constructing an R^2 regression-like measure for HLM. Their objective for undertaking this task is to offer a model fit statistic that will be useful to "non-HLM minded researchers" (p.17) and to make HLM more user friendly. The first approach is the easiest of the three to understand and does in some sense relate to regression. The formula is $1 - \frac{\sigma^2_{error}}{\sigma^2_{total}}$. σ^2_{error} is calculated as it would be in regression with $\sum(Y_{ij} - \hat{Y}_{ij})^2$. σ^2_{total} is the sum of all of the squared residuals calculated by $\sum(Y_{ij} - \tilde{Y}_j)^2$where $\tilde{Y}_j$ is the random estimate for the group. The calculation for $\tilde{Y}_j$ was not described, and these variance estimates are biased. The second method proposed by Roberts and Monaco (2006) requires the use of a Gaussian probability density function in the equations which would provide an R^2 that is a function of the probability that a certain value is included given the model. The third approach, referred to as the Group Initiated R^2 is based on Weighted Least Squares. Foundational to this approach is the idea that each group has an R^2 measurement which is weighted based on the probability and precision of the group estimate. The total model R^2 from this approach is the averaged weighted least squares R^2 measurement for each group. Calculating each Roberts and Monaco's (2006) R^2, model fit measurements is a challenge and is arguably still not equivalent to the R^2 measurement from regression.

As a more straightforward approach, Gagné and Furlow (2009) have proposed a measure of R^2 for HLM that utilizes a regression approach to calculate R^2. In regression, R^2 can be calculated by multiplying each predictor-criterion correlation by its

corresponding standardized beta coefficient. The products for all the combinations are totaled and the result is R^2. Standardized beta coefficients in regression can be obtained by

$$\beta_k = b_k * \frac{\sigma_k}{\sigma_Y}, \quad (14)$$

where *k* is a predictor variable, β is the beta coefficient, σ_Y is the standard deviation for the outcome variable, and σ_k is the standard deviation for the predictor variable.

Gamma coefficients in HLM can be standardized in the same way that b-values are used to obtain standardized beta coefficients; simply substitute γ_k for b_k (Hox, 2002). Gagné and Furlow (2009) designate the resulting standardized coefficient as Б, the capital, Cyrillic letter beh. With the Бs, R^2 can be calculated with the formula

$$R^2 = Б_1 r_{Y1} + Б_2 r_{Y2} + Б_3 r_{Y3} + \ldots + Б_k r_{Yk}, \quad (15)$$

where r is the predictor-criterion correlation, the correlation between a given predictor variable and the outcome variable. This R^2, given the identical nature and derivation to regression, makes it easy for many people familiar with regression to interpret and implement. Gagné and Furlow (2009) repeatedly mention that the Б derived R^2 is not intended as a replacement for level or component approaches to measures of R^2, but that it is intended as another option to be considered along with others when evaluating explained variance. Likewise, as an overall measurement of model fit, like Gagné and Furlow's R^2 provides additional option for cross-validating HLM.

One of the areas this study intends to explore is the degree to which cross-validity shrinkage differs under various model fit measures. The variance of the intercept and slope are central to the above described component and level measures of explained variance; under these approaches, the measurement R^2 is by definition linked to the

variance. The importance of variance to measures of model fit is the impetus for the inclusion of centering in this study.

Centering in HLM is an often overlooked factor that affects the variance of the model. If the level-1 predictor variables are transformed from their raw score form, the most common transformation performed is centering. Centering is most typically discussed in terms of the impact it has on the interpretation of the coefficients. For purposes of this study, the specific interpretation of the coefficients is not a concern. Of interest is the impact centering has on shrinkage due to the effect centering has on the estimation of variance.

There are four primary ways to center level-1 variables in HLM analysis; no centering (raw score), grand-mean, group-mean, and centering around a value of interest. Centering around a value of interest can vary from study to study, so for purposes of this study, the focus will be on the other two approaches. A general understanding of each approach and the impact centering has on variance will be conceptually useful.

Centering is typically performed for one of three reasons. The first is that the research question of interest can be better addressed by centering the variables so that the coefficients are interpreted in a way that is pertinent to the question. An example of this would be group-mean centering to focus on context effects. The second reason is that, in some situations, the interpretation of the intercept is not meaningful with the original measurement metric of the variable. A familiar example is the SAT. In a model with no level-2 predictors, the intercept is the predicted outcome for Y when the level-1 predictors equal 0. The minimum score on the SAT is 400, so the interpretation of the intercept as the predicted outcome, $\hat{Y}$, when SAT is 0 is somewhat meaningless as it is

not a possible scenario. Finally, centering is recommended by some methodologists to aid in obtaining stable parameter estimates and convergence (Hox, 2002; Raudenbush & Bryk, 2002).

When analysis with raw scores will not work, grand-mean centering is often employed. Grand-mean centering is a linear transformation performed on a variable. As the name implies, with this approach to centering, a single value, the mean across all values of *X*, is subtracted from each of the *X* values. The HLM level-1 equation with a grand-mean-centered X value is

$$Y_{ij} = b_{0j} + b_{1j}\left(X_{ij} - \bar{X}..\right) + r_{ij}, \quad (16)$$

where $\bar{X}..$ is the grand mean for the *X* variable in the overall sample. Since the grand-mean is the mean for the entire sample, subtracting it from each X_{ij}has the benefit of maintaining the relationship of each data point to the rest, so the data points maintain their relative positions to each other in terms of order and degree. After grand-mean centering, the estimated slope coefficients will be unchanged, but the intercept will change. The intercept after grand-mean centering can be thought of as the average adjusted mean for cluster *j* as well as the expected value of *Y* when the value of the *X* value is equal to the mean of *X* (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002b).

Group-mean centering is conceptually and procedurally very similar to grand mean centering but the implications are very different. Group-mean centering transforms the *X* variable by subtracting from X_{ij} the mean value of X for the group to which X_{ij} belongs. The level-1 equation then becomes

$$Y_{ij} = b_{0j} + b_{1j}\left(X_{ij} - \bar{X}_{.j}\right) + r_{ij}. \quad (17)$$

Since the value subtracted from eachX_{ij} varies based on group membership, the relationship between the X_{ij} values is not preserved as they are with grand-mean centering. The order and degree of difference of the values has been altered. The between-group differences have been removed, as each *X*-value is a deviation score reflecting the relationship between it and other cases in the same group. In essence, the model has been changed and a new variable, with a mean of 0 for each group, has been introduced. With this "new" variable, all scores are relative to the means, so if persons A and B both had scores 10 points below their respective group means, the scores are equal after group-mean centering even though their actual scores may be appreciably different.

The implications of grand- and group-mean centering to the intercept and slope variance are not intuitive or easily understood. Raudenbush and Bryk (2002) provide an explanation that offers some insight. Variance components are estimated through iterations of an expectation-maximization algorithm (EM) which is built upon the measurement of empirical Bayes (EB) residuals. EB residuals are calculated from error in the expected values given the data, previous error measures, and the model parameters for the given iteration. Centering changes EB residuals. The primary result of this change is that grand-mean centered models will experience slope homogenization (Raudenbush & Bryk, 2002). Slope homogenization occurs because the adjusted means created by grand-mean centering necessitate extrapolation outside of the data to arrive at intercepts for groups comprised of higher or lower values on the *X* variable. Because of the extrapolation, the estimates are less reliable, and less reliable estimates result in intercepts pulled in the direction of the center of the distribution. With the intercept pulled toward the center, slopes are also pulled. Slopes that would be flat for the more extreme groups

are made steeper which tends to homogenize and underestimate the variance. Groups in the center of the distribution are less impacted because they already tend toward the center of the distribution.

Group-mean centering does have some impact on the estimation of variance components, but not to the same extent as grand-mean centering. If the intercept has low reliability, the variance estimated for it will decrease, with the extent of the decrease determined by how low the reliability is. Similarly, the variance of the slope will also decrease if reliability for the slope parameter is low. Raudenbush and Bryk (2002) state that the decrease in the intercept will not have much impact on the slope unless there is a high correlation between the two.

The estimation of variance components is impacted by not only centering, but also by sample size, as sample size plays a role in the stability and accuracy of the estimates. Therefore, sample size needs to be taken into consideration in planning not only this study, but all HLM studies. In determining sample size conditions for this simulation study, two areas of research are relevant: sample size as it relates to HLM and sample size as it pertains to cross-validating regression models. Within each of the two areas of research, there is a plethora of research addressing sample size as it relates to effect size, power analysis, cluster sampling, and budget constraints. While all of these are important, the focus of this study is on cross-validating, so the remainder of this review will explore research related to the impact of sample size on the estimation of coefficients and variance in HLM contexts and on shrinkage in cross-validation.

A review of HLM sample size literature presents the selection of an optimal sample size within several different contexts. As mentioned, sample size can be

contextualized into various optimal sample sizes: for power, for small standard errors, for parameter estimation, for significance testing. Unfortunately, the optimal sample size may differ based on which aspect is under consideration. Overriding the statistical implications of optimal sample size are often circumstances such as budgets, time, and ethics, making "sample size a constrained optimization problem" (Afshartous, 1995, p.4). The researcher must determine the optimal sample size based on the aspects most important to the research question, and then, given the constraints, determine the sample size that is as close to optimal as possible.

An objective of this study is to explore the impact of sample size on the cross-validation of HLM. Given the question of interest, it seems clear that the most important implications of sample size relate to the estimation of parameters. Researchers exploring sample size with HLM consistently take budgetary constraints into account by assuming a fixed budget and varying the number of subjects at level-1 and the number of groups at level-2 to order stay within budget while attempting to optimize sample size for stable parameter estimates. General guidelines that apply for regression also apply to HLM models. Snijders and Bosker (1999) point out that within constraints, sample size should be selected such that the standard errors for the parameters of interest are minimized. When others factors are held constant, increases in sample size lead to decreases in standard errors. Because of the cluster sampling that inherently accompanies HLM and estimation procedures, a general guideline is that the level-2 sample size should be larger than would typically be recommended for a single-level regression study with the same number of predictor variables.

Several simulation studies (Kreft, as cited in Hox, 2002; Mok, 1995; Maas & Hox, 2005;Van der Leeden, Busing, & Meijer, 1997) have explored the question of whether a larger level-1 or level-2 sample size impacts the accuracy of parameter estimates. It seems clear that for accuracy of obtaining estimates of the coefficients and the variances, more groups with fewer individuals per group is a better option than more individuals within fewer groups for a given total sample size. There is, however, less consensus on the number of groups needed for accurate parameter estimates. Kreft proposed a series of guidelines, of which many exist for determining sample size for regression equations. These rules are based on the component of interest; the 30/30 rule, 30 groups with 30 subjects when the interest is in fixed parameters; the 50/20 rule, 50 groups with 20 subjects per group when the cross-level interaction is of particular interest; when variance and covariance are the focus, the 100/10 rule. Along similar lines, Van der Leeden, Busing and Meijer (1997) found that when level-2 variance components are of interest, at least 100 groups are needed for accurate estimates. In a simulation study conditioning on sample size, slope and intercept correlations, and varied intra-class correlations, Van der Leeden and Busing (as cited in Kreft, 2007) found that when the number of groups is less than 300, variance components are consistently underestimated. In direct contrast, Brown and Draper (2000) suggested that acceptable variance estimates can be achieved with 6-12 groups and 48 groups show good estimates. It should be noted that there are several different estimation techniques that can be implemented to arrive at the variance estimates, and it is clear that different approaches can make a difference in the optimal sample sizes required.

In an often cited article, Mok (1995) frames the question of sample size in a different way. She looks at the bias and consistency of parameter estimates by comparing estimates by total sample sizes (number of groups x number of subjects) across different distributions of the total sample size: level-1 larger than level-2; level-2 larger than level-1; and equal sizes. An example would be 40 students in 10 schools. This was compared to designs with an equal number of level-1 and level-2 units, 20 students in 20 schools, and designs with a greater number of level-2 units than level-1 units; 10 students in 40 schools, but with a total sample size of 400. For each total sample size, there were at least 2 points for comparison. Mok (1995) found that for slope and intercept coefficients, bias decreased as total sample size increased and in cases where total sample size is less than 800, designs with fewer level-2 units than level-1 units showed more bias. Estimates of the level-2 variance components show more bias when the number of level-1 units is greater than the number of level-2 units. Increasing total sample size reduces the bias of designs with more level-1 units than level-2 units. Estimates of the level-1 variance component of all designs are equivalent when the total sample size is greater than 4000. The bias in the level-1 variance decreases quickly when going from a sample size of 25 to 600, but after 600 the decrease levels off; samples with more level 1-units than level-2 units are less stable. The overall conclusion is that when possible, have more groups than subjects per group and larger total sample sizes are preferred.

Sample size recommendations for HLM are one way to approach sample size for this study, but this study is also about cross-validating HLM, so sample size can also be explored from previous cross-validation research. Since research on cross-validation is embedded in regression, methods of selecting sample size for regression get intertwined

with the cross-validation sample size research. A perfect example is Osborne's (2000) work. In looking at sample sizes needed to minimize shrinkage with double cross-validation, he makes use of a guideline for sample size calculation. Many authors have tried to provide an easy guide for applied researchers as to how big a sample needs to be for regression. Some of the more well-known guidelines are reviewed in Green's (1991) effort to add a power component to a ratio approach. Osborne makes use of the ratio approach to sample size where a specified number of subjects are needed per predictor variable in the model. Selecting five different sample sizes of 5, 15, 40, 100, and 400 subjects per predictor, Osborne uses a large national database to randomly sample subjects, construct a model, and double cross-validate. His findings suggest that explained variance in the sample is overestimated when compared to ρ^2 when the ratio of subjects to predictors is less than 100:1. In terms of comparing cross-validity coefficients, it appears that shrinkage becomes minimal at a 40:1 ratio of subjects to predictors.

Approaching the cross-validation sample size from a different perspective, Algina and Keselman (2000) conducted a simulation study to determine the sample size needed for shrinkage to be at or below a specific amount of acceptable shrinkage in 95% of the replications. They outlined four levels of shrinkage; .025, .05, .075, and .1, used in the simulation study. Varying the number of predictor variables, the size of R^2 , and sample sizes, they simulated 5000 replications of shrinkage estimates which were used as the distribution of shrinkages given the specific combination of conditions. For each combination, the proportion of shrinkages below the specific levels was then determined. The smallest sample size was determined for each combination of predictor, R^2 , and shrinkage level such that the probability of obtaining a shrinkage greater than the

specified amount was .05 or less. Algina and Keselman's findings indicate that as the number of predictors increased, *N* also had to increase to maintain the .95 probability of remaining under the specified shrinkage level. To obtain a shrinkage of .05 with an R^2 equal to .35 and two predictor variables, the sample size needed is 60. Dropping the acceptable shrinkage to .025 while maintaining the same R^2 and number of predictors, the *N* required increases to 100. Similarly, with an R^2 of .35 and eight predictor variables, the *N* required for a maximum shrinkage of .05 is 190, whereas with a maximum shrinkage of .025, the *N* needed becomes 360. The challenge with implementing Algina and Keselman's (2000) optimal sample size recommendations is that R^2 must be known or guessed. Their results show that in guessing, it is important to be conservative in specifying a R^2 because the optimal sample size recommended will be too small if the speculated R^2 is larger than the actual R^2.

In responding to the lack of cross-validated research findings published in journals in the field of business and policy strategy, St. John and Roth (1999), using sample size, number of predictors, and R^2, substantiate that research findings regularly perceived as highly predictive may be seen as appreciably weaker after being cross-validated. Enough information is provided in journal articles that, using a shrinkage formula, St. John and Roth were able to calculate R^2_{cv} for many of the predictive models presented in journals pertaining to business and policy strategy. They found that in studies with a sample size of 90 or less and four or fewer predictor variables, the R^2 decreased by 33% when cross-validated. In studies with four or fewer predictors and a sample size over 250, the R^2 shrunk by only 7%. Sample size makes a large difference in the amount of shrinkage that occurs in R^2. The formula St. John and Roth used was the

Drasgow, Dorans, and Tucker (1979) version of the Browne (1975) formula. A quick glance at this formula, and all formulas that purport to estimate the squared cross-validity coefficient, reveals the central role sample size plays in the accuracy of cross-validation. Adjustments to the measured R^2 based on sample size constitute the vast majority of formulas.

As this review explains, cross-validation studies are not always as easy as one might expect. Couple the intricacies of cross-validation with a complicated approach like HLM and problems as well as unanswered questions abound. Despite the challenges, HLM, like regression, can be used for the prediction of an outcome variable, so the need to cross-validate is clear. What is absent in the literature is guidance on how best to cross-validate HLM. This simulation study will begin to address basic but important questions regarding the conditions that may affect the cross-validation HLM. Sample size, predictor correlations, model fit, and centering may all have an impact on cross-validation within an HLM framework, so this study intends to investigate the impact of each.

CHAPTER 3

METHOD

A cross-validation simulation study was conducted to investigate the impact of sample size, predictor-criterion correlations, centering, and R^2 approaches on shrinkage. The specific factors examined in this study include: four different sample sizes for level-1, four sample sizes for level-2, two sets of predictor-criterion correlations, three approaches of measuring "explained" variance, and two centering approaches. Initially, three centering approaches were planned, but after a preliminary analysis it was determined that grand-mean centering and a raw score approach would produce identical results, so this study employed two approaches to centering, group-mean centering and raw scores from a grand-mean-centered population. All levels were fully crossed resulting in 64 conditions which are listed in Table 1. The 64 conditions resulted in 256 shrinkage measurements for comparison.

Table 1

Study Conditions

Level-1 Sample Size

1. 10
2. 15
3. 20
4. 30

Level-2 Sample Size

1. 10
2. 15
3. 20
4. 30

Predictor-Criterion Correlations

1. ρ_{xy} = .3, ρ_{wy} = .4, and $\rho_{y(wx)}$ = .3
2. ρ_{xy} = .4, ρ_{wy} = .3, and $\rho_{y(wx)}$ = .3

Centering Approaches

1. Raw
2. Group- mean

Shrinkage Measurements (3 approaches)

1. Snijders & Bosker
    a. Level-2 R_{cv}^2-R^2
2. Raudenbush & Bryk
    a. Intercept variance R_{cv}^2-R^2
    b. Slope R_{cv}^2-R^2
3. Gagné and Furlow
    a. R_{cv}^2-R^2

Conditions

Sample size. Sample sizes were selected with the intention that this study would be relevant to applied researchers and the sample size constraints they frequently face. Extreme sizes are often not practical and were avoided in this simulation. The conditions for the level-1 sample sizes were 10, 15, 20, and 30. At level-2, the sample sizes were also 10, 15, 20, and 30. The specific sample sizes were selected such that a number of conditions existed in which level-1 would be larger than level-2, and several level-2 sample sizes would be larger than level-1. The common recommendation from HLM research related to sample size is that larger sample sizes at level-2 are beneficial in achieving stable and less biased parameter estimates. In order to test that idea with cross-validation, the design of this study required situations where the larger sample size is at level-1. Additionally, the design of this study allowed for comparisons of both large and small total sample sizes and the various level-1 and level-2 size allocations of the total sample size.

Centering. This study included two of the main types of centering for cross-sectional studies. In practical application, the centering approach employed in a given study is determined by the research questions of interest. It was, therefore, pertinent to determine if each centering approach cross-validates similarly under the varying conditions of sample size, predictor-criterion correlations, and the “explained” variance options.

Predictor-Criterion Correlations. The predictor-criterion correlations used in this study consisted of the correlation between the X and Y variables alternating between .3 and .4. The correlation between the W and Y variables was assigned the value of .4 when

ρ_{xy}= .3, and ρ_{wy}= .3 when ρ_{xy}= .4. The correlation between Y and the product term, $\rho_{Y(WX)}$, was fixed at .3.

Shrinkage. To obtain shrinkage measurements, this study utilized the two commonly employed measures of explained variance for HLM, Snijders and Bosker (1994, 1999) and Raudenbush and Bryk (2002). This study also included a proposed measure of overall model explained variance, Gagné and Furlow's (2009) R^2. The Gagné and Furlow approach was selected over Roberts and Monaco's (2006) approaches because the standardized coefficients used by Gagné and Furlow renders their approach to R^2 in HLM identical to calculating the regression R^2, and is, therefore, more accessible to researchers less familiar with HLM. For ease of reading, the following abbreviations will be used to denote the R^2 approaches under consideration: Raudenbush and Bryk's (2002) R^2 from intercept variance, R^2_{RB0}; Raudenbush and Bryk's (2002) R^2 from slope variance, R^2_{RB1}; Snijders and Bosker's (1994, 1999) R^2 from level-2 variance, R^2_{SB}; and Gagné and Furlow's (2009) R^2, R^2_{GF}.

Data for each condition were generated in SAS/IML. Once generated, the data were split to obtain the estimation and validation samples. In order to split the data, the number of groups for a condition was doubled and then the groups were divided such that the first half of the groups constituted the estimation sample and the second half formed the validation sample. While splitting the data for cross-validation is questioned by the likes of Murphy (1983) who warns of the silliness in comparing a sample of idiosyncratic data to itself, this study was not concerned as much with sampling problems that can exist within samples or about how reflective the R^2 is of the population. The research questions were limited to the impact of different design attributes on cross-validation. While much

research with HLM is conducted under less than ideal research designs, this study will simulate random data so that conclusions can be drawn about the impact of the various conditions under a simulated ideal situation. Additionally, since data are split in this study, all replications should cross-validate better than a traditional study, so this study, within an "ideal" setting, can compare which approaches validate better and worse. Therefore, splitting the data is acceptable with this intention.

After the data were generated and split, PROC MIXED was used to estimate the model parameters for the estimation sample baseline model and conditional models, and R^2s were calculated. The conditional model selected was a random-intercepts, random-slopes model. The validation sample was then used to estimate the parameters for its unconditional model. Coefficients from the estimation sample's conditional model were then applied to the validation sample, and the R_{cv}^2s were calculated. With the R^2s and the R_{cv}^2s, shrinkage can be tabulated. This process was replicated 1000 times for each condition, and average shrinkages per condition were then compared.

CHAPTER 4

RESULTS

Each R^2shrinkage measurement presented is the average shrinkage for the 1,000 replications within each condition. Since each of the R^2 measures considered in this study is defined differently and shrinkage in cross-validation is a measurement of the change in the R^2 values, shrinkage measurements lose meaning when not provided in relation to the original R^2 measurements. Therefore, in much of this chapter, shrinkage will be considered as the percent change in the estimation sample R^2. Shrinkage of 19% would indicate that the difference between the estimation sample $R^2$and the validation sample R^2 decreased by 19% of the value of the estimation sample R^2. In addition to framing shrinkage as a percentage of the estimation sample R^2, the analysis of the impact of the various conditions will be examined from within each R^2 approach. First, an overview of the R^2 and shrinkage measurements across all the conditions is appropriate.

Table 2 provides the range of the R^2 measures for the estimation and validation samples for each approach to explained variance. Notably there are several occurrences of negative R^2 values in the validation samples (i.e., for the cross-validity coefficients) for the majority of the R^2 approaches. The R^2_{GF} is the only type of R^2 measures that does not have a negative value. This is as opposed to R^2_{SB} in which 52 of the 64 conditions averaged a negative R^2 cross-validity coefficient measure.

Table 2

R^2 Values for Estimation and Validation Samples in Selected Approaches

	R^2_{RB0}_E	R^2_{RB0}_V	R^2_{RB1}_E	R^2_{RB1}_V	R^2_{SB}_E	R^2_{SB}_V	R^2_{GF}_E	R^2_{GF}_V
Min Avg. R^2	0.4044	-7.1098	0.3738	-0.9229	-0.0574	-1.0506	0.3374	0.2526
Max Avg. R^2	0.6087	0.5013	0.4701	0.3369	0.3120	0.1662	0.3714	0.3252
Count R^2< 0	0	31	0	21	1	52	0	0

Note. _E = Estimation R^2, _V = Cross-validity coefficient.

While the R^2 amounts show the variability of the measures, the real interest in a cross-validation study is the change in $R^2$from the estimation sample to the validation sample. The range of average shrinkage and the magnitude of those estimates are displayed in Table 3. The magnitude of the shrinkages resulting from the vastly different R^2s is made clear by the shrinkage percentage range. A brief survey of the table reveals that in comparison, the R^2_{SB} results in a relatively large range for the magnitude of shrinkage and the R^2_{GF} results in a relatively narrow range for the magnitude of shrinkage.

Table 3

Range of Average Shrinkage Estimates

	R^2_{RB0}	R^2_{RB1}	R^2_{SB}	R^2_{GF}
Min. Shrinkage	0.0995	0.1246	0.1458	0.0161
Max. Shrinkage	7.5142	1.3167	0.9932	0.1125
Min. Shrinkage %	16.57	27.00	-1729.82	4.72
Max. Shrinkage %	1858.11	334.34	10643.32	30.29

The remainder of this chapter will focus on the impact of the conditions under investigation by exploring their impact on shrinkage within each of the R^2 measures of explained variance. Each analysis will contain two tables. The first table will focus on

shrinkage within each of the conditions but will do so by total sample size. This table will be useful in the analysis of shrinkage across total sample size along with the allocations of the total sample size into level-1 and level-2 units. The impact of the predictor-criterion correlations and centering can also be assessed with the layout of this table. The second table will focus on the change in shrinkage across level-1 and level-2 sample sizes holding the other conditions constant.

Raudenbush & Bryk's (2002) R_2^2

Raudenbush & Bryk's (2002) approach to explained variance involves multiple R^2 measures, one for each component in the model. This study investigated the level-2 explained variance measures, R_{RB0}^2 and R_{RB1}^2. In considering the shrinkage estimates forR_{RB0}^2, it is notable that in nearly every case (15 out of the 16) increasing level-2 sample size while holding level-1 sample size constant yielded less shrinkage as reflected in the "Shrinkage %" column in Table 4. Similar results are found when comparing shrinkage across level-1 sample sizes within a level-2 sample size. In 15 of the 16 combinations, larger level-1 sample sizes are associated with less shrinkage.

Table 4

Shrinkage Estimates for R^2_{RB0} by Total Sample Size

Sample Size			Shrinkage %		Correlations	
Total	L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}
100	10	10	230.09	88.57	0.4	0.3
150	15	10	65.33	70.37	0.4	0.3
150	10	15	61.10	56.93	0.4	0.3
200	20	10	47.00	46.68	0.4	0.3
200	10	20	52.90	47.77	0.4	0.3
225	15	15	41.24	43.57	0.4	0.3
300	30	10	45.06	34.94	0.4	0.3
300	10	30	43.43	39.34	0.4	0.3
300	20	15	33.28	27.65	0.4	0.3
300	15	20	33.43	31.41	0.4	0.3
400	20	20	29.20	26.73	0.4	0.3
450	30	15	25.76	21.45	0.4	0.3
450	15	30	28.94	27.23	0.4	0.3
600	30	20	18.48	16.88	0.4	0.3
600	20	30	22.54	20.95	0.4	0.3
900	30	30	17.04	16.57	0.4	0.3
100	10	10	1858.11	358.38	0.3	0.4
150	15	10	264.29	247.52	0.3	0.4
150	10	15	240.88	210.80	0.3	0.4
200	20	10	216.67	356.68	0.3	0.4
200	10	20	194.61	166.59	0.3	0.4
225	15	15	178.34	168.83	0.3	0.4
300	30	10	161.50	174.43	0.3	0.4
300	10	30	155.08	150.67	0.3	0.4
300	20	15	156.02	142.83	0.3	0.4
300	15	20	157.54	141.19	0.3	0.4
400	20	20	134.28	122.69	0.3	0.4
450	30	15	126.99	132.11	0.3	0.4
450	15	30	122.12	124.24	0.3	0.4
600	30	20	115.01	109.66	0.3	0.4
600	20	30	108.57	109.07	0.3	0.4
900	30	30	99.02	98.65	0.3	0.4

Shrinkage decreased the most within level-1 combinations when the level-2 sample size increased from 10 to 15. Under grand mean centering with $\rho_{xy} = .4$, $\rho_{wy} = .3$, shrinkage decreased more when moving from level-2 sample sizes of 10 to 15 and 20 to 30 than when moving from 15 to 20 (see Table 5). This holds true for only half the grand-mean centered values with $\rho_{xy} = .3$, $\rho_{wy} = .4$. The trend (7 of the 8) with group-mean centered conditions is that with each increase of level-2 sample size, the magnitude of the shrinkage decrease is less and less substantial regardless of the predictor-criterion correlations.

When changing level-1 sample size within a given level-2 sample size, the greatest decrease in shrinkage is no longer consistently associated with an increase from 10 to 15. Grand-mean-centered data display the greatest decrease in shrinkage when moving from a level-1 sample size of 10 to 15. This, however, is true for only four of the eight combinations under group-mean centering. For the other four cases, three are under the $\rho_{xy} = .3$, $\rho_{wy} = .4$ predictor-criterion correlation, and two of the four show the greatest change in moving from a level-1 sample size of 15 to 20.

The impact of the predictor-criterion correlations with R^2_{RB0} is clearly evident in Table 4. Shrinkage tends to be less with $\rho_{xy} = .3$, $\rho_{wy} = .4$, and the trend within each predictor-criterion correlation is downward with larger total sample sizes. When looking at a specific total sample size and varying the allotment of the total sample across level-1 and level-2, shrinkage is smaller when level-1 sample sizes are larger than level-2 sizes under the $\rho_{xy} = .3$, $\rho_{wy} = .4$. The reverse tends to be true when $\rho_{xy} = .4$, $\rho_{wy} = .3$: Smaller shrinkage is associated with larger level-2 sizes.

Table 5

% Change in Shrinkage Estimates for R^2_{RB0}

Sample Size		% Change across n_{L2}		Correlations		Sample Size		% Change across n_{L1}	
L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}	L2	L1	Grand	Group
10	10			0.4	0.3	10	10		
	15	71.85	28.57	0.4	0.3		15	70.87	16.36
	20	11.27	15.55	0.4	0.3		20	29.26	33.44
	30	16.26	15.57	0.4	0.3		30	1.84	25.07
	10			0.3	0.4		10		
	15	86.13	36.30	0.3	0.4		15	85.27	27.87
	20	15.65	19.19	0.3	0.4		20	16.86	-42.28
	30	19.65	8.90	0.3	0.4		30	25.67	53.09
15	10			0.4	0.3	15	10		
	15	34.22	35.35	0.4	0.3		15	31.91	24.30
	20	17.95	26.95	0.4	0.3		20	19.45	37.57
	30	12.80	11.39	0.4	0.3		30	21.99	20.92
	10			0.3	0.4		10		
	15	28.83	29.11	0.3	0.4		15	24.40	19.72
	20	9.78	16.71	0.3	0.4		20	12.64	16.17
	30	21.45	8.83	0.3	0.4		30	19.08	9.29
20	10			0.4	0.3	20	10		
	15	25.07	39.34	0.4	0.3		15	37.06	34.52
	20	10.37	-0.29	0.4	0.3		20	11.98	14.27
	30	22.76	21.22	0.4	0.3		30	37.85	37.50
	10			0.3	0.4		10		
	15	25.23	58.23	0.3	0.4		15	19.14	17.26
	20	12.07	12.56	0.3	0.4		20	14.87	11.99
	30	17.14	8.43	0.3	0.4		30	13.90	11.50
30	10			0.4	0.3	30	10		
	15	40.45	36.00	0.4	0.3		15	34.43	31.29
	20	28.55	20.72	0.4	0.3		20	22.04	23.77
	30	5.28	0.04	0.4	0.3		30	23.76	20.65
	10			0.3	0.4		10		
	15	18.60	19.24	0.3	0.4		15	20.94	17.21
	20	6.45	14.68	0.3	0.4		20	10.21	11.61
	30	12.53	6.60	0.3	0.4		30	9.12	9.72

A final observation related to shrinkage and the application of R^2_{RB0} can be made. In reviewing Table 4, it appears that the centering approach employed impacts shrinkage. The trend is not obvious because many of the grand-mean and group-mean shrinkage estimates are within 10 percentage points of each other, but 78% of the estimates from grand-mean centering data are greater than those from group-mean centered data. The seven instances in which group-mean centering resulted in larger shrinkage estimates are predominantly found with $\rho_{xy} = .4$, $\rho_{wy} = .3$. Most of the instances with less than a 10 percentage point difference between group- and grand-mean centering occur in the $\rho_{xy} = .3$, $\rho_{wy} = .4$ category. There are seven instances within $\rho_{xy} = .4$, $\rho_{wy} = .3$, where the grand- and group-mean shrinkage estimates differ by less than 10 percentage points. Of these, five appear in conjunction with the largest total sample sizes.

Raudenbush & Bryk's (2002) $R^2_{2_1}$

As with R^2_{RB0} shrinkage, it appears that in using R^2_{RB1} smaller shrinkage estimates are associated with larger level-2 sample sizes for a given level-1 sample size. Under grand-mean centering, larger level-1 sample sizes for a given level-2 sample size are associated with less shrinkage. This is not the case with group-mean centering. Under group-mean centering, there are instances when a larger level-1 sample size results in a higher shrinkage estimate. These instances are found only when the sample size for level 2 is 10 or 15 (see Table 6).

Table 6

Shrinkage Estimates for R^2_{RB1} by Total Sample Size

Sample Size			Shrinkage %		Correlations	
Total	L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}
100	10	10	183.29	243.05	0.4	0.3
150	15	10	127.63	124.42	0.4	0.3
150	10	15	168.84	150.01	0.4	0.3
200	20	10	91.38	151.97	0.4	0.3
200	10	20	113.88	109.66	0.4	0.3
225	15	15	82.47	90.18	0.4	0.3
300	20	15	62.56	58.12	0.4	0.3
300	15	20	67.73	67.85	0.4	0.3
300	30	10	65.25	80.80	0.4	0.3
300	10	30	98.55	101.04	0.4	0.3
400	20	20	43.90	51.44	0.4	0.3
450	30	15	41.56	43.34	0.4	0.3
450	15	30	56.87	58.26	0.4	0.3
600	30	20	33.97	32.49	0.4	0.3
600	20	30	41.10	44.36	0.4	0.3
900	30	30	30.73	27.66	0.4	0.3
100	10	10	334.34	189.04	0.3	0.4
150	15	10	126.06	124.62	0.3	0.4
150	10	15	150.27	149.98	0.3	0.4
200	20	10	101.89	164.30	0.3	0.4
200	10	20	112.77	126.60	0.3	0.4
225	15	15	77.09	75.49	0.3	0.4
300	20	15	60.15	128.97	0.3	0.4
300	15	20	63.75	64.63	0.3	0.4
300	30	10	66.24	70.56	0.3	0.4
300	10	30	98.85	97.41	0.3	0.4
400	20	20	45.87	47.26	0.3	0.4
450	30	15	41.71	37.57	0.3	0.4
450	15	30	56.34	57.69	0.3	0.4
600	30	20	35.76	33.15	0.3	0.4
600	20	30	42.35	40.11	0.3	0.4
900	30	30	28.11	27.00	0.3	0.4

Within level-1 combinations, six of the eight grand-mean centered conditions show the largest reduction in shrinkage when increasing the level-2 sample size from 10 to 15. In five of the eight combinations, the magnitude of the change in shrinkage decreases with each increase of the level-2 sample size. In comparison, under group-mean centering five of the eight combinations show the largest shrinkage reduction when moving from a level-2 sample size of 10 to 15, but only three of the combinations reflect the pattern of orderly decrease when level-2 sample size increases (see Table 7).

When varying level-1 sample size within a given value of level-2 sample size, 13 of the 16 combinations have the largest decreases in shrinkage when moving from a level-1 sample size of 10 to 15. Under grand-mean centering fewer than half of the eight combinations present a pattern where each increase in level-1 sample size yields a decrease in the magnitude of the change in shrinkage. Only one level-2, group-mean centering condition combination reflects a decreasing magnitude of change for each increase in the level-1 sample size. This is the level-2 size of 15 with $\rho_{xy} = .3$, $\rho_{wy} = .4$. The other seven condition combinations reflect a smaller percent change in shrinkage when moving from a level-1 sample size of 15 to 20 than in moving from 20 to 30.

Table 7

% Change in Shrinkage Estimates for R^2_{RB1}

Sample Size		% Change across n_{L2}		Correlations		Sample Size		% Change across n_{L1}	
L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}	L2	L1	Grand	Group
10	10			0.4	0.3	10	10		
	15	-1.1	31.3	0.4	0.3		15	25.1	48.8
	20	26.6	25.5	0.4	0.3		20	26.8	-22.1
	30	13.5	3.8	0.4	0.3		30	27.4	46.8
	10			0.3	0.4		10		
	15	51.7	4.7	0.3	0.4		15	62.0	34.1
	20	21.4	13.1	0.3	0.4		20	15.3	-31.8
	30	10.4	21.8	0.3	0.4		30	35.3	57.1
15	10			0.4	0.3	15	10		
	15	30.0	19.3	0.4	0.3		15	48.1	39.9
	20	15.3	22.5	0.4	0.3		20	23.5	35.6
	30	15.1	12.3	0.4	0.3		30	33.0	25.4
	10			0.3	0.4		10		
	15	32.1	36.6	0.3	0.4		15	46.5	49.7
	20	15.7	14.3	0.3	0.4		20	20.7	-70.8
	30	9.6	6.4	0.3	0.4		30	31.8	70.9
20	10			0.4	0.3	20	10		
	15	26.9	59.1	0.4	0.3		15	40.2	38.1
	20	29.9	9.8	0.4	0.3		20	36.7	24.2
	30	5.3	12.7	0.4	0.3		30	21.4	36.8
	10			0.3	0.4		10		
	15	36.4	15.0	0.3	0.4		15	42.6	48.9
	20	23.7	62.3	0.3	0.4		20	28.3	26.9
	30	5.3	13.6	0.3	0.4		30	21.7	29.8
30	10			0.4	0.3	30	10		
	15	32.5	43.6	0.4	0.3		15	41.3	42.3
	20	17.7	23.1	0.4	0.3		20	29.4	23.8
	30	6.2	12.2	0.4	0.3		30	22.1	37.7
	10			0.3	0.4		10		
	15	33.0	46.4	0.3	0.4		15	42.1	40.8
	20	12.4	9.8	0.3	0.4		20	24.9	30.5
	30	19.2	16.8	0.3	0.4		30	33.2	32.7

Shrinkage, as it relates to total sample size with R^2_{RB1}, is smaller when total sample sizes are larger. There are exceptions to this general trend. Most of the exceptions occur under group-mean centering and under the predictor-criterion correlation $\rho_{xy} = .4$, $\rho_{wy} = .3$. The majority of the exception also occur when the level-2 allocation of the total sample size is larger than the level-1 allocation (see Table 6).

Larger level-2 sample sizes for a specific total sample size are associated with increased amounts of shrinkage. This is true for all of the group-mean centered instances and true for 9 out of 12 of the grand-mean cases. In looking closer at the total sample size of 300, the fluctuations in the shrinkage magnitude are exemplified. With a total sample size of 300 and predictor-criterion correlations of $\rho_{xy} = .3$, $\rho_{wy} = .4$, under both the group- and grand-mean centering, the comparison of each sample size combination reflects a larger shrinkage with the larger level-2 sample size, with the exception of the level-2 of 10 and 15. With each level-2 increase after 15, shrinkage increased. This holds true for the grand-mean-centered shrinkage under $\rho_{xy} = .4$, $\rho_{wy} = .3$. Shrinkage for group-mean-centered instances of 300 with the same predictor-criterion correlation is larger with a level-2 sample size of 10 to 15, decreases when level-2 is 20 and once again larger when sample size is 30.

In comparing the effect of centering on shrinkage, 17 of the 32 comparisons show larger shrinkage with group-mean centering than with grand-mean centering. There is no apparent trend as to the conditions that are related to the larger shrinkage for group-mean centering with this R^2 approach. Of the 17 instances where the group-mean demonstrates larger shrinkage, 10 of those occur with a predictor-criterion correlation of $\rho_{xy} = .3$, $\rho_{wy} = .4$; no sample size trend is evident. In 22 instances, the shrinkage estimate between the

two approaches differed by less than five percentage points, and in 12 cases shrinkage differed by less than 2.

Additionally, there are no substantial differences between the shrinkage estimates under the two different predictor criterion correlations. With both predictor-criterion correlations, shrinkage can range from 100 percentage points between the correlations to less than one percentage point. There are a few exceptions, but for the majority of the instances, the predictor-criterion correlations differ by 10 or fewer percentage points. The interaction of predictor-criterion correlation with sample size, as already discussed, impacts shrinkage.

Snijders and Bosker's (1994) R_2^2

Using Snijders and Bosker's (1994) level-specific approach to R^2, R_{SB}^2, shrinkage estimates were smaller within level-1 size conditions when level-2 sample sizes were larger. Shrinkage is also less with larger level-1 sample sizes for a given level-2 sample size (see Table 8). There is one instance with R_{SB}^2 where a negative shrinkage is produced. Negative shrinkage indicates that the coefficients predicted the outcome better in the validation sample than in the sample from which they were estimated. The negative shrinkage occurs in the level-1, level-2 size combination of 10/10, under group-mean centering with predictor-criterion correlations of $\rho_{xy} = .4$, $\rho_{wy} = .3$.

Table 8

Shrinkage Estimates for R^2_{SB} by Total Sample Size

Sample Size			Shrinkage %		Correlations	
Total	L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}
100	10	10	364.20	521.81	0.4	0.3
150	15	10	206.46	245.56	0.4	0.3
150	10	15	238.45	325.69	0.4	0.3
200	20	10	141.82	155.63	0.4	0.3
200	10	20	203.90	273.22	0.4	0.3
225	15	15	137.72	162.91	0.4	0.3
300	20	15	99.90	101.25	0.4	0.3
300	15	20	113.48	132.05	0.4	0.3
300	30	10	112.39	107.03	0.4	0.3
300	10	30	174.08	220.70	0.4	0.3
400	20	20	86.91	91.71	0.4	0.3
450	30	15	72.08	70.36	0.4	0.3
450	15	30	99.54	114.80	0.4	0.3
600	30	20	54.40	55.25	0.4	0.3
600	20	30	71.69	77.86	0.4	0.3
900	30	30	46.74	50.74	0.4	0.3
100	10	10	4005.87	-1729.82	0.3	0.4
150	15	10	1016.05	2616.67	0.3	0.4
150	10	15	1058.71	10643.32	0.3	0.4
200	20	10	747.13	1452.21	0.3	0.4
200	10	20	680.28	4359.51	0.3	0.4
225	15	15	552.58	1146.06	0.3	0.4
300	20	15	440.74	696.06	0.3	0.4
300	15	20	446.82	1022.80	0.3	0.4
300	30	10	482.74	656.11	0.3	0.4
300	10	30	601.26	2359.40	0.3	0.4
400	20	20	381.00	566.33	0.3	0.4
450	30	15	351.58	484.11	0.3	0.4
450	15	30	360.09	713.54	0.3	0.4
600	30	20	290.03	382.56	0.3	0.4
600	20	30	296.63	451.79	0.3	0.4
900	30	30	250.69	325.98	0.3	0.4

In 15 of the 16 instances, the greatest decrease in shrinkage occurs when the level-2 sample size is increased from 10 to 15 for a given level-1 sample size. The magnitude of the shrinkage decrease when moving to the each higher level-2 sample size fluctuates under grand-mean centering (see Table 9). With increases of the level-2 sample size from 10 to 15 and 20 to 30, shrinkage decreases were greater than the shrinkage decrease in moving from 15 to 20. When the data are group-mean centered, the magnitude of the change in shrinkage within level-1 combinations was reduced with each increase in level-2 sample size. The exception to this is the level-1 sample size of 20 where both combinations display the same fluctuating pattern as grand-mean centering.

Within level-2 combinations, it is apparent that with each increase of level-1 sample size, shrinkage decreases. The greatest decrease was found in moving from a level-1 size of 10 to 15 in nearly all the grand-mean centering combinations, but only the case for the group-mean centered combinations when the level-2 size was 20 or 30. The magnitude of the decrease in shrinkage when moving to the next higher level-1 size was rarely orderly. Under grand-mean centering the large decrease when moving from a level-1 size of 10 to 15, was followed by a less substantial decrease in shrinkage when moving from a 15 to a 20. In moving from a 20 to a 30 level-1 sample size, however, the decrease is much more substantial than the 15 to 20 increase. Under group-mean centering, a pattern of magnitude change with each level-1 increase is not apparent.

Table 9

% Change in Shrinkage Estimates for R^2_{SB}

Sample Size		% Change across n_{L2}		Correlations		Sample Size		% Change across n_{L1}	
L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}	L2	L1	Grand	Group
10	10			0.4	0.3	10	10		
	15	13.0	11.1	0.4	0.3		15	24.9	24.5
	20	6.7	8.6	0.4	0.3		20	23.0	25.3
	30	10.9	8.2	0.4	0.3		30	6.7	17.6
	10			0.3	0.4		10		
	15	15.6	12.4	0.3	0.4		15	15.9	9.0
	20	6.1	7.7	0.3	0.4		20	6.2	8.9
	30	10.3	6.2	0.3	0.4		30	13.6	15.2
15	10			0.4	0.3	15	10		
	15	21.7	17.8	0.4	0.3		15	32.4	30.1
	20	12.0	16.4	0.4	0.3		20	17.8	32.7
	30	8.8	7.7	0.4	0.3		30	20.4	17.9
	10			0.3	0.4		10		
	15	14.0	19.2	0.3	0.4		15	14.3	16.1
	20	8.2	12.1	0.3	0.4		20	9.7	16.2
	30	15.8	4.6	0.3	0.4		30	17.3	6.1
20	10			0.4	0.3	20	10		
	15	16.4	26.0	0.4	0.3		15	36.2	36.0
	20	11.1	0.2	0.4	0.3		20	17.0	19.7
	30	14.9	13.4	0.4	0.3		30	33.7	34.5
	10			0.3	0.4		10		
	15	17.2	25.7	0.3	0.4		15	16.2	20.1
	20	11.5	6.1	0.3	0.4		20	13.0	10.5
	30	12.5	7.9	0.3	0.4		30	10.5	12.0
30	10			0.4	0.3	30	10		
	15	28.7	26.2	0.4	0.3		15	34.7	35.7
	20	26.0	20.4	0.4	0.3		20	22.5	24.6
	30	9.9	4.0	0.4	0.3		30	29.9	27.4
	10			0.3	0.4		10		
	15	20.8	17.8	0.3	0.4		15	21.4	18.7
	20	4.3	12.0	0.3	0.4		20	9.6	13.6
	30	10.3	5.7	0.3	0.4		30	8.3	9.9

Total sample size, under R^2_{SB}, impacts shrinkage such that as total sample size increases, shrinkage decreases. This holds true within the centering and predictor-criterion correlation condition combinations. Varying the predictor-criterion correlation between $\rho_{xy} = .3$, $\rho_{wy} = .4$ and $\rho_{xy} = .4$, $\rho_{wy} = .3$ results in vastly different magnitudes of shrinkage. The shrinkage for sample sizes and centering conditions within the predictor-criterion correlation of $\rho_{xy} = .4$, $\rho_{wy} = .3$ is substantially larger than shrinkage under the other predictor-criterion correlation condition.

While shrinkage decreases as total sample size increases, the degree of shrinkage differs based on the level-1 and level-2 allocation of the total sample size. There were 24 sample size combinations in which the total sample size was equal to that in at least one other level-1, level-2 combination. In 23 of the 24 level-1 and level-2 combinations, shrinkage for a given total sample is larger when level-2 sample size is larger (see Table 8). With a total sample size of 300, in three of the four predictor-criterion correlation and centering combinations, shrinkage initially decreased when the level-2 sample size changes from 10 to 15, but then increases as the level-2 changes from 15 to 20 and 20 to 30. In the fourth of the four combinations for 300, the group-mean centered $\rho_{xy} = .4$, $\rho_{wy} = .3$ condition reflects a shrinkage that increases with a level-2 size change of 10 to 15 and continues to climb with each increase of the level-2 sample size.

As is clear from Table 8, the centering approach employed also impacts shrinkage. In 28 of the 32 centering comparisons, shrinkage under the group-mean centering is larger than under grand-mean centering. The predictor-criterion correlation of $\rho_{xy} = .4$, $\rho_{wy} = .3$ reflects 14 of the 16 group-mean shrinkage estimates as larger than the grand-mean. The two exceptions are the total sample size of 100 and the total sample

size of 150 with the level-2 sample size of 10. There was not another common condition to explain the two exceptions to the larger group-mean centering trend in the $\rho_{xy} = .3$, $\rho_{wy} = .4$ condition. Seven of the grand- and group-mean centering estimates differ by less than 10 percentage points and those instances are found in the $\rho_{xy} = .4$, $\rho_{wy} = .3$ condition with large total sample sizes.

Shrinkage estimates under different predictor-criterion correlations conditions vary by extreme amounts (see Table 9). The difference between the estimates ranges from 203 to 10,000 percentage points with most differing by several hundred points. The combination $\rho_{xy} = .4$, $\rho_{wy} = .3$ in all but one instance is associated with larger shrinkage estimates than $\rho_{xy} = .3$, $\rho_{wy} = .4$.

Gagné and Furlow's (2009) R^2

With few exceptions, for a given level-1 sample size, shrinkage under R^2_{GF} is smaller when level-2 sample sizes are larger. There are five exceptions when a higher level-2 sample size is associated with a larger amount of shrinkage (see Table 10). These exceptions are found within both grand- and group mean centering conditions and within both predictor-criterion correlations, but three of the five occur with a level-1 sample size 30. In comparing shrinkage across level-1 sample sizes, larger level-1 sample sizes are, in general, associated with smaller shrinkage estimates. There are, however, more exceptions when comparing across level-1 sizes as opposed to above where five exceptions were found. There are 16 instances when a larger level-1 sample size is associated with a larger shrinkage estimate. There does not appear to be a pattern as to the conditions associated with the instances of increased shrinkage.

Table 10

Shrinkage Estimates for R^2_{GF} by Total Sample Size

Sample Size			Shrinkage %		Correlations	
Total	L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}
100	10	10	18.41	18.65	0.4	0.3
150	15	10	17.54	16.12	0.4	0.3
150	10	15	11.15	12.41	0.4	0.3
200	20	10	14.48	15.83	0.4	0.3
200	10	20	10.42	9.44	0.4	0.3
225	15	15	12.08	10.91	0.4	0.3
300	20	15	9.99	8.76	0.4	0.3
300	15	20	8.98	7.94	0.4	0.3
300	30	10	15.73	14.56	0.4	0.3
300	10	30	6.72	6.72	0.4	0.3
400	20	20	9.21	9.86	0.4	0.3
450	30	15	11.15	10.12	0.4	0.3
450	15	30	5.76	5.88	0.4	0.3
600	30	20	6.14	6.46	0.4	0.3
600	20	30	4.72	5.83	0.4	0.3
900	30	30	6.48	5.82	0.4	0.3
100	10	10	30.29	27.87	0.3	0.4
150	15	10	27.53	29.20	0.3	0.4
150	10	15	26.06	26.86	0.3	0.4
200	20	10	29.81	27.69	0.3	0.4
200	10	20	24.70	24.46	0.3	0.4
225	15	15	25.33	25.58	0.3	0.4
300	20	15	26.38	24.58	0.3	0.4
300	15	20	25.66	23.69	0.3	0.4
300	30	10	27.09	27.41	0.3	0.4
300	10	30	22.01	23.19	0.3	0.4
400	20	20	24.06	24.15	0.3	0.4
450	30	15	23.54	23.61	0.3	0.4
450	15	30	21.14	23.06	0.3	0.4
600	30	20	23.87	24.00	0.3	0.4
600	20	30	21.73	21.68	0.3	0.4
900	30	30	22.03	22.21	0.3	0.4

For a given level-1 sample size, over half, 10 of the 16 combinations, show the largest decreases in shrinkage occurs when moving from a level-2 sample size of 10 to 15. The pattern of change for a given level-1 sample size fluctuates in both direction and degree as level-2 sample size increases (see Table 11). Holding level-2 sample size constant and examining the magnitude and direction of the change in shrinkage with increases in level-1 sample size reveals that the magnitude of the change is much smaller in general than that found across level-2 samples sizes holding level-1 sample size constant. Under R^2_{GF}, the largest change in shrinkage for a given level-2 sample size is evenly distributed between the three level-1 sample size changes with 6 occurrences associated with the increase from a level-1 sample size of 10 to 15, and 5 occurrences when moving from 15 to 20 and 5 when increasing from 20 to 30.

Under R^2_{GF}, level-2 sample size favorably impacts the magnitude of shrinkage within a given total sample size. The larger the level-2 sample size allocation of the total size, the less shrinkage occurs. The most comprehensive example of this is within the total sample size of 300 with four level-1, level-2 size combinations. Each increase of level-2 sample size for a total size of three hundred is consistently met with a decrease in shrinkage.

Table 11

% Change in Shrinkage Estimates for R^2_{GF}

Sample Size		% Change across n_{L2}		Correlations		Sample Size		% Change across n_{L1}	
L1	L2	Grand	Group	ρ_{WY}	ρ_{XY}	L2	L1	Grand	Group
10	10			0.4	0.3	10	10		
	15	41.5	35.9	0.4	0.3		15	6.1	15.8
	20	5.6	24.3	0.4	0.3		20	18.1	0.2
	30	36.7	28.8	0.4	0.3		30	-7.3	9.4
	10			0.3	0.4		10		
	15	18.4	5.3	0.3	0.4		15	12.0	-5.8
	20	4.9	10.1	0.3	0.4		20	-8.7	6.2
	30	13.1	6.6	0.3	0.4		30	10.1	0.2
15	10			0.4	0.3	15	10		
	15	32.0	32.5	0.4	0.3		15	-9.2	11.3
	20	26.1	28.4	0.4	0.3		20	18.2	20.1
	30	36.9	25.8	0.4	0.3		30	-12.7	-16.0
	10			0.3	0.4		10		
	15	9.2	15.5	0.3	0.4		15	2.1	5.7
	20	-1.0	8.5	0.3	0.4		20	-4.4	3.3
	30	19.4	3.0	0.3	0.4		30	12.2	4.1
20	10			0.4	0.3	20	10		
	15	32.0	46.0	0.4	0.3		15	14.5	16.0
	20	7.7	-13.3	0.4	0.3		20	-2.2	-26.4
	30	49.9	41.7	0.4	0.3		30	34.7	35.6
	10			0.3	0.4		10		
	15	12.8	13.0	0.3	0.4		15	-4.0	3.9
	20	10.3	2.5	0.3	0.4		20	7.3	-3.0
	30	10.4	11.2	0.3	0.4		30	0.4	0.3
30	10			0.4	0.3	30	10		
	15	28.6	30.8	0.4	0.3		15	14.9	12.5
	20	46.5	37.1	0.4	0.3		20	18.7	0.7
	30	-6.8	9.3	0.4	0.3		30	-39.2	-0.2
	10			0.3	0.4		10		
	15	14.9	16.4	0.3	0.4		15	3.5	0.2
	20	-1.8	-1.4	0.3	0.4		20	-2.9	5.7
	30	9.1	8.2	0.3	0.4		30	-1.2	-3.1

Shrinkage differs little under grand-mean and group-mean centering (see Table 10). In comparing across the 32 grand- and group-centered instances, 50% of the comparisons reflect the group-mean shrinkage estimate larger than the grand-mean estimate. Of the 32 grand-/group-mean centered comparisons, 30 of the pairs have shrinkage estimates that differ by less than two percentage points, with many showing a less than a one percentage point difference. The two instances that differed by more than two percentage points, differed by less than three percentage points.

In viewing shrinkage by predictor-criterion correlations with R^2_{GF}, the difference in the magnitude of shrinkage between the $\rho_{xy} = .3$, $\rho_{wy} = .4$ and $\rho_{xy} = .4$, $\rho_{wy} = .3$ is immediately apparent (see Table 10). The smallest shrinkage magnitude under the $\rho_{xy} = .4$, $\rho_{wy} = .3$ condition is larger than the largest shrinkage estimate under the $\rho_{xy} = .3$, $\rho_{wy} = .4$ predictor-criterion correlation condition. All of the instances where the magnitude of shrinkage is less than 10% occur under the predictor-criterion correlation condition of $\rho_{xy} = .3$, $\rho_{wy} = .4$ with larger level-2 sample sizes. The five largest shrinkage magnitudes occurred in the predictor-criterion correlation condition of $\rho_{xy} = .4$, $\rho_{wy} = .3$. The difference between shrinkage estimates under the two predictor-criterion correlations range from 9.22 to 17.73 percentage points.

Comparison ofR^2 Approaches

The impact of sample size was generally consistent across the R^2 approaches. With a handful of exceptions, as level-2 sample size increases or as level-1 sample size increases shrinkage became smaller. Shrinkage in comparison to changes in total sample size was also consistent across R^2 approaches. As total sample size increases, shrinkage decreases. The allocation of the total sample size to level-1 and level-2 samples sizes did

produce differing results from the approaches. With R^2_{RB0}, R^2_{RB1}, and R^2_{SB}, shrinkage was less for a given total sample size when the level-1 sample size was greater than the level-2 sample size. As the exception, under R^2_{GF}, shrinkage was less when the level-2 sample size was greater than the level-1 sample size.

The pattern of change when moving across level-1 or level-2 sample sizes is different between R^2 approaches and also varies with centering. The increase from a sample size of 10 to 15 has the largest impact on shrinkage for R^2_{RB0} and R^2_{RB1}. That change in sample size also has the largest impact on R^2_{SB} for increases in level-2 sample size, and under grand-mean centering when level-1 sample size changes. R^2_{GF} differs from the other approaches in that under grand-mean centering for a given level-1 sample size, only half the combinations reflect the change from a level-2 sample size of 10 to 15 as the largest change and six of the eight combinations under group-mean centering reflect the 10 to 15 change as the largest change. Across level-1 sample sizes for a given level-2 sample size, 6 of 16 combinations under R^2_{GF} show the change from a level-1 sample size of 10 to 15 is the largest shrinkage change.

The magnitude of the change with each increase in sample size within the combinations also differs based on R^2 approach and centering. In some cases, with each increase in sample size, the change in shrinkage decreases in an ordered pattern. With R^2_{SB}, under group-mean centering for a given level-1 sample size, six of the eight instances display a decreasing, orderly pattern to the change in shrinkage for increases in level-2 sample size. R^2_{RB0} also displays a similar orderly pattern under the same conditions. The grand-mean conditions for both R^2_{RB0} and R^2_{SB} have a pattern of change that fluctuates between larger and smaller degrees of decrease. On the other hand, for a

fixed level-1 sample size, half of the 16 occurrences with R^2_{RB1} demonstrate an orderly change pattern, and R^2_{GF} have only three instances with a fixed level-1 sample size where increases in level-2 result in an orderly change pattern. None of the R^2 approaches, under group or grand-mean-centering, display an orderly pattern of change across level-1 sample size increases for a given level-2 sample size.

In comparison to other approaches, R^2_{GF} displays the most occurrences of negative shrinkage change within sample size combinations. Across level-2 sample sizes, there are five instances where the magnitude of shrinkage increases when moving to a larger sample size. Across level-1 sample sizes there are 14 negative shrinkage changes. In comparison, R^2_{RB0} contains one negative shrinkage change for across level-1 sample size changes and one across level-2 changes. With R^2_{RB1}, there is one occurrence of negative shrinkage change across level-1 sample size and three negative changes across the level-2 sample size changes. Negative shrinkage changes were not found with R^2_{SB}.

The impact of centering and predictor-criterion correlations varies depending on the R^2 approach. Under R^2_{GF} and R^2_{RB1}, the differences between grand-mean and group-mean centering were negligible in most cases. Group-mean centering under R^2_{SB} is associated with larger shrinkage estimates whereas smaller shrinkage estimates are found with group-mean centering under R^2_{RB0}. The impact of predictor-criterion correlations is more consistent. While the range is large, most of the differences between the predictor-criterion correlations of $\rho_{xy} = .4$, $\rho_{wy} = .3$ and of $\rho_{xy} = .3$, $\rho_{wy} = .4$ are relatively close under R^2_{RB1}. With R^2_{RB0}, R^2_{SB}, and R^2_{GF}, however, the difference is clear that the correlation combination of $\rho_{xy} = .4$, $\rho_{wy} = .3$ results in higher shrinkage estimates for the three R^2 approaches.

CHAPTER 5

DISCUSSION

The purpose of this study was to explore the impact of sample size, predictor-criterion correlations, centering, and different forms of explained variance, R^2, on cross-validation shrinkage. This chapter provides a summary of the findings related to each of the four factors investigated. Limitations of the study as well as the implications and thoughts for future research will be presented.

In considering sample size, it is important to remember R^2 measures have not previously been used for cross-validating HLM, therefore, the only sample size research to draw from is that related to cross-validating regression models and research related to the estimation of HLM parameters. In general, it seems that the common finding is that larger sample sizes at level-2 are more important than the level-1 sample size for parameter estimation with HLM. This study, however, does not find that to be entirely applicable when cross-validating HLM, via the R^2 approaches used. This study supports a conclusion that shrinkage decreases with both larger level-2 sample sizes and level-1 sample sizes, holding other conditions constant. Total sample size, and level-1 sample size in comparison to level-2 sample size is also important. It is also the case that the larger the total sample size, the less shrinkage will occur. In three of the four approaches, however, shrinkage was typically less for a given total sample size when the level-1 sample size was larger than the level-2. This is contrary to what would be expected given the emphasis placed on the importance of level-2 sample size for parameter estimation

and statistical power considerations (e.g., Hox, 2002; Kreft, 2007; Mok, 1995; Van der Leeden, Busing, & Meijer, 1997). The only approach that conforms to the importance of larger level-2 sample size within a given total sample size is the Gagné and Furlow (2009) approach.

With most level-2 sample size recommendations ranging from 100 to thousands of groups (Hox, 2002; Kreft, 2007; Mok, 1995; Van der Leeden, Busing, & Meijer, 1997), Brown and Draper (2000) suggest that having 6 – 12 groups allows for acceptable variance estimates to be achieved. The results from this study indicate that even though Brown and Draper suggest acceptable variance can be obtained with fewer than 10 groups, shrinkage estimates obtained from a group size of 10 are rarely acceptable.

Interestingly, shrinkage with the R^2_{GF} approach somewhat mimics St. John and Roth's (1999) findings. St. John and Roth looked at sample size and regression cross-validation shrinkage of published research. Their findings indicate that in cross-validating regression with samples sizes of less than 90 and four or fewer predictors, the R^2 can be expected to shrink 33%. With samples of 250 or more and four or fewer predictors, shrinkage can be expected to be approximately 7%. In this study, the R^2_{GF} shrinkages with the predictor-criterion correlation of ρ_{xy} = .3, ρ_{wy} = .4 and a total sample size of 300 or more ranges from 4.72% to 11.15%, excluding the 300 total sample size with a group size of 10. This study does not have a sample size of 90 for comparison purposes, but the smallest sample size is 100 and the largest shrinkage at 100 with a ρ_{xy} = .4, ρ_{wy} = .3 is 30.29% with the R^2_{GF}.

The impact of predictor-criterion correlations is easy to see in this study. In general, shrinkage across sample sizes is smaller when the relationship of the level-2

variable with the outcome variable is slightly stronger than the relationship of the level-1 variable with the outcome variable. Specifically, shrinkage was less with three of the four R^2 shrinkage approaches when $\rho_{xy} = .3$, $\rho_{wy} = .4$. It is, however, not possible to put these results in relation to previous research as no cross-validation research has employed predictor-criterion correlations as a condition in the study.

The impact of centering in cross-validating HLM is a quandary. Centering was included in this study because centering is frequently employed in applied studies and variance estimates under grand-mean centering are likely negatively biased. This was not the case. Shrinkage estimates derived from group-mean centered data were smaller than those from grand-mean centered data with the R^2_{RB0}. Conversely, results from R^2_{SB} indicate shrinkage is larger with group-mean centered data. The remaining two approaches, R^2_{GF} and R^2_{RB1}, point to no shrinkage differences between grand- and group-mean centered data. One would presume that even if the R^2 measures differed due to centering that the difference would have a consistent impact across the approaches. At this time, explaining these contradictory findings between approaches is challenging. It is possible that the explanation for the centering differences is simply that some R^2measures do not lend themselves to cross-validation.

Shrinkage has to fall within a reasonable range in order for a researcher to determine that the application of the model to other samples of the population is appropriate. A known downside of using the R^2_{RB0} and R^2_{RB1}as well as R^2_{SB} is the possibility of obtaining negative R^2s. When calculating shrinkage, the validation R^2_{cv} is subtracted from the estimation R^2, $R^2 - R^2_{cv}$. The negative R^2_{cv}s in this study produced a sizable shrinkage magnitude as a negative R^2_{cv} is added to the estimation R^2 and that sum

is then divided by R^2 resulting in many cases of shrinkage over 100% of the estimation R^2. This means that the explained variance has decreased so much in going from the estimation sample to the validation sample that whatever variance was explained before is no longer explained.

The degree to which negative R^2s occurred in this study is surprising, especially with the Snijders and Bosker's approach since they formulated their R^2 approach to avoid negative R^2 values. While their R^2 was not formulated to avoid negative cross-validation coefficients, conceptually, it should be the same as avoiding negative R^2s. Even though explained variance should less with R^2_{cv}, negative explained variance is hard to comprehend whether within cross-validation or estimation. A negative R^2 is equivalent to decreasing the explained variance by adding predictor variables to the model. This never happens with the regression R^2 or R^2_{cv} measures. Gagné and Furlow's (2009) R^2 avoids this pitfall of negative R^2s by using an explained variance approach identical to regression where explained variance is an additive process with the addition of each variable contributing to the variance explained.

Given that all shrinkage measurements within each condition were calculated based on the same data, the usefulness of the models is apparent. In attempting to determine the degree to which a model works in a second sample from a population, shrinkage figures such as 80% and 220% are difficult to comprehend, especially in light of shrinkage measures for the same data that are 10% or 27%. As mentioned, the R^2_{GF} approach provided what appear to be useful estimates. While R^2_{RB1} provides shrinkage estimates more tangible and meaningful than the remaining two R^2 approaches, it provides information only about how well one piece of the model cross-validates and

says nothing about the entire model's applicability. This is not to say that knowing the shrinkage magnitude for one piece of the model is not useful. In some situations and with some research questions, it might be the most valuable information, but this study investigated cross-validating HLM with the applied researcher in mind, and it seems likely that the applicability of the entire model is more likely to be of interest.

The fact that this study in no way attempts to determine how to use the pieces of the component or level specific approaches to R^2 to arrive at an overall assessment of how well a model cross-validates could be viewed as a limitation of the study. As just mentioned, with the applied researcher in mind, the interest is likely on the applicability of the whole model. Trying to do so, however, would be premature as until this point, it was unknown how the separate pieces would cross-validate. As HLM grows in application it seems that the issue of model fit will increasingly become a question that will need to be resolved. As with all simulation studies, there are always more conditions that could be examined and additional comparisons that could be made, but the constraints of time and resources always exist and limit what can be done. Several of these constrained curiosities are suggestions for future research.

Given the constraints, this study does provide some groundwork for future researchers and some guidance for applied researchers interested in cross-validating HLM results. The overriding implication is that the Raudenbush and Bryk (2002) and Snijders and Bosker (1994) R^2 approaches have limited usefulness when attempting to cross-validate. The unpredictably high rate of negative cross-validity coefficients makes the usefulness of the metric questionable. If the researcher is particularly interested in the explained variance solely at level-2 or with a level-2 component, and opts to cross-

validate with R^2_{RB0}, R^2_{RB1}, and R^2_{SB}, the implication of sample size, predictor-criterion correlations, and centering need to be kept in mind as informed research design decisions can make the best out of a less than ideal approach.

For any of the approaches investigated, as sample size increases, shrinkage will decrease, so the advice to select a larger sample size when feasible holds in cross-validating HLM. In distributing the total sample size across level-1 and level-2, recommendations about the importance of a larger level-2 than level-1 does not consistently apply when cross-validating. Only when employing the Gagné and Furlow (2009) R^2 approach does a larger level-2 consistently result in less shrinkage. With the other approaches, a larger level-1 appears to be more important. At the same time, in many of the conditions considered, increasing a level-1 or level-2 sample size while holding the other level constant does not always result in a substantial decrease in shrinkage. This is true especially of grand-mean centering. Increasing the level-2 size, from 10 to 15 or 20 to 30 may very well be worth the cost in terms of shrinkage decreased, but increasing from 15 to 20 may not be worth the shrinkage reduction.

The applied researcher also has to make decisions about centering. This study indicates that when cross-validating, there is an intriguing relationship between centering, R^2 approach, and predictor-criterion correlations. Until more is understood about this relationship, it would serve the applied researcher to keep in mind when opting for a centering option that the R^2 method employed, combined with centering can impact the magnitude of shrinkage. A predictor-criterion correlation that is stronger for the level-2 variable than the level-1 variable ($\rho_{xy} = .3$, $\rho_{wy} = .4$) instead of the reverse ($\rho_{xy} = .4$, $\rho_{wy} =$

.3) may mitigate a portion of the shrinkage differences between the two shrinkage approaches somewhat when differences exist.

A follow-up study to investigate the relationship between centering approach, predictor correlations, and R^2 may be useful as the results in this study were unexpected, and therefore seems to warrant additional research. While it is logical that variance would be different given a centering approach, one would think the difference created by the centering approach would carry forward consistently to each of the R^2 approaches. This area of exploration, however, is a secondary priority in terms of future research. Given the importance of cross-validating, model fit questions are more pressing.

The results of this study suggest that using the two most commonly accepted measures of model fit, Raudenbush and Bryk (2002) and Snijders and Bosker (1994), are of questionable use in cross-validating when the applicability of level specific and component specific measures of model fit are of interest and of no use at this point when questions about the applicability of the entire model are in question. Future research exploring and developing model fit measures for the entire model would be beneficial. Gagné and Furlow (2009) have proposed one approach that was employed in this study, but other approaches exist and still others could be developed and all opened to critique in order for a commonly agreed upon measure to emerge. Some methodologists may disagree with this suggestion and argue that the strength of HLM is in the ability to partition variance by the level, and they are correct. At the same time, quantifying an overall model measure of explained variance does not diminish the added value of HLM in being able to partition the variance over multiple levels.

Future research could also involve additional design characteristics that would explore the shrinkage impact of unbalanced data sets as it is rare that an applied researcher gets to work with balanced data. Additional exploration of larger sample sizes and sample increases in consistent increments at level-1 and level-2 would provide insight into how much reduction in shrinkage is gained at different points and under different conditions for minimal increases in size. Differing levels of sample size in the validation and estimation samples under different conditions would aid in determining effective methods of data splitting since more data is better in an HLM analysis. This study conditioned on predictor-criterion correlations; adding multicollinearity conditions might be useful given the work of applied researchers and the potential consequences to shrinkage of interpredictor correlations. Finally comparing shrinkage with different variance estimation procedures would be beneficial in that different software packages use different estimation procedures and it is likely that this to may impact cross-validation.

References

Afshartous, D. (1995, April). Determination of sample size for multilevel model design. Paper presented at the meeting of the American Educational Research Association, San Francisco, Ca. Retrieved July 7, 2010 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.4318

Afshartous, D., & de Leeuw, J. (2004). An application of multilevel model prediction to NELS:88. *Behaviormetrika*,*31*(1), 43-66.

Algina, J., & Keselman, H. J. (2000). Cross-validation sample sizes. *Applied Psychological Measurement*, *24*, 173-179.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79.

Astin, A. W., & Denson, N. (2009). Multi-campus studies of college impact: Which statistical method is appropriate? *Research in Higher Education*, *50*, 354-367.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology, 28*, 79-87.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108-132.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*,

391-420.

Cattin, P. (1980a). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65*, 407-414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of a regression model. *Psychological Bulletin, 87*, 63-65.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2*, 595-607.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. *Journal of Marketing Research, 24*, 271-279.

Cotter, K. L., & Raju, N. S. (1982). An evaluation of formula-based population squared cross-validity estimates and factor score estimates in prediction. *Educational and Psychological Measurement, 42*, 493-519.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10*, 94-96.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. *Applied Psychological Measurement, 3*, 387-399.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121-138.

Gagné, P., & Furlow, C. (2009). *Applying multiple regression's R^2 to hierarchical linear models*. Unpublished manuscript.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70*, 320-328.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499-510.

Hox, J. (1998). Multilevel modeling: When & why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York, NY: Springer Verlag. Retrieved August 15, 2010 from http://joophox.net/publist/whenwhy.pdf

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. London: Lawrence Erlbaum.

Kreft, I. G. G., & de Leeuw, J. (2007). *Introducing multilevel modeling.* London: Sage.

Kromrey, J. D., & Hines, C. V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. *Journal of Experimental Education, 64*, 240-267.

Kurtz, A. K. (1948). A research test of the Rorschach Test. *Personnel Psychology: A Journal of Applied Research, 1*, 41-51.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology, 22*, 45-55.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

Mitchell, T. W., & Klimoski, R. J. (1986). Estimating the validity of cross-validity estimation. *Journal of Applied Psychology, 71*, 311-317.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5-11.

Mox, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modeling Newsletter*, *7*(2), 11-15.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology*, *36*(1), 111-118.

Murphy, K. R. (1984). Cost-benefit considerations in choosing among cross-validation methods. *Personnel Psychology*, *37*(1), 15-22.

Osborne, J. W. (2000). Prediction in multiple regression. *Practical Assessment, Research, & Evaluation*, *7*(2). Retrieved October 17, 2009 from http://PAREonline.net/getvn.asp?v=1&n=2

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights. *Applied Psychological Measurement, 23(2)*, 99-115.

Raudenbush, D. S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). London: Sage.

Roberts, J. K., & Monaco, J. P. (2006, April). Effect size measures for the two-level linear multilevel model. Presented at the American Educational Research Association. Retrieved from http://www.hlm-online.com/papers/HLM_effect_size.pdf

Schmitt, N. (1982). Formula estimation of cross-validated multiple correlation. (ERIC Document Reproduction No. 227 137)

Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics, 19,* 415-428.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research, 22,* 342-363.

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

St. John, C. H., & Roth, P. L. (1999). The impact of cross-validation adjustments on estimates of effect size in business policy and strategy research. *Organizational Research Methods, 2,* 157-174.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 111-147.

van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1997, April). Bootstrap methods for two-level models. Paper presented at the Multilevel Conference, Amsterdam. Retrieved May 23, 2010 from http://www.soziologie.uni-halle.de/langer/buecher/mehrebenen/literatur/busing1997.pdf

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education, 69,* 203-224.

APPENDIXES

APPENDIX A

Grand-mean Centered SAS Code

```
%let corrwx=0;
%MACRO RUNHLM;
%let seed1=10050001; %let seed2=20050001;
%let groups=40; /* Level-2 n */
%let n_per=10; /* Level-1 n */
%let t00=61.9285714285714; %let t11=0.24190848214286;
%let corrwy=.4;
%let corrxy=.3;
%let corrwxy=.3;
%let reps=3000;
%let goal=1000;
%let count=0;
%let tally=0;

proc datasets library=work;
  delete squared_stuff;
run;
PROC PRINTTO FILE='C:\dissertation\HLMOUT.TXT' NEW;
PROC PRINTTO LOG='C:\dissertation\HLM_05.TXT' NEW;

%do rep=1 %to &reps;

proc iml;
preds=3;
data_err=repeat(0,&groups*&n_per,7);
pred_sd=repeat(0,preds,preds);
gammas=repeat(0,1+preds,1);
errorvar=289;
poppredc={1.00 &corrwx 0.00,
&corrwx 1.00 0.00,
              0.00 0.00 1.00};
corrwthy={&corrwy, &corrxy, &corrwxy};
standgam=inv(poppredc)*corrwthy;
mu={90, 0, 0, 0};
pred_var={441, 256, 0};
  pred_var[3,1]=pred_var[1,1]*pred_var[2,1]*(1 + poppredc[2,1]**2);
```

```
y_var=(&t00+&t11*pred_var[2,1]+errorvar)/(1-standgam`*poppredc*standgam);
y_sd=sqrt(y_var);

do p=1 to preds;
  pred_sd[p,p]=sqrt(pred_var[p,1]);
  gammas[p+1,1]=standgam[p,1]*y_sd/pred_sd[p,p];
end;

prdcdata=repeat(0,&groups*&n_per,3); /* w x wx */
grpdata=normal(repeat(&seed1,&groups,3)); /* w u0 u1 */
  grpdata[,2]=sqrt(&t00)*grpdata[,2];
  grpdata[,3]=sqrt(&t11)*grpdata[,3];

persdata=normal(repeat(&seed2,&groups*&n_per,2)); /* x rij */
  persdata[,2]=sqrt(errorvar)*persdata[,2];

varcovar=pred_sd*poppredc*pred_sd;
  mu[4,1]=mu[2,1]*mu[3,1]+varcovar[2,1];
  gammas[1,1]=mu[1,1]-gammas[2,1]*mu[2,1]-gammas[3,1]*mu[3,1]-
gammas[4,1]*mu[4,1];
  cd=root(varcovar);

do i=1 to &groups;
  do j=1 to &n_per;
    prdcdata[(i-1)*&n_per+j,1]=grpdata[i,1]; /* w */
    prdcdata[(i-1)*&n_per+j,2]=persdata[(i-1)*&n_per+j,1]; /* x */
    prdcdata[(i-1)*&n_per+j,3]=grpdata[i,1]*persdata[(i-1)*&n_per+j,1]; /* wx */
    data_err[(i-1)*&n_per+j,4]=grpdata[i,2];  /* u0 */
    data_err[(i-1)*&n_per+j,5]=grpdata[i,3]; /* u1 */
    data_err[(i-1)*&n_per+j,6]=persdata[(i-1)*&n_per+j,2]; /* rij */
    data_err[(i-1)*&n_per+j,7]=i;
  end;
end;

prdcdata=prdcdata*cd;
prdcdata[,3]=prdcdata[,3]+mu[4,1];
data_err[,1]=prdcdata[,1]; data_err[,2]=prdcdata[,2]; data_err[,3]=prdcdata[,3];

gammas=gammas`;
create allgs from gammas;
append from gammas;

create combine from data_err;
append from data_err;
quit;
```

```
data d1; set allgs; keep g00 g01 g10 g11;
g00=col1; g01=col2; g10=col3; g11=col4;
run;

/*CREATE Y FOR EACH PARTICIPANT USING COMBINED MODEL*/
DATA CREAT_Y; SET COMBINE; if _N_=1 then set d1; drop col1-col6;
  w1=col1; x1=col2; wx=col3; u0=col4; u1=col5; rij=col6; site=col7;
  Y = G00 + G10*X1 + G01*W1 + G11*wx + u0 + u1*x1 + RIJ;
RUN;

/* divide generated data into two sets */
%let dsid=%sysfunc(open(work.CREAT_Y,in));
%let nobs=%sysfunc(attrn(&dsid,nobs));
%if &dsid > 0 %then %let rc=%sysfunc(close(&dsid));
data estim valid;
set CREAT_Y;
if _n_ le &nobs/2
then output estim;
else if _n_ GT &nobs/2
then output valid;
run;

/* Estimation unconditional*/
ods listing;
ods output covparms=est_Utau;
ods output solutionf=est_Ugam;

proc mixed data = ESTIM covtest;
class site;
model y=x1 /ddfm=residual solution cl notest;
random intercept x1 / sub=site type=un;
run;

data covs_UE;
set est_Utau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cue;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
```

```
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';
keep t00est t11est t10est var1est;

array se(4) t00se t11se t10se var1se;
worked_cue=4;
do i=1 to 4;
  if se(i)=. then worked_cue=worked_cue-1;
end;
call symput("worked_cue", worked_cue);
keep worked_cue;
run;

%if &worked_cue = 0 %then %let worked_pue=0;
%else %do;
data parms_UE;
set est_Ugam;
retain g00est g10est g00se g10se worked_pue;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='x1';

array se(2) g00se g10se;
worked_pue=2;
do i=1 to 2;
  if se(i)=. then worked_pue=worked_pue-1;
end;
call symput("worked_pue", worked_pue);
keep g00est g10est worked_pue;
run;
%end; /* worked_cue check */

/*  Estimation Conditional */
ods listing;
ods output covparms=est_Ctau;
ods output solutionf=est_Cgam;

proc mixed data = ESTIM covtest;
class site;
model y=x1 w1 wx/ddfm=residual solution cl notest;
random intercept x1 /sub=site type=un;
```

```
run;

data covs_CE;
set est_Ctau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cce;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';

array se(4) t00se t11se t10se var1se;
worked_cce=4;
do i=1 to 4;
  if se(i)=. then worked_cce=worked_cce-1;
end;
call symput('worked_cce', worked_cce);
keep t00est t11est t10est var1est worked_cce;
run;

%if &worked_cce = 0 %then %let worked_pce=0;
%else %do;
data parms_CE;
set est_Cgam;
retain g00est g10est g01est g11est g00se g10se g01se g11se worked_pce;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='w1' then do;
  g01est=estimate; g01se=stderr;
end;
if effect='wx' then do;
  g11est=estimate; g11se=stderr;
end;
if effect='wx';
```

```
array se(4) g00se g10se g01se g11se;
worked_pce=4;
do i=1 to 4;
  if se(i)=. then worked_pce=worked_pce-1;
end;
call symput('worked_pce', worked_pce);
keep g00est g10est g01est g11est worked_pce;
run;
%end; /* worked_cce check */

/*  VALIDATION Unconditional  */
ods listing;
ods output covparms=val_UVtau;
ods output solutionf=val_UVgam;

proc mixed data = valid covtest;
class site;
model y=x1 /ddfm=residual solution cl notest;
random intercept x1/sub=site type=un;
run;

data covs_UV;
set val_UVtau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cuv;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';
keep t00est t11est t10est var1est;
array se(4) t00se t11se t10se var1se;
worked_cuv=4;
do i=1 to 4;
  if se(i)=. then worked_cuv=worked_cuv-1;
end;
call symput("worked_cuv", worked_cuv);
keep worked_cuv;
```

```
run;

%if &worked_cuv = 0 %then %let worked_puv=0;
%else %do;
data parms_UV;
set val_UVgam;
retain g00est g10est g00se g10se worked_puv;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='x1';
array se(2) g00se g10se;
worked_puv=2;
do i=1 to 2;
  if se(i)=. then worked_puv=worked_puv-1;
end;
call symput("worked_puv", worked_puv);
keep g00est g10est worked_puv;
run;
%end; /* worked_puv check */

%let
tally=&worked_cue+&worked_pue+&worked_cce+&worked_pce+&worked_cuv+&wor
ked_puv;

%if &tally=20 %then %do;
%let count=&count+1;
data estim2iml; set estim;
drop col7 g00 g01 g10 g11 u0 u1 rij site;
run;

data valid2iml; set valid;
drop col7 g00 g01 g10 g11 u0 u1 rij site;
run;

proc iml;
use estim2iml; read all into est_data; * w x wx y;
use valid2iml; read all into val_data; * w x wx y;
use parms_ce; read all into gammas;
use covs_UE; read all into UE_vars;
use covs_CE; read all into CE_vars;
use covs_UV; read all into UV_vars;
```

```
n1=&n_per; n2=&groups/2; n=n1*n2;
gamma00=gammas[1,1]; gamma01=gammas[1,2];
gamma10=gammas[1,3]; gamma11=gammas[1,4];
est_vcov=repeat(0,4,4); est_sds=repeat(0,4,4);
val_vcov=repeat(0,4,4); val_sds=repeat(0,4,4);
crossval_taus=repeat(0,1,3); *tau00 tau11 tau10;
est_RB_Rsq=repeat(0,1,2); crossval_RB_Rsq=repeat(0,1,2);
right_bs=repeat(0,n2,2); pred_bs=repeat(0,n2,2); u=repeat(0,n2,2);
pred_ys=repeat(0,n1*n2,1);

esterr_b=repeat(0,n2,1);
esterr_w=repeat(0,n,1);
crosserr_b=repeat(0,n2,1);
crosserr_w=repeat(0,n,1);
est_ybar=repeat(0,n2,1);
val_ybar=repeat(0,n2,1);

storstuf=repeat(0,1,9);

do j=1 to n2;
  w=val_data[(j-1)*n1+1,1];
  y=val_data[(j-1)*n1+1:j*n1,4];
  val_x=val_data[(j-1)*n1+1:j*n1,2];
  x=repeat(1,n1,1)||val_x;
  right_bs[j,]=(inv(x`*x)*x`*y)`;
  pred_bs[j,1]=gamma00+gamma01*w;
  pred_bs[j,2]=gamma10+gamma11*w;
  u[j,1]=pred_bs[j,1]-right_bs[j,1];
  u[j,2]=pred_bs[j,2]-right_bs[j,2];
  pred_ys[(j-
1)*n1+1:j*n1,1]=gamma00+gamma01*w+gamma10*val_x+gamma11*w*val_x;

  est_y=est_data[(j-1)*n1+1:j*n1,4];
  est_x=est_data[(j-1)*n1+1:j*n1,2];
  est_ybar[j,1]=est_y[:];
  est_xbar=est_x[:];
  esterr_b[j,1]=est_ybar[j,1]-gamma10*est_xbar;
  esterr_w[(j-1)*n1+1:j*n1,1]=est_y-gamma10*est_x;
  val_ybar[j,1]=y[:];
  val_xbar=val_x[:];
  crosserr_b[j,1]=val_ybar[j,1]-gamma10*val_xbar;
  crosserr_w[(j-1)*n1+1:j*n1,1]=y-gamma10*val_x;
end;

/* Compute R² via behs for each sample */
do p=1 to 4;
```

```
  do q=1 to 4;
    est_vcov[p,q]=(est_data[,p]-est_data[:,p])`*(est_data[,q]-est_data[:,q])/(n-1);
    val_vcov[p,q]=(val_data[,p]-val_data[:,p])`*(val_data[,q]-val_data[:,q])/(n-1);
  end;
  est_sds[p,p]=sqrt(est_vcov[p,p]);
  val_sds[p,p]=sqrt(val_vcov[p,p]);
end;
est_corr=inv(est_sds)*est_vcov*inv(est_sds);
val_corr=inv(val_sds)*val_vcov*inv(val_sds);
est_beh_Rsq=(inv(est_corr[1:3,1:3])*est_corr[1:3,4])`*est_corr[1:3,4];
val_beh_Rsq=(inv(val_corr[1:3,1:3])*val_corr[1:3,4])`*val_corr[1:3,4];

/* Compute the crossvalidity coefficient and square it */
crossval=(pred_ys[,1]-pred_ys[:,1])`*(val_data[,4]-val_data[:,4])/
        sqrt((pred_ys[,1]-pred_ys[:,1])`*(pred_ys[,1]-pred_ys[:,1])*
                    (val_data[,4]-val_data[:,4])`*(val_data[,4]-val_data[:,4]));
crossval_sq=crossval*crossval;

/* Compute the Raudenbush & Bryk "R²" for t00 and t11 */
do i=1 to 2;
  est_RB_Rsq[1,i]=(UE_vars[1,i]-CE_vars[1,i])/UE_vars[1,i];
  crossval_taus[1,i]=(u[,i]-u[:,i])`*(u[,i]-u[:,i])/(n2-1);
  crossval_RB_Rsq[1,i]=(UV_vars[1,i]-crossval_taus[1,i])/UV_vars[1,i];
end;

crossval_taus[1,3]=(u[,1]-u[:,1])`*(u[,2]-u[:,2])/(n2-1);

/* Compute the Snijders & Bosker level-2 R² */
est_var_ybar=(est_ybar[,1]-est_ybar[:,1])`*(est_ybar[,1]-est_ybar[:,1])/(n2-1);
val_var_ybar=(val_ybar[,1]-val_ybar[:,1])`*(val_ybar[,1]-val_ybar[:,1])/(n2-1);
est_var_y=(est_data[,4]-est_data[:,4])`*(est_data[,4]-est_data[:,4])/(n-1);
val_var_y=(val_data[,4]-val_data[:,4])`*(val_data[,4]-val_data[:,4])/(n-1);
est_xbar_all=est_data[:,2];
val_xbar_all=val_data[:,2];
est_var_err_b=(esterr_b[,1]-esterr_b[:,1])`*(esterr_b[,1]-esterr_b[:,1])/(n2-1);
est_var_err_w=(esterr_w[,1]-esterr_w[:,1])`*(esterr_w[,1]-esterr_w[:,1])/(n-1);
cross_var_err_b=(crosserr_b[,1]-crosserr_b[:,1])`*(crosserr_b[,1]-crosserr_b[:,1])/(n2-1);
cross_var_err_w=(crosserr_w[,1]-crosserr_w[:,1])`*(crosserr_w[,1]-crosserr_w[:,1])/(n-
1);

est_SB_23p=CE_vars[1,1]+2*est_xbar_all*CE_vars[1,3]
       +CE_vars[1,2]*(est_xbar_all*est_xbar_all+est_var_err_b+est_var_err_w/n1)
       +CE_vars[1,4]/n1;
est_SB_rsq2=1-(est_SB_23p/est_var_ybar);
cross_SB_23p=crossval_taus[1,1]+2*val_xbar_all*crossval_taus[1,3]
```

```
+crossval_taus[1,2]*(val_xbar_all*val_xbar_all+cross_var_err_b+cross_var_err_w/n1)
        +UV_vars[1,4]/n1;
cross_SB_rsq2=1-(cross_SB_23p/val_var_ybar);

storstuf[1,1]=est_beh_Rsq; storstuf[1,2]=val_beh_Rsq;
storstuf[1,3]=crossval_sq;
storstuf[1,4]=est_RB_Rsq[1,1]; storstuf[1,5]=est_RB_Rsq[1,2];
storstuf[1,6]=crossval_RB_Rsq[1,1]; storstuf[1,7]=crossval_RB_Rsq[1,2];
storstuf[1,8]=est_SB_rsq2; storstuf[1,9]=cross_SB_rsq2;
create sendout from storstuf;
append from storstuf;
quit;

data save_rep; set sendout;
proc append base=squared_stuff; run;

%end; /* If tally=20 */
%else %let count=&count;

%if &count <&goal %then %do;
  %LET seed1=&seed1+2;
  %LET seed2=&seed2+2;
%end;

%else %if &count = &goal %then %let rep=&reps;

%end; /* Replication loop */

proc iml;
count=&count;
print count;
use squared_stuff; read all into getmeans;
themeans=repeat(0,1,ncol(getmeans));
do i=1 to ncol(getmeans);
  themeans[1,i]=getmeans[:,i];
end;
shrink=j(1,8,0);
shrink[1,1]=themeans[1,1]-themeans[1,3];              /*beh - cv beh*/
shrink[1,2]=(themeans[1,1]-themeans[1,3])/themeans[1,1]; /*%change_beh*/
shrink[1,3]=themeans[1,4]-themeans[1,6];              /*RB t00 - cv t00*/
shrink[1,4]=(themeans[1,4]-themeans[1,6])/themeans[1,4]; /*%change RBt00)*/
shrink[1,5]=themeans[1,5]-themeans[1,7];              /*RB t11 -cv t11*/
shrink[1,6]=(themeans[1,5]-themeans[1,7])/themeans[1,5]; /*%change RB t11)*/
shrink[1,7]=themeans[1,8]-themeans[1,9];
shrink[1,8]=(themeans[1,8]-themeans[1,9])/themeans[1,8];
```

```
themeans=themeans||shrink;
print themeans;
file 'C:\dissertations\Tracy\results005.dat';
do r=1 to 1;
  do c=1 to ncol(themeans);
    put(themeans[r,c]) +1 @;
  end;
  put;
end;
closefile 'C:\dissertation\results005.dat';
quit;
%MEND RUNHLM;
%runhlm;
```

APPENDIX B

Group-mean Centered SAS Code

```
%let corrwx=0;
%MACRO RUNHLM;
%let seed1=10370001; %let seed2=20370001;
%let groups=40; /* Level-2 n */
%let n_per=10; /* Level-1 n */
%let t00=61.9285714285714; %let t11=0.24190848214286;
%let corrwy=.4;
%let corrxy=.3;
%let corrwxy=.3;
%let reps=3000;
%let goal=1000;
%let count=0;
%let tally=0;

proc datasets library=work;
  delete squared_stuff;
run;
PROC PRINTTO FILE='C:\dissertation\HLMOUT.TXT' NEW;
PROC PRINTTO LOG='C:\dissertation\HLM_05.TXT' NEW;

%do rep=1 %to &reps;

proc iml;
preds=3;
data_err=repeat(0,&groups*&n_per,7);
pred_sd=repeat(0,preds,preds);
gammas=repeat(0,1+preds,1);
errorvar=289;
poppredc={1.00 &corrwx 0.00,
&corrwx 1.00 0.00,
              0.00 0.00 1.00};
corrwthy={&corrwy, &corrxy, &corrwxy};
standgam=inv(poppredc)*corrwthy;
mu={90, 0, 0, 0};
pred_var={441, 256, 0};
  pred_var[3,1]=pred_var[1,1]*pred_var[2,1]*(1 + poppredc[2,1]**2);
y_var=(&t00+&t11*pred_var[2,1]+errorvar)/(1-standgam`*poppredc*standgam);
```

```
y_sd=sqrt(y_var);

do p=1 to preds;
  pred_sd[p,p]=sqrt(pred_var[p,1]);
  gammas[p+1,1]=standgam[p,1]*y_sd/pred_sd[p,p];
end;

prdcdata=repeat(0,&groups*&n_per,3); /* w x wx */
grpdata=normal(repeat(&seed1,&groups,3)); /* w u0 u1 */
  grpdata[,2]=sqrt(&t00)*grpdata[,2];
  grpdata[,3]=sqrt(&t11)*grpdata[,3];

persdata=normal(repeat(&seed2,&groups*&n_per,2)); /* x rij */
  persdata[,2]=sqrt(errorvar)*persdata[,2];

/* Group-mean centering  */
do j=1 to &groups;
  x=persdata[(j-1)*&n_per+1:j*&n_per,1];
  x=x-x[:,1];
  persdata[(j-1)*&n_per+1:j*&n_per,1]=x;
end;

varcovar=pred_sd*poppredc*pred_sd;
  mu[4,1]=mu[2,1]*mu[3,1]+varcovar[2,1];
  gammas[1,1]=mu[1,1]-gammas[2,1]*mu[2,1]-gammas[3,1]*mu[3,1]-
gammas[4,1]*mu[4,1];
  cd=root(varcovar);

do i=1 to &groups;
  do j=1 to &n_per;
    prdcdata[(i-1)*&n_per+j,1]=grpdata[i,1]; /* w */
    prdcdata[(i-1)*&n_per+j,2]=persdata[(i-1)*&n_per+j,1]; /* x */
    prdcdata[(i-1)*&n_per+j,3]=grpdata[i,1]*persdata[(i-1)*&n_per+j,1]; /* wx */
    data_err[(i-1)*&n_per+j,4]=grpdata[i,2];  /* u0 */
    data_err[(i-1)*&n_per+j,5]=grpdata[i,3]; /* u1 */
    data_err[(i-1)*&n_per+j,6]=persdata[(i-1)*&n_per+j,2]; /* rij */
    data_err[(i-1)*&n_per+j,7]=i;
  end;
end;

prdcdata=prdcdata*cd;
prdcdata[,3]=prdcdata[,3]+mu[4,1];
data_err[,1]=prdcdata[,1]; data_err[,2]=prdcdata[,2]; data_err[,3]=prdcdata[,3];

gammas=gammas`;
create allgs from gammas;
```

```
append from gammas;

create combine from data_err;
append from data_err;
quit;

data d1; set allgs; keep g00 g01 g10 g11;
g00=col1; g01=col2; g10=col3; g11=col4;
run;

/*CREATE Y FOR EACH PARTICIPANT USING COMBINED MODEL*/
DATA CREAT_Y; SET COMBINE; if _N_=1 then set d1; drop col1-col6;
  w1=col1; x1=col2; wx=col3; u0=col4; u1=col5; rij=col6; site=col7;
  Y = G00 + G10*X1 + G01*W1 + G11*wx + u0 + u1*x1 + RIJ;
RUN;

/* divide generated data into two sets */
%let dsid=%sysfunc(open(work.CREAT_Y,in));
%let nobs=%sysfunc(attrn(&dsid,nobs));
%if &dsid > 0 %then %let rc=%sysfunc(close(&dsid));
data estim valid;
set CREAT_Y;
if _n_ le &nobs/2
then output estim;
else if _n_ GT &nobs/2
then output valid;
run;

/* Estimation unconditional*/
ods listing;
ods output covparms=est_Utau;
ods output solutionf=est_Ugam;

proc mixed data = ESTIM covtest;
class site;
model y=x1 /ddfm=residual solution cl notest;
random intercept x1 / sub=site type=un;
run;

data covs_UE;
set est_Utau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cue;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
```

```
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';
keep t00est t11est t10est var1est;

array se(4) t00se t11se t10se var1se;
worked_cue=4;
do i=1 to 4;
  if se(i)=. then worked_cue=worked_cue-1;
end;
call symput("worked_cue", worked_cue);
keep worked_cue;
run;

%if &worked_cue = 0 %then %let worked_pue=0;
%else %do;
data parms_UE;
set est_Ugam;
retain g00est g10est g00se g10se worked_pue;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='x1';

array se(2) g00se g10se;
worked_pue=2;
do i=1 to 2;
  if se(i)=. then worked_pue=worked_pue-1;
end;
call symput("worked_pue", worked_pue);
keep g00est g10est worked_pue;
run;
%end; /* worked_cue check */

/*  Estimation Conditional */
ods listing;
ods output covparms=est_Ctau;
```

```
ods output solutionf=est_Cgam;

proc mixed data = ESTIM covtest;
class site;
model y=x1 w1 wx/ddfm=residual solution cl notest;
random intercept x1 /sub=site type=un;
run;

data covs_CE;
set est_Ctau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cce;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';

array se(4) t00se t11se t10se var1se;
worked_cce=4;
do i=1 to 4;
  if se(i)=. then worked_cce=worked_cce-1;
end;
call symput('worked_cce', worked_cce);
keep t00est t11est t10est var1est worked_cce;
run;

%if &worked_cce = 0 %then %let worked_pce=0;
%else %do;
data parms_CE;
set est_Cgam;
retain g00est g10est g01est g11est g00se g10se g01se g11se worked_pce;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='w1' then do;
```

```
  g01est=estimate; g01se=stderr;
end;
if effect='wx' then do;
  g11est=estimate; g11se=stderr;
end;
if effect='wx';

array se(4) g00se g10se g01se g11se;
worked_pce=4;
do i=1 to 4;
  if se(i)=. then worked_pce=worked_pce-1;
end;
call symput('worked_pce', worked_pce);
keep g00est g10est g01est g11est worked_pce;
run;
%end; /* worked_cce check */

/*  VALIDATION Unconditional  */
ods listing;
ods output covparms=val_UVtau;
ods output solutionf=val_UVgam;

proc mixed data = valid covtest;
class site;
model y=x1 /ddfm=residual solution cl notest;
random intercept x1/sub=site type=un;
run;

data covs_UV;
set val_UVtau;
retain t00est t11est t10est var1est t00se t11se t10se var1se worked_cuv;
if COVPARM='UN(1,1)' then do;
  t00est=estimate; t00se=stderr;
end;
if covparm='UN(2,2)' then do;
  t11est=estimate; t11se=stderr;
end;
if covparm='UN(2,1)' then do;
  t10est=estimate; t10se=stderr;
end;
if covparm='Residual' then do;
  var1est=estimate; var1se=stderr;
end;
if covparm='Residual';
keep t00est t11est t10est var1est;
array se(4) t00se t11se t10se var1se;
```

```
worked_cuv=4;
do i=1 to 4;
  if se(i)=. then worked_cuv=worked_cuv-1;
end;
call symput("worked_cuv", worked_cuv);
keep worked_cuv;
run;

%if &worked_cuv = 0 %then %let worked_puv=0;
%else %do;
data parms_UV;
set val_UVgam;
retain g00est g10est g00se g10se worked_puv;
if effect='Intercept' then do;
  g00est=estimate; g00se=stderr;
end;
if effect='x1' then do;
  g10est=estimate; g10se=stderr;
end;
if effect='x1';
array se(2) g00se g10se;
worked_puv=2;
do i=1 to 2;
  if se(i)=. then worked_puv=worked_puv-1;
end;
call symput("worked_puv", worked_puv);
keep g00est g10est worked_puv;
run;
%end; /* worked_puv check */

%let
tally=&worked_cue+&worked_pue+&worked_cce+&worked_pce+&worked_cuv+&wor
ked_puv;

%if &tally=20 %then %do;
%let count=&count+1;
data estim2iml; set estim;
drop col7 g00 g01 g10 g11 u0 u1 rij site;
run;

data valid2iml; set valid;
drop col7 g00 g01 g10 g11 u0 u1 rij site;
run;

proc iml;
use estim2iml; read all into est_data; * w x wx y;
```

```
use valid2iml; read all into val_data; * w x wx y;
use parms_ce; read all into gammas;
use covs_UE; read all into UE_vars;
use covs_CE; read all into CE_vars;
use covs_UV; read all into UV_vars;

n1=&n_per; n2=&groups/2; n=n1*n2;
gamma00=gammas[1,1]; gamma01=gammas[1,2];
gamma10=gammas[1,3]; gamma11=gammas[1,4];
est_vcov=repeat(0,4,4); est_sds=repeat(0,4,4);
val_vcov=repeat(0,4,4); val_sds=repeat(0,4,4);
crossval_taus=repeat(0,1,3); *tau00 tau11 tau10;
est_RB_Rsq=repeat(0,1,2); crossval_RB_Rsq=repeat(0,1,2);
right_bs=repeat(0,n2,2); pred_bs=repeat(0,n2,2); u=repeat(0,n2,2);
pred_ys=repeat(0,n1*n2,1);

esterr_b=repeat(0,n2,1);
esterr_w=repeat(0,n,1);
crosserr_b=repeat(0,n2,1);
crosserr_w=repeat(0,n,1);
est_ybar=repeat(0,n2,1);
val_ybar=repeat(0,n2,1);

storstuf=repeat(0,1,9);

do j=1 to n2;
  w=val_data[(j-1)*n1+1,1];
  y=val_data[(j-1)*n1+1:j*n1,4];
  val_x=val_data[(j-1)*n1+1:j*n1,2];
  x=repeat(1,n1,1)||val_x;
  right_bs[j,]=(inv(x`*x)*x`*y)`;
  pred_bs[j,1]=gamma00+gamma01*w;
  pred_bs[j,2]=gamma10+gamma11*w;
  u[j,1]=pred_bs[j,1]-right_bs[j,1];
  u[j,2]=pred_bs[j,2]-right_bs[j,2];
  pred_ys[(j-
1)*n1+1:j*n1,1]=gamma00+gamma01*w+gamma10*val_x+gamma11*w*val_x;

  est_y=est_data[(j-1)*n1+1:j*n1,4];
  est_x=est_data[(j-1)*n1+1:j*n1,2];
  est_ybar[j,1]=est_y[:];
  est_xbar=est_x[:];
  esterr_b[j,1]=est_ybar[j,1]-gamma10*est_xbar;
  esterr_w[(j-1)*n1+1:j*n1,1]=est_y-gamma10*est_x;
  val_ybar[j,1]=y[:];
  val_xbar=val_x[:];
```

```
  crosserr_b[j,1]=val_ybar[j,1]-gamma10*val_xbar;
  crosserr_w[(j-1)*n1+1:j*n1,1]=y-gamma10*val_x;
end;

/* Compute R² via behs for each sample */
do p=1 to 4;
  do q=1 to 4;
    est_vcov[p,q]=(est_data[,p]-est_data[:,p])`*(est_data[,q]-est_data[:,q])/(n-1);
    val_vcov[p,q]=(val_data[,p]-val_data[:,p])`*(val_data[,q]-val_data[:,q])/(n-1);
  end;
  est_sds[p,p]=sqrt(est_vcov[p,p]);
  val_sds[p,p]=sqrt(val_vcov[p,p]);
end;
est_corr=inv(est_sds)*est_vcov*inv(est_sds);
val_corr=inv(val_sds)*val_vcov*inv(val_sds);
est_beh_Rsq=(inv(est_corr[1:3,1:3])*est_corr[1:3,4])`*est_corr[1:3,4];
val_beh_Rsq=(inv(val_corr[1:3,1:3])*val_corr[1:3,4])`*val_corr[1:3,4];

/* Compute the crossvalidity coefficient and square it */
crossval=(pred_ys[,1]-pred_ys[:,1])`*(val_data[,4]-val_data[:,4])/
        sqrt((pred_ys[,1]-pred_ys[:,1])`*(pred_ys[,1]-pred_ys[:,1])*
                (val_data[,4]-val_data[:,4])`*(val_data[,4]-val_data[:,4]));
crossval_sq=crossval*crossval;

/* Compute the Raudenbush & Bryk "R²" for t00 and t11 */
do i=1 to 2;
  est_RB_Rsq[1,i]=(UE_vars[1,i]-CE_vars[1,i])/UE_vars[1,i];
  crossval_taus[1,i]=(u[,i]-u[:,i])`*(u[,i]-u[:,i])/(n2-1);
  crossval_RB_Rsq[1,i]=(UV_vars[1,i]-crossval_taus[1,i])/UV_vars[1,i];
end;

crossval_taus[1,3]=(u[,1]-u[:,1])`*(u[,2]-u[:,2])/(n2-1);

/* Compute the Snijders & Bosker level-2 R² */
est_var_ybar=(est_ybar[,1]-est_ybar[:,1])`*(est_ybar[,1]-est_ybar[:,1])/(n2-1);
val_var_ybar=(val_ybar[,1]-val_ybar[:,1])`*(val_ybar[,1]-val_ybar[:,1])/(n2-1);
est_var_y=(est_data[,4]-est_data[:,4])`*(est_data[,4]-est_data[:,4])/(n-1);
val_var_y=(val_data[,4]-val_data[:,4])`*(val_data[,4]-val_data[:,4])/(n-1);
est_xbar_all=est_data[:,2];
val_xbar_all=val_data[:,2];
est_var_err_b=(esterr_b[,1]-esterr_b[:,1])`*(esterr_b[,1]-esterr_b[:,1])/(n2-1);
est_var_err_w=(esterr_w[,1]-esterr_w[:,1])`*(esterr_w[,1]-esterr_w[:,1])/(n-1);
cross_var_err_b=(crosserr_b[,1]-crosserr_b[:,1])`*(crosserr_b[,1]-crosserr_b[:,1])/(n2-1);
cross_var_err_w=(crosserr_w[,1]-crosserr_w[:,1])`*(crosserr_w[,1]-crosserr_w[:,1])/(n-
1);
```

```
est_SB_23p=CE_vars[1,1]+2*est_xbar_all*CE_vars[1,3]
        +CE_vars[1,2]*(est_xbar_all*est_xbar_all+est_var_err_b+est_var_err_w/n1)
        +CE_vars[1,4]/n1;
est_SB_rsq2=1-(est_SB_23p/est_var_ybar);
cross_SB_23p=crossval_taus[1,1]+2*val_xbar_all*crossval_taus[1,3]

+crossval_taus[1,2]*(val_xbar_all*val_xbar_all+cross_var_err_b+cross_var_err_w/n1)
        +UV_vars[1,4]/n1;
cross_SB_rsq2=1-(cross_SB_23p/val_var_ybar);

storstuf[1,1]=est_beh_Rsq; storstuf[1,2]=val_beh_Rsq;
storstuf[1,3]=crossval_sq;
storstuf[1,4]=est_RB_Rsq[1,1]; storstuf[1,5]=est_RB_Rsq[1,2];
storstuf[1,6]=crossval_RB_Rsq[1,1]; storstuf[1,7]=crossval_RB_Rsq[1,2];
storstuf[1,8]=est_SB_rsq2; storstuf[1,9]=cross_SB_rsq2;
create sendout from storstuf;
append from storstuf;
quit;

data save_rep; set sendout;
proc append base=squared_stuff; run;

%end; /* If tally=20 */
%else %let count=&count;

%if &count <&goal %then %do;
  %LET seed1=&seed1+2;
  %LET seed2=&seed2+2;
%end;

%else %if &count = &goal %then %let rep=&reps;

%end; /* Replication loop */

proc iml;
count=&count;
print count;
use squared_stuff; read all into getmeans;
themeans=repeat(0,1,ncol(getmeans));
do i=1 to ncol(getmeans);
  themeans[1,i]=getmeans[:,i];
end;
shrink=j(1,8,0);
shrink[1,1]=themeans[1,1]-themeans[1,3];              /*beh - cv beh*/
shrink[1,2]=(themeans[1,1]-themeans[1,3])/themeans[1,1]; /*%change_beh*/
shrink[1,3]=themeans[1,4]-themeans[1,6];              /*RB t00 - cv t00*/
```

```
shrink[1,4]=(themeans[1,4]-themeans[1,6])/themeans[1,4]; /*%change RBt00)*/
shrink[1,5]=themeans[1,5]-themeans[1,7];                /*RB t11 -cv t11*/
shrink[1,6]=(themeans[1,5]-themeans[1,7])/themeans[1,5]; /*%change RB t11)*/
shrink[1,7]=themeans[1,8]-themeans[1,9];
shrink[1,8]=(themeans[1,8]-themeans[1,9])/themeans[1,8];
themeans=themeans||shrink;
print themeans;
file 'C:\dissertations\Tracy\results037.dat';
do r=1 to 1;
  do c=1 to ncol(themeans);
    put(themeans[r,c]) +1 @;
  end;
  put;
end;
closefile 'C:\dissertation\results037.dat';
quit;

%MEND RUNHLM;
%runhlm;
```