UNIVERSITY OF CALIFORNIA

Los Angeles

Testing Non-Nested Multilevel Models

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Andrew Lawrence Moskowitz

2017

ABSTRACT OF THE DISSERTATION

Testing Non-nested Multilevel models

by

Andrew Lawrence Moskowitz

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2017

Professor Jennifer Lynn Krull, Co-Chair

Professor Craig Kyle Enders, Co-Chair

Comparing theories represented by statistical models is central to psychological research. Historically, comparisons between so called "non-nested" models have been error prone in the absence of a null hypothesis test. Recent research by Levy and Hancock and Merkle, You, and Preacher has extended Vuong's Likelihood Ratio Test of non-nested models to Structural Equation Models (SEMs). A notable omission of recent work is the extension of Vuong's test to the case of multilevel regression- a common approach for modeling longitudinal or grouped data. This dissertation leverages the similarities between SEMs and multilevel models to extend Vuong's test to the multilevel framework. The logic of Vuong's test as it relates to multilevel regression was explored and a SAS macro developed to facilitate the comparison between two models known to be non-nested a priori. The ability of Vuong's test to select the true or "best" model was compared to that of information criteria in three simulation studies reflecting

scenarios in which non-nestedness is commonly encountered in multilevel regression: non-nested covariate sets, level 1 residual covariance structures, and functional forms. Selection rates of the incorrect models were also examined. Vuong's test showed almost no incorrect model selection across all scenarios, although its power to select the correct model was generally modest. Model comparisons among information criteria tended to be more sensitive than Vuong's test but also selected the incorrect model more often. Finally, Vuong's test was applied to three real data sets comparing competing models in the same scenarios as the simulation studies. Implications and recommendations for use are discussed.

The dissertation of Andrew Lawrence Moskowitz is approved.

Matthew S. Fritz

Bruce F. Chorpita

Craig Kyle Enders, Co-Chair

Jennifer Lynn Krull, Co-Chair

University of California, Los Angeles

2017

For Mom and Dad

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my advisor and dissertation committee chair, Jennifer Krull, for her guidance over the last five years. Thank you for the spirited debates in our weekly meetings and encouraging me to think independently as a scholar. I also owe a great deal to Bruce Chorpita. Thank you Bruce, you have helped me to navigate the academic minefield and at times have emerged as my strongest advocate. Thank you and the rest of the "Chorpitans" for welcoming me into your lab. Thank you also to the other members of my dissertation committee, Dr. Craig Enders who served as co-chair and Dr. Matthew Fritz whose interest in this project reinforced its importance.

I owe an enormous amount my fiancé Becca and "Bruce the Batdog" whose warmth, positivity, and love continues to give me something to look forward to every day. Thank you for putting up with all of the stress and anxiety. Finally, I would like to extend my deepest gratitude to my parents Mark and Faith Moskowitz. Without their unconditional love, support, and blind encouragement I simply would not have accomplished this work. You taught me that with hard work and focus I could achieve anything. This accomplishment is a result of the wonderful life and opportunities you have afforded me and this honor belongs to you as much as it belongs to me. Thank you from the bottom of my heart.

I would like to acknowledge the National Institute of Drug Abuse, National Institute of Health for financial support and training through grant DA007272-22 to the University of California Los Angeles. The opinions expressed in this dissertation are my own and do not represent the views of the National institute of Drug Abuse or the National Institute of Health.

Vita

EDUCATION

2006 B.A., Psychology/Criminology, University of Miami (FL)

2013 M.A., Quantitative Psychology, University of California, Los Angeles

PUBLICATIONS

Moskowitz, A. L., Krull, J. L., Trickey, K. A., & Chorpita, B. F. (2017). Quality Vs. Quantity: Assessing behavior change over time. *Journal of Psychopathology and Behavioral Assessment.* 1-20.

Hser, Y-I., Huang, D., Saxon, A., Woody, G., Moskowitz, A. L., Matthews, A. G., & Ling, W. (2017) Distinctive trajectories of opioid use over an extended follow-up of patients in a multisite trial on Buprenorphine + Naloxone and Methadone. *Journal of Addiction Medicine, 11*(1), 63-69.

Tsai, K. H., Moskowitz A. L., Lynch, R. E., Daleiden E., Mueller C. W., Krull, J. L., & Chorpita, B. F. (2016). Do treatment plans matter? Moving from recommendations to action. *Journal of Clinical Child & Adolescent Psychology.* 1-7.

Park, A. L., Moskowitz, A. L., & Chorpita, B. F. (2016). Community-based providers' selection of practices for children and adolescents with comorbid mental health problems. *Journal of Clinical Child and Adolescent Psychology,* 1-12.

Moreno, P. I., Moskowitz, A. L., Ganz, P. A., & Bower, J. E (2016). Positive affect and inflammatory activity in breast cancer survivors: Examining the role of affective arousal. *Psychosomatic Medicine. 78* (5). 532-541.

Tsai, K. H., Moskowitz, A. L.. Brown, T. E., Park, A. L., & Chorpita, B. F. (2015). Interpreting progress feedback to guide clinical decision-making in children's mental health services. *Administration and Policy in Mental Health and Mental Health Services Research,* 1-8.

Fisher, M. H., Moskowitz, A. L., & Hodapp, R. M. (2013). Differences in social vulnerability among individuals with Autism Spectrum Disorder, Williams Syndrome, and Down Syndrome. *Research in Autism Spectrum Disorders,* 7(8), 931-937.

Favazza, P. C., Siperstein, G. N., Zeisel, S., Odom, S. L., Sideris, J. H. & Moskowitz, A. L. (2013) Young athletes program: Impact on motor development. *Adaptive Physical Activity Quarterly*, 30*(3)*, 235-253.

Fisher, M. H., Moskowitz, A. L., & Hodapp, R. M. (2012). Vulnerability and experiences related to social victimization among individuals with intellectual and developmental disabilities. *Journal Of Mental Health Research In Intellectual Disabilities*, *5*(1), 32-48.

Chapter 1: Introduction

Model evaluation and comparison has long been of central importance to psychology. As early as Charles Spearman, and certainly earlier, scholars hotly debated their theories against those of their contemporaries. In a well-known disagreement over the nature of intelligence, Spearman, Sir Godfrey Thomson, and colleagues argued over the existence of a general intelligence factor and its implications. While the state of statistical methodology in the early 1900s was nascent at best, these types of debates were central to the evolution of psychology as a science. It wasn't until more advanced methods were developed later in the century that competing theories could be evaluated empirically.

As methodologies and supporting technologies advanced, it became easier to build comprehensive models with the ability to handle more complex data. Personal computers have allowed researchers to develop iterative algorithms for estimation to make the process of estimating these advanced models mathematically tractable and as a result, Maximum Likelihood (ML) estimation and its derivatives have emerged as the preferred method. Broadly, ML estimation converges on parameter values that have the highest probability of reproducing data in a particular sample; that is to say they maximize the likelihood of a given sample via the estimated parameters. The likelihood is quite literally the joint probability of observing the collective data given the parameters, however, in practice this maximization typically occurs over the joint *log likelihoods* for mathematical convenience. Intuitively, the log likelihood provides a measure of model quality. However, as will be discussed subsequently, the log likelihood is not an *absolute* measure of model fit and therefore cannot be evaluated on its own. In certain scenarios the log likelihoods of two models can be compared and their difference evaluated.

Fitting the Simple Regression Model

Estimators that appear simple on the exterior may actually be classified as ML and thus produce a model-based log likelihood. One such estimator in which the log likelihood metric is often overlooked is Ordinary Least Squares (OLS) regression. Because simple and intuitive evaluation measures are available and analytic solutions are easily computed, it is often unnecessary to approach estimation from an iterative likelihood perspective. Probably the most well-known model evaluation metric in OLS regression is the coefficient of determination or "R^2". Regularly conceptualized as the percentage of variance in the criterion that a model explains, R^2 is bounded between 0 and 1 and increases monotonically as regressors are added to the model. Thus a simple increase in R^2 is not necessarily instructive to determine if the addition of a variable improves the predictive power of a model beyond expectation.

In an attempt to allow for a more evaluative measure of model fit in which additional regressors are not guaranteed to improve the model, an adjusted version of R^2 was proposed. The adjusted R^2 is one of the simplest *penalized* criteria in model evaluation. Specifically, the ratio of explained to total variance is weighted, or penalized, by a proportion of degrees of freedom. With the addition of a new predictor variable, adjusted R^2 will only increase if the variable adds more predictive power than would be expected by chance. Therefore the model incurs a penalty when variables that do not improve its fit are added to the model.

Because adjusted R^2 must be used for comparative purposes and is only valid for nested comparisons, it is necessary to also have a metric that may be used to compare *non-nested* models. While non-nestedness will be discussed in detail later in this chapter, briefly two models are considered non-nested if one cannot be obtained by algebraically manipulating the other (e.g., setting one or more coefficients in the more general model to zero). To facilitate such

comparisons, a class of metrics known as information criteria are available, however, their calculation requires the log likelihood of the model and its transformation to the deviance which can be computed by

$$\ln(\mathcal{L}) = \left(-\frac{1}{2}n(\ln(2\pi) + \ln(SS_e) - \ln(n) + 1)\right) \quad (1)$$

$$-2\ln(\mathcal{L}) = n(\ln(2\pi) + \ln(SS_e) - \ln(n) + 1) \quad (2)$$

where SS_e is the sum of squared residuals (Gagné & Dayton, 2002). The two most common information criteria, Akaike's Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978), can be represented using the deviance as

$$AIC = -2\ln(\mathcal{L}) + 2(k_m + 2) \quad (3)$$

$$BIC = -2\ln(\mathcal{L}) + \ln(n)\,(k_m + 2) \quad (4)$$

where k_m is the number of predictor variables in model *m* and a constant of 2 is added to account for the intercept and residual variance. Like the adjusted R^2, AIC and BIC penalize more complex models with predictors that do not meaningfully contribute to the model. Models with the lowest values of these criteria are selected as the "best" model under their respective assumptions. I expand on the theory behind common information criteria including AIC and BIC in subsequent sections.

While R^2 measures do not necessarily map directly on to some of the more elaborate statistical frameworks (e.g., multilevel modeling), information criteria can be calculated for any model estimated under maximum likelihood. As a result they are widely used to compare non-nested models in a rudimentary fashion. Typically, information criteria are simply compared to one another in an absolute sense without any acknowledgement of sampling variability in the estimate. A generalizability problem also exists with model selection based on AIC and BIC in

that a model selected in one sample may not be selected in another sample. Bootstrapping methods (e.g., Bollen & Stine, 1992; Shang & Cavanaugh, 2008a, 2008b; Shibata, 1997) used to create empirical confidence intervals around AIC and BIC have recently gained attention, however, their large computational burden, difficulty to implement, and nonstandard interpretation have limited their adoption in the literature (Merkle, You, & Preacher, 2015). To facilitate model comparisons it would be desirable to create an easy to use and interpretable null hypothesis test for comparing non-nested models.

Vuong (1989) proposed such a test for non-nested models which has recently garnered increased attention in psychology through applications to structural equation models (SEMs) by Levy and Hancock (2007, 2011) and Merkle et al. (2015). Although closely related to SEM, multilevel models contain several nuances that complicate the use of many model evaluation methods utilized in SEM. For instance, the lack of a population model hinders computations of a measure of absolute model fit. Furthermore, the presence of random effects in the multilevel model complicates implementation of the Merkle et al. methodology in current software packages. Because no test for non-nested hypotheses exists in the multilevel literature and bootstrapping may be untenable or difficult to implement in multilevel data (Hox, 2010), I adopt Vuong's Likelihood Ratio Test for non-nested models for use in the multilevel framework. I begin by describing how multilevel models are estimated using maximum likelihood and expand on the likelihood's relation to information theory, particularly as it was defined by Kullback and Leibler (1951). I then describe in detail the theory behind information criteria commonly used in multilevel model selection, specifically AIC, AIC_C, and BIC. Subsequently I discuss a null hypothesis test applicable for nested multilevel models, the likelihood ratio test (LRT), and continue to describe its extension to the non-nested case via Vuong, Merkle et al., and Levy and

Hancock. I conclude the introduction with a short review of the literature surrounding multilevel model selection based on information criteria.

Estimation in Multilevel Models

Under the assumptions of normality and Gauss-Markov, estimation by ordinary least squares is the best unbiased linear predictor of the outcome *y*. One such assumption, namely that error terms for different observations are uncorrelated, is inherently violated in multilevel data; by their nature observations of the same group or person are more correlated with one another than those of other groups or persons. As a result when data are estimated under ordinary least squares thereby ignoring the correlation among observations and violating the independence assumption, standard errors are downwardly biased and alpha levels are inflated (Cohen, Cohen, West, & Aiken, 2002; Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). To account for the correlation among observations, random coefficient(s), or rather their variances, are estimated in multilevel models to allow for the similarity within groups to be accounted for. Ordinary least squares does not provide a mechanism by which to estimate this additional variance term and thus maximum likelihood estimation must be employed to estimate the additional variances.

As a class of estimators, maximum likelihood results in coefficients that *maximize* the *likelihood* of the observed data given the model. Ordinary least squares is a maximum likelihood estimator for the regression problem when the assumptions are met. The type of maximum likelihood used in multilevel regression problems employs an iterative process to find estimates that produce predicted values closest to the observed data. Two types of maximum likelihood estimators, full maximum likelihood (also known as full information maximum likelihood, FIML, FML, or simply maximum likelihood) and restricted maximum likelihood (REML or

RML), are commonplace in multilevel model estimation. Essentially, FML and REML differ in *what* they are maximizing the likelihood of and how they treat the fixed effects. FML operates on the data directly but assumes the fixed effects to be known and does not correct for them in the degrees of freedom during calculations (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). As a result, variance components are downwardly biased unless the sample is sufficiently large, in which case a bias still exists, it is just vanishingly small. REML, on the other hand, conditions its estimates on the fixed part of the model and performs estimation on the residuals (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). REML accounts for the estimation of the fixed effects, but because of this, comparisons across models are only valid for those that differ *exclusively* in random effects. If there is any difference in fixed effect structure between two models, they cannot be compared via REML criteria. Asymptotically, these methods are equivalent and because FML can be used to compare models which differ in either their fixed and/or random parts, I will focus on full maximum likelihood for the duration of this dissertation.

Using vector notation the combined form of the multilevel model can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}, \tag{5}$$

where **y** is an n x 1 vector of responses, $\boldsymbol{X}$ is a n x q design matrix for fixed effects, $\boldsymbol{\beta}$ is a q x 1 vector of fixed effects, $\boldsymbol{Z}$ is an n x p random effects design matrix, $\boldsymbol{\gamma}$ is a p x 1 vector of random effects parameters, and $\boldsymbol{\epsilon}$ is an n x 1 vector of residuals. For the general linear mixed model we assume that $\boldsymbol{y}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\epsilon}$ are all normally distributed and $\boldsymbol{\gamma}$, and $\boldsymbol{\epsilon}$ have means of 0 and respective variance covariance matrices $\boldsymbol{\tau}$ and $\boldsymbol{R}$. Furthermore, we assume that $\boldsymbol{\tau}$ and $\boldsymbol{R}$ are independent. A common assumption for cross-sectional data in multilevel regression is that $\boldsymbol{R} = \sigma^2\mathbf{I}_n$ where $\mathbf{I}_n$ is a n x n identity matrix resulting in a diagonal (i.e., independent) matrix for Level 1 residuals. In longitudinal data, it is common for a variety of structures to be specified for $\boldsymbol{R}$ and their fit

evaluated with LRTs if two structures are nested. A central goal of this dissertation is to present a test that is capable of comparing two non-nested structures.

The variance in **y** is given by the equation $\boldsymbol{V = Z\,\tau\,Z` + R}$. $\boldsymbol{V}$ is an n x n block diagonal matrix with each block representing an independent Level 2 group. Ultimately, a likelihood function is maximized with respect to the parameters in $\boldsymbol{\tau}$ and **R**. For maximum likelihood estimation in the multilevel model this function is

$$l(\boldsymbol{\tau}, \boldsymbol{R}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{r`V^{-1}r} - \frac{n}{2}\log(2\pi), \tag{6}$$

where $\boldsymbol{r = y - X(X`V^{-1}X)^{-}X`V^{-1}y}$. Essentially **r** represents residual values from a generalized least squares perspective.

Another common interpretation of the likelihood is the Kullback-Leibler (K-L) divergence; that is, how distant the estimated model is from the true data generating model. Based on the theory proposed by Kullback (1959) the difference in information between models can be defined by the multidimensional integral

$$I(f, g) = \int f(x) log\left(\frac{f(x)}{g(x|\theta)}\right) dx, \tag{7}$$

where *I(f,g)* is the K-L divergence, *f(x)* is the true data generating function, and $g(x|\theta)$ is the approximating function[1]. Assuming that both $f(x)$ and $g(x|\theta)$ are known, it is possible to calculate the exact K-L divergence. However, if one or both are unknown, as is typically the case in psychological research, it is possible to relax the assumption that both functions are known by calculating relative distance. Burnham and Anderson (1998) show that Equation 7 can be written equivalently as

$$I(f,g) = \int f(x)\log\big(f(x)\big)\,dx - \int f(x)\log\big(g(x|\theta)\big)\,dx \quad (8)$$

illustrating that *I(f,g)* amounts to the expectations of the true and approximating functions *f(x)* and $g(x|\theta)$ with respect to the distribution *f*

$$I(f,g) = E_f\big[\log\big(f(x)\big)\big] - E_f\big[\log\big(g(x|\theta)\big)\big]. \quad (9)$$

In any model of the same outcome, the first term, $E_f[\log\big(f(x)\big)]$, is a constant and can be ignored. Eliminating this constant, we are left with the *relative* distance from the hypothesized (approximating) model to the truth $-E_f\big[\log\big(g(x|\theta)\big)\big]$. That this statistic is relative as opposed to absolute is an important property. In reality, we do not know the true data generating model nor do we know the parameters of the approximating model $g(x|\theta)$. Thus, it is necessary to estimate the parameters of the data generating model and instead examine the estimated divergence between the approximating model and the truth. Instead of using the expected K-L divergence between the approximating model and truth, we use expected estimated K-L divergence (Burnham & Anderson, 1998). The maximized log likelihood is often used as a basis for the expected estimated K-L divergence, but is known to have upward bias, especially when the ratio between estimated parameters and sample size is large (Akaike, 1974). As a result, a number of researchers have proposed several bias corrections, some of which have been more widely adopted than others. The most commonly used, and those of central importance in this dissertation, are AIC, and the small sample Akaike's Information Criterion Corrected (AIC_C).

AIC and AIC_C

AIC and AIC_C were developed in an attempt to reduce the upward bias of K-L based statistics. Because the likelihood (and by extension, the log-likelihood) is an upwardly biased estimate of the K-L discrepancy, Akaike (1974) proposed applying a correction equal to twice the number of estimated parameters, *k*. AIC is computed as

$$AIC = -2\log\left(\mathcal{L}\left(\hat{\theta}\middle|y\right)\right) + 2k \quad (10)$$

where $\log\left(\mathcal{L}\left(\hat{\theta}\middle|y\right)\right)$ is the maximized log likelihood (i.e., the upwardly biased K-L discrepancy) and k is the number of estimated parameters in the model. The factor of -2 results from the commonly known result in which multiplying a ratio of two log likelihoods by -2 results in a statistic that is asymptotically distributed as chi-square under certain assumptions (Raudenbush & Bryk, 2002; Singer & Willett, 2003). This quantity is also known as a *deviance* and is the crucial element in the likelihood ratio test. The factor of -2 is also distributed to the correction term and changes its sign positive.

AIC_C has an additional correction intended to improve performance of AIC in small samples, or more precisely when the ratio of estimated parameters to sample size is large. When sample size is small or *k* is particularly large, the correction term ($\frac{n}{n-k-1}$) will increase and the statistic will be calculated as

$$AIC_C = -2\log\left(\mathcal{L}\left(\hat{\theta}\middle|y\right)\right) + 2k\left(\frac{n}{n-k-1}\right). \quad (11)$$

As sample size grows relative to the number of estimated parameters the ratio will approach 1 and AIC_C and AIC will be equivalent. Burnham and Anderson (1998) suggest that AIC_C should be preferred over AIC when there are fewer than 40 observations per estimated parameter.

AIC and AIC_C can be used to qualitatively rank models that are both nested and non-nested. Because the information criteria are relative quantities, their absolute values cannot be interpreted directly and instead the differences of criteria between two models should be evaluated. Thus, it is not possible to determine if any of the models are necessarily plausible candidates for the true model but rather which model has the most supporting evidence. Raftery (1995) provided a rough rule of thumb for evaluating models based on differences in AIC (or

AIC_C). Specifically, a difference between 0-2 is considered "weak" evidence and therefore neither model should be eliminated from consideration. A difference of 2-6 is considered "positive" and lends slight evidence that one model is better than the other. Differences between 6-10 suggest "strong" evidence in favor of the model with the smaller information criterion and a difference greater than 10 provides very strong evidence (Sterba & Pek, 2012).

While Akiake's Information Criteria and its correction are both philosophically and practically intuitive, a major limitation is their failure to account for any sort of variability in estimates. Recall that to calculate the statistic we use the *expected estimated K-L discrepancy* which itself is subject to sampling variability. As a result, simple comparisons between AIC or AICc over repeated samples may not converge on the best model. Additionally, there is no reference distribution on which to evaluate differences in AIC and thus it is not possible to indicate which model is best with any certainty. Bootstrapping methods to create confidence intervals for AIC and related statistics have recently become more popular (Bollen & Stine, 1992; Merkle et al., 2015; Millsap, 2010; Müller, Scealy, & Welsh, 2013; Pornprasertmanit, Wu, & Little, 2013; Preacher & Merkle, 2012; Shang & Cavanaugh, 2008a, 2008b), however due to their relative difficulty to implement, most researchers continue to compare simple differences in information criteria.

BIC

The Schwarz Information Criterion, commonly referred to as the Bayesian Information Criterion, is another commonly used method for model selection. Similar to the AIC, the BIC uses a penalty term, but rather than a fixed proportion, the penalty is scaled by a transformation of the sample size,

$$BIC = -2\log\left(\mathcal{L}\left(\hat{\theta}|y\right)\right) + k(\ln(N)). \quad (12)$$

As the sample size increases, the penalty term for the number of parameters will continue to increase and ultimately become much larger than 2, the coefficient for the penalty in AIC. Thus, in large samples, BIC will tend to prefer more parsimonious models than AIC (that is, when the reduction in the deviance is less than ln(N)).

BIC has been shown to perform reasonably well in avoiding overfitting due to the large penalty for extra parameters, however in multilevel models the penalty term is somewhat ambiguous. In multilevel models the choice for N can be unclear; even for the simplest models there are two choices for N, sample size at Level 1 or at Level 2. Disagreement surrounds the selection of the correct sample size to use in BIC for mixed models and this controversy extends to the output of common statistical programs (McCoach & Black, 2008). For instance, SAS Proc MIXED uses the number of independent sampling units (i.e., highest level sample size) to determine the penalty. This position is supported by Hox (2010), Singer and Willett (2003), and Raftery (1995). Alternatively, SPSS and R use the total number of Level 1 units, a position advocated by Snijders and Bosker (2012). Others advocate for calculating an effective sample size (e.g., Delattre, Lavielle, & Poursat, 2014; Jones, 2011). To examine BIC performance I will examine the BIC based on the total number of independent sampling units to coincide with the logic and sample size of Vuong's LRT.

While on the surface the BIC seems fairly similar to AIC, differing only by the magnitude of the penalty term, they differ drastically in philosophy. Recall that AIC is an estimate of K-L discrepancy in which the main goal is to choose a model that *approximates* the truth. In fact, from the perspective of AIC (and other K-L based measures) it is assumed that no true model can be estimated because it is too complex (Burnham & Anderson, 1998). Given a set

of candidate models, a researcher should be able to distinguish the closest approximation to the truth, even if it is a truly poor model in absolute terms.

BIC, on the other hand, is not based on the K-L discrepancy and instead is rooted in Bayesian philosophy as an approximation to the Bayes factor (Weakliem, 2016). It is assumed that a true model exists, it is part of the candidate set, and the goal of analysis is its identification (Burnham & Anderson, 1998). Additionally, the impetus to develop BIC arose from the desire to find a *consistent* estimator, one that will converge on the "correct" number of parameters as sample size increases (Bozdogan, 1987). Burnham and Anderson (1998) note that the assumptions surrounding BIC are particularly absurd. Assuming that the researcher *knows* the true model is only a small step from the assumptions that a simple true model exists and is in the candidate set. Thus they have questioned the BIC's utility. They also note that Monte Carlo studies that have been conducted with the BIC tend to adhere to these untenable assumptions and as a result are not very informative. Following their suggestion, one part of this dissertation will examine candidate model sets excluding the true model (Burnham & Anderson 1998; p. 287).

Likelihood Ratio Test

The likelihood ratio test (LRT) has a long history of testing nested models. Virtually every textbook on data analysis topics from multiple regression to multilevel modeling to structural equation modeling has sections at least touching on the topic (Cohen et al., 2002; Hox, 2010; Kline, 2015; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 2012). As noted above, the deviance ($-2\log\left(\mathcal{L}\left(\hat{\theta}|y\right)\right)$) is a measure of lack of fit between the model and the data. This quantity lacks an absolute meaning and must be evaluated relative to another model's deviance. Comparing two nested models typically involves a simple calculation with a known referent chi-square distribution. Procedures for testing two non-nested models are

computationally more involved and not widely known. When two models are nested they are said to be related to one another in such a way that the more parsimonious model (i.e., restricted model) can be derived from the larger model (i.e., full model) via a set of constraints. Typically when testing models in their fixed effects, nesting occurs by constraining the effects of certain variables to 0. More common in nested random effects structures (although still permissible in fixed effects) are equality constraints where variances are constrained to be equal. A full explanation of nested models and examples will be provided in a subsequent section.

To conduct the LRT for nested models, one must simply estimate the full and restricted models and subtract the full model's deviance from the restricted. Because the full model will always fit better than the restricted model, (i.e., the restricted model will have a higher deviance) the resulting quantity characterizes the decrease in fit of the restricted model relative to the full model. The deviance difference is then evaluated on a chi-square distribution with degrees of freedom equal to the difference in the number of estimated parameters of the two models. A significant result indicates that the restricted model fits more poorly and thus the more complete model should be preferred.

Echoing the views of Raftery (1995), McCoach and Black (2008) suggest that the largest drawback of the LRT is that it can *only* be used to test nested models. While this widely held belief has proliferated the use of information criteria to compare two dissimilar (i.e., non-nested) models, it is ultimately unfounded. Vuong (1989) developed a two-step likelihood ratio test to compare two non-nested models. The first step determines the nestedness of two candidate models and the second compares the difference in likelihood ratios of the two candidate models to the weighted variance of the *individual-specific* (i.e., independent unit) log likelihood ratios. When models are not nested, the limiting distribution of the deviance difference is no longer

distributed as a Chi-square with degrees of freedom equal to the difference in model parameters and instead can be evaluated on a standard normal distribution. Vuong's original exposition focused on linear regression models, however, Rivers and Vuong (2002) later expanded it to test non-nested time series models and most recently Merkle et al. (2015) applied the test to non-nested structural equation models. In the following sections I revisit Vuong's original logic and its adaptation to structural equation modeling and in the following chapter I explain how the test can be applied to multilevel models from both mathematical and practical perspectives.

Vuong's (1989) Likelihood Ratio Test

The LRT as proposed by Vuong (1989) is actually a two-step procedure where, in the first step, a researcher tests if models are indistinguishable, nested, non-nested, or overlapping and in the second step the hypothesis test is performed comparing the fit between two distinguishable (i.e., not indistinguishable) models. The initial test, often referred to as the test for distinguishability, is necessary to determine the limiting distribution of the test statistic. To conduct this test, variances of the differences in individual-specific log likelihoods are evaluated on weighted mixtures of chi-squares (Golden, 2000; Levy & Hancock, 2007; Merkle et al., 2015). Models that are deemed indistinguishable do not need to be tested, as they are equivalent in the population. As noted by Merkle et al. (2015), if one knows a priori that models are nested, non-nested, or overlapping, one can proceed directly to the appropriate LRT.

Before formally defining the test of distinguishability, it is important to understand the possible relationships among models. While Levy and Hancock (2007) discuss model relationships in the context of mean and covariance structures, they also provide a more general visualization of possible relationships. I recreate this here and provide corresponding situations in multilevel models. The four panels in Figure 1.1 reflect the potential relationships between

two models. In panel (i) the two models are completely overlapping such that their relationship to the data is equivalent. Levy and Hancock relate completely overlapping models to those that produce the same mean and covariance structures regardless of population. However, in the multilevel context, two completely overlapping models may be those involving variable transformations. For instance, if model A was fit with a raw variable indicating number of drinks per month over the last year in a population with substance use disorder and model B were fit with that same variable grand mean centered, the two models could be considered completely overlapping as the transformation would not change the model fit in the population. From the figure, it is obvious that each model fits no better than the other and thus no hypothesis test should be conducted.

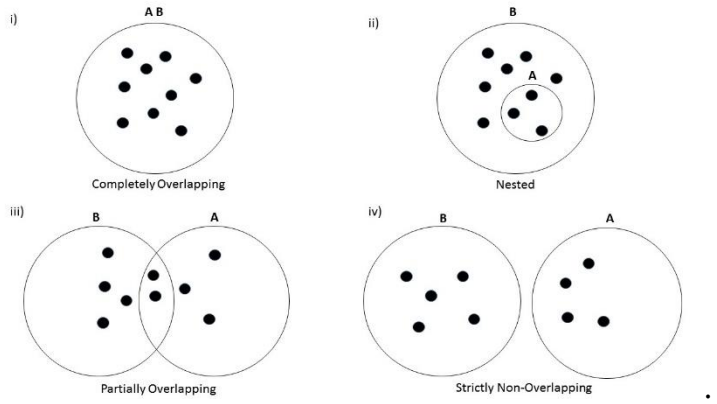


Figure 1.1 Graphical depiction of possible relationships between two models. (i) Completely Overlapping (ii) Nested (iii) Partially Overlapping (iv) Strictly Non-Overlapping

Panels ii, iii, and iv depict models that are distinguishable in the population but have very different relationships to one another. Panel ii shows model A nested within model B. These models are typical in multilevel analyses and are often used for multiparameter significance tests of fixed effects or variance components via nested LRT. For instance, model B contains predictor variables X_1, X_2 … , X_n. Model A is fit to the same outcome but only includes

predictors $X_1, X_2, \ldots, X_{n-2}$ and as a result model A has two fewer parameters to estimate. Model A is considered nested in Model B because Model B can be transformed into Model A by imposing a set of restrictions, namely setting coefficients of X_n and X_{n-1} to zero.

Panel iii shows partially overlapping models. In this scenario, two similar models with slight differences are fit to data. Levy and Hancock (2007) explain these as models that "share some distributions but each contains unique distributions" (p. 47). A more intuitive explanation is provided by Merkle et al. (2015): Figure 1.2 displays a two factor model each with three indicators. Let model A represent the case in which there is an additional path from latent variable η_1 to X_4 (denoted by the dashed line and labeled "A") and model B represent the case in which there is an additional path from latent variable η_2 to X_3 (denoted by the dotted line and labeled "B"). These models are considered overlapping because their predictions and fit statistics will be the same in populations where paths A and B are both zero (Merkle et al., 2015). A similar situation can be found in multilevel models in which there exists a common covariate set across models but competing substantive variables driving a theory (or vice versa). For instance, in a study of the effect of positive affect on inflammation in breast cancer patients (Moreno, Moskowitz, Ganz, & Bower, 2016), we had to make a qualitative decision as to which of two highly related Level 2 covariates to use, fatigue severity or fatigue interference, holding the rest of the model constant.

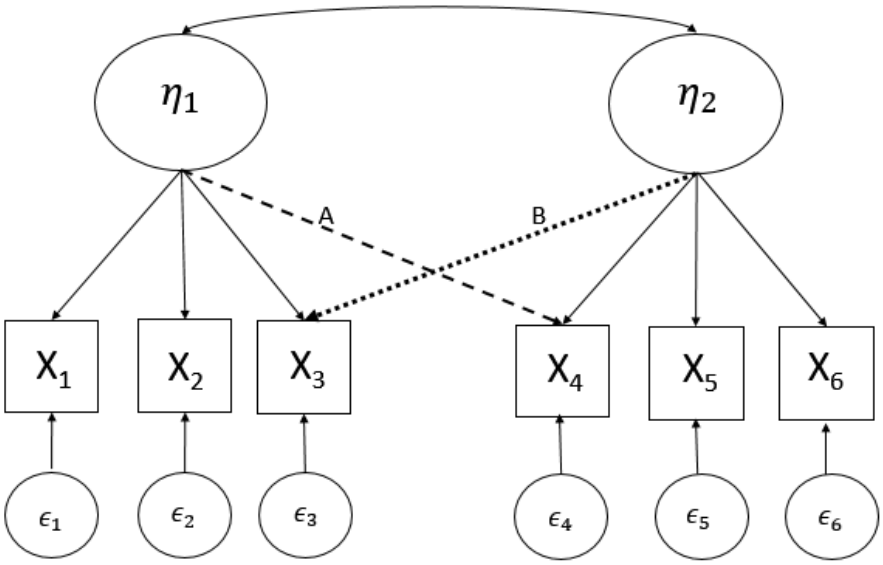


Figure 1.2. Two factor model overlapping with paths A and B.

The final panel depicts the case in which two models are strictly non-overlapping. Models which are non-overlapping will never result in the same fit indices or model implied moments (Merkle et al., 2015). In other words, "if no set of constraints allows for a solution, the models are completely non-overlapping" (Levy & Hancock, 2007). While it is sometimes difficult to think of a case in which models would be truly non-overlapping in the population, differing functional forms of growth is one instance where these types of models are common. Consider the equations for simple power and exponential growth curves,

$$Power: y = \beta_0(Time^{\beta_1}) + e \tag{13}$$

$$Exponential: y = \beta_0\left(e^{\beta_1 * Time}\right) + e. \tag{14}$$

In equation 13 β_0 reflects the value of y at time 1 and β_1 controls the concavity of the function (Timmons & Preacher, 2015). In equation 14 β_0 is the intercept (i.e., when time = 0) and β_1 is the exponential growth rate. These models contain the same number of parameters, however, their forms of growth are markedly different as are the definitions of their intercept analogues

(β_0). In these cases it would be possible to skip the test of distinguishability and proceed directly to Vuong's LRT for non-nested models.

Vuong's LRT assumes that in addition to i.i.d. observations, second derivatives of the likelihood function exist, ML estimates are unique and not on the boundary, and the variance in the individual likelihood ratios between full and restricted models is non-zero (Golden, 2000; Merkle et al., 2015; Vuong, 1989). Under the assumption of the null hypothesis, the two candidate models cannot be distinguished in the population and thus should not be tested against one another. Merkle et al. (2015) formalize this test with null and alternative hypotheses

$$H_0: \omega_*^2 = 0 \tag{15}$$

$$H_1: \omega_*^2 > 0 \tag{16}$$

respectively. An estimate of ω_*^2 is given by

$$\widehat{\omega}_*^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{f_A(x_i;\hat{\theta}_A)}{f_B(x_i;\hat{\theta}_B)}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{f_A(x_i;\hat{\theta}_A)}{f_B(x_i;\hat{\theta}_B)}\right]^2 \tag{17}$$

where $f_A(x_i;\hat{\theta}_A)$ is the individual-specific likelihood function for model A and the individual-specific likelihood for model B is defined similarly (Golden, 2000; Levy & Hancock, 2007; Merkle et al., 2015; Vuong, 1989). Multiplying $\widehat{\omega}_*^2$ by the number of cases (n) produces a statistic that can be evaluated on a weighted chi-square distribution with the weights arising from the squared eigenvalues of the second derivative and information matrices from the candidate models (Merkle et al., 2015; Vuong, 1989). If the models are indistinguishable then the variability in the individual log likelihood ratios should be close to zero and the null hypothesis retained. Such models should not be candidates for further comparison. Conversely, rejecting the null hypothesis would indicate that the variability in individual log likelihoods across the two

models is non-zero allowing for a statistical test between models (Merkle et al., 2015). Because of this test's reliance on non-standard output and complex computations, Levy and Hancock (2007) took an analytic approach to testing distinguishability originally proposed by Raykov and Penev (1999). Briefly, this method requires the user to set up systems of equations to determine if parameter matrices are transformations of one another via linear algebra. Even though it does not rely on software output, the algebraic method is not easily implemented by applied researchers. Additionally, it lacks a certain generality inherent to Vuong's original proposal and does not consistently perform well (Merkle et al., 2015). As a result, Merkle and You (2014) and Merkle et al. (2015) have created software in which these tests are implemented for SEMs.

Upon establishing that models are in fact distinguishable, researchers can implement a nested or Vuong's non-nested LRT. Formally, when performing this test a researcher poses the hypotheses

$$H_0: E\left[l\left(\widehat{\theta_A}; x_i\right)\right] = E\left[l\left(\widehat{\theta_B}; x_i\right)\right] \tag{18}$$

$$H_{1A}: E\left[l\left(\widehat{\theta_A}; x_i\right)\right] > E\left[l\left(\widehat{\theta_B}; x_i\right)\right] \tag{19}$$

$$H_{1B}: E\left[l\left(\widehat{\theta_A}; x_i\right)\right] < E\left[l\left(\widehat{\theta_B}; x_i\right)\right] \tag{20}$$

where expected values of each likelihood constitute the expected K-L distance discussed earlier. The null hypothesis H_0 posits that the K-L distances between the truth and models A and B are equal. The alternative hypothesis H_{1A} posits that the expected likelihood of model A is greater (i.e., closer to the truth in K-L discrepancy) than the likelihood of model B and therefore model A should be preferred. Conversely, H_{1B} suggests that the expected likelihood of model B is greater than model that of model A and therefore model B should be preferred. Conclusions drawn from this hypothesis test are that either model A should be preferred to model B, model B

should be preferred to model A, or there is not sufficient evidence to support one model over the other.

To test these hypotheses for non-nested distinguishable models the statistic

$$LR_{AB} = n^{-\left(\frac{1}{2}\right)} \sum_{i=1}^{n} \log \frac{f_A\left(x_i; \hat{\theta}_A\right)}{f_B\left(x_i; \hat{\theta}_B\right)} \xrightarrow{d} N(0, \omega_*^2) \tag{21}$$

can be used under the assumption of H_0. The statistic is computed by first extracting the log likelihoods for each *observation* under the two candidate models A and B. The log of the likelihood ratio for each case is then summed over the number of observations, divided by its square root and evaluated on a normal distribution (Golden, 2000; Merkle et al., 2015). Golden (2000) and Merkle et al. (2015) show that this result can also be used to test nested models under an alternative limiting distribution, however, the LRT for nested models is not of interest here.

While Vuong's LRT has been expanded on to accommodate a variety of common economic models (e.g., incompletely specified models, a variety of estimators, alternative model selection procedures, and nonlinear dynamic data [Rivers and Vuong 2002]; time series models, [Golden 2000]) its use has until recently been restricted to single level univariate models. The fairly technical treatment of the test in the economic literature, in addition to its reliance on non-standard output from statistical packages, has created substantial barriers to its adoption outside of economics. Levy and Hancock (2007, 2011), Merkle and You (2014), and Merkle et al. (2015) have all contributed to expanding the test's use to SEMs and presenting it to substantive researchers in psychology. A notable exception has been its application to multilevel models, both latent and otherwise. Although the assumptions of Vuong's test, specifically i.i.d. observations, may imply that the test is not suitable for multilevel data, recent discussions in the literature paint a different picture.

Multilevel Models and SEM

The proliferation of multilevel SEM (MSEM) and latent growth curve models have continued to blur the lines between SEM and multilevel models (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005; Mehta & West, 2000). In virtually all scenarios, multilevel models can be reparametrized in the SEM framework (Curran, 2003; Mehta & Neale, 2005). Although it is usually more convenient to work in one framework over another depending on the nuances of research questions and data, the parallels between MLM and SEM can be leveraged to facilitate the application of Vuong's LRT to multilevel models.

To illustrate, consider the case of a multilevel growth curve model and a latent growth curve model. For 500 persons each with 4 time points, data are usually structured in one of two ways: "long" as is typically utilized in multilevel models or "wide" as is typical in structural equation models. Examples of these data types can be found on Table 1.1. While data formats do not make a model, in this context they serve to help one orient to the necessary unit of analysis. In multilevel models, the unit of analysis is the observation. Correspondingly, each row in the long dataset represents an observation, to whom it belongs is indexed by another column, and time is indexed by yet another column. In SEM the unit of analysis is the individual and as such each row corresponds to an individual with each observation represented by a new variable. For the multilevel model, consider the standard unconditional linear growth model of the form

$$\text{Level 1: } y_{ti} = \beta_{0i} + \beta_{1i} Time_{ti} + e_{ti} \tag{22}$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + u_{0i} \tag{23}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \tag{24}$$

where observation y at time t is nested within person i. β_{0i} and β_{1i} are person-specific intercept and slope parameters, respectively, and γ_{00} and γ_{10} are the corresponding population estimates.

u_{0i} and u_{1i} are the random effects for the person-level intercept and slope parameters, respectively. Finally, e_{ti} is the time-point-specific error term. Typically, assumptions are placed on the random effects and residuals

$$e_{ti} \sim N(0, \sigma^2 \boldsymbol{I}_n) \tag{25}$$

$$u_{.i} \sim N(0, \boldsymbol{\tau}). \tag{26}$$

The random effects and residuals are assumed to be independent from one another as well as from any fixed effects. Although homogeneity of variance (i.e., $\sigma^2 \boldsymbol{I}_n$) is initially assumed at Level 1, this assumption can be and often is relaxed in longitudinal data. The covariance structure at Level 2 is typically unstructured, but once again this is not generally required.

Table 1.1 Data structure to estimate as a mixed model (left) or latent growth curve (right).

ID	Time	Y
1	1	8
1	2	7
1	3	6
1	4	5
2	1	10
2	2	9
2	3	7
2	4	4
3	1	8
3	2	7
3	3	8
3	4	6
.	.	.
.	.	.
.	.	.
500	1	6
500	2	3
500	3	4
500	4	5

ID	Time 1	Time 2	Time 3	Time 4
1	8	7	6	5
2	10	9	7	4
3	8	7	8	6
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
500	6	3	4	5

Curran (2003) rewrites the above equations using matrices to facilitate the comparison between multilevel models and SEM. He writes the Level 1 equation from the multilevel model (here equation 22) as

$$\boldsymbol{y_i} = \boldsymbol{X\beta_i} + \boldsymbol{r_i} \tag{27}$$

Where the first column of the design matrix $\boldsymbol{X}$ is a vector of 1s to indicate an intercept and (with a single predictor time) the second column contains the individual-specific measures of time for each observation. The Level 2 equation can be rewritten as

$$\boldsymbol{\beta_i} = \boldsymbol{\Gamma} + \boldsymbol{u_i}. \tag{28}$$

for the unconditional example model (i.e., with no Level 2 predictors). Although it is not commonly explicated in discussions of multilevel models, the model does imply a mean and covariance structure

$$\boldsymbol{\mu_y} = \boldsymbol{X\Gamma} \tag{29}$$

$$\boldsymbol{\Sigma_{yy}} = \boldsymbol{X\tau X'} + \boldsymbol{\Sigma_r}, \tag{30}$$

Where $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma_r}$ are Level 2 and Level 1 covariance matrices, respectively (Curran, 2003). A SEM can be defined similarly as

$$\boldsymbol{y_i} = \boldsymbol{\Lambda\eta_i} + \boldsymbol{\epsilon_i} \tag{31}$$

$$\boldsymbol{\eta_i} = \boldsymbol{\mu} + \boldsymbol{\zeta_i} \tag{32}$$

With implied mean and covariance structures as

$$\boldsymbol{\mu(\theta)} = \boldsymbol{\Lambda\mu} \tag{33}$$

$$\boldsymbol{\Sigma(\theta)} = \boldsymbol{\Lambda\Psi\Lambda'} + \boldsymbol{\Theta_{\epsilon'}}. \tag{34}$$

As Curran explains, Equations 33 and 34 parallel Equations 29 and 30 where the matrices in one can be substituted for the other such that

$$\boldsymbol{X} = \boldsymbol{\Lambda} \tag{35}$$

$$\boldsymbol{\tau} = \boldsymbol{\Psi} \tag{36}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{r}} = \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \tag{37}$$

$$\boldsymbol{\Gamma} = \boldsymbol{\mu}. \tag{38}$$

Furthermore, while in multilevel models time is a predictor variable included in matrix $\boldsymbol{X}$, it is fixed as a factor loading in matrix $\boldsymbol{\Lambda}$ in SEM.

While the above explanation discusses growth models exclusively, the equivalence between multilevel and SEM extends to the more general case as well. Mehta and Neale (2005) explain in detail the notion of univariate multilevel models as multivariate "unilevel" models and provide a number of examples fitting a series of increasingly complex multilevel models as confirmatory factor analysis (CFA) models. Doing so requires a shift in perspective away from the typical multilevel model to an alternate unit of analysis. Normally when fitting multilevel models it is typical to consider each outcome as indicative of each person's score within a group. In SEM, each "construct is assessed with *p* equivalent and exchangeable tests." (Mehta & Neale, 2005 p. 264). The shift in perspective comes from viewing individuals within groups as indicators of a construct (or deviations from a mean) rather than individual sampling units. In a random intercept model, the CFA partitions variance as between group (i.e., common variance) and within group (i.e., unique variance) similar to the variance partitioning in multilevel models. Additionally, unique variances are conditionally independent based on the common factor just as within group observations are independent conditional on group membership. Random slopes are easily included by the addition of other latent variables with mean structures and factor loadings fixed to the variables' observed values.

Similarities between multilevel models and SEM have motivated some researchers to develop fit indices for both frameworks in tandem. Specifically, Sterba and Pek (2012) proposed three diagnostics for individual influence on model selection. That is, they developed statistics that quantified individual *case* contributions to chi squares, AIC, BIC, and their differences. To equate the multilevel and SEM approaches they explicitly defined a *case* as "the highest level unit in an analysis." While this may seem counterintuitive in the multilevel context where we typically think of each lower level observation as a case, the distinction in SEM is obvious. I will show in Chapter 2 specifically why cases are conceptualized as the highest level of a multilevel analysis via the mechanics of maximum likelihood estimation and why the individual observation (i.e., Level 1) contributions cannot be obtained. However, based on the evidence provided by Curran (2003), Bauer (2003), Mehta and West (2000), Mehta and Neale (2005), and the proof of concept in Sterba and Pek's (2012) case contribution statistics, it follows that extension of Vuong's (1989) LRT should generalize from SEM to multilevel modeling as well.

Despite the equivalence between multilevel models and SEM, fitting models using one framework over another might provide a more elegant solution in practice. Further, while much of what is possible in multilevel models may be achievable in SEM, it may still be more intuitive, in specification and interpretation, to estimate models in the multilevel framework. For instance, categorical outcomes can be included in both modeling frameworks, however, only multilevel models use a link function that facilitates estimation (Curran, 2003). Multilevel models also benefit from straightforward specification of Level 1 covariance structures, interaction effects, nonlinear effects, and multiple random effects. Finally, SEM models require that the effects of predictors be separated into their between and within group (or individual) effects. While such orthogonality facilitates interpretation, it may not be realistic in practice.

Although these are only a few of the reasons that a researcher might prefer multilevel models over SEM, and there exist several more occasions where the preference would be reversed, it is clear that multilevel modeling continues to provide utility for the applied researcher and thus a test for non-nested models in this framework would be useful.

Studies of AIC/BIC in multilevel models

Without the option of a null hypothesis statistical test (NHST) to compare non-nested multilevel models, researchers have historically been restricted to simple comparisons among information criteria. Although numerous options exist, all with different penalty functions, the most commonly used in multilevel models are AIC, AIC_C, and BIC. Motivating the development of so many information criteria (e.g., Akaike, 1974; Bozdogan, 1987; Hannan & Quinn, 1979; Hurvich & Tsai, 1989; Schwarz, 1978) is that the *best* information criterion (and by extension the best penalty function) to use for model selection remains an open question (Vallejo, Ato, & Valdés, 2008; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011; Vallejo, Tuero-Herrero, Núñez, & Rosário, 2014).

Most studies tend to agree that while there is no information criterion that rises above the rest in every scenario, AIC and AIC_C are consistently among the top performing (Dimova, Markatou, & Talal, 2011; Pu & Niu, 2006; Vallejo et al., 2008; Vallejo et al., 2011; Vallejo et al., 2014). Despite the inconclusive nature of many simulation studies exploring information criteria in multilevel models, researchers have found that many of the same factors that contribute to power in NHST also contribute to correct model selection.

Perhaps the finding most consistent with the multilevel power literature is that the greatest effect on model selection comes from the sample size at the highest level (Vallejo et al., 2011; Vallejo et al., 2014; Wang & Schaalje, 2009). In several studies, researchers found that

information criteria improved in selecting the correct model based on differences in fixed effects as the sample size at the highest level increased (Vallejo et al., 2011; Vallejo et al., 2014; Wang & Schaalje, 2009). When testing differences in random effects, however, increases in the number of observations per individual had substantial impact on the performance of the information criteria (Vallejo et al., 2014). Finally, different sample-size-based penalties offered the best performance in AIC_C and BIC. It is recommended that the upper level sample size be used as a penalty for the BIC, whereas the total number of observations should be used for the AIC_C penalty term (Vallejo et al., 2011).

ICC also played a role in the accuracy of model selection, albeit a smaller one than sample size. When ICC values were small (roughly .1) the information criteria performed better than when ICC values were high, but only when random effects were uncorrelated (Vallejo et al., 2014). Similarly, BIC outperformed AIC and AIC_C when random effects were uncorrelated, but in the presence of correlated random effects AIC and AIC_C more consistently preferred the correct model. Differences in performance between information criteria conditional on the correlation between random effects is a fairly novel finding. Generally in power studies, the relationship between intercept and slope variances is considered inconsequential. Perhaps, the association among random effects plays an important role in the performance of information criteria.

Finally, contributing to the debate surrounding the use of information criteria in model selection is whether the analysis should be conducted under FML or REML. In addition to the debate over *which* effects, fixed or random, can be tested, the dichotomy raises questions (and produces new indices) regarding the quantities that should be included in the penalty terms. Several studies (e.g., Vallejo et al., 2008; Vallejo et al., 2011; Vallejo et al., 2014; Wang &

Schaalje, 2009) indicated that information criteria tended to perform better when models were estimated under REML as opposed to FML. Furthermore, this finding held even when models differed in *fixed effects.* It follows logically that the information criteria perform well when comparing random effects under REML as the variance components estimates are less biased. However, the reason for improved performance of information criteria in selecting the fixed effects under REML remains an open research question.

Still, with the plethora of information criteria available and the uncertainty that surrounds their performance, researchers continue to try to correct for apparent biases. Pu and Niu (2006) developed a generalized version of AIC and BIC for multilevel models called the GIC which intended to allow for increased flexibility in penalty functions. Their study found that the new criteria performed well for the selection of fixed effects but not random effects. Greven and Kneib (2010) derived conditional and marginal corrections for the AIC but failed to study its performance. Analytically it appears to be unbiased, however, the conditions in which that comes to fruition are undetermined.

Other researchers take an empirical approach to correcting information criteria. A number of studies (Kitagawa & Konishi, 2010; Shang & Cavanaugh, 2008a, 2008b; Shibata, 1997) use bootstrapping to correct bias in the estimate of K-L discrepancy. Bootstrapping is an attractive option for information criteria in that the technique can be applied to many different types of models and requires very few assumptions, in general. In multilevel models, bootstrapping is somewhat of an enigma however, in that typical approaches tend to make it difficult to recreate the Level 2 variance structure (Goldstein, 2011). The fully parametric bootstrap is typically preferred if model assumptions are accepted (Goldstein, 2011). When model assumptions are not accepted, the modified residual bootstrap (Goldstein, 2011) is preferred. Although useful,

bootstrapping is rarely applied because of substantial computing time needed to create the bootstrapped samples (e.g., Dimova, 2011).

Shang and Cavanaugh (2008a, 2008b) formalize two bootstrapped approaches to adjust bias in the AIC for multilevel models. The results presented in the papers are based on the parametric bootstrap assuming normality of error terms, however semiparametric and nonparametric bootstrap approaches requiring fewer assumptions are possible. The authors found that the bootstrapped criteria outperform AIC without bootstrapping in selecting the best model, especially when sample sizes are small. Importantly, the bootstrapped statistics previously examined do not address differences in random effects structures, only fixed effects.

It is clear from this brief review that model selection based on information criteria is still very much an open question. Despite the recent disparagement of NHST, it would be desirable to have a statistic that reports with some degree of certainty that one model should be preferred over another. Thus, I propose the following dissertation.

Project Aims

In this dissertation I first discuss in detail how Vuong's LRT extends to the multilevel framework by exploring how changes in model specification manifest in the likelihood. In order to provide a comprehensive explanation of how model differences common to those seen in non-nested candidate models affect the likelihood, I will discuss different covariate sets including differences in fixed effects at Level 1 and Level 2, non-nested Level 1 random effect structures, and non-linear functional forms of growth. The following three chapters explore the performance of Vuong's test in the three common types of non-nested models listed above and evaluate its performance relative to AIC, AIC_C, and BIC. I will conclude with three empirical examples, one

in which fixed effects are non-nested, another comparing functional forms, and the other in which Level 1 random effect structures are non-nested.

Chapter 2: Vuong's Test for Multilevel Models

In this chapter I build the argument for applying Vuong's test for non-nested models to the multilevel framework. To this end, I begin by reintroducing Vuong's likelihood ratio statistic and the multilevel log likelihood and explain computationally why cases at Level 2 must be the unit of analysis. I then examine how differences in model specification that might cause two models to be non-nested manifest in the log likelihood regardless of the level at which they occur; first for fixed effects in covariate sets, then for Level 1 random effects, and finally for non-linear trends. I walk through a short example illustrating subtle differences in the computations of individual log likelihoods for non-linear models specifically. Finally, I show empirically that the individual log likelihoods and the results of Vuong's Test are the same for both multilevel and latent variable models for growth models where observations are ordered (non-exchangeable) and when individuals are nested within groups (exchangeable observations). Through induction this equality serves as the basis on which Vuong's test for multilevel models can be extended to more complex cases when one modeling framework might be preferred to the other.

Vuong's Likelihood Ratio

As outlined above, Vuong's test for non-nested models is effectively a z-test contrasting the individual-specific likelihood ratios of two competing models. For two models, A and B, each case's contribution to the log-likelihood is calculated and the differences across models computed. The variance of these differences is scaled by $\frac{n-1}{n}$ resulting in $\widehat{\omega}^2$, where n is equal to the number of individual cases, which is then included in the equation of the test statistic itself,

$$Vuong_{LR} = \left(\frac{1}{\sqrt{n}}\right) * \frac{\sum(indLL_{m1} - indLL_{m2})}{\sqrt{\widehat{\omega}^2}}. \quad (39)$$

Vuong's LR statistic is then evaluated on a standard normal distribution. If Vuong's LR falls beyond the 95% confidence interval about 0, then the model with the greater log likelihood is preferred. That is, the model with the lower K-L discrepancy is thought to be a better approximation of the truth. Integral to calculating this test statistic are the individual case contributions to the log likelihood, as the variance of their differences between models, as well as the differences themselves, are used in the likelihood ratio statistic.

Multilevel Log Likelihood

Estimating a mixed model via maximum likelihood can be accomplished with a number of estimators. Two common approaches used to solve the linear mixed model problem are penalized likelihood (PLS; used in the lme4 package in R) and generalized least squares (GLS; used in SAS). Although Vuong's test can be applied to any maximum likelihood estimator (Merkle et al., 2015), in this study I embrace the GLS approach due to its implementation in SAS and previous work explicating the extraction of the case-wise likelihood using SAS's Proc IML (Mistler, 2013). Recall the log-likelihood expression provided in equation 6

$$l(\boldsymbol{\tau}, \boldsymbol{R}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} - \frac{n}{2}\log(2\pi),$$

.

and the equations for $\boldsymbol{V}$ and $\boldsymbol{r}$,

$$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\tau}\boldsymbol{Z}' + \boldsymbol{R},$$

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.$$

It is worth reiterating that $\boldsymbol{V}$ represents the total variance in outcome $\boldsymbol{y}$ attributable to the random effects and is n x n block diagonal with each block representing an independent case (i.e., Level 2 group). $\boldsymbol{Z}$ is the n x p known random effects design matrix, $\boldsymbol{\tau}$ is the p x p random effects variance covariance matrix, and $\boldsymbol{R}$ is the n x n Level 1 residual variance matrix. Additionally, $\boldsymbol{r}$ is an n x 1 vector of residual values from a generalized least squares perspective.

$\boldsymbol{R}$ and $\boldsymbol{r}$ are distinct quantities with the former referring to a (co)variance matrix and the latter a vector of residual values. Because *cases* (i.e., Level 2 units) are independent of one another, the joint likelihood of the Level 2 units can be calculated by the product of the cases' individual likelihoods. However, working with joint likelihoods on a scale larger than only the smallest datasets quickly causes computational issues. Instead it is more practical and common to use the log likelihood, which is additive.

Individuals as Cases

Because the likelihood calculation involves the determinant of the variance-covariance matrix of the outcome, **V**, it requires that the input be a square matrix. Generally, matrix $\boldsymbol{V}$ is a square, block diagonal matrix and is used in its entirety during model estimation. However, due to the independence of the individual cases it is possible to compute the individual-specific log likelihoods by executing the computations iteratively for each block, since each block is also square. If instead the "individual cases" were thought of in a more traditional multilevel sense, with the unit of analysis at Level 1, it would be impossible to calculate the determinant because $\boldsymbol{V}$ would be a scalar and an alternative estimation approach would be required. Further, treating the Level 1 units as the individual cases would fail to preserve the complexity of the multilevel data; specifically, the information shared among members of the same group would be lost as there is no place to express the covariance among units. The assumption of independence at the greatest sampling unit implies that no information should be lost from treating each block in $\boldsymbol{V}$ by itself, as cells in the n x n matrix outside of the blocks are necessarily zero under the assumption of conditional independence. Therefore, by treating each Level 2 unit as an observation it is possible to calculate $l(\boldsymbol{\tau}, \boldsymbol{R})$ for each case individually using the above equations sacrificing only computational time.

It is also possible to think about the level of a "case" intuitively. Considering the variance covariance structure of multilevel data, for a random intercept only model we can assume each block has a structure of

$$\boldsymbol{\Sigma_n} = \begin{bmatrix} \tau + \sigma^2 & \tau & \tau & \tau & \tau \\ \tau & \tau + \sigma^2 & \tau & \tau & \tau \\ \tau & \tau & \tau + \sigma^2 & \tau & \tau \\ \tau & \tau & \tau & \tau + \sigma^2 & \tau \\ \tau & \tau & \tau & \tau & \tau + \sigma^2 \end{bmatrix} \tag{39}$$

and $\boldsymbol{V}$ is made up of independent blocks of $\boldsymbol{\Sigma}$ such that,

$$\boldsymbol{V} = \begin{bmatrix} \Sigma_1 & 0_{\dots} & 0_n \\ 0_{\vdots} & \ddots & 0_{\vdots} \\ 0_n & 0_{\dots} & \Sigma_n \end{bmatrix}. \tag{40}$$

First, focusing on matrix $\boldsymbol{\Sigma}$ it is easy to understand why treating a single data point as the unit of analysis would be problematic for the multilevel model. Assuming for a moment that the likelihood was not dependent on calculating the determinant and thus did not require a square matrix rather than a scalar, it would appear that focusing on a single data point (i.e., a single diagonal element of matrix $\boldsymbol{\Sigma}$) would ignore information regarding the relationships among data within group. Furthermore, even if there were some way to account for interrelatedness of individual observations, absent a block diagonal structure (and related observations scattered around the design matrix) computational time would inevitably be sacrificed as relationships among many more observations would need to be estimated and evaluated. Instead, the block diagonal structure of the $\boldsymbol{V}$ matrix imposes an intuitively and computationally satisfying solution.

Effects of Interest and the Likelihood

While the logic of treating Level 2 units as "cases" is sound, it is still important to understand analytically how different specifications of the multilevel model might affect the likelihood calculations. I now explain how differences in the three specifications of models where non-nestedness is explored in this dissertation (different covariate sets, Level 1 residual covariance matrices, and non-linear models) enter the likelihood resulting in differential model fit.

Covariate sets. The mechanics of how non-nested fixed effects alter the likelihood are straightforward. Every time a predictor is added or replaced in the model, the values in the corresponding column(s) of $\boldsymbol{X}$ are changed. In actuality $\boldsymbol{X}$ does not discriminate among the levels at which different variables enter the model and thus the effect, at least computationally, of changing the fixed effects structure of the model is the same regardless of level. Differences in $\boldsymbol{X}$ are localized to only the second term in the likelihood via $\boldsymbol{r}$.

Level 1 Covariance. In the example of $\boldsymbol{\Sigma}$ above, $\boldsymbol{R}$ is implied to be diagonal with covariances among observations resulting only from the Level 2 units, however, alternative structures can be easily incorporated into the likelihood. For instance, if instead the structure were autoregressive with a lag of 1 $\boldsymbol{\Sigma}$ would take the form

$$\boldsymbol{\Sigma_n} = \begin{bmatrix} \tau+\sigma^2 & \tau+(\sigma^2*\rho) & \tau+(\sigma^2*\rho^2) & \tau+(\sigma^2*\rho^3) & \tau+(\sigma^2*\rho^4) \\ \tau+(\sigma^2*\rho) & \tau+\sigma^2 & \tau+(\sigma^2*\rho) & \tau+(\sigma^2*\rho^2) & \tau+(\sigma^2*\rho^3) \\ \tau+(\sigma^2*\rho^2) & \tau+(\sigma^2*\rho) & \tau+\sigma^2 & \tau+(\sigma^2*\rho) & \tau+(\sigma^2*\rho^2) \\ \tau+(\sigma^2*\rho^3) & \tau+(\sigma^2*\rho^2) & \tau+(\sigma^2*\rho) & \tau+\sigma^2 & \tau+(\sigma^2*\rho) \\ \tau+(\sigma^2*\rho^4) & \tau+(\sigma^2*\rho^3) & \tau+(\sigma^2*\rho^2) & \tau+(\sigma^2*\rho) & \tau+\sigma^2 \end{bmatrix}. \quad (41)$$

As each block of $\boldsymbol{\Sigma_n}$ is of the same dimensions as $\boldsymbol{Z\tau Z`}$ for a specific Level 2 unit, any estimable Level 1 covariance structure can be seamlessly included into the likelihood calculation. As the error structure better approximates the data generating process, the likelihood of the data given

the model would be expected to increase and the K-L discrepancy decrease indicating a better fitting model. It is worth pointing out that the structure of $\boldsymbol{R}$ may be particularly important given its place in the likelihood. Not only does the Level 1 error matrix $\boldsymbol{R}$ influence $\boldsymbol{V}$ which shows up explicitly in both the first and second term of the likelihood, but it also helps to define the residual vector $\boldsymbol{r}$. Substituting the equation for **V** into **r**, the expanded form becomes:

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'(\boldsymbol{Z}\boldsymbol{\tau}\boldsymbol{Z}' + \boldsymbol{R})^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'(\boldsymbol{Z}\boldsymbol{\tau}\boldsymbol{Z}' + \boldsymbol{R})^{-1}\boldsymbol{y}. \quad (42)$$

Again, $\boldsymbol{R}$ (in addition to the random effects design matrix $\boldsymbol{Z}$ and Level 2 covariance matrix $\boldsymbol{\tau}$) is accounted for when calculating the residuals.

Turning attention now to the total residual variance, $\boldsymbol{V}$ takes the familiar structure of independent and identically distributed data with variances (here blocks) on the diagonal and zeros on the off diagonal. This independence at Level 2 allows for log likelihoods to be added without the increased computational complexity imposed by relationships among observations (i.e., covariances). Although the observations appear to be independent, and are assumed to be at the highest sampling unit, each block of $\boldsymbol{\Sigma}$ contains information about the variability between and within groups thus satisfying the multilevel problem.

Nonlinear Models. Estimating nonlinear models in SAS uses a marginal likelihood which by necessity requires integration over the random effects (SAS "SAS Institute. SAS OnlineDoc 9.1.3," 2002-2005). One method that has shown promise in efficiently and accurately approximating these integrals is the First Order Approximation (Beal & Sheiner, 1988). This integral approximation leverages a one-term Taylor series expansion on the mean (i.e., fixed effects) structure and evaluates it at the average of the random effects (i.e., $u_i = 0$, SAS Institute 2002-2005) to reduce computational burden (Lindstrom & Bates, 1990). Furthermore, because the one-term Taylor series expansion is taken about the random effects it results in a likelihood

function where the random effects design matrix, $\boldsymbol{Z}$, enters the equation as the first derivative of the mean structure with respect to each random effect evaluated at its average (i.e., when $u_i = 0$). Substituting this new $\boldsymbol{Z}$ matrix into the normal likelihood (assuming that the outcome is still normally distributed) provides individual-specific log likelihoods that can be used in computation of Vuong's test for non-nestedness.

Effectively this changes computation as follows. First, the researcher must define the functional form(s) that they wish to test. After defining the equation, they must compute the partial derivatives of that equation with respect to each random effect. There are a number of ways to arrive at these derivatives. Of course, researchers with advanced knowledge (and substantial time) can work to compute these partial derivatives analytically. A more practical method, however, is to use one of the more widely available mathematical packages that can be used to find the partial derivatives such as Matlab or Wolfram Alpha. Certain versions of these applications are capable of directly computing the Jacobian. Once partial derivatives for the random effects have been obtained, a researcher can substitute necessary variable values into the obtained derivatives and with the results create a new random effects design matrix $\boldsymbol{Z^*}$. Using $\boldsymbol{Z^*}$ instead of $\boldsymbol{Z}$ in the equations for $\boldsymbol{V}$ will allow for accurate computations of the individual-specific log likelihoods.

To provide a concrete example I use an exponential growth model with a random intercept and slope of the form

$$y_{ij} = (\gamma_{00} + u_{0j}) * e^{\gamma_{10} * Time_{ij} + u_{1j} * Time_{ij}} + e_{ij}. \quad (43)$$

Assuming each individual is assessed weekly for five weeks with a baseline measurement indicated by a time value of zero, the fixed effects design matrix $\boldsymbol{X}$ would be of the form

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 7 \\ 1 & 14 \\ 1 & 21 \\ 1 & 28 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 28 \end{bmatrix}.$$

Note that the second column of $\boldsymbol{X}$ (i.e., the time values) is conceptualized here as days. Matrices $\boldsymbol{\tau}$ and $\boldsymbol{R}$ would be of the standard forms discussed above; $\boldsymbol{\tau}$ would have dimension 2 x 2 and $\boldsymbol{R}$ would be N x N diagonal. To calculate individual-specific contributions to the log likelihood, a single block of $\boldsymbol{R}$ would be 5 x 5 in this scenario. To create the $\boldsymbol{Z^*}$ matrix, it is necessary to find the partial derivatives of equation 43 with respect to u_{0j} and u_{1j},

$$\left[e^{\gamma_{10}*Time+u_{1j}*Time}, Time * e^{\gamma_{10}*Time+u_{1j}*Time} * (\gamma_{00} + u_{oj})\right]. \quad (44)$$

Setting $u_i = 0$ we are left with the equations to calculate our final $\mathbf{Z^*}$ matrix

$$\left[e^{\gamma_{10}*Time}, Time * e^{\gamma_{10}*Time} * (\gamma_{00})\right], \quad (45)$$

and using the arbitrary values of $\gamma_{00} = 26.73$ and γ_{10} = -.00037 the resulting values of the $\mathbf{Z^*}$ matrix are

$$\boldsymbol{Z}^* = \begin{bmatrix} 1 & 0 \\ .9974 & 186.63 \\ .9948 & 372.29 \\ .9922 & 556.99 \\ .9897 & 740.73 \\ 1 & 0 \\ \vdots & \vdots \\ .9897 & 740.73 \end{bmatrix}.$$

Substituting $\boldsymbol{Z^*}$ for $\boldsymbol{Z}$ into the log likelihood equation will produce the correct log likelihoods.

With an understanding of how changes in various parts of the model specification manifest in the likelihood of multilevel models, I will now establish that the same likelihood quantities, test statistics, and conclusions can be obtained whether using multilevel regression or

latent variable models. In doing so, I make an inductive argument that Vuong's test for non-nested models can be used in more complex cases where incorporating certain model features (i.e., nonlinear model parameters) are applied in multilevel regression more easily.

Empirical Illustrations

Fundamentally, multilevel regression and latent variable modeling are two approaches by which to solve the same problem. Multilevel regression has an appealing intuition for longitudinal problems in that equations can, and often are, thought of for each level. In Level 1, observations over time comprise the unit of analysis and specific response values result from deviations around each person's individual mean and individual time slope (in the case of random slopes). At Level 2, each person's mean (and slope) is thought to deviate around a population value. The Level 1 model naturally maps onto our typical intuition about regression while the Level 2 model accommodates the differences across individuals. By this logic, each observation is conditionally independent with respect to their group membership assuming there are no other (unmodeled) shared traits among them.

On the other hand, a latent variable modeling perspective maintains the traditional single level approach. Observations at each time point are considered indicators of random latent variables representing intercepts and slopes. The means of these latent variables represent the population regression coefficients and their (co)variances the variability and relatedness among the parameters. In more general latent variable models, the loading of each indicator on the latent variable is estimated and interpreted as a regression coefficient representing the strength of the relationship between the latent factor and the indicator. In a growth model, these paths are typically fixed to values representing indicators' relationships with the intercept (with a vector of ones) and a pre-specified functional form (e.g., a linearly increasing quantification of time for a

linear growth model). Functionally, these methods can be parameterized to result in the same solutions (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005; Mehta & West, 2000). Still, there are instances where intuition, specification, and software capabilities would cause researchers to prefer multilevel regression over a latent variable model to accomplish a given task.

Multilevel regression may be preferred in scenarios where observations within groups are unordered (such as individuals within groups) and thus observations become exchangeable. That is, the order of observations within a group is arbitrary whereas in a growth model order is paramount. While the only difference between a single Level 1 predictor multilevel regression with time nested within people and one with people nested within groups is the inclusion of a "time" variable and its requisite ordering, the difference between a single Level 1 predictor latent growth model and an intercept only model as a latent variable model is more complex. Latent variable models for individuals within groups require an unappealing shift in perspective where individuals become "indicators" of some latent variable. To fit a random intercept only model, a researcher would have to apply equality constraints across paths to match the multilevel specification. In a simple unconditional model, a researcher would have to constrain all of the residual variances to equality, however, as additional Level 1 variables without random slopes enter the model, their effect at each time point must be specified separately and those paths also constrained to equality. Adding random Level 1 variables complicates specification further with additional paths to be specified and constrained.

In other situations easily handled by the multilevel framework, latent variable models are either difficult or impossible to compute. Non-linear models, for instance, where random effects enter the equation non-linearly are omitted from or prohibitively difficult to specify in all of the standard latent variable software. While certain transformations and software-specific "tricks"

may make it possible to find solutions for these models, interpreting an already complex model becomes even more difficult in the best case and models may still not be estimable in the worst. I apply Vuong's test for non-nested models to these nonlinear models in Chapter 5.

To apply Vuong's test to multilevel models, I have written a SAS macro that calculates the individual-specific log-likelihoods and conducts the statistical test for pairs of candidate models known to be non-nested a priori (Appendix A). I now use this macro in an example to establish that the theoretical equivalence between multilevel models and latent variable models manifests appropriately in Vuong's test's calculation. To do so, I fit equivalent multilevel and latent variable models and compare candidates pairwise with the newly written macro and the nonnest2 (Merkle et al., 2015) package, respectively, with the expectation their results are identical. In showing the equivalence in Vuong's test across frameworks for this simple case, I provide justification for the application of Vuong's test to more complex instances of non-nestedness in multilevel regression.

Data Generation. Non-Exchangeable Observations. To demonstrate the use of Vuong's test and its equivalence in multilevel and latent variable frameworks, I simulated data under a two-level random intercept and slope linear growth model with three binary predictors at Level 2. Specifically, these models were of the form

$$\begin{aligned} level\ 1{:}\ y_{ij} &= \beta_{0j} + \beta_{1j}Time + e_{ij} \\ level\ 2{:}\ \beta_{0j} &= \gamma_{00} + \gamma_{01}W_1 + \gamma_{02}W_2 + \gamma_{03}W_3 + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \begin{bmatrix} \boldsymbol{u_j} \\ e_{ij} \end{bmatrix} &\sim N\begin{pmatrix} 0, & \boldsymbol{\tau} \\ 0, & \sigma^2 \end{pmatrix} \end{aligned} \tag{46}$$

The intercept, time, and residual variance were normally distributed with means of zero and variances of 4, 1 and 3, respectively, resulting in a residual-ICC of .5 after accounting for the random time effect. Fixed effects coefficients were selected to produce large (.8), medium (.5), and small (.2) standardized d-type effect sizes for W1, W2, and W3 respectively. Data were generated for 200 individuals with 6 waves of data and uniform assessment schedules. Three models were fit to these data differing in only the Level 2 predictor, however, in addition to the random intercept and Level 2 fixed effect they also included a random time slope, and a linear fixed effect for time.

Exchangeable Observations. To illustrate the cases with exchangeable observations, data were generated under a two-level random intercept model with three binary predictors at Level 2. Specifically, the model was of the form

$$\begin{aligned} Level\ 1: y_{ij} &= \beta_{0j} + e_{ij} \\ Level\ 2: \beta_{0j} &= \gamma_{00} + \gamma_{01}W_1 + \gamma_{02}W_2 + \gamma_{03}W_3 + u_{0j} \end{aligned} \tag{47}$$

Both the intercept and residual variances, u_{0j} and e_{ij} respectively, were normally distributed with means of zero and variances of 1 and 4 resulting in an ICC of .2. Again, fixed effects coefficients were selected to produce large, medium, and small d-type effect sizes for W1, W2, and W3, respectively. Under these specifications, data were simulated for 50 groups consisting of 20 individuals. Three models were fit, each with a single predictor, and 3 pairwise model comparisons conducted to test all possible model differences.

Analysis. Each of the three models for exchangeable and non-exchangeable observations was fit as a multilevel regression using SAS Proc MIXED with maximum likelihood estimation and Satterthwaite degrees of freedom and as a latent variable model using Mplus. Models in both frameworks were estimated with unstructured Level 2 covariance matrices. Individual (i.e.,

Level 2) log likelihoods were calculated using the developed SAS Macro and the LIKELIHOOD option in Mplus for each model in each program. Individual log likelihoods were then examined within models across frameworks to ensure that their estimation was equivalent (within rounding error). Subsequently, pairwise comparisons of models within framework were calculated using Vuong's test (in SAS for multilevel models and the nonnest2 package for R [for latent variable models; Merkle et al., 2015]) and their outcomes compared. To fit a latent variable model equivalent to the multilevel regression with exchangeable observations, the paths from the latent intercept to each indicator were constrained to one, similar to the latent growth curve model above. Additionally, residual variances were once again constrained to be equal, however, this assumption can be relaxed in both models.

Results

Growth Model (Non-Exchangeable Observations)

Fitting a growth model with non-exchangeable observations is a canonical example of the equivalence between multilevel regression and latent variable modeling. Data were generated under the above specifications, and pairwise model comparisons made for models fit with a random linear time effect at Level 1 and different Level 2 predictors. That is, a model with predictors "W1" and "time" was compared to a model with "W2" and "time" and a model of "W3" and "time". The "W2" model was also compared with "W3".

Fitted models produced identical parameter estimates as well as identical likelihood estimates in the aggregate within rounding error. This result comes as no surprise considering the well-established equivalence of the two model parameterizations in the literature (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005; Mehta & West, 2000). Beyond the equivalent parameter estimates, it was necessary to check for the equivalence of the individual specific contributions

to the (log) likelihood across frameworks. The individual-specific likelihoods for the first 10 individuals in each model are presented in Table 2.1.

Table 2.1 Case Wise Likelihoods for Growth Models

	W1 Model		W2 Model		W3 Model	
Case #	MLM	SEM	MLM	SEM	MLM	SEM
1	-13.993	-13.993	-13.032	-13.032	-13.236	-13.236
2	-16.337	-16.337	-16.395	-16.395	-16.485	-16.485
3	-14.012	-14.012	-14.128	-14.128	-14.363	-14.363
4	-13.351	-13.351	-13.351	-13.351	-13.361	-13.361
5	-13.290	-13.290	-12.928	-12.928	-13.131	-13.131
6	-13.224	-13.224	-13.132	-13.132	-12.996	-12.996
7	-13.728	-13.728	-14.009	-14.009	-13.834	-13.834
8	-13.410	-13.410	-12.998	-12.998	-13.224	-13.224
9	-15.567	-15.567	-15.625	-15.625	-15.715	-15.715
10	-13.942	-13.942	-13.340	-13.340	-13.618	-13.618

Overall, the individual-specific likelihoods are identical with some deviation occurring after the third decimal place. These minute differences are maintained in the test statistics (Table 2.2), however, what small differences there are can be attributed to rounding conventions specific to each program. When rounding was constrained to be equivalent across programs, so too were the likelihood ratios and test statistics. Qualitatively, the small differences in rounding made no difference as the conclusions remained the same across programs and frameworks. Comparing the model with predictor W1 and the model with W2, Vuong's test found no preference for either model (p_{w1} = .176). A similar, but less extreme result occurred when comparing the models with W2 and W3 (p_{w2} = .057). Finally, Vuong's test showed a preference for the model including W1 as a predictor over W3 (p_{w1} = .017). When specifying the number of places for rounding, p-values were absolutely identical across programs.

Table 2.2 Vuong's Test for Growth Models fit as MLM and SEM

Model	LR		Test Stat		Prob A > B		Prob B > A	
	MLM	SEM	MLM	SEM	MLM	SEM	MLM	SEM
W1-W2	5.104	5.108	.931	.932	.176	.176	.824	.824
W1-W3	9.988	9.982	2.128	2.127	.017	.017	.983	.983
W2-W3	4.884	4.874	1.580	1.577	.057	.057	.943	.943

Individuals within Groups (Exchangeable Observations).

Models were fit using the same methods as growth models omitting the fixed and random effects of time. The individual-specific log likelihoods for the first 10 cases for the models of exchangeable observations can be found in Table 2.3. Once again, each model's individual-specific log likelihoods were identical across parameterizations to the third decimal place. These quantities were then used in SAS or R (dependent on the modeling framework) to conduct pairwise comparisons between non-nested models via Vuong's test. Once again, results were identical.

Similar to the above example, rounding conventions may produce slight aberrations across programs (and subsequently frameworks), however, these rounding errors are due to floating point truncation native to each program and can again be ignored. Fixing rounding conventions to be the same across programs yielded entirely identical results.

Table 2.3 Case Wise Likelihoods for Models of Individuals Nested within Groups

	W1 Model		W2 Model		W3 Model	
Case #	MLM	SEM	MLM	SEM	MLM	SEM
1	-41.570	-41.570	-41.640	-41.640	-41.404	-41.404
2	-41.370	-41.370	-41.238	-41.238	-41.152	-41.152
3	-40.267	-41.267	-39.563	-39.563	-39.447	-39.447
4	-45.488	-45.488	-45.397	-45.397	-45.582	-45.582
5	-41.931	-41.931	-42.655	-42.654	-41.716	-41.716
6	-48.448	-48.448	-48.704	-48.704	-47.721	-47.721
7	-46.717	-46.717	-45.917	-45.917	-45.842	-45.842
8	-43.031	-43.031	-43.692	-43.692	-42.824	-42.824
9	-40.649	-40.649	-41.140	-41.140	-40.411	-40.411
10	-48.191	-48.191	-48.053	-48.053	-48.784	-48.784

Table 2.4 shows the results for the pairwise comparisons among models with exchangeable observations. Results indicated that no model should be preferred over any other with all p-values exceeding the traditional target significance of .05. Considering that the standardized effect sizes were equal to those in the previous example, these results beg the

question of how differences in characteristics typical of individuals nested within groups (i.e., smaller Level 2 sample sizes and ICCs) contribute to the ability of Vuong's test to detect a difference in KL-distance between models. Given what we know about the effects of sample size on the power to detect effects in the hypothesis testing literature (e.g., Hox, 2010; Maas & Hox, 2004, 2005; Raudenbush & Liu, 2000, 2001; Scherbaum & Ferreter, 2009), it is likely that Level 2 sample size is highly influential on the power of Vuong's test. However, this notion will be explored empirically in the following chapters.

Table 2.4 Vuong's Test for Models of Individuals Nested within Groups in MLM and SEM

	LR		Test Stat		Prob A > B		Prob B > A	
Model	MLM	SEM	MLM	SEM	MLM	SEM	MLM	SEM
W1-W2	1.814	1.814	.751	.751	.226	.226	.774	.774
W1-W3	3.650	3.651	1.288	1.289	.099	.099	.901	.901
W2-W3	1.836	1.837	.548	.549	.292	.292	.708	.708

In this chapter I have shown the equivalence of Vuong's test across SEM and MLM for simple models when observations were both exchangeable and non-exchangeable. In the following chapters I extend Vuong's test to more complex multilevel models including those where certain features (e.g., nonlinear parameters) may be more easily accommodated by MLM than latent variable models and evaluate its performance relative to current model selection techniques (i.e., information criteria comparison). These studies will explore how differences in study design (e.g., Level 1 and Level 2 sample size, ICC, effect size, etc.) affect the ability of Vuong's test to differentiate among candidate models where pairs of multilevel regressions are non-nested in their covariates, Level 1 error structures, or functional forms.

Chapter 3: Testing Non-nested Covariate Sets

In this chapter I examine the effectiveness of Vuong's Likelihood Ratio test to detect the "best" model from pairs of candidates in cases when non-nestedness manifested as different covariate sets at Level 1 or Level 2. While non-nestedness in covariates is not exceedingly common in the literature or exclusive to multilevel models, it is an important first step in establishing Vuong's test as a useful tool in the multilevel case as well as understanding differences in its behavior, if any, when non-nestedness occurs at a particular level. Although relatively few studies may report competing non-nested candidate models, they routinely arise early in the research process. For instance, in Moreno, Moskowitz, Ganz, and Bower (2016) two related variables were candidates for inclusion in the model, fatigue severity and fatigue interference. Because fatigue interference was not related to the outcome, only fatigue severity was included in the models but authors were unable to test whether the models with either variable fit the data differently. In this case, Vuong's test would have been helpful in determining if there was a significant difference in model fit when one covariate was included over the other.

To explore the performance of Vuong's test when covariate sets were non-nested, a simulation study was conducted comparing the model selection guided by Vuong's test compare to selection guided by information criteria. Additionally, analyses explored how Level 1 sample size, Level 2 sample size, effect size, ICC, and correlations among random effects might affect the ability of Vuong's test to detect the best model.

Method

Data Generation

To generate multilevel data on which to fit models, within- and between group correlation matrices were specified to define the relationships between all variables at Level 1 and Level 2. In the Level 1 correlation matrix, the relationship between the outcome, y, and Level 1 predictors, time, X1, X2, and X3 were defined with correlations of .2, .2, .3, and .4 respectively. Predictors were correlated with one another at .3. Level 2 variables were included in the Level 1 correlation matrix with zero vectors indicating no correlation with the outcome at Level 1 nor any correlation with Level 1 predictors. The Level 2 correlation matrix included the same Level 1 sub-matrix used to define Level 1 as well as correlations among Level 2 variables and their relationship with the outcome. Level 2 predictors, W1, W2, and W3 were correlated with the outcome at .2, .3, and .4, respectively. Relationships among Level 2 predictors matched those relationships among Level 1 predictors with a correlation of .3. Despite the inclusion of the between group relationships of the Level 1 variables they remained uncorrelated with Level 2 covariates. These correlation matrices were used to define the regression coefficients of y on the predictors.

After defining the regression coefficients of the predictors, values for each predictor were generated. All variables, fixed and random, were generated to have a mean of zero. Variances of Level 1 predictors were defined by the ICC whereas the variances of the Level 2 predictors were set to 1. The random intercept variance was set such that residual ICC was .4 in the small ICC condition and .7 in the large ICC condition. That is, the remaining ICC when the random effects portion of the model was fully specified. Thus the intercept variance was set to a value of either .666 or 2.3333. The intercept-slope covariance was defined by a pre-specified correlation

between random intercept and slope (0, .2, or .4) and the random slope variance was always half of the intercept variance. Level 1 predictor variables were generated from a multivariate normal distribution using the RANDNORMAL module in SAS Proc IML under the specifications defined above. Values for the Level 1 residual were drawn from a random normal distribution with mean zero and its total variance (1) adjusted for the covariates. Level 2 predictors were generated in the same fashion as the Level 1 fixed effects from a multivariate normal distribution defined by the between groups correlation matrix. Level 2 residuals were also generated in the same way with a covariance matrix defined by the desired random effects matrix resulting from the specified ICC and the correlation among random effects.

To generate the outcome y, predictors were combined according to the following equation:

$$y_{ij} = \gamma_{time\ .j} time_{ij} + \gamma_{x1} X_{ij}^{(1)} + \gamma_{x2} X_{ij}^{(2)} + \gamma_{x3} X_{ij}^{(3)} + \gamma_{w1} W_{.j}^{(1)} +$$

$$\gamma_{w2} W_{.j}^{(2)} + \gamma_{w3} W_{.j}^{(3)} + u_{int\ .j} + time\ u_{time\ .j} + e_{ij}$$

where superscripts define different variables. X variables refer to Level 1 covariates whereas W variables refer to Level 2 covariates. The intercept in the data generating model was set to zero and as a result was omitted from the above equation. After creating the outcome variable, predictors and the outcome were written to a SAS dataset.

Several models were then fit to the resulting data set using SAS Proc MIXED and a macro developed for this dissertation to conduct Vuong's test on each pair of models. Proc MIXED was run using maximum likelihood estimation (method = ML) and Satterthwaite degrees of freedom, although the degrees of freedom do not affect the subsequent analyses. The macro for conducting Vuong's test can be found in the appendix. First, six pairwise model comparisons were conducted using only Level 1 predictors. Models were compared each with a

single predictor (e.g., X1 v X2, X1 v X3, and X2 v X3) as well as models with two predictors compared with the excluded predictor (e.g., X1 v X2 & X3, X2 v X1 & X3, and X3 v X1 & X2). The same comparisons were made for the Level 2 covariates. Every estimated model contained time as a common predictor. Both results of Vuong's test comparing these models and each model's information criteria were collected and saved for analysis.

A factorial design was used to generate data from every combination of 3 Level 1 sample sizes, 3 Level 2 sample sizes, 2 ICCs, and 3 random effects correlations. Effect size was manipulated within condition with each pair of model comparisons manifesting a different effect size. One-thousand replications were generated for each of the 54 study conditions resulting in 54,000 unique data sets. Twelve models were run on each of these 54,000 data sets resulting in 648,000 hypothesis tests.

Sample Size. Sample sizes at both Level 1 and Level 2 were chosen to approximate the range of sample sizes used in longitudinal research. At Level 2, samples of 50, 100, and 200 were chosen to represent individuals to be measured repeatedly over time whereas 5, 13, or 25 Level 1 observations were simulated as a representation of studies of various lengths. The upper limits of these sample sizes were selected as they have been shown empirically to approach the ceiling of statistical power (Hox, 2010; Maas & Hox, 2004, 2005; Raudenbush & Liu, 2000, 2001; Scherbaum & Ferreter, 2009).

ICC. ICC was manipulated via the Level 2 intercept variance to achieve residual ICC values of .4 and .7. Residual ICC values are conceptualized here as the proportion of remaining variance attributable to the intercept after accounting for the random (and fixed) effects of time. Because longitudinal models are rarely estimated without a random time effect, I estimate it and remove it from the equation when calculating ICC. That is, the proportion of variance attributed

to the intercept parameter was 40% or 70% of the total variance in the outcome excluding the random slope variance (and the intercept slope covariance) and was calculated as $\frac{\tau_{00}}{(\tau_{00}+\sigma^2)}$. If the ICC were calculated without modeling the random time effect, it would appear greater than intended.

Effect Size. Effect size was conceptualized as the difference in variance explained between the two candidate models under consideration. For example, if a model including covariate X1 explained 6% of the variance in Y and the model with covariate X2 explained 10% of the variance in Y, the effect size for the comparison between the two candidate models would be 4%. An empirical effect size was generated using a very large data set (500 observations for 10,000 Level 2 units). Effect sizes for all covariate sets are displayed on Table 3.1.

Preliminary analyses indicated that in addition to effect size, the difference in the number of parameters in the candidate models may influence the probability of Vuong's test selecting the best model. Thus an indicator variable was coded to identify conditions in which candidate models contained the same or different numbers of parameters. Within these groups, I ranked the effect sizes as "small", "medium", and "large" and refer to them as such throughout this chapter. These designations have no relation to Cohen's (1977) widely accepted small, medium, and large effect sizes nor is it meant to equate effect sizes across number of parameter groups. Referring to these effects as "small", "medium", and "large" is purely for convenience.

Table 3.1. Effect Size conditions.

Level 1 Comparison	Diff Variance Explained	Level 2 Comparison	Diff Variance Explained
X1 vs X2	4.2%	W1 vs W2	3.7%
X2 vs X3	6.8%	W2 vs W3	6.7%
X1 vs X3	11.0%	W1 vs W3	10.4%
X1 X2 vs X3	5.9%	W1 W2 vs W3	5.6%
X1 X3 vs X2	7.3%	W1 W3 vs W2	7.1%
X2 X3 vs X1	14.1%	W2 W3 vs W1	12.8%

Random Effects Correlation. The random effects correlation was manipulated directly as the covariance between intercept and slope variances. Once Level 2 intercept variances were determined to achieve the desired ICCs, covariances were computed to achieve the desired intercept-slope correlation. Following the methodology set out by Vallejo et al. (2014), one of the only studies to report an impact of intercept-slope covariance on the performance of model selection, covariances were determined that resulted in correlations among random effects of 0, .2, and .4.

Data Analysis

Once model comparisons were conducted and results retained, each replication was classified depending on whether the "best" model was detected. The true data generating mechanism—a full model with time, three Level 1 predictors, three Level 2 predictors, a random intercept, and a random slope—was never among the candidates. As a result, the "best model" was defined as the model with the greatest combined effect on the outcome. Because all predictors were equally correlated with one another, had the same mean, the same variance (within level), and differed only in their relationship with y (but were in the same direction), the differences between combinations of regression coefficients should be sufficient to determine the expected best model. That is, summing the scaled magnitudes of estimated effects and calculating the difference between two models should indicate which model would be expected to fit the data best. This logic maps nicely onto the concept of K-L divergence as well. If the outcome, y, is a function of six predictor variables, the candidate model with the largest absolute value of the total effect should result in the best predictions, the greatest likelihood, and the smallest K-L divergence. However, to map results onto a more intuitively appealing and

generalizable metric, the best model is discussed by the amount of variance its predictors explained.

Results of each model selection procedure were then tabled and compared on correct model selection rate and incorrect model selection rate. Special consideration was also given to the non-significance rate of Vuong's test as the ability of the test to fail to find evidence in support of either model rather than potentially selecting the incorrect model as a result of a forced choice is an appealing property of the method. Although model selection is not technically a classification procedure, I use the terms correct classification and misclassification to describe when Vuong's test or information criteria select the correct or incorrect model, respectively.

To facilitate comparison among misclassification rates of Vuong's test and information criteria, a sensitivity analysis was conducted to determine the difference from a constant proportion that would be detected using a binomial test with a power of .8 and 1000 observations per cell. The misclassification rate of Vuong's test was vanishingly small (at most 1%; Table 3.3) for Level 1 covariates and only slightly larger (at most 3%; Table 3.6) for Level 2 covariates. Using G*Power 3.1.5 and a constant misclassification proportion of .03 the maximum misclassification of Vuong's test in this study the sensitivity analysis determined that there was enough power to detect a difference in percentages of 1.7% at the canonical .05 significance level. Conservatively rounding this difference to 2%, it was reasonable to assume that any misclassification of 5% or greater can be considered significantly poorer performance than Vuong's test.

After comparing the performance of Vuong's test that of information criteria logistic regression was then used to determine the factors that contributed to Vuong's test's capacity to

detect the best model. Using SAS Proc LOGISITIC, main effects were explored first to describe the general behavior of Vuong's test. Next, all possible interactions among study factors were included in a full model and the significance of each omnibus test evaluated. One by one, higher order effects were removed until only significant differences among groups remained in the model. Non-significant lower order effects were retained if they were qualified by a higher order interaction. Initially, some logistic regressions were inestimable due to a high degree of separation in the outcome. To account for the near perfect separation ("quasi-separation") Firth's method of penalized likelihood was employed to facilitate fitting the logistic regressions (Heinze & Schemper, 2002). Throughout analyses, study factors were treated categorically.

Results

Level 1 Covariates

Selecting the Correct Model. To compare the results of Vuong's test with the performance of information criteria I first examined the rates at which each model selection procedure selected the correct model. As the trends are generally the same across ICC conditions and random coefficients correlations with only small differences between model selection rates (the correct model is selected with slightly higher frequency in the small ICC condition and a small random effects correlation) I present the correct classification rates for the small ICC condition with no random effects correlation on Table 3.2. A complete set of model selection tables (correct classification, misclassification, and non-significance rates of Vuong's test) across ICCs and random effects correlations can be found in the supplementary material. Shaded cells indicate conditions in which the empirical power of Vuong's test to detect the best model was above .8.

When Level 1 sample size contained 25 observations Vuong's test power for Vuong's test was above .8. As Level 1 sample size decreased, the power of Vuong's test tended to decrease as well. When the number of Level 1 observations was 13, Vuong's test was underpowered when differences between models were small or medium when the number of parameters in the models was the same and when Level 2 sample size was small. When the number of parameters was unequal across models and effect size was small, Vuong's test remained underpowered when Level 2 sample size was small. When Level 1 sample size was only 5 observations, Vuong's test selected the correct model in only when effect size was large or Level 2 sample size was large and effect size was medium.

Information criteria tended to select the correct model in almost every case when Level 1 sample size was 13 observations or more. Even when the number of Level 1 observations was only 5, information criteria still tended to select in at least 85% of replications in all conditions. While these results might suggest that information criteria should be used over Vuong's test when comparing non-nested fixed effects at Level 1, examining misclassification of information criteria suggest that in certain cases the decision may not be as clear.

Table 3.2 Correct model selection rates for Level 1 covariate sets when ICC = .4 and $\tau_{01} = 0$

			L1 SS											
# of Params	%Var Explained	L2SS	5				13				25			
			Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	4.2%	50	32%	89%	89%	89%	71%	99%	99%	99%	93%	100%	100%	100%
		100	49%	97%	97%	97%	91%	100%	100%	100%	100%	100%	100%	100%
		200	81%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	6.8%	50	37%	91%	91%	91%	71%	99%	99%	99%	94%	100%	100%	100%
		100	62%	97%	97%	97%	93%	100%	100%	100%	100%	100%	100%	100%
		200	83%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	11%	50	82%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		100	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Unequal	5.9%	50	27%	91%	91%	94%	55%	98%	98%	98%	85%	100%	100%	100%
		100	47%	95%	95%	96%	83%	100%	100%	100%	98%	100%	100%	100%
		200	69%	99%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%
	7.3%	50	45%	90%	90%	88%	81%	99%	99%	99%	97%	100%	100%	100%
		100	71%	97%	97%	97%	97%	100%	100%	100%	100%	100%	100%	100%
		200	90%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	14.1%	50	95%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Misclassification. Misclassification rates for Vuong's test and the information criteria can be found on Table 3.3 for the same condition as above. Vuong's test selected the incorrect model in at most 1% of the replications. Results from the sensitivity analysis suggest that any misclassification in the information criteria greater than 5% can be considered significantly poorer performance than Vuong's test.

Despite the uniformly high correct classification rates of the information criteria, Table 3.3 shows that when Level 2 sample size is 50 and Level 1 sample size is 5, information criteria consistently selected the incorrect model in over 10% of cases. As either Level 1 or Level 2 sample size increased, so did the propensity of information criteria to select the best model.

The benefit of Vuong's test can best be observed when differences in fit among candidate models are difficult to detect. Table 3.4 displays the non-significance rates for Vuong's tests. Highlighted cells indicate conditions in which information criteria select the incorrect model in a significantly greater proportion of replications. While Vuong's test would not necessarily provide any insight into which model should be preferred in these scenarios, its capacity to return a null result in the case where there is not enough information to determine which model fits the data better is an attractive aspect of the test. Having compared the performance of Vuong's test to that of information criteria, I now explore differences in the power of Vuong's test to detect the best model between study factors.

Table 3.3 Incorrect model selection rates for Level 1 covariate sets when ICC = .4 and $\tau_{01} = 0$

			L1SS											
			5				13				25			
# of Params	%Var Explained	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	4.20%	50	0%	11%	11%	11%	0%	1%	1%	1%	0%	0%	0%	0%
		100	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	6.80%	50	0%	9%	9%	9%	0%	1%	1%	1%	0%	0%	0%	0%
		100	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	11%	50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Unequal	5.90%	50	0%	9%	9%	6%	0%	3%	3%	2%	0%	0%	0%	0%
		100	0%	5%	5%	4%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	7.30%	50	0%	10%	10%	12%	0%	1%	1%	1%	0%	0%	0%	0%
		100	0%	3%	3%	4%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	14.10%	50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 3.4 Non Significance rates of Vuong's test for Level 1 covariate sets when ICC = .4 and $\tau_{01} = 0$

			L1SS		
# of Params	%Var Explained	L2SS	5	13	25
Equal	4.20%	50	68%	29%	7%
		100	51%	9%	0%
		200	19%	0%	0%
	6.80%	50	63%	29%	6%
		100	39%	7%	0%
		200	17%	0%	0%
	11%	50	19%	0%	0%
		100	2%	0%	0%
		200	0%	0%	0%
Unequal	5.90%	50	72%	45%	15%
		100	53%	17%	2%
		200	31%	1%	0%
	7.30%	50	55%	19%	3%
		100	29%	3%	0%
		200	10%	0%	0%
	14.10%	50	5%	0%	0%
		100	0%	0%	0%
		200	0%	0%	0%

Power. Logistic regression was employed to determine the factors that impact the power of Vuong's likelihood ratio test to select the best model. Tests of main effects can be found in Table 3.5. Results indicated significant effects of Level 2 sample size, Level 1 sample size, effect size, and ICC. As Level 2 sample size, Level 1 sample size, or effect size increased so did power for Vuong's test to select the best model. Additionally, power was greater when ICC was small. The significant main effect for ICC was such that the probability of detecting the best model when ICC was .7 was only 94% as high as when ICC was .4. There was no significant main effect for random effects correlation nor was there a main effect for equality of parameters.

Table 3.5. Omnibus Test for Main effects Predicting Power to detect Non-nestedness of Level 1 covariate sets.

Effect	DF	χ^2
Level 2 Sample Size	2	20208.22*
Level 1 Sample Size	2	33191.65*
ICC	1	21.12
Effect Size	2	18279.73*
Ran Eff Correlation	2	1.30
# Parameters Equality	1	.63

* $p < .0001$.

Results indicated a significant 4-way interaction between Level 2 sample size, Level 1 sample size, effect size, and equality of parameters, $\chi^2(8) = 20.1641$, $p = .0097$. This relationship is illustrated on Figure 3.1 where the left column contains figures where the number of parameters among candidates are equal and the right column contains figures where the number of parameters among candidates are unequal. Each row contains figures for a different effect size and each line on individual panels represents a different Level 1 sample size. Finally, Level 2 sample size is represented on the horizontal axis and the power to detect the best model is represented on the vertical axis. Omnibus tests for each lower order effect for equal and unequal numbers of parameters are displayed on Table 3.6.

Omnibus tests of simple effects were largely similar when the sample was split by differences in the number of parameters in candidate models. Only the significance tests of the Level 1 sample size by ICC interaction and the random effects correlation by ICC interaction differed across parameter equality groups. Follow up testing examining the difference in the Level 1 sample size by ICC interaction across parameter equality groups confirmed that there was no significant difference in the Level 1 sample size by ICC interaction, $\chi^2(2) = .1090$, $p = .95$. Follow up analyses also confirmed no significant differences in the ICC by correlation interaction across equality groups, $\chi^2(2) = 1.09$, $p = .59$ there was no difference in the ICC by random effects correlation

Table 3.6 Omnibus tests for lower order interactions for equal and unequal numbers of parameters.

Effect	DF	Equal	Unequal
Level 2 Sample Size	2	1774.62***	1346.34***
Level 1 Sample Size	2	1369.35***	1087.13***
Effect Size	2	1808.78***	1971.20***
Correlation	2	.24	5.37
ICC	1	1.27	3.83
Level 2 Sample Size * Level 1 Sample Size	4	360.57***	500.45***
Level 2 Sample Size * Effect Size	4	239.89***	166.79***
Level 2 Sample Size * ICC	2	4.99	5.43
Level 1 Sample Size * Effect Size	4	149.89***	138.16***
Level 1 Sample Size * Correlation	4	2.32	7.58
Level 1 Sample Size * ICC	2	2.05	10.20**
ICC * Effect Size	2	6.44*	8.50*
ICC * Correlation	2	2.22	14.54***
Level 2 Sample Size * Level 1 Sample Size * Effect Size	8	36.73***	64.89***

*p < .05 **p < .01 ***p < .001

Across all panels, power increased monotonically as either Level 1 or Level 2 sample size increased until it reached its asymptote. The same can be said for effect size. As the differences between models grew, power to detect that difference grew as well. As expected, there were diminishing returns in power as it approached its maximum. That is, as power neared 100%, adding an observation at either Level 1 or Level 2 offers a smaller improvement in power.

Power also tended to be greater when models had unequal numbers of parameters than when they had the same number of parameters, except when effect size was small. Power in the small effect size condition was greater when the number of parameters in candidate models was equal compared to when they were unequal. While differences in the magnitude of effect sizes across equal and unequal numbers of parameters in candidate models would suggest that an interaction should exist between effect size and number of parameters, this effect was in the unexpected direction. That is, the difference in variance explained between models in the small sample size condition when the number of parameters between models was unequal was 5.9%

whereas there was only a difference of 4.2% variance explained when the number of parameters were equal.

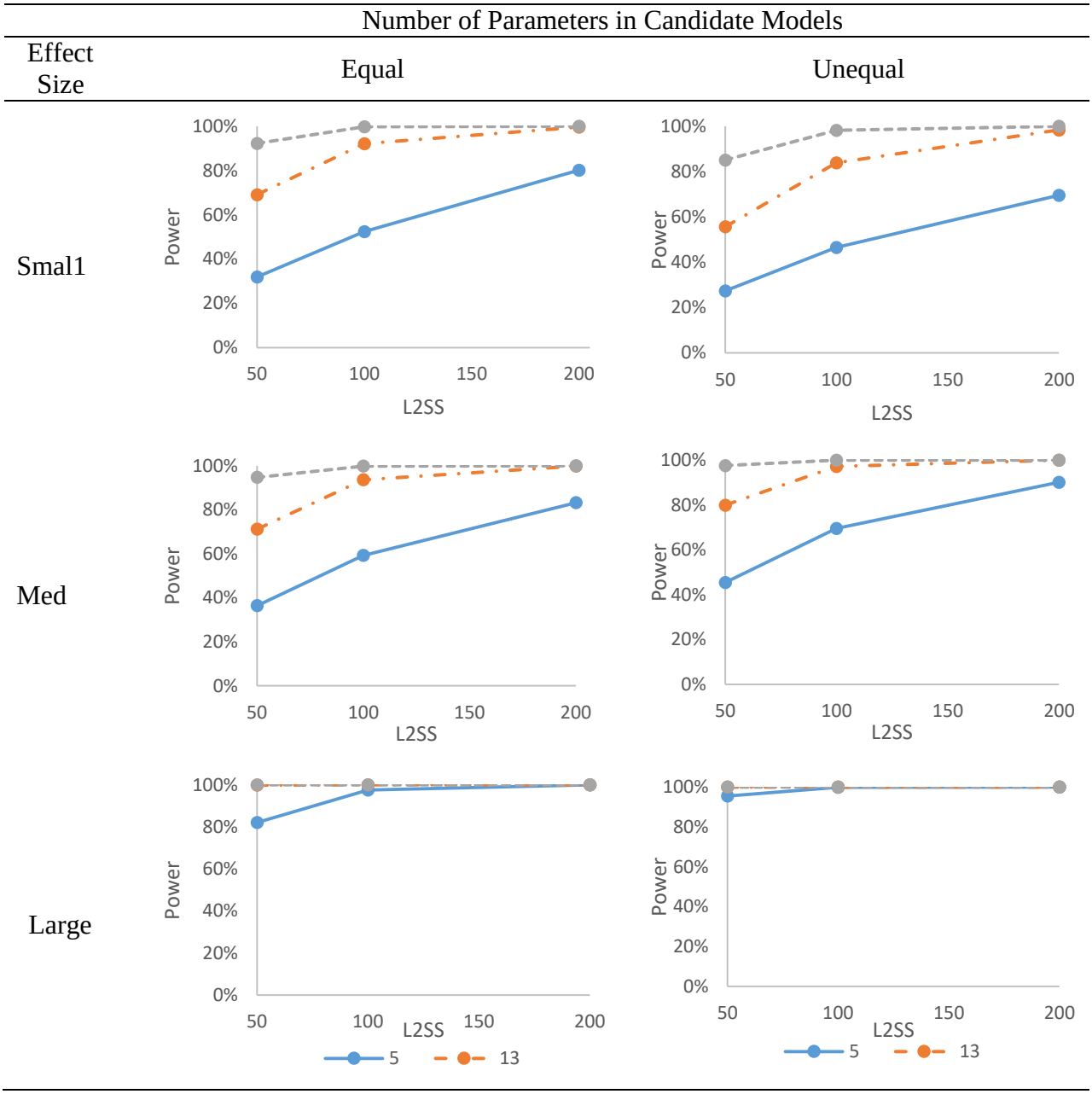


Figure 3.1 Level 1 Sample size x Level 2 Sample size x Effect Size x Number of Parameters Interaction

In addition to the complex effect decomposed above, there were several significant effects not qualified by the 4-way interaction. Specifically, there were significant interactions

between Level 1 sample size, ICC, and random effects correlation (χ^2 (4)= 9.6737, p = .0463), ICC and effect size (χ^2 (2)= 12.6408, p = .0018), and Level 2 sample size and ICC (χ^2 (2)= 9.0038, p = .0111). Although significant, the Level 1 sample size by ICC by random effects correlation interaction was qualitatively trivial. Power increased only slightly faster as Level 1 sample size increased when the random effects correlation was zero compared to when it was non-zero. The Level 2 sample size by ICC interaction was also trivial; differences in power between ICC conditions at each effect size differed only by 1-2%.

Table 3.7 displays the predicted power values for each effect size at both levels of ICC. Generally, there was less power for Vuong's test to detect the best model when ICC was large than when ICC was small. However, as effect size increased, the difference in power between the two ICC conditions increased as well. In the small effect size condition the difference in power between ICCs is 1.7%, which increased to 2.4% at the medium effect size. When effect size was large, the difference in predicted power across ICCs was 4.5%.

Table 3.7 Predicted power of detecting the best model for the ICC x Effect Size interaction

ICC	Effect Size		
	Small	Medium	Large
.4	32.3%	36.9%	82.7%
.7	30.6%	34.5%	78.1%
Difference	1.7%	2.4%	4.5%

Level 2 Covariates

Selecting the Correct Model. To compare the results of Vuong's test with the performance of information criteria when non-nestedness occurred in the Level 2 covariates, I examined the rates at which model selection procedures selected the correct model. Because there was a reasonably large effect for ICC (see section on power below) on the power of Vuong's test to detect the best model, Tables 3.8 and 3.9 present the correct classification rates of Vuong's test and information criteria for the small and large ICCs, respectively. A full set of tables including correct model selection rates, misclassification rates, and Vuong's test's non-significance rates for each random effects correlation at each ICC can be found in the supplementary material. Shaded cells indicate conditions in which the empirical power of Vuong's test to detect the best model was above .8.

Results on Tables 3.8 and 3.9 indicated that Vuong's test was generally underpowered when non-nestedness occurred at Level 2. When ICC was small, Vuong's test never reached adequate power to detect the best model. In the large ICC condition, Vuong's test only achieved power of .8 in the conditions where there was a large effect and the number of parameters in candidate models was unequal, Level 2 sample size was large, and Level 1 sample size was at least 13. The rest of the conditions left Vuong's test underpowered. In the worse cases (small sample sizes at Level 1 and Level 2, small effect size with equal number of parameters) power of Vuong's test to detect the best model was about 4%.

Table 3.8 Correct model selection rates for Level 2 covariate sets when ICC = .4 and $\tau_{01} = 0$

			L1SS											
			5				13				25			
# of Params	%Var Explained	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	3.70%	50	4%	61%	61%	61%	4%	65%	65%	65%	4%	65%	65%	65%
		100	6%	73%	73%	73%	7%	75%	75%	75%	9%	76%	76%	76%
		200	13%	81%	81%	81%	15%	83%	83%	83%	19%	82%	82%	82%
	6.70%	50	6%	65%	65%	65%	8%	67%	67%	67%	9%	69%	69%	69%
		100	10%	69%	69%	69%	12%	73%	73%	73%	13%	76%	76%	76%
		200	20%	80%	80%	80%	19%	84%	84%	84%	23%	85%	85%	85%
	10.40%	50	8%	79%	79%	79%	10%	80%	80%	80%	12%	82%	82%	82%
		100	17%	87%	87%	87%	23%	90%	90%	90%	29%	93%	93%	93%
		200	44%	96%	96%	96%	52%	96%	96%	96%	55%	98%	98%	98%
Unequal	5.60%	50	2%	69%	70%	83%	5%	71%	71%	82%	5%	73%	73%	84%
		100	6%	72%	73%	86%	7%	74%	75%	85%	8%	77%	77%	88%
		200	14%	79%	79%	88%	12%	82%	82%	92%	14%	82%	82%	90%
	7.10%	50	9%	56%	54%	62%	12%	57%	57%	58%	12%	62%	62%	54%
		100	13%	64%	63%	55%	15%	70%	70%	53%	18%	72%	72%	57%
		200	26%	78%	78%	65%	26%	83%	83%	72%	30%	85%	85%	74%
	12.80%	50	17%	70%	69%	53%	20%	77%	77%	62%	23%	79%	79%	65%
		100	33%	88%	88%	71%	42%	92%	92%	82%	49%	94%	94%	85%
		200	67%	98%	98%	94%	73%	98%	98%	96%	77%	99%	99%	98%

Table 3.9 Correct model selection rates for Level 2 covariate sets when ICC = .7 and $\tau_{01} = 0$

			L1SS											
			5				13				25			
# of Params	%Var Explained	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	3.70%	50	6%	67%	67%	67%	5%	68%	68%	68%	5%	69%	69%	69%
		100	9%	77%	77%	77%	10%	78%	78%	78%	10%	76%	76%	76%
		200	16%	84%	84%	84%	22%	84%	84%	84%	19%	86%	86%	86%
	6.70%	50	7%	70%	70%	70%	9%	70%	70%	70%	9%	69%	69%	69%
		100	11%	74%	74%	74%	13%	75%	75%	75%	18%	79%	79%	79%
		200	21%	85%	85%	85%	24%	84%	84%	84%	25%	86%	86%	86%
	10.40%	50	11%	82%	82%	82%	13%	82%	82%	82%	15%	84%	84%	84%
		100	27%	91%	91%	91%	30%	91%	91%	91%	35%	93%	93%	93%
		200	55%	97%	97%	97%	60%	99%	99%	99%	62%	99%	99%	99%
Unequal	5.60%	50	4%	74%	75%	84%	5%	71%	71%	83%	5%	72%	72%	82%
		100	6%	75%	75%	85%	8%	75%	76%	86%	12%	79%	79%	88%
		200	14%	81%	81%	89%	15%	81%	81%	89%	17%	83%	83%	89%
	7.10%	50	11%	62%	61%	55%	13%	61%	61%	54%	12%	62%	61%	52%
		100	16%	71%	71%	58%	19%	72%	72%	57%	23%	75%	75%	61%
		200	29%	86%	86%	75%	30%	84%	84%	73%	32%	87%	87%	76%
	12.80%	50	25%	79%	78%	64%	25%	79%	79%	66%	29%	84%	84%	70%
		100	48%	93%	93%	84%	51%	94%	94%	85%	56%	96%	96%	88%
		200	78%	99%	99%	97%	83%	99%	99%	98%	85%	99%	99%	98%

Correct classification rates of information criteria are also presented on Tables 3.8 and 3.9. Overall, information criteria performed worse when non-nestedness occurred at Level 2 than when it occurred at Level 1. Correct classification rates dropped as low as 52% in certain conditions and only rose above 80% when effect size or Level 2 sample size was large and models had unequal numbers of parameters. While these classification rates might seem better than Vuong's test they come at a cost: when information criteria were not selecting the correct model, they were selecting the *incorrect* model. As a result, the metric of .8 for the "power" of information criteria was an inadequate bar on which to evaluate information criteria as it would imply that a 20% error rate would be acceptable. Given that information criteria must make a decision about which model to choose, it must be held to a much higher standard so as to not lead researchers to make incorrect conclusions. Examining the misclassification rates of information criteria, instead of their correct classification rates, provides insight into how these moderate correct classification rates can be worrisome.

Misclassification. Misclassification rates for the same conditions described above can be found on Tables 3.10 and 3.11. In either ICC, Vuong's test selected the wrong model in at most 3% of replications. Misclassification to this degree was only observed in cases where effect size comparing two models with unequal numbers of parameters was small, Level 1 sample size was small or medium, and Level 2 sample size was small. In all other conditions, Vuong's test chose the incorrect model in less than 3% of replications.

Table 3.10 Incorrect model selection rates for Level 2 covariate sets when ICC = .4 and $\tau_{01} = 0$

			L1SS											
			5				13				25			
# of Params	%Var Explained	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	3.70%	50	1%	39%	39%	39%	1%	35%	35%	35%	1%	36%	36%	36%
		100	0%	27%	27%	27%	0%	25%	25%	25%	1%	24%	24%	24%
		200	0%	19%	19%	19%	0%	17%	17%	17%	0%	18%	18%	18%
	6.70%	50	1%	35%	35%	35%	1%	33%	33%	33%	1%	32%	32%	32%
		100	1%	31%	31%	31%	1%	27%	27%	27%	0%	24%	24%	24%
		200	1%	20%	20%	20%	1%	16%	16%	16%	0%	15%	15%	15%
	10.40%	50	0%	21%	21%	21%	0%	20%	20%	20%	1%	18%	18%	18%
		100	0%	13%	13%	13%	0%	10%	10%	10%	0%	7%	7%	7%
		200	0%	4%	4%	4%	0%	4%	4%	4%	0%	2%	2%	2%
Unequal	5.60%	50	2%	31%	30%	17%	3%	29%	29%	18%	2%	28%	27%	16%
		100	1%	28%	27%	14%	1%	26%	26%	15%	1%	23%	23%	12%
		200	1%	21%	21%	13%	1%	18%	18%	8%	0%	18%	18%	10%
	7.10%	50	0%	45%	46%	38%	1%	43%	43%	42%	1%	38%	38%	47%
		100	0%	37%	37%	45%	0%	30%	30%	48%	0%	28%	28%	43%
		200	1%	22%	22%	35%	0%	17%	17%	29%	0%	15%	15%	26%
	12.80%	50	0%	30%	31%	47%	0%	23%	23%	38%	0%	21%	21%	35%
		100	0%	12%	12%	29%	0%	8%	8%	19%	0%	6%	6%	15%
		200	0%	2%	2%	6%	0%	2%	2%	4%	0%	1%	1%	3%

Table 3.11 Incorrect model selection rates for Level 2 covariate sets when ICC = .7 and $\tau_{01} = 0$

			L1SS											
			5				13				25			
# of Params	%Var Explained	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
Equal	3.70%	50	1%	34%	34%	34%	1%	32%	32%	32%	1%	31%	31%	31%
		100	1%	23%	23%	23%	0%	22%	22%	22%	1%	24%	24%	24%
		200	0%	16%	16%	16%	0%	16%	16%	16%	0%	14%	14%	14%
	6.70%	50	2%	30%	30%	30%	1%	30%	30%	30%	1%	31%	31%	31%
		100	1%	26%	26%	26%	1%	25%	25%	25%	1%	21%	21%	21%
		200	0%	15%	15%	15%	0%	16%	16%	16%	1%	14%	14%	14%
	10.40%	50	0%	18%	18%	18%	0%	18%	18%	18%	0%	16%	16%	16%
		100	0%	9%	9%	9%	0%	9%	9%	9%	0%	7%	7%	7%
		200	0%	3%	3%	3%	0%	2%	2%	2%	0%	2%	2%	2%
Unequal	5.60%	50	3%	26%	26%	16%	3%	29%	29%	18%	2%	28%	28%	18%
		100	1%	25%	25%	15%	2%	25%	25%	14%	2%	21%	21%	12%
		200	1%	19%	19%	12%	1%	19%	19%	11%	1%	17%	17%	11%
	7.10%	50	1%	38%	39%	45%	1%	39%	39%	46%	1%	38%	39%	48%
		100	0%	29%	29%	42%	1%	28%	28%	43%	1%	25%	25%	39%
		200	0%	14%	15%	26%	0%	16%	16%	27%	0%	13%	13%	24%
	12.80%	50	0%	22%	22%	36%	0%	21%	21%	34%	0%	16%	16%	31%
		100	0%	7%	7%	16%	0%	6%	6%	15%	0%	4%	5%	12%
		200	0%	1%	1%	3%	0%	1%	1%	2%	0%	1%	1%	2%

Following the sensitivity analysis described above, any misclassification in the information criteria above 5% could be considered significantly poorer performance than Vuong's test. In all but two conditions (when effect size and Level 2 sample size are both large) sample size information criteria select the incorrect model significantly more often than Vuong's test. In the worst cases, model selection based on information criteria were almost no better than a coin flip with misclassification occurring in 45%-46% of replications.

Table 3.12 displays the non-significance rates of Vuong's test when the random effects correlation was 0 and the ICCs were small and large. Shaded cells represent cases in which information criteria selected the incorrect model in a significant proportion of replications. The table was almost completely shaded; in virtually every condition information criteria selected the incorrect model in a significant proportion of cases. Furthermore, this misclassification was not marginally significant, but drastically so. Even when misclassification began approaching non-significance, the incorrect model was selected in about 10%-20% of replications when following information criteria. Although Vuong's test was underpowered and did not select the correct model, it also did not select the incorrect model. Instead, Vuong's test provided a non-significant result indicating that there was no evidence that one model fit the data better than the other. That is, the models fit the data equally well.

Table 3.12 Non-significance rates for non-nested Level 2 covariate sets when $\tau_{01} = 0$

			ICC					
			0.4			0.7		
			L1 SS					
# of Params	%Var Explained	L2SS	5	13	25	5	13	25
Equal	3.70%	50	95%	96%	95%	93%	94%	94%
		100	94%	93%	91%	91%	90%	90%
		200	87%	85%	81%	84%	78%	81%
	6.70%	50	94%	91%	91%	91%	90%	90%
		100	90%	88%	87%	88%	86%	81%
		200	79%	80%	77%	78%	76%	75%
	10.40%	50	92%	89%	88%	89%	87%	85%
		100	83%	77%	72%	73%	70%	65%
		200	56%	49%	45%	45%	40%	39%
Unequal	5.60%	50	96%	93%	93%	93%	93%	93%
		100	93%	92%	91%	92%	90%	86%
		200	85%	87%	86%	85%	84%	82%
	7.10%	50	91%	88%	87%	88%	86%	87%
		100	87%	85%	82%	83%	81%	77%
		200	73%	74%	70%	70%	69%	67%
	12.80%	50	83%	80%	77%	76%	75%	71%
		100	67%	58%	51%	52%	49%	44%
		200	33%	27%	23%	22%	17%	16%

Power. The same approach was utilized to test the factors that influence the performance of Vuong's test when non-nestedness manifests in covariates at Level 2. Initial analyses indicated main effects for all of the factors (Table 3.13). The effects of Level 1 sample size, Level 2 sample size and effect size all trended in the same direction -- increases in any of these factors significantly increased statistical power. ICC, however, had a different effect when non-nestedness occurred at Level 2 than when it occurred at Level 1; a larger ICC led to greater power. When ICC was large, the best model was selected 1.42 times more often than when ICC was small. The effect of correlation and the number of parameters in the candidates also had

significant main effects for non-nestedness in Level 2 covariates. As the correlation among random effects increased, so too did the probability of Vuong's test selecting the best model. Finally, power was greater when the number of parameters among candidates was unequal compared to when it was equal. In Vuong's test where candidates had an unequal number of parameters, power was 1.68 times greater than when candidates had an equal number of parameters. These effects were qualified by higher order interactions.

Table 3.13. Omnibus Test for Main effects Predicting Power to detect Non-nestedness of Level 2 covariate sets.

Effect	DF	χ^2
Level 2 Sample Size	2	21443.6246
Level 1 Sample Size	2	1136.2607
ICC	1	1392.3033
Effect Size	2	30958.0268
Ran Eff Correlation	2	95.6522
# Parameters Equality	1	2993.3995

Note: All p values < .0001.

The final model examining the effects of design factors on the probability of detecting the best model when differences occurred in the Level 2 fixed effects contained no 6-, 5-, or 4- way interactions. A number of three-way interactions emerged as significant (Table 3.14). To probe these interactions further, data were first split on ICC as it was the most common factor among the interactions.

Table 3.14 Omnibus tests of significant three-way interactions affecting the power of Vuong's test when models are non-nested in Level 2 covariates.

Effect	df	χ^2
ICC* Effect Size * Unequal	2	6.21*
Level 2 Sample Size* Effect Size * Unequal	4	125.37***
Level 2 Sample Size*Level 1 Sample Size* Correlation	8	37.92***
Level 2 Sample Size*Level 1 Sample Size*ICC	4	14.66**
Level 2 Sample Size*ICC* Effect Size	4	11.18*
Level 1 Sample Size*ICC* Effect Size	4	13.27*
Level 1 Sample Size*ICC* Correlation	4	25.01***

*p < .05 **p < .01 ***p < .001

Results of omnibus tests for lower order effects in the small and large ICC conditions are shown in Table 3.15. Significance tests were largely the same across ICC conditions with the exceptions of the Level 2 sample size by random effects correlation interaction and the main effect of random effects correlation. When ICC was large, the Level 2 sample size by correlation interaction is nonsignificant whereas when ICC was small it was significant. The opposite occurred for the main effect of correlation; when ICC was small, the main effect of correlation was non-significant whereas it was significant when ICC was large.

Table 3.15 Omnibus Tests of Simple Effects at Each Level of ICC

Effect	DF	ICC = .4	ICC = .7
Level 2 Sample Size	2	445.49***	432.10***
Level 1 Sample Size	2	12.57**	3.16
Effect Size	2	141.43***	298.05***
Correlation	2	2.24	12.39**
Unequal	1	.01	.02
Level 2 Sample Size * Level 1 Sample Size	4	11.58*	15.12**
Level 2 Sample Size * Effect Size	4	275.07***	361.42***
Level 2 Sample Size * Correlation	4	10.68*	7.69
Level 1 Sample Size * Effect Size	4	78.96***	20.73***
Level 1 Sample Size * Correlation	4	4.28	9.26
Effect Size * Unequal	2	96.66***	117.19***
Level 2 Sample Size * Unequal	2	7.03*	16.97***
Level 2 Sample Size * Effect Size * Unequal	4	21.02***	73.68***
Level 2 Sample Size * Level 1 Sample Size * Correlation	8	24.54**	21.40**

*p < .05 **p < .01 ***p < .001

Table 3.16 shows the predicted probabilities of the effect size by equality of parameters interactions in the small and large ICC conditions. Overall, power increased at a faster rate when ICC was large compared to when ICC was small. When the effect size was small, power was uniformly small in all cases. However, as effect size increased power increased at a greater rate when ICC was large. Given the inequality among differences in percent variance explained across the parameter equality groups, this effect was not unexpected. Percent variance explained was always higher in conditions where the number of parameters across models were different.

Table 3.16 Predicted Probabilities from Effect Size x ICC x Equality of Parameters Interaction

	ICC			
	.4		.7	
	Equal	Unequal	Equal	Unequal
Small	3%	3%	5%	5%
Medium	6%	9%	8%	12%
Large	7%	16%	12%	23%

Interactions involving Level 1 sample size were generally too small to be qualitatively interesting. Although they were statistically significant, interactions involving Level 1 sample size tended to result in effects changing in magnitude of no more than a few percent. Generally, effects slightly increased as Level 1 sample size increased. Despite their significance, effects were so small that researchers would gain almost nothing in terms of power to detect differences in models non-nested at Level 2 by increasing Level 1 sample size. As a result they were omitted from discussion.

Figure 3.2 shows the two-way interaction between effect size and Level 2 sample size at small and large ICCs. Comparing these two panels it can be observed that power differences between Level 2 sample size conditions were slightly greater in in the high ICC condition, and as effect size increased those differences become more pronounced. For instance, when ICC was low, the difference between large and small Level 2 sample sizes for a big effect size was 37%. When ICC was high, that difference grew to 42%. The same comparisons when Level 2 sample size was small resulted in differences of 11% and 12%, respectively.

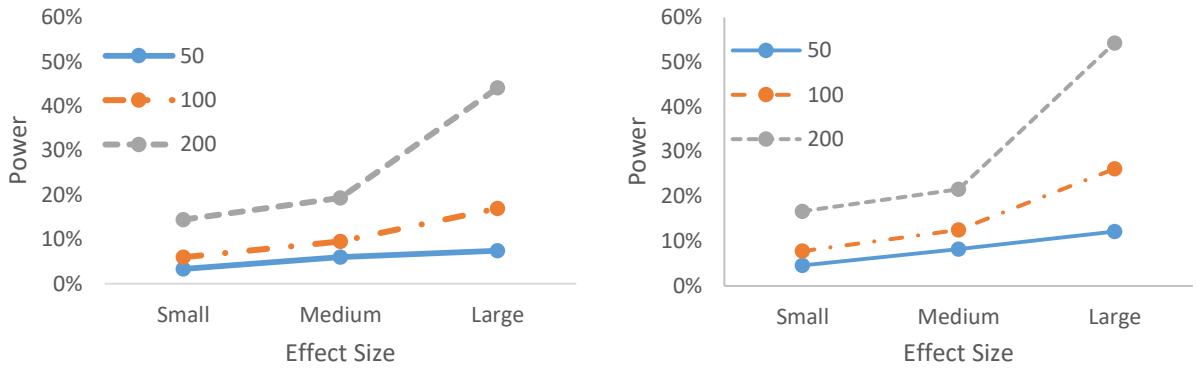


Figure 3.2 Effect Size by Level 2 Sample Size Interaction in Small ICC (Left) and Large ICC (Right) Conditions

Two three way interactions remained that did not depend on ICC but did include Level 2 sample size: Level 2 sample size by effect size by equality of parameters and Level 2 sample size by Level 1 sample size by random effects correlation (explored because it contained Level 2 sample size). The Level 2 sample size by effect size by equality of parameters interaction can largely be explained by the inequality of effect sizes across number of parameter groups. Because effect sizes were uniformly larger when the number of parameters among candidate models was unequal, the effect size by Level 2 sample size interaction should be greater when the number of parameters was unequal. Finally, results indicated that there were no qualitative differences (< 2%) in the random effects correlation by Level 2 sample size interaction across Level 1 sample sizes.

Discussion

In this study I have compared the performance of Vuong's test to that of information criteria when attempting to select between two candidate models non-nested in their fixed effects. When comparing the performance of Vuong's test to the performance of information criteria to select the best model in non-nested Level 1 covariates, Vuong's test tended to be sufficiently powered to detect the correct effect in a large proportion of conditions specifically

when effect size was large or there were at least 13 observations at Level 1. Information criteria also performed rather well when examining non-nestedness in Level 1 covariates. In fact, information criteria only selected the wrong model in a significant proportion of replications when Level 1 and Level 2 sample sizes were small.

However, when choosing between models non-nested in their Level 2 covariates the information criteria selected the wrong model much more frequently whereas Vuong's test was rarely ever significant in favor of the wrong model. Even though Vuong's test rarely ever reached adequate power when detecting differences in models non-nested at Level 2, rather than select the incorrect model it failed to reject the null hypothesis. The null result of Vuong's test suggested that there is insufficient evidence in support of one model over the other. Alternatively, it implies that both models fit the data equally well.

To say that Vuong's test outperformed information criteria by failing to reject the null hypothesis rather than selecting the correct model begs the question "What do you do when Vuong's test fails to reject the null hypothesis?" Depending on the purpose of the model comparisons and the underlying theories, researchers can follow a number of different paths. First, should one model be more complex than another (e.g., contain more complex interaction terms or more variables in general) logic would dictate selection of the more parsimonious model. That is, if model fit is not improved by adding additional, or a more complex set, of predictors, the simpler theory should be preferred.

Alternatively, one could use Vuong's test as a pseudo-diagnostic for the information criteria. If Vuong's test prefers one model over another, a researcher can be confident that a commensurate difference in information criteria is in fact detecting the best model. If Vuong's test fails to converge on the same conclusion one would draw from comparing information

criteria, researchers might still proceed with the model preferred by information criteria but with caution, understanding that while the information criteria are more sensitive to differences in non-nested models, they might have chosen incorrectly. To confirm their theory more research would be needed. Ultimately, researchers will need to continue with the model fitting process even in the presence of a non-significant result from Vuong's test. However, it is my hope that its results are considered when discussing the preferred model and conclusions are tempered accordingly.

I have also explored the factors that contribute to the power of Vuong's test to select the best model between pairs of candidates. Overall, Level 1 sample size, Level 2 sample size, and effect size affected power as expected. Power increased with increases in either factor. Effects occurred in the opposite direction when the main effect of ICC was tested for non-nested Level 1 covariates compared to non-nestedness at Level 2. When non-nestedness occurred at Level 1, there was less power for Vuong's test to detect a significant effect when ICC was large compared to when ICC was small. However, more power was observed with a larger ICC when covariates were non-nested at Level 2. The effect of ICC when models are non-nested at Level 2 is counter to what is typically known about the effect of ICC on power (Hox, 2010). This anomalous result may be related to the fact that Vuong's test is conducted on the case wise (or individual specific) log likelihoods and more variability at the case level is advantageous. Further work is needed, however, to explain this anomalous result.

The random effects correlation exhibited main effects only when non-nestedness occurred at Level 2. As the random effects correlation increased, so did power to detect the best model. Power increased monotonically with effect size within equality of parameters groups. Power increases when the number of parameters were unequal appeared to be greater than when

parameters among candidates were equal, however, equality and effect size were confounded and effects may be attributable to a greater proportion of variance explained when there were unequal numbers of parameters. A single anomalous result where power was lower despite having a larger effect size in the small effect conditions motivated this dichotomization. All of these effects, both at Level 1 and Level 2, were qualified by higher order interactions.

Main effects of factors affecting power when models were non-nested in Level 1 covariates were qualified by a 4-way Level 1 sample size by Level 2 sample size by effect size by unequal number of parameters interaction. The Level 1 by Level 2 sample size interaction tended to behave similarly across effect sizes and equality of parameters, however, in the small effect size with unequal numbers of parameters (where the anomalous power rates surfaced) the greatest difference in power was seen when comparing the small and medium Level 2 sample sizes at the small Level 1 sample size. Otherwise, this interaction was largely driven by changes in the effects of variables expected as power approached 100%.

Interactions not qualified by the 4-way interaction also occurred when parameters were non-nested at Level 1. Specifically, a three-way interaction between Level 1 sample size, ICC, and random effects correlation, ICC and effect size, and Level 2 sample size and ICC all emerged as significant. However, only the ICC by effect size interaction appeared to make a qualitative difference in power. At larger effect sizes, the difference in power between large and small ICCs increased.

There were also several three-way interactions when non-nestedness occurred at Level 2, most of which were conditional on ICC. These interactions were such that simple two-way effects tended to increase in magnitude when ICC was greater. Additionally, the difference in the effects between ICC conditions tended to be larger when either Level 2 sample size was large or

effect size was large, depending on the interaction. For instance, a greater difference in power between the unequal and equal number of parameters conditions for the large effect size when ICC was large. A similar effect occurred for the Level 2 sample size by effect size by ICC condition; the increase in power from a medium to large effect in all Level 2 sample sizes was greater when ICC was large compared to when ICC was small, however, the difference was amplified when Level 2 effect size was large.

While non-nested covariate sets as depicted in this simulation do not constitute the most interesting application of Vuong's test, compelling cases require only a small departure from models described here. For instance, a researcher may desire to understand a curvilinear relationship and be forced to decide between two growth curves: one with a quadratic trend and the other defined through a piecewise function. Although these two forms could be used to understand a curvilinear trend in the data, their parametrization would be a case of non-nested fixed effects. As a result, their treatment would be no different than in this chapter. Assuming non-nestedness only occurs in the fixed effects, the only difference to the likelihood would occur in the fixed effects design matrix which would affect only the residual term, **r**. The residual vector **r** is also the only aspect of the likelihood affected by non-nestedness in this study. Therefore it would be reasonable to assume similar behavior of Vuong's test with other non-nested fixed effects. Detecting differences in truly nonlinear forms is a specific and more complex case addressed in Chapter 5.

Cross level interactions, particularly those involving time, are often of particular interest when fitting multilevel models to longitudinal data. That is, the motivating research question is typically not to describe the effects of participant characteristics at baseline, but the effects of change over time. Exploring the power of Vuong's test to detect non-nested differences in these

cross-level interactions is a logical extension for this research in the future. Open questions surround the effects of ICC, sample sizes at different levels, as well as the differences in the number of parameters between candidate models since these factors do not necessarily result in the same effects across levels. By nature, the cross-level interaction contains effects at multiple levels and therefore the effects of different study factors remains open for inquiry.

Limitations

While this study exhibits for the first time the behavior of Vuong's test in multilevel regression and compares its performance to that of information criteria, it employs a straightforward model. Rarely are models as simple as the one utilized in this study (i.e., no cross level interactions nor are there interactions with time) and outlined above are a number of extensions for the proposed method. Proving Vuong's test's utility and comparing its performance to the current standard, however, served a necessary preliminary function. Beyond the model form utilized by the present study, other limitations bound its generalization. First, the present study continues to explore the properties of Vuong's test, only in the context of growth curves. However, as discussed in chapters 1 and 2 the results of this study should apply directly to multilevel models of individuals within groups. An additional study should be undertaken exploring the power of Vuong's test and performance relative to information criteria in study designs common to multilevel models with exchangeable observations (e.g., larger Level 1 sample size, smaller Level 2 sample size, and smaller ICC). Finally, given that there is currently no known way to operationalize the difference between (e.g., small, medium, or large true differences in likelihoods) the effect size component used in this study was developed somewhat ad hoc. While the rudimentary effect size used herein was sufficient for describing the general

behavior of Vuong's test, it would be instructive for future studies to establish a well-defined and empirically vetted measure of model difference.

Despite these limitations, this study served as a first step in understanding the behavior of Vuong's test over a variety of Level 1 sample sizes, Level 2 sample sizes, ICCs, effect sizes, and random effects correlations in longitudinal multilevel models. Additionally, Vuong's test was shown to serve its purpose in cases when comparisons among information criteria did not perform well. That is, when information criteria chose the wrong model at a high rate, Vuong's test failed to reject the null hypothesis. However, when Vuong's test selected the best model with sufficient power, information criteria also tended to select the best model. Thus at this early stage, Vuong's test might be used as more of a diagnostic test to determine the probable accuracy of information criteria rather than being used independently. In the following chapters, I extend Vuong's test to contexts where multilevel modeling is the more obvious preferred approach: non-nested Level 1 residual covariance structures and non-nested non-linear forms of growth.

Chapter 4: Testing Non-nested Level 1 Covariance Structures

In this chapter I explore the performance of Vuong's test to detect the correct model when comparing two candidates non-nested in their Level 1 residual covariance structures. While most multilevel models are specified using a default identity structure, autoregressive or Toeplitz structures are attractive alternatives for longitudinal or time series data (Kwok, West, & Green, 2007). In fact, autoregressive models now enjoy regular use in psychology through adoption in longitudinal structural equation models (Bollen & Curran, 2004, 2006). Ferron, Dailey, and Yi (2002) and Keselman, Algina, Kowalchuk, and Wolfinger (1998) both studied the performance of AIC and BIC in selecting the optimal covariance structure. Their results indicated that these methods were only marginally accurate and as a result more accurate methods are needed (Kwok, West, & Green, 2007). This study intends to serve that purpose.

Method

Data generation

Using SAS Proc IML, a data generation program similar to that used to generate data for non-nested covariate sets generated data with varying Level 1 covariance structures. The fixed effect for time was specified such that for every unit increase in time, the outcome would increase by .2. Time was represented by an integer variable and was generated with the maximum value defined by the Level 1 sample size for the condition with a value of zero indicating the baseline measurement. As a result time ranged from zero to either four, twelve, or twenty-four. Level 2 random effects were specified such that the residual random intercept variance would result in residual ICCs of either .4 or .7 for properly specified models. The random slope variance was defined as half of the random intercept variance and a covariance specified to obtain a correlation between slope and intercept of .4.

Level 1 residual covariance structures were generated using the "Toeplitz" function in SAS Proc IML. With the covariance matrices specified, the RANDNORMAL command was used to generate a matrix of multivariate normal variables with means of zero and the desired covariance matrix. This matrix was reshaped into a single n x p vector to facilitate data generation.

After writing the ID, outcome, and time variables to a dataset, Vuong's test for the two candidate models was conducted with a SAS Macro created for this study. The macro was used to estimate two multilevel models using SAS Proc MIXED under maximum likelihood estimation and Satterthwaite degrees of freedom; one model was fit using an autoregressive structure with lag 1 (AR(1)) and the other was fit using a Toeplitz(3) (TOEP(3)) structure. All models included fixed effects for time and random Level 2 intercepts and time slopes in addition to the structures imposed on the Level 1 residual covariance matrix. The results of both mixed models were parsed to obtain the necessary quantities to calculate the individual log-likelihoods. Once individual log-likelihoods were calculated, it was possible to conduct Vuong's test and save the output for analysis.

A factorial design was used to generate data from every combination of 3 Level 1 sample sizes, 3 Level 2 sample sizes, 2 ICCs, and 4 different true models representing a "small" and "large" effect of the Level 1 covariance structure. As will be discussed below, small and large are relative terms comparing the amount of misfit introduced; they do not refer to any accepted effect size scale. One thousand replications were generated for each of the 72 study conditions resulting in 72,000 unique datasets for analysis.

Sample size. The same sample sizes used in the study examining non-nested covariate sets were used to examine non-nested residual error structures as well. Data were generated for 50, 100, or 200 individuals at Level 2 each with 5, 13, or 25 observations at Level 1.

ICC. Two values of ICC were included in this study, .4 and .7. Because longitudinal models require a time effect by definition and this effect is typically assumed to vary in multilevel models, ICC was conceptualized as a residual ICC after accounting for the time slope and the Level 1 residual covariance structure. That is, the specified ICCs denote the residual ICC estimated when the Level 1 and Level 2 covariance structures are both properly specified (i.e., a time slope at Level 2 and the proper structure at Level 1).

Level 1 Covariance Structures: Effect Size. Following the logic set forth by Kwok, West, and Green (2007), autoregressive and Toeplitz structures were utilized to evaluate the utility of Vuong's test for non-nestedness when non-nesting manifests in Level 1 covariance structures. Both the autoregressive and the Toeplitz structures provide a mechanism by which Level 1 residuals can adhere to a structure, however, the Toeplitz structure can provide more flexibility than the autoregressive model. Figure 4.1 shows the general specification of the Toeplitz and autoregressive structures. Here, the Toeplitz structure contains 3 non-zero bands and is referred to as Toeplitz(3) or TOEP(3) and the autoregressive structure is autoregressive with a lag of 1, AR(1).

$$TOEP(3) = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 & 0 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ 0 & \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix} AR(1) = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Figure 4.1 General Forms of Toeplitz(3) and Autoregressive(1) Structures

As can be seen from the above structures, the Toeplitz structure estimates a variance and a unique covariance for each off diagonal band not indexed by a zero. Each band is estimated independently and can take any admissible value regardless of the band that preceded it. As a result there are as many parameters estimated at Level 2 as there are non-zero bands in the matrix. The AR(1) structure requires a stricter form. That is, the covariance changes at a specific rate defined by the parameter ρ. At each successive band, the rate at which the variance changes decreases exponentially, however, the exponentiated parameter is constant across bands. Thus in the AR(1) structure only a single parameter besides the variance need be estimated to fit the Level 1 residual variance structure.

The parameters chosen for the four Level 1 residual variance structures were based on those of Kwok, West, and Green (2007). The "small" covariance structure was defined by diagonal bands decreasing with an autoregressive parameter of .5 for both the Toeplitz and AR(1) structures, whereas the "large" covariance structure was based on off diagonal bands decreasing with an autoregressive parameter of .7. An example of this specification for the small effect size condition is provided below. While Kwok, West, and Green used a large autoregressive parameter of .8, the same value in this simulation resulted in matrices with negative eigenvalues that were not invertible. As such, the largest coefficient that was found to be invertible (.7) was used instead. Additionally the TOEP(3) model with a large Level 1 sample size produced negative eigenvalues when non-zero bands followed an autoregressive pattern. As a result a small degree of misspecification was added to the large TOEP(3) condition with the third band taking a value of .3 rather than .49 as would be dictated by an autoregressive pattern.

An effect size was created by the degree of misfit between the two Level 1 covariance structures. For instance, when the true variance structure was the small Toeplitz(3) structure,

misfit for an estimated AR(1) structure was created by the number of zeroes in the matrix (and a smaller third band in the large TOEP(3) condition). For example, the Toeplitz(3) and AR(1) matrices below show the true covariance structure for the small effect conditions when Level 1 sample size was small.

$$TOEP(3) = \begin{bmatrix} 1 & .5 & .25 & 0 & 0 \\ .5 & 1 & .5 & .25 & 0 \\ .25 & .5 & 1 & .5 & .25 \\ 0 & .25 & .5 & 1 & .5 \\ 0 & 0 & .25 & .5 & 1 \end{bmatrix} \quad AR(1) = \begin{bmatrix} 1 & .5 & .25 & .125 & .0625 \\ .5 & 1 & .5 & .25 & .125 \\ .25 & .5 & 1 & .5 & .25 \\ .125 & .25 & .5 & 1 & .5 \\ .0625 & .125 & .25 & .5 & 1 \end{bmatrix}$$

Sources of misfit are identified by the triangular regions in the above matrices for the small effects and include the dotted region when effects are large. When the true model is TOEP(3), the AR(1) structure will not fit the data properly in an attempt to accommodate the zero covariances in the fourth and fifth bands. Conversely, when the true model is AR(1), the TOEP(3) model will be unable to accommodate any nonzero covariance beyond the first two off diagonals. These differences are expected to increase when the non-zero off diagonal elements are larger and there is more misspecification between candidate models.

Parametrizing effect size in this way confounds the effects of Level 1 sample size and discrepancies in the Level 1 residual covariance matrix. By increasing Level 1 sample size I increase the number of off diagonal zeroes for the TOEP(3) structure and additional covariance is estimated in the AR(1) structure. Therefore the effects of these factors cannot be truly differentiated. Although this approach implies a Level 1 sample size by effect size interaction, it is not without precedence. Ferron, Dailey, and Yi (2002) used precisely this approach when examining the effects of misspecifying the Level 1 residual covariance structure in two Level models. While the effects of the individual study factors may be confounded, the choice was made to introduce this confounding in an effort to keep models consistent across Level 1 sample sizes.

Data Analysis

To evaluate the performance of Vuong's test its propensity to select the correct model was compared to that of information criteria. First, tables are presented comparing the correct model selection (correct classification) of Vuong's test to the correct classification of information criteria. Second, results where the incorrect model was chosen (misclassification) by Vuong's test is contrasted with the misclassification rate of information criteria. Conditions where information criteria select the incorrect model significantly more often than Vuong's test are highlighted. Misclassification of Vuong's test was at most only slightly over 1%, and as a result a conservative estimate of 2% was used to determine the power to detect a significant difference in proportions. G*Power 3.1.5 was used to calculate the difference from a constant proportion of 2% that could be detected with a power of .8 and a total sample size of 973. Although 1000 replications were generated for each condition, occasional convergence issues necessitated a small number of cases be discarded. The smallest remaining condition contained 973 replications. This number was conservatively used for all differences. The results of the sensitivity analysis indicated that a difference of 1.4% (from 2%) could be detected with a power of .8 and as a result any misclassification of at least 4% could be considered a significant degree of misclassification for the information criteria. Finally, the non-significance rates of Vuong's test were highlighted to illustrate the advantage of Vuong's test in cases where information criteria are error prone.

Logistic regression was then used to determine the effect of Level 1 sample size, Level 2 sample size, ICC, and effect size on the ability of Vuong's test to detect the correct model. Contrary to chapter 3 where the true data generating process was not among candidates, this study seeks to select the true model. Using SAS PROC Logistic with Firth's penalized

likelihood, a main effects model was estimated to understand the general behavior of Vuong's test in the context of non-nested Level 1 covariance structures and across study factors described above. After defining the general trends in the effects of Vuong's test across study factors, a full logistic regression model was estimated with all possible higher order interactions. Starting at the highest level, non-significant effects were removed one-by-one based on their significance level. Non-significant lower order effects were retained if their components contributed to a higher order interaction. Throughout analyses, all study factors were treated categorically.

Results

Because results across the true Toeplitz and autoregressive candidates are not directly comparable they are analyzed independently. Within each true structure, however, both effect sizes were analyzed concurrently and included as a study factor. I first discuss results describing Vuong's test when the true data generating process contains a TOEP(3) structure for the Level 1 residual covariance matrix before transitioning to the true AR(1) structures. Discussions will first contrast the performance of Vuong's test with that of the information criteria beginning with an examination of the differences in the methods' propensity to select the correct models and then their propensity to select the incorrect model which is referred to here as misclassification. I then provide results from analyses examining differences in how study factors affect the power of Vuong's test to detect the correct model.

True Toeplitz Model

Selecting the Correct Model. To compare the results of Vuong's test with the performance of information criteria I first examined the rates at which each model selection procedure selected the true model. Table 4.1 provides rates of correct model selection for Vuong's test and information criteria. Cells are shaded to indicate conditions in which the power

of Vuong's test to detect the true data generating process is at or above .8. As can be seen in the table, when there are at least 13 observations at Level 1, Vuong's test almost always has enough power to detect to true model. The only exceptions to the uniformly high power rates occur when Level 2 sample size is small the effect size is small, and Level 1 sample size has 13 observations. Still, power rates remain over .7.

When the power of Vuong's test is high, so is the correct model selection rate of the information criteria. Comparisons among information criteria select the correct model in almost every case when Level 1 sample size is 13 observations or more. Slight deviations from 100% correct model selection occur in conditions where Vuong's test loses power as well, however, correct classification rates remain over 95% for information criteria.

Selecting the correct model becomes more tenuous when Level 1 sample size is small, especially in the small effect size condition. When effect size was large, Vuong's test and information criteria both tended to perform well, however, Vuong's test was slightly underpowered when Level 2 sample size was small and ICC was large. In the small effect size condition, performance was suspect. Vuong's test was exceedingly underpowered when effect size was small and there were only 5 Level 1 observations. When Level 2 sample size was also small, the power of Vuong's test was in the single digits. Even when sample size reached 200, the power of Vuong's test was at most 40%.

In these conditions when the power of Vuong's test was exceedingly bad, information criteria also tended to perform poorly. When Level 2 sample size was small, information criteria performed no better than randomly selecting from the two candidates. In fact, BIC performed worse than random selection. As Level 2 sample size increased, the performance of information criteria improved with AIC and AICc reaching correct model selection rates of over 90% when

Level 2 sample size was 200. BIC on the other hand, continued to perform poorly when Level 2 sample size was large; it selected the correct model just over 70% of the time.

Table 4.1 Correct Model Selection Rates for True Toeplitz Models

			L1SS											
			5				13				25			
ICC	Effect	L2SS	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC
0.4	Toep(3) .5	50	7%	52%	50%	26%	75%	99%	99%	97%	97%	100%	100%	100%
		100	16%	71%	70%	43%	95%	100%	100%	100%	100%	100%	100%	100%
		200	40%	91%	91%	70%	100%	100%	100%	100%	100%	100%	100%	100%
	Toep(3) .7	50	82%	99%	98%	93%	100%	100%	100%	100%	100%	100%	100%	100%
		100	99%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
0.7	Toep(3) .5	50	6%	46%	44%	25%	74%	98%	98%	97%	98%	100%	100%	100%
		100	13%	70%	69%	43%	96%	100%	100%	100%	100%	100%	100%	100%
		200	36%	93%	92%	72%	100%	100%	100%	100%	100%	100%	100%	100%
	Toep(3) .7	50	73%	97%	97%	90%	100%	100%	100%	100%	100%	100%	100%	100%
		100	95%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Misclassification. Vuong's test may not appear to add substantial value over information criteria when differences between models are easily detected. It can, however, outperform information criteria when differences between models are not as obvious. Table 4.2 contains the misclassification rates of Vuong's test and the information criteria. At most, Vuong's test selects the incorrect model in 1% of replications. Based on the sensitivity analysis discussed above, any misclassification greater than 4% in information criteria can be considered significantly greater than the misclassification of Vuong's test.

With respect to misclassification, Vuong's test performs quite well. When Level 1 sample size is at least 13, Vuong's test never selects the incorrect model. Even when Level 2 sample size and effect size were small, Vuong's test did not ever select the incorrect model. While not technically significant, it is notable that when Level 2 sample size and effect size were both small, there was still a small degree of misclassification for the information criteria. To be absolutely positive of a correct result, samples were required to be rather large.

When Level 1 sample size was small, Vuong's test selected the incorrect model in at most 1% of cases. While Vuong's test, AIC, and AICc generally performed well when effect size was large, BIC continued to select the incorrect model with significant frequency when Level 2 sample size was small. The biggest benefit of Vuong's test can be seen in the small effect size conditions.

When effect size was small, information criteria selected the incorrect model in a significant number of replications regardless of Level 2 sample size. In the smallest conditions, AIC and AICc chose the incorrect model roughly in roughly half of the replications whereas BIC chose the incorrect model in about three-quarters of replications. As effect size increased, model

selection tended to improve across all methods. However, even at the largest Level 2 sample sizes BIC was still selecting the incorrect model in about 30% of cases.

The benefit of Vuong's test can be best observed when differences in fit among candidate models are difficult to detect. Table 4.3 displays the non-significance rates of Vuong's test. Shaded cells indicate conditions in which information criteria perform significantly worse than Vuong's test. Worth noting in this table is the large values in the first column representing the small Level 1 sample size. When information criteria are selecting the correct model at rates worse than chance, Vuong's test produces a null result indicating that there is insufficient evidence to prefer one model over another. Although ambiguous, this null result would be preferable to making an incorrect decision. Thus it could be argued that Vuong's test is most beneficial, when information criteria are performing at their worst.

Table 4.3 Non-significance Rates of Vuong's Test

			L1SS		
ICC	Effect	L2SS	5	13	25
0.4	Toep(3) .5	50	93%	25%	3%
		100	83%	5%	0%
		200	60%	0%	0%
	Toep(3) .7	50	17%	0%	0%
		100	1%	0%	0%
		200	0%	0%	0%
0.7	Toep(3) .5	50	94%	26%	2%
		100	87%	4%	0%
		200	64%	0%	0%
	Toep(3) .7	50	27%	0%	0%
		100	4%	0%	0%
		200	0%	0%	0%

Table 4.2 Incorrect Model Selection Rates for True Toeplitz Models

			L1SS											
			5				13				25			
ICC	Effect	L2SS	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC
0.4	Toep(3) .5	50	0%	48%	50%	74%	0%	1%	1%	3%	0%	0%	0%	0%
		100	1%	29%	30%	57%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	9%	9%	30%	0%	0%	0%	0%	0%	0%	0%	0%
	Toep(3) .7	50	1%	1%	2%	7%	0%	0%	0%	0%	0%	0%	0%	0%
		100	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0.7	Toep(3) .5	50	0%	54%	56%	75%	0%	2%	2%	3%	0%	0%	0%	0%
		100	0%	30%	31%	57%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	7%	8%	28%	0%	0%	0%	0%	0%	0%	0%	0%
	Toep(3) .7	50	0%	3%	3%	10%	0%	0%	0%	0%	0%	0%	0%	0%
		100	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Power. Results of a logistic regression with Firth's penalized likelihood indicated significant omnibus tests for the effects of Level 1 sample size ($\chi^2(2) = 4441.75$, $p < .0001$), Level 2 sample size ($\chi^2(2) = 1507.79$, $p < .0001$), ICC ($\chi^2(1) = 23.68$, $p < .0001$), and effect size ($\chi^2(1) = 3793.50.75$, $p < .0001$). As sample size at both Level 1 and Level 2 increased so did the power of Vuong's test to detect the correct model. Similarly, power was greater when the non-zero covariances among residuals were large than when they were small. When ICC was large the power of Vuong's test to detect the correct model was reduced. Specifically, Vuong's test was 1.27 times less likely to detect a significant effect in favor of the correct candidate when ICC was .7 compared to when it was .4. While these main effects describe the behavior of Vuong's test on average, they were qualified by higher order interactions.

A final model with non-significant higher order terms trimmed indicated a significant three way interaction for Level 1 sample size, Level 2 sample size, and effect size ($\chi^2(4) = 26.53$, $p < .0001$). Figure 4.2 shows the Level 1 sample size by Level 2 sample size interaction for the small and large effect size conditions. In the large effect size condition (right) it is clear that power was at its maximum for all but the smallest sample sizes. When Level 1 and Level 2 sample size were both small, power was above 80%, however, power at medium and large Level 2 sample sizes was at or are extremely close to 100% even when Level 1 sample size is small. When Level 1 sample size reaches 13 time points, even the small Level 2 sample size had reached maximum power. In contrast the small effect size condition (left) illustrates that increasing Level 1 sample size from 5 to 13 has a greater effect on power when Level 2 sample size is 100 than when it is 50 or 200. As the medium and large Level 2 sample sizes reach maximum power at the Level 1 sample size of 13, the improvement in power gained in the 50

observation Level 1 sample size condition is greater than the larger Level 1 sample sizes, as would be expected.

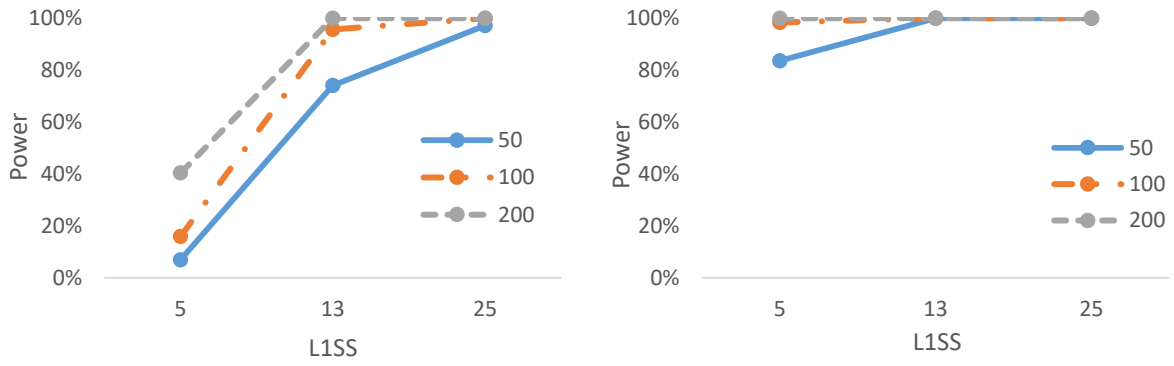


Figure 4.2 Level 1 Sample Size by Level 2 Sample Size Interaction in the Small (Left) and Large (Right) Effect Size Conditions.

Two lower order interactions not qualified by the three-way Level 1 sample size by Level 2 sample size by effect size interaction also emerged as significant: the Level 1 sample size by ICC interaction ($\chi^2(2) = 6.87$, $p = .032$) and the ICC by effect size interaction, $\chi^2(1) = 14.05$, $p = .0002$. Despite their significance, both of these interactions yield qualitatively uninteresting results. When Level 1 sample size is small, the small ICC condition has marginally greater power than when ICC is large. When Level 1 sample size is large the relationship is reversed; the large ICC marginally outperforms the small ICC condition. The ICC by effect size interaction suggests that at larger effect sizes, the difference in power between large and small ICCs becomes larger. Although significant, the difference was fairly trivial as the difference in power between small and large ICC conditions was only 2% greater than the difference between ICCs when effect size was small.

True AR(1) Model

Selecting the Correct Model. Following the approach taken to compare the performance of Vuong's test and information criteria when the true data generating process was Toeplitz, Table 4.4 displays the rates at which Vuong's test and information criteria select the correct model. Conditions in which Vuong's test had power greater than .8 are highlighted in the table. Vuong's test tended to achieve power of .8 when Level 1 sample size and effect size were both large. In cases when the effect size was small but Level 1 and Level 2 sample size was large, power still remained around 90%. The only instances in which Vuong's test achieved adequate power with a Level 1 sample size below 25 was when effect size and Level 2 sample size were both large and there were 13 observations at Level 1. In all other cases, Vuong's test failed to reach adequate power.

Information criteria, on the other hand, tended to perform rather well when the true data generating mechanism was AR(1). In every condition AIC and AICc selected the correct model over 80% of the time, whereas BIC selected the correct model in over 94% of replications in all but three conditions. Although there was never a case in the AR(1) conditions when information criteria performed exceedingly poorly, high correct model selection rates are only part of the story.

Misclassification. Misclassification for Vuong's test and information criteria are presented on Table 4.5. Once again, Vuong's test selects the incorrect model in at most 1% of replications. While information criteria also tended to perform well when Level 1 sample size was large, especially when effect size was also large, there was a stark contrast in performance in the small and medium Level 1 sample sizes.

Table 4.4 Correct Model Selection Rates for True Autoregressive Models

			L1SS										
			5				13				25		
ICC	Effect	L2SS	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC
0.4	AR(1) .5	50	1%	86%	87%	97%	7%	84%	85%	94%	30%	95%	95%	97%
		100	0%	80%	81%	96%	17%	90%	90%	97%	58%	99%	99%	100%
		200	0%	83%	84%	98%	36%	97%	97%	99%	89%	100%	100%	100%
	AR(1) .7	50	1%	82%	84%	96%	40%	97%	97%	99%	96%	100%	100%	100%
		100	1%	83%	83%	96%	69%	98%	98%	99%	100%	100%	100%	100%
		200	2%	83%	83%	97%	96%	100%	100%	100%	100%	100%	100%	100%
0.7	AR(1) .5	50	0%	84%	84%	95%	9%	85%	85%	94%	32%	94%	94%	97%
		100	0%	82%	83%	97%	16%	88%	88%	96%	60%	99%	99%	99%
		200	0%	82%	82%	97%	35%	95%	95%	98%	88%	100%	100%	100%
	AR(1) .7	50	1%	81%	82%	94%	37%	96%	96%	98%	97%	100%	100%	100%
		100	1%	80%	81%	96%	69%	99%	99%	100%	100%	100%	100%	100%
		200	1%	81%	81%	98%	95%	100%	100%	100%	100%	100%	100%	100%

Table 4.5 Incorrect Model Selection Rates for True Autoregressive Models

			L1SS											
			5				13				25			
ICC	Effect	L2SS	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC	VT	AIC	AICc	BIC
0.4	AR(1) .5	50	0%	14%	13%	3%	1%	16%	15%	6%	0%	5%	5%	3%
		100	1%	20%	19%	4%	0%	10%	10%	3%	0%	1%	1%	0%
		200	1%	17%	16%	2%	0%	3%	3%	1%	0%	0%	0%	0%
	AR(1) .7	50	0%	18%	16%	4%	0%	3%	3%	1%	0%	0%	0%	0%
		100	0%	17%	17%	4%	0%	2%	2%	1%	0%	0%	0%	0%
		200	1%	17%	17%	3%	0%	0%	0%	0%	0%	0%	0%	0%
0.7	AR(1) .5	50	1%	16%	16%	5%	1%	15%	15%	6%	0%	6%	6%	3%
		100	1%	18%	17%	3%	0%	12%	12%	4%	0%	1%	1%	1%
		200	1%	18%	18%	3%	0%	5%	5%	2%	0%	0%	0%	0%
	AR(1) .7	50	1%	19%	18%	6%	0%	4%	4%	2%	0%	0%	0%	0%
		100	1%	20%	19%	4%	0%	1%	1%	0%	0%	0%	0%	0%
		200	1%	19%	19%	2%	0%	0%	0%	0%	0%	0%	0%	0%

Information criteria performed generally well in the medium sample size conditions when effect size was large. By the significance standard set by the afore mentioned sensitivity analysis, AIC and AICc perform significantly worse than Vuong's test when ICC was high, and Level 2 sample size was small when effect size was large and 13 observations occupied Level 1. In conditions where effect size was small, AIC and AICc consistently selected the incorrect model in a significant proportion of cases, except when Level 2 sample size was large and ICC was small. BIC performed slightly better than the other information criteria only significantly misclassifying Level 2 sample size was small and ICC was small or Level 2 sample size was small and ICC was large.

The greatest advantage of Vuong's test was observed when there were few Level 1 units. With 5 observations at Level 1, AIC and AICc consistently selected the incorrect model in over 15% of replications with the only exception being the small effect size with a small Level 2 sample size and small ICC where misclassification dropped to 14% and 13% for AIC and AICc, respectively. BIC generally performed well. Although significant misclassification occurred in some conditions, it never rose above 6% and tended to do so only when Level 2 sample size was small. Cells are shaded to indicate conditions in which information criteria select the incorrect model in a significant proportion of cases.

Although when Level 1 sample size was small Vuong's test rarely ever selected the correct model, it also rarely selected the incorrect model. Non-significance rates from Vuong's test can be seen on Table 4.6. Although these high rates of non-significance when the number of Level 1 units was small are not ideal, they are preferable to the large misclassification rates exhibited by information criteria.

Table 4.6 Non-Significance of Vuong's Test

			L1SS		
ICC		L2SS	5	13	25
0.4	AR(1) Small	50	99%	92%	70%
		100	99%	83%	42%
		200	99%	64%	11%
	AR(1) Large	50	99%	61%	4%
		100	99%	31%	0%
		200	97%	5%	0%
0.7	AR(1) Small	50	99%	90%	68%
		100	99%	84%	40%
		200	99%	65%	12%
	AR(1) Large	50	98%	63%	3%
		100	98%	31%	0%
		200	98%	6%	0%

Power. Logistic regression with Firth's penalized likelihood was used again to determine the effects of study factors on the power of Vuong's test to select the correct candidate. A main effects model was first estimated to determine the average trends of study factors affecting power. Results indicated significant main effects for Level 1 sample size ($\chi^2(2) = 5990.02$, $p < .0001$), Level 2 sample size($\chi^2(2) = 2828.59$, $p < .0001$), and effect size ($\chi^2(1) = 4659.56$, $p < .0001$). As expected, as sample size increases at either level, power increases. Effect size also maintained the expected positive relationship with power. ICC, on the other hand, was non-significant, $\chi^2(1) = .056$, $p = .8136$. There was no difference in power on average between low ICC and high ICC. These main effects were qualified by higher order interactions.

A full model testing all of the interactions between study factors resulted in no interactions of an order greater than two nor was there any effect of ICC whatsoever. The final model included main effects for Level 1 sample size, Level 2 sample size, and effect size as well

as the two-way interactions among the three factors. The Level 1 sample size by Level 2 sample size interaction ($\chi^2(4) = 153.36$, $p < .0001$) is displayed on Figure 4.3 with values estimated values collapsed across effect sizes. The interaction was such that as Level 1 sample size increased, the difference in power between Level 2 sample sizes increased to a point and began decrease as power approached the asymptote. When Level 1 sample size was small, power was at its minimum and there was no difference between Level 2 sample sizes. As Level 1 sample size increased to 13, the difference in power between Level 2 sample sizes increased as well. As Level 1 sample size increased further to 25 observations, power differences between Level 2 sample sizes began to decrease. Should Level 1 sample size increase beyond 25 observations it would be expected that there be no difference in power across Level 2 sample sizes as it reaches its maximum.

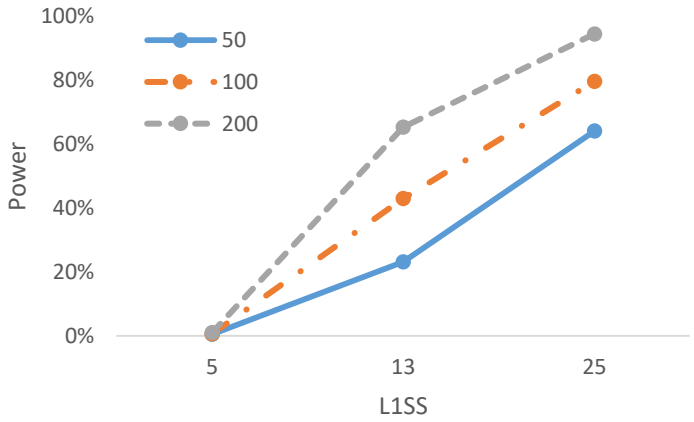


Figure 4.3 Level 1 Sample Size by Level 2 Sample Size Interaction Averaged Over Level 2 Sample Sizes

The Level 1 sample size by effect size interaction is displayed on Figure 4.4 averaged across Level 2 sample sizes. As Level 1 sample size increased the difference in power between the large and small effect size conditions once again grew initially, before decreasing as the large effect size condition reached its upper limit. As mentioned earlier, this interaction is in large part

a product of the confounding between effect size and Level 1 sample size. As Level 1 sample size increased, the effect size inherently increased as there were more zero elements in the comparison Toeplitz(3) candidate.

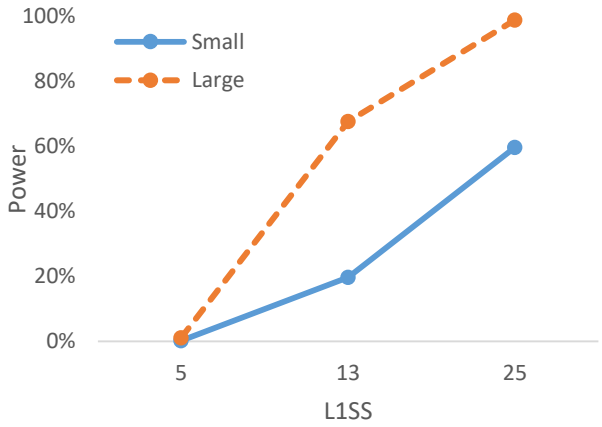


Figure 4.4 Level 1 Sample Size by Effect Size Interaction Averaged Over Level 2 Sample Sizes

Finally, the Level 2 sample size by effect size interaction were averaged over Level 1 sample sizes and are presented in Figure 4.5. While there was a large difference in the overall power between large and small effect sizes, this figure reflects the same pattern seen above: as power becomes high, increases in sample size tend to have less of an effect on power. When there is still room to improve power (i.e., it is low), increases in sample size have a greater effect on power.

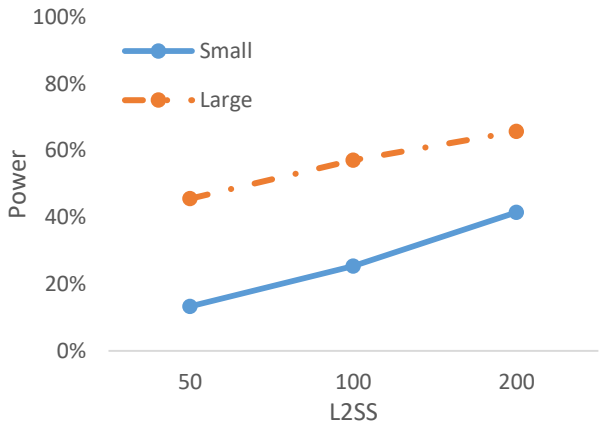


Figure 4.5 Level 2 Sample Size by Effect Size Interaction Averaged Across Level 1 Sample Sizes

Discussion

In this study I have examined the factors that contribute to the power of Vuong's test to detect the correct model when non-nestedness occurs in the Level 1 residual covariance structures. To examine this situation, I conducted a simulation study in which data generation included either a TOEP(3) structure where three parameters were estimated or an AR(1) structure estimating a residual variance and an autoregressive parameter. Within each structure two "effect sizes" were generated: one in which the off diagonal elements were based on a value of .7 and another based on a value of .5. The TOEP(3) structure was specified to have the same values as the autoregressive structure for the non-zero elements in the small condition with slightly more misspecification in the third band as required to remain invertible. However, beyond those elements misfit was introduced by restricting the covariances to zero. Alternatively, the autoregressive structures were fully autoregressive and contained non-zero values for every element.

Across data generating processes, Vuong's test outperformed information criteria in the sense that it rarely selected the incorrect candidate. In the case that Vuong's test was unable to detect the correct candidate it almost always failed to reject the null hypothesis. Information

criteria, on the other hand, exhibited exceedingly high misclassification rates. Rates as large as 74% were seen for BIC in Toeplitz models, indicating that in such conditions BIC is three times as likely as AIC or AICc to lead a researcher to select the wrong candidate. Information criteria also performed poorly, albeit to a lesser extent, in small sample sizes for the true Autoregressive models. This behavior provides insight into the tangible benefits of Vuong's test over tests using information criteria. To understand these benefits I explore the behavior of the information criteria themselves.

As was discussed in Chapter 1, information criteria are comprised of the deviance, or negative two times the log likelihood, and some penalty term. The calculation of the penalty term is what differentiates each information criterion. SAS calculates the information criteria using the following equations:

$$AIC\colon\ -2l + 2d$$

$$AICc\colon\ -2l + \frac{2dn^*}{n^* - d - 1} \tag{49}$$

$$BIC\colon\ -2l + dln(n)$$

where d is the number of parameters in the model (fixed and random), n* is the total sample size (t*i), and n is the number of Level 2 units. Essentially, these information criteria penalize the likelihood for model complexity where models with more parameters are considered to be more complex. From the results of this chapter it was obvious that the information criteria perform worse when Vuong's test is especially underpowered. That is to say that when the difference is so small that Vuong's test is almost never able to detect the difference between models, information criteria select the wrong model at a greater rate. This effect appears to be amplified when the number of parameters in the true model are greater than in the alternative. Information criteria performed especially badly when the difference between candidate models was small and

the true data generating process is more complex than the other candidate. Consider the case when the true data generating process was Toeplitz(3). Even through the true data model was among the candidates, because it was more complex than the alternative and there was a small difference between the two candidates in model fit, the information criteria prefer the simpler model. Furthermore, the information criteria with the more extreme penalty (i.e., BIC) selected the simpler and incorrect candidate more often. Thus, relying on information criteria to inform model selection is particularly problematic when the true data generating process is more complex than other candidates. Again, results suggest that information criteria were untrustworthy when the difference between models was small (e.g., when Vuong's test cannot detect a difference). This is precisely the case in which significance testing for model fit would be desirable.

This theory is supported when examining the performance of information criteria in the autoregressive models. Information criteria continue to display significant misclassification rates except with large sample size or effect size. However, in this case BIC outperformed AIC and AICc. Because the AR(1) model was more parsimonious than the TOEP(3) model, the larger penalty term worked in favor of selecting the correct model when it was in fact less complex than the other candidate.

While understanding the performance of information criteria is important, this study highlights exactly why their use is problematic in practice. Because data for this study were simulated and the true models known a priori, it was possible to know whether the correct model was being selected when using information criteria. In applications of this method, researchers would not know which model is the true or best model. Similarly, researchers would not know a priori the effect size for the models they are comparing and as a result would not know if

information criteria will lead them to the correct conclusion. Despite its lack of power in some of the conditions tested here, Vuong's test provided a safer alternative for researchers to test differences between two models. The results presented in this chapter suggest that significant tests of non-nested models can be trusted with near certainty and Vuong's test provides a real and conservative alternative to testing models non-nested in their Level 1 residual covariance structures.

Power followed the expected trends for both the true Toeplitz and true autoregressive models. When either Level 1 sample size, Level 2 sample size, or effect size increased so did the power of Vuong's test to detect a significant effect. There was a significant main effect for ICC when the true data generating process was Toeplitz, but not when it was Autoregressive. Power of Vuong's test tended to decrease in high ICCs when the true data generating process was Toeplitz. This effect appears to be driven largely by a single condition (i.e., large effect size when Level 1 and Level 2 sample sizes are small). At all other conditions the difference in power between small and large ICCs for the true Toeplitz structure was negligible.

While this study illustrated the ability of Vuong's test to detect the true Level 1 residual covariance matrix and studied its behavior across Level 1 sample size, Level 2 sample size, ICC, and effect size, it is limited by a number of factors. Most notably, effect size and Level 1 sample size were confounded. For instance, when the true data generating process was AR(1), the amount of covariance unable to be captured by the zero elements in the TOEP(3) candidate increased with the number of time points, not just the size of the autoregressive parameter. While this confounding limits the interpretability of the findings, namely that improvements in power thought to result from increases in Level 1 sample size might be attributed to the increased degree of misfit, it reflects situations that are encountered in the real world. Still, it would be

ideal to control for this confounding in future studies as more is learned about the power of Vuong's test. Additional confounding occurred between the number of parameters in the models and performance, particularly when examining performance of information criteria. In the previous study where non-nestedness occurred for Level 1 covariate sets, effects of other study factors differed depending on the number of parameters in candidate models. In the present study there was no condition in which the candidates had the same number of parameters and as a result, the added flexibility of the more complex Toeplitz model may have influenced the results. While it would have been possible to estimate a Toeplitz model with only two bands, it seemed a larger transgression to introduce additional misfit by restricting additional parameters to zero. Still, building on this work, it would be instructive to determine if model complexity significantly affects the performance of Vuong's test when testing Level 1 residual covariance structures, as it did when examining Level 1 covariate sets. Kwok, West, and Green (2007) chose to compare non-nested Level 1 covariance structures with the same numbers of parameters (i.e., AR(1) and Toeplitz(2)), however, they did not examine model selection based on information criteria, only bias and type 1 error rate of the fixed effects. Ferron, Dailey, and Yi (2002) examined the effects of misspecification on the ability of information criteria to select the best model using methodology similar to that in the current study. They compared the "identity" structure where a single residual variance is estimated for all time points and the autoregressive structure. Their results showed that correct identification rates were "unacceptable" specifically when the number of Level 1 units was small.

Addressing these confounds inherent to testing alternative Level 1 residual covariance structures would be a logical first step in future research exploring the power of Vuong's test in multilevel models. To advance the understanding of Vuong's test further, it would be instructive

to determine how different degrees and types of misspecification affect the performance of the test overall. Kwok, West, and Green (2007) found that non-nestedness in the Level 1 covariance structure (which they referred to as "general misspecification") was associated with overestimation of the Level 2 variances, over-estimations of the standard errors of fixed effects growth parameters, and decreased statistical power. Type-1 error for detecting the fixed growth parameters was not negatively affected. While Kwok, West, and Green only examined the effect of non-nestedness when the number of parameters among candidate models was equal, it is possible that there are combined effects between complexity and non-nestedness. A test of this hypothesis specifically would be informative.

Type-1 error rate is another area in which there is room to build on this research. The goal of this study was to first establish Vuong's test as a viable option for testing non-nested Level 1 covariance structures and compare its performance to model selection based on information criteria. While the current results of small effect sizes might suggest that type-1 error rate would be well within the nominal .05 limits, it would be useful to examine the type-1 error rates of Vuong's test, and the performance of information criteria, when there is no difference in model fit between two candidates non-nested in their Level 1 covariance structures. Doing so would require two different parameterizations of the Level 1 covariance structure that fit the data equally well. However, I know of no currently available method that can accomplish this task.

In this study I have established Vuong's test as a useful option for comparing two candidate models non-nested in their Level 1 covariance structures. Results indicated that Vuong's test rarely, if ever, suggests the wrong model be preferred over the true data generating model. Choosing the correct model based on comparisons among information criteria is more error prone. Misclassification rates of information criteria reached as high as 75% when the true model was

more complex than the alternative and there was a small degree of misfit. Thus, when the true model is unknown, as is the case in applied research, Vuong's test allows researchers to be more cautious in selecting the correct model and provides more accurate and reliable model selection.

Chapter 5: Testing Non-nested Functional Forms

The last context in which Vuong's test for non-nested multilevel models was explored was competing non-linear functional forms. Often the first step in the modeling process, understanding the shape of a curve is essential to appropriately describing trends in the data. Typically, researchers will examine general trends in their data at the outset of modeling through a number of means such as graphical representation (e.g., scatter plots), correlations, and other descriptive statistics (e.g., means and variances) across time points. These methods can, and often do provide a rough approximation of the shape of the phenomenon under study. With this information in hand, functional forms resembling the observed data can be fit and evaluated.

While linear growth is a simple and widely used function, it fails to capture some of the more curvilinear relationships often studied in psychology. From language development (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991) to randomized controlled trials (Weisz et al., 2012), there are many instances when linear growth curves might not be sufficient to fully address the research question. A common remedy for curvilinear data is the quadratic growth curve model. In addition to the linear trend, a quadratic term is included to capture differences in the rate of change at each time point. Although quadratic (and higher order) trends are sometimes sufficient for understanding curvilinear data, their use implies several assumptions. One such assumption of the quadratic model is that at some point growth changes directions, although this directional shift may not occur in the functional space being modeled. Additionally, a quadratic term suggests that growth occurs infinitely. This assumption is untenable when there is an absolute ceiling or floor inherent to the process under study (e.g., behavior). Therefore, despite being an improvement over the linear model, issues still arise when fitting and interpreting a higher order growth curve.

Some of the issues encountered with quadratic growth curves can be mitigated by truly non-linear functions. For instance, the exponential decay model decreases monotonically until it reaches an asymptote. This form solves the problem of changes in direction and predicting values outside of the admissible range (e.g., increasing to infinity). Additionally, when a data generating process is truly non-linear, and "non-quadratic" by extension, non-linear models may provide a more natural context in which to interpret results. Not only is it a possibility that the model fit the data better empirically, but also that the true non-linear form maps onto theory more naturally. The present study examines the performance of Vuong's test when selecting from different functional forms.

Although the entirety of this study is examining the performance of Vuong's test in detecting different functional forms, it is split into two parts. The first examines the power and performance of Vuong's test in multilevel non-linear growth curves when the growth parameter is fixed. In restricting growth to a fixed effect I was able to explore additional factors that might affect the performance of Vuong's test which created estimation issues when growth was random. In the second part of the chapter I examine how a restricted set of study factors might influence Vuong's test when there is also a random growth parameter. Although the non-linear models explored here do not have the same meaning of the "slope" parameter in linear models, I refer to the growth parameter as a "slope" for the remainder of the chapter.

Random Intercept Only Models

Method

Data Generation

Using SAS Proc IML, a data generation program was written to create data that followed one of two non-linear forms of decay: exponential and power (see below). Fixed and random effects

were loosely based on results from Cudeck and Harring (2007) and Timmons and Preacher (2015). In all models, the intercept or intercept analogue was set to a value of 27, the empirical intercept found by Cudeck and Harring. Initial slope values for each model were then set in decrements of .2 with one condition of each model set to a value of -.10. Initial results indicated a ceiling effect in power, specifically when the true data generating process followed a power function. Therefore intermediary conditions were generated with slope values between the two smallest slope parameters. I omit the original conditions where power is almost perfect.

The random intercept variance was specified to produce a residual ICC of either .86, .70, or .50 with a residual Level 1 variance of 10. Both the intercept and residual variances were normally distributed with means of zero. Additionally the large intercept variance and the Level 1 residual variance were both based on Cudeck and Harring (2007). An interval scale time variable was also created dependent on Level 1 sample size. Having specified values for all data generating parameters and a time variable, an outcome was calculated according to the true model of the current condition. True values were saved to be used for start values during model fitting. Through a process of trial and error, start values facilitating model fitting in the alternative non-linear candidate were specified as well.

Four models were fit to each dataset: linear, quadratic, exponential, and power. Linear and quadratic models were fit using SAS Proc MIXED with maximum likelihood estimation and Satterthwaite degrees of freedom. Non-linear models were estimated using SAS Proc NLMIXED with the "FIRO" estimation method and start values as defined above. Requisite output (i.e., parameter estimates and fit statistics) from all models was saved to be used in the calculation of Vuong's test or model fit analyses.

Using output from the fitted models, individual log likelihoods were computed for each model and Vuong's Test conducted for each pair of non-nested candidates. While Merkle et al., (2015) suggest that multiple models might be tested with an F-type statistic, I maintain the pairwise approach in this study. Thus, the following results refer to comparisons among the exponential, power, linear, and quadratic models for several effect sizes in true exponential or power data generating functions.

Sample Size. Level 2 sample sizes were consistent with the previous two studies. Data for 50, 100, and 200 individuals were generated. When the number of Level 1 units became too large (i.e., 25) a significant degree of non-convergence and fatal errors occurred when fitting the non-linear models. This difficulty, combined with uniformly high power in the previous studies, dictated that Level 1 sample size be adjusted. Samples of 5, 9, or 13 observations were generated for each Level 2 unit.

ICC. Three values of residual ICC were included in this study: .5, .7, and .86. The large residual ICC of .86 was included as it was the residual ICC in Cudeck and Harring (2007). The residual ICC of .7 was included to maintain continuity across the other studies. For the same reason, I attempted to include an ICC of .4, however, convergence issues prohibited the use of any random intercept variance less than 10. As a result, the small residual ICC condition was .5.

Effect Size. Similar to the previous studies, effect size was conceptualized as the degree of misfit between two candidate models and the data generating process. While observing the degree of misfit in linear models with non-nested covariate sets or non-nested Level 1 residual covariance structures was fairly intuitive, quantifying the degree of misfit among non-linear models was somewhat opaque. Figure 5.1 displays the estimated curves for the power and exponential models when Level 1 sample size is small in the true exponential model conditions.

Figure 5.2 displays the same curves when the true model was a power function. While these curves exhibit similar properties (e.g., monotonically decreasing and zero asymptote) they may differ in fit considerably.

Examining Figure 5.1 more closely, misfit between the exponential (black) and power (grey) curves can be observed from the degree, or lack thereof, of overlap. Because the true model is exponential, the black lines representing the exponential model do reasonably well at approximating the true curve. Therefore, misfit between the true data generating process and the estimated model was always larger for power and can be observed from the difference in the corresponding grey and black lines. Interestingly, misfit seemed to have a curvilinear relationship with the slope. When slope was small as in the top two lines, there was a relatively small degree of misfit between the exponential and power models. As slope increased and the curve of the true function changed, that difference appeared to become larger. However, when the slope became sufficiently large the difference between the two models was reduced, as in the lowest two pairs of curves on Figure 5.1.

This phenomenon can be explained by the rate of change of the true model. When the slope parameter was small (top pair of models), the trend was close to linear and able to be approximated by both models reasonably well. When the slope parameter was sufficiently large, as begins to be the case in the bottommost pair of lines on Figure 5.1, the function approached the asymptote so quickly that there was little room for misfit. The same can be observed for the true power models on Figure 5.2, however it was not possible to consistently simulate data and estimate models with large enough slopes that decreases in misfit would be observed. Regardless, misfit, and power by extension, should *not* increase monotonically with increases in slope.

Effect size was manipulated by altering the slope parameter in each true model. Given the true model forms

$$power: y_{ij} = (\gamma_{00} + u_{0j}) * time_{ij}^{\gamma_{10}} + e_{ij}, \quad (50)$$

and,

$$exponential: y_{ij} = (\gamma_{00} + u_{0j}) e^{\gamma_{10} * time_{ij}} + e_{ij} \quad (51)$$

the γ_{10}parameter was used to manipulate the effect size. The largest value for the power model was based on the slope used in the simulation study performed by Timmons and Preacher (2015). The application analyzed by Cudeck and Harring (2007) provided the large effect value for the exponential decay model. From each of these slopes smaller slope values were chosen to provide a sufficient gradient in model misfit. Initially four conditions were chosen for each true model: -,86, -.65, -.45, and -.10 for the power models and -.776, -.55, -.35, and -.10 for the exponential models. After observing ceiling effects for the power models and the non-monotonic changes in power over effect sizes in the exponential models, an intermediate slope value of -.25 was generated for the power model and -.25 and -.15 were generated for the exponential model.

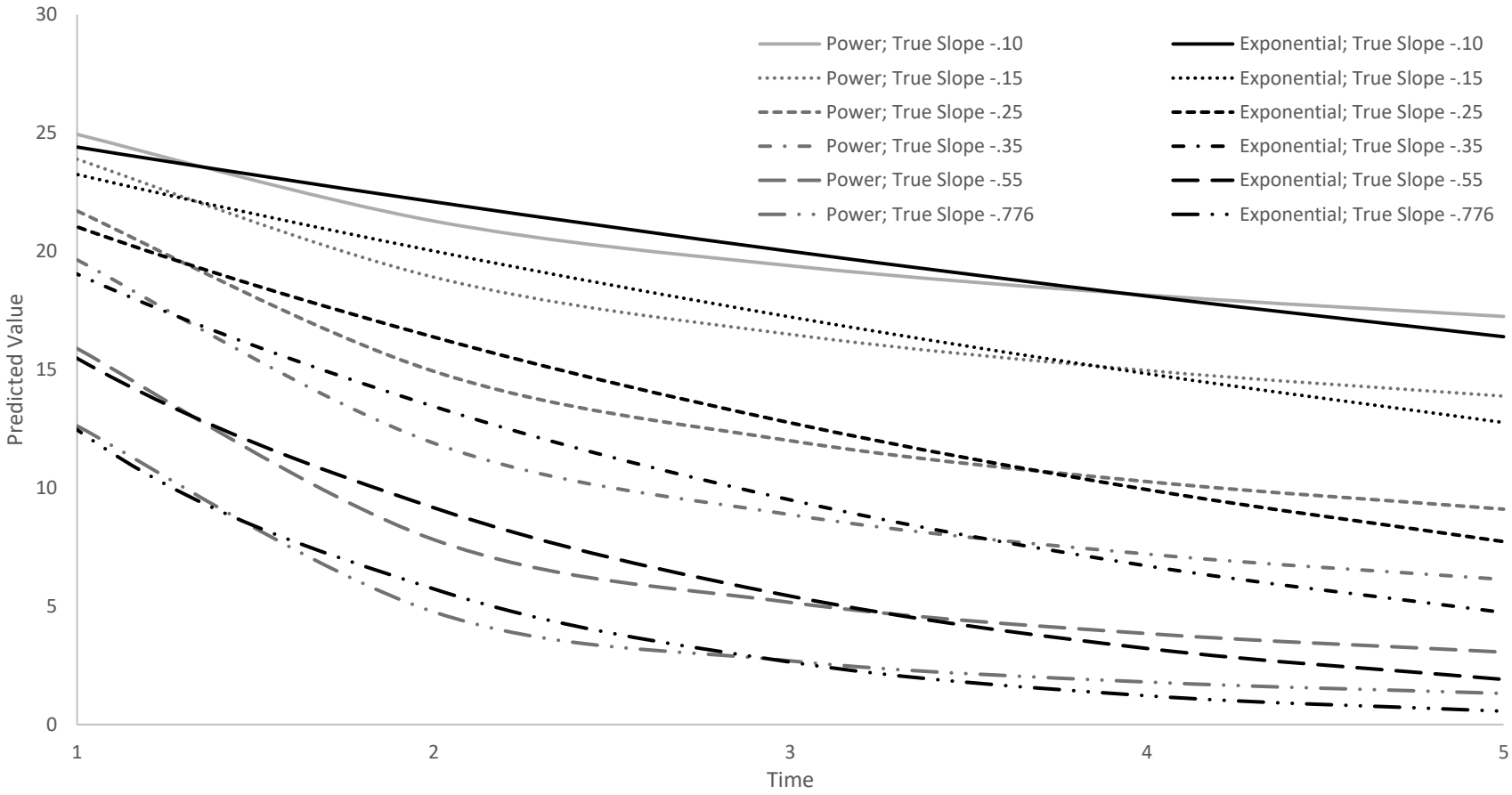


Figure 5.1 Predicted Values for Exponential and Power Models for True Exponential Model

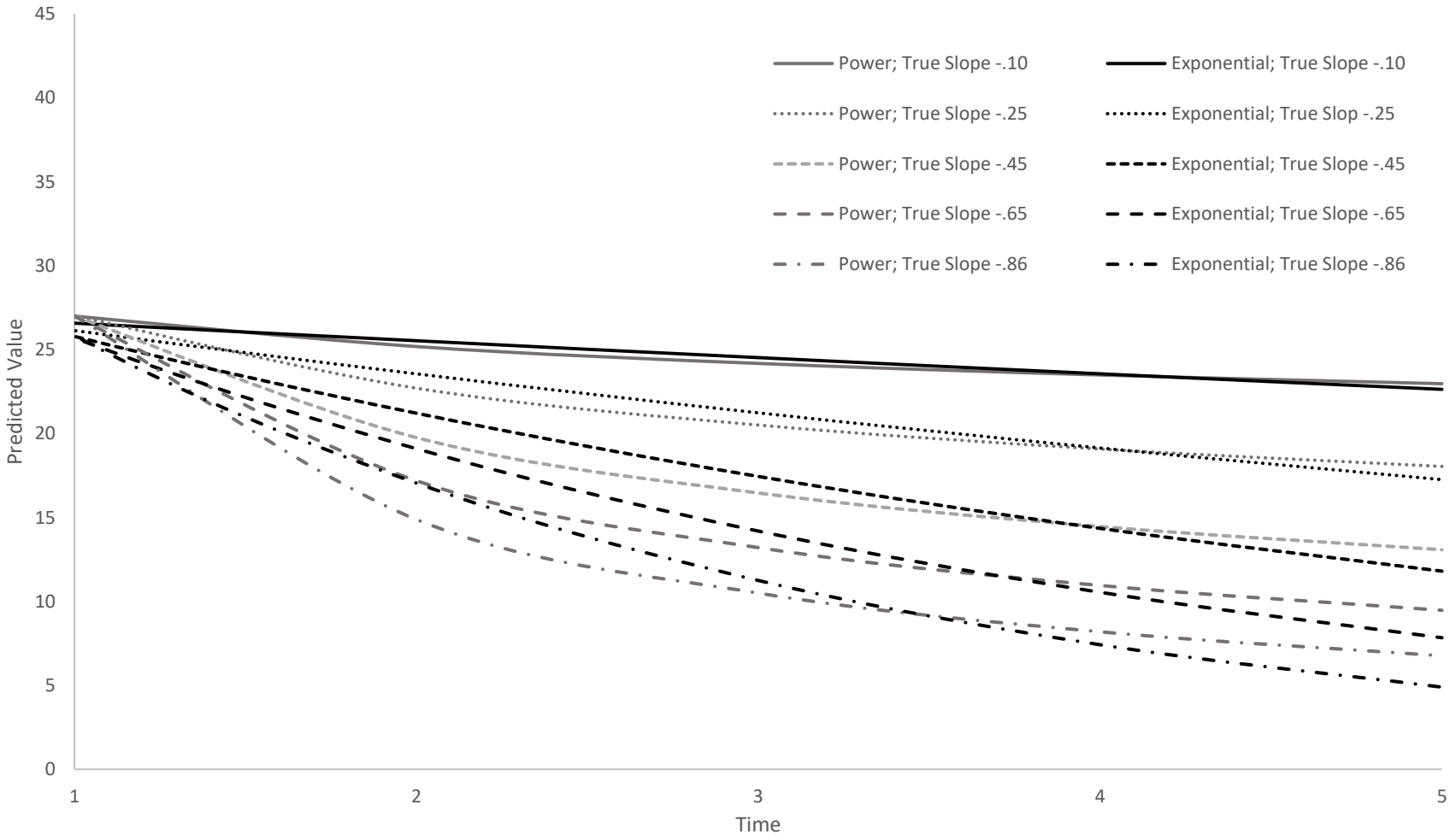


Figure 5.2 Predicted Values for Exponential and Power Models for True Power Model

Data Analysis

After comparing all models and retaining the results, a variable was coded for each replication to indicate whether the true model was selected by Vuong's test and/or information criteria. Only the selection of true models was included in this study because of ambiguity surrounding the definition of the "best" model when candidates did not include the true data generating process. That is, when the true data generating process was exponential, there was no definitive way of determining if a power or quadratic model should fit the data better in the population.

The rates of correct model selection (correct classification) of Vuong's test were compared to the correct classification of information criteria. After examining when both Vuong's test and information criteria selected the correct model, the misclassification rates of Vuong's test and information criteria were examined. Vuong's test selected the incorrect model in at most 2% of replications in a single anomalous cell. To provide a context for comparison, G*Power 3.1.5 was used to determine the difference from a constant proportion that would be detected with a binomial test with a power of .8 and a minimum of 918 observations per cell. Although 1000 replications were generated for each condition, convergence issues rendered some replications unusable. Using the maximum misclassification of Vuong's test of 2% as the referent proportion, sensitivity analysis indicated that there was enough power to detect a difference of 1.5%. Thus, any misclassification of information criteria 4% or greater could be considered significant. After exploring contexts in which information criteria perform significantly more poorly than Vuong's test, non-significance rates of Vuong's test are discussed for comparison.

After comparing the performance of Vuong's test and information criteria, logistic regression was used to determine the degree to which different study factors impacted the ability of Vuong's test to detect the correct candidate. To facilitate the discussion of Vuong's test in the context of non-linear models, I use the term "correct classification rate" for Vuong's test in lieu of "power" so as to not confuse the power function with statistical power. Using SAS Proc LOGISTIC with Firth's penalized likelihood, main effects models were explored to understand the general behavior of Vuong's test in the context of non-linear candidates. Next, logistic regression analyses with all interactions among study factors were run and the significance of each omnibus test evaluated. Non-significant higher order interactions were removed from the model one by one and non-significant lower order effects were retained if they were qualified by a significant higher order interaction. Throughout analyses all study factors were treated categorically.

Results

True Exponential Decay Models.

Selecting the Correct Model. Table 5.1 shows the empirical correct classification rates (i.e. power) for Vuong's test and the information criteria comparing the true exponential model to either a linear, quadratic, or power model across all effect sizes, Level 1 sample sizes, and Level 2 sample sizes when ICC was small. Because of an unexpected positive effect of ICC (see below) Table 5.2 was also included to display correct classification rates in the large ICC condition. Correct classification rates tended to be greater at larger ICCs such that only a small number of conditions remained underpowered; specifically when both Level 1 and Level 2 sample size were small and slopes were strong or weak (as opposed to moderate). Vuong's test was underpowered in the large ICC condition when Level 1 sample size was greater than 5 only

when the data generating model had a strong slope, Level 2 sample size was small and the alternative model was a power function. Information criteria performed uniformly well in the high ICC condition with correct classification rates rarely deviating from 100%. Shaded cells identify conditions in which the correct classification rate of Vuong's test to select the correct model was greater than .8.

In the small ICC condition, Vuong's test almost always achieved adequate power to select the true exponential decay model when there were at least 9 observations at Level 1. The rare instances in which Vuong's test was underpowered and Level 1 sample size was 13 occurred when Level 2 sample size was small and comparisons were made between the true exponential model and the power function when the slope was strongest, or the exponential and quadratic models in the two weakest slope conditions. Fewer conditions achieved correct classification rates of .8 when there were 9 observations at Level 1. Specifically, when the slope was weak it became more difficult for Vuong's test to differentiate the exponential model from the quadratic model at larger Level 2 sample sizes and distinguish from the linear model when Level 2 sample size was small.

When Level 1 sample size was small, the rate at which the true exponential model was selected dropped considerably. Exploring first the comparisons of the true exponential model with the linear model, when the slope of the true model was weak and had the least amount of curvature, Vuong's test had the most difficulty selecting the true exponential model. As slope increased, and curvature was introduced into the data generating process the correct classification rate of Vuong's test increased monotonically until it reached 100% in the strongest slope condition. When comparing the true exponential model to either a quadratic or a power model, the relationship between slope and correct classification was more complex.

In small Level 1 sample sizes the correct classification rate of Vuong's test tended to increase as slope became larger when the alternative model was quadratic. However, at moderate slopes the correct classification rate improved very little. Specifically, when slope increased from -.15 to -.25, there was a substantial increase in correct classification when the Level 2 sample size was 200. At smaller Level 2 sample sizes the increase was more modest. This correct classification rate remained fairly constant until slope reached -.78 where another drastic increase was observed. It is likely that when there were few Level 1 units and a moderate slope, that is, the data generating process was non-linear but not yet reaching the asymptote, the quadratic model was flexible enough to accommodate the curvature of the exponential form within the functional space.

Finally, when comparing the power model to the true exponential function the expected curvilinear relationship between slope coefficient and correct classification emerged. However, this effect was attenuated by increases in Level 2 sample size as correct classification rates approached the asymptote. Correct classification rates for the small and medium Level 2 samples size conditions formed almost the perfect inverted-U and can be seen on Figure 5.3. Correct classification rates were greater when the slope of the true model was moderate and attenuated when slopes were extreme. When Level 2 sample size was large, Vuong's test selected the correct model at a higher rate overall. At medium Level 2 sample sizes, the correct classification rate was greater than .8 for all but the most extreme slopes. As Level 2 sample size decreased to 50, adequate correct classification rates were only achieved for the most moderate slopes.

Information criteria tended to perform well when Level 1 sample size was at least 9. Commensurate with Vuong's test, the correct model selection rate decreased at the small Level 1 sample size, however, in most slope conditions comparisons between information criteria still

tended to perform well. Information criteria displayed the worst performance when comparing the exponential model to the linear model in the weakest slope condition. When Level 2 sample size was small, comparisons among information criteria tended to select the correct model in only 76% of cases. Although information criteria appeared to perform well across slopes even when Level 1 sample size was small, the advantage of Vuong's test over comparisons among information criteria can best be observed through the misclassification rates.

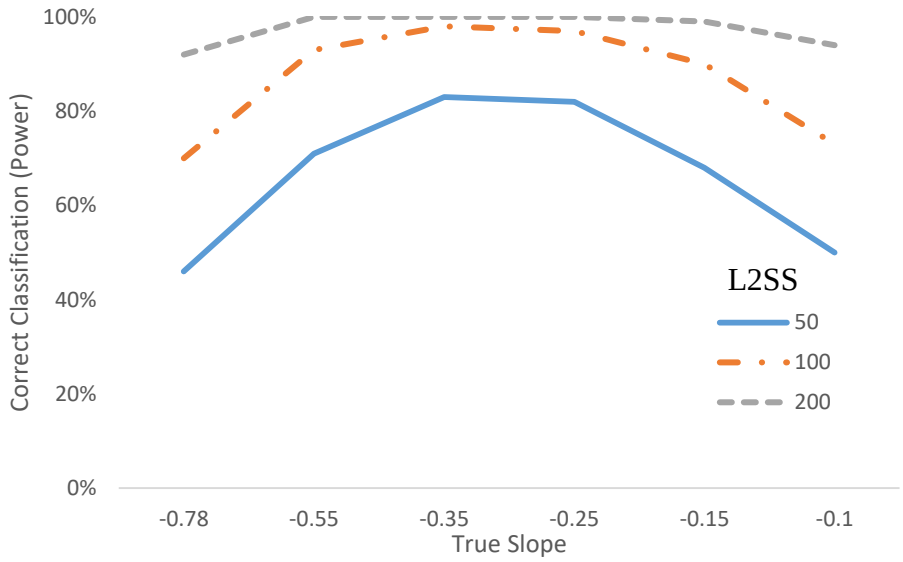


Figure 5.3. Correct Classification Rates Comparing Power Model and True Exponential Model

Table 5.1 Correct Model Selection Rates for True Exponential Decay Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	19%	88%	88%	90%	97%	100%	100%	100%	100%	100%	100%	100%
-.78	Quad	100	40%	95%	95%	97%	100%	100%	100%	100%	100%	100%	100%	100%
		200	68%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	46%	93%	93%	93%	67%	98%	98%	98%	70%	98%	98%	98%
	Pwr	100	70%	99%	99%	99%	91%	100%	100%	100%	91%	100%	100%	100%
		200	92%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%
		50	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	12%	86%	87%	92%	88%	100%	100%	100%	100%	100%	100%	100%
-.55	Quad	100	26%	94%	94%	96%	100%	100%	100%	100%	100%	100%	100%	100%
		200	55%	98%	98%	99%	100%	100%	100%	100%	100%	100%	100%	100%
		50	71%	99%	99%	99%	94%	100%	100%	100%	99%	100%	100%	100%
	Pwr	100	93%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	10%	89%	89%	96%	62%	99%	99%	99%	99%	100%	100%	100%
-.35	Quad	100	30%	95%	95%	98%	91%	100%	100%	100%	100%	100%	100%	100%
		200	59%	99%	99%	100%	99%	100%	100%	100%	100%	100%	100%	100%
		50	83%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.1(cont'd) Correct Model Selection Rates for True Exponential Decay Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	65%	98%	98%	98%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	90%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	11%	89%	89%	95%	46%	97%	97%	98%	89%	100%	100%	100%
-.25	Quad	100	30%	95%	95%	98%	82%	100%	100%	100%	100%	100%	100%	100%
		200	61%	99%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%
		50	82%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	35%	88%	88%	88%	96%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	54%	96%	96%	96%	100%	100%	100%	100%	100%	100%	100%	100%
		200	78%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
		50	11%	88%	88%	95%	42%	96%	96%	98%	72%	100%	100%	100%
-.15	Quad	100	26%	91%	92%	97%	73%	99%	99%	100%	97%	100%	100%	100%
		200	47%	97%	97%	99%	96%	100%	100%	100%	100%	100%	100%	100%
		50	68%	98%	98%	98%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	90%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	19%	76%	76%	76%	71%	99%	99%	99%	99%	100%	100%	100%
	Linear	100	33%	88%	88%	88%	94%	100%	100%	100%	100%	100%	100%	100%
		200	49%	95%	95%	95%	100%	100%	100%	100%	100%	100%	100%	100%
		50	8%	82%	83%	93%	31%	96%	96%	98%	63%	99%	99%	99%
-.10	Quad	100	17%	88%	88%	96%	61%	99%	99%	100%	92%	100%	100%	100%
		200	31%	93%	93%	98%	88%	100%	100%	100%	100%	100%	100%	100%
		50	50%	95%	95%	95%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	73%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
		200	94%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.2 Correct Model Selection Rates for True Exponential Decay Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	70%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.78	Quad	100	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	51%	94%	94%	94%	67%	98%	98%	98%	72%	99%	99%	99%
	Pwr	100	71%	99%	99%	99%	90%	100%	100%	100%	95%	100%	100%	100%
		200	93%	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.55	Quad	100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	74%	99%	99%	99%	96%	100%	100%	100%	99%	100%	100%	100%
	Pwr	100	95%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	94%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.35	Quad	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	84%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.2 (cont'd) Correct Model Selection Rates for True Exponential Decay Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	89%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.25	Quad	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	83%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	85%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	78%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.15	Quad	100	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	69%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	91%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	66%	97%	97%	97%	99%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	89%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	57%	98%	98%	99%	98%	100%	100%	100%	100%	100%	100%	100%
-.10	Quad	100	85%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	53%	96%	96%	96%	100%	100%	100%	100%	100%	100%	100%	100%
	Pwr	100	80%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Misclassification. Misclassification rates for Vuong's test and information criteria are presented on Table 5.3. I present only the table for the small ICC condition here because misclassification rates in the large ICC condition were negligible. Across all conditions, Vuong's test rarely if ever selected the incorrect model. When Level 1 sample size reached 9 or above, there was little to no misclassification for Vuong's test or information criteria under any study conditions. Misclassification only reached the significant 4% Level for information criteria in very specific conditions (comparing the quadratic model to the true exponential model, slope was no stronger than -.15, Level 1 sample size was 9, and Level 2 sample size was 50). Even under these specific conditions, misclassification was only barely significant (4-5%) and only reached significant levels for AIC and AICc. Because the true data generating model had fewer parameters than the alternative quadratic model, BIC misclassified at a lesser rate.

When Level 1 sample size was small, there tended to be more misclassification when model selection was based on information criteria. Comparisons to the quadratic model consistently produced significant misclassification rates when Level 2 sample size was small or medium when using AIC or AICc. Again, BIC exhibited better performance than AIC or AICc when the alternative model was quadratic because of the extra parameter of the quadratic model encouraged the selection of the more parsimonious true model. When comparing the true exponential model to the linear model significant misclassification only arose for the weakest slope conditions (-.15, -.10). When the exponential slope was at its weakest, the linear model was wrongfully selected in almost a quarter of replications when Level 2 sample size was small. Information criteria tended to have difficulty selecting the correct model when the slope was weak across all comparisons. Finally, when comparing the true exponential model to the power model, information criteria only misclassified in a significant proportion of replications when

Level 2 sample size was small and slopes were in the extremes. Even then misclassification was still relatively infrequent never exceeding 7%.

Non-significance rates for Vuong's test can be found on Table 5.4 where cells are shaded to indicate conditions in which information criteria selected the incorrect model in a significant proportion of replications. The shaded cells indicate that significant misclassification from information criteria occurred when Level 1 sample size was small. Furthermore, incorrectly selecting a quadratic model appeared to have occurred across the range of slope coefficients whereas linear models were only wrongfully selected when the slope was weak and power models incorrectly selected when the slope was in the extremes. As expected significant misclassification via information criteria only occurred when Vuong's test was unable to determine which model fit the data best.

Table 5.3 Incorrect Model Selection Rates for True Exponential Decay Models when Residual ICC = .50

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	13%	12%	10%	0%	0%	0%	0%	0%	0%	0%	0%
-.78	Quad	100	0%	5%	5%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	7%	7%	7%	0%	2%	2%	2%	0%	2%	2%	2%
	Pwr	100	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	14%	13%	8%	0%	0%	0%	0%	0%	0%	0%	0%
-.55	Quad	100	0%	6%	6%	4%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	2%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	Pwr	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	11%	11%	4%	0%	1%	1%	1%	0%	0%	0%	0%
-.35	Quad	100	0%	5%	5%	2%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	Pwr	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.3 (cont'd) Incorrect Model Selection Rates for True Exponential Decay Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	0%	2%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	11%	11%	5%	0%	3%	3%	2%	0%	0%	0%	0%
-.25	Quad	100	0%	5%	5%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	Pwr	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	12%	12%	12%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	4%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	12%	12%	5%	0%	4%	4%	2%	0%	1%	1%	0%
-.15	Quad	100	0%	9%	8%	3%	0%	1%	1%	0%	0%	0%	0%	0%
		200	0%	3%	3%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	2%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%
	Pwr	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	24%	24%	24%	0%	1%	1%	1%	0%	0%	0%	0%
	Linear	100	0%	13%	13%	13%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	5%	5%	5%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	18%	17%	7%	1%	5%	5%	3%	0%	1%	1%	1%
-.10	Quad	100	0%	12%	12%	4%	0%	1%	1%	1%	0%	0%	0%	0%
		200	0%	7%	7%	2%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	5%	5%	5%	0%	0%	0%	0%	0%	0%	0%	0%
	Pwr	100	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.4 Non-significance rates for Vuong's Test for the True Exponential Decay Model

			ICC								
			0.5			0.7			0.86		
			L1SS								
Slope	Alt Model	L2SS	5	9	13	5	9	13	5	9	13
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	81%	3%	0%	75%	1%	0%	30%	0%	0%
-.78	Quad	100	60%	0%	0%	45%	0%	0%	4%	0%	0%
		200	32%	0%	0%	13%	0%	0%	0%	0%	0%
		50	54%	33%	30%	51%	32%	29%	49%	33%	28%
	Pwr	100	31%	9%	9%	31%	12%	9%	29%	10%	6%
		200	8%	1%	1%	7%	1%	0%	7%	1%	0%
		50	1%	0%	0%	1%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	88%	12%	0%	68%	5%	0%	12%	0%	0%
-.55	Quad	100	74%	0%	0%	34%	0%	0%	1%	0%	0%
		200	46%	0%	0%	7%	0%	0%	0%	0%	0%
		50	29%	6%	2%	28%	4%	1%	26%	4%	1%
	Pwr	100	7%	0%	0%	8%	0%	0%	5%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	12%	0%	0%	5%	0%	0%	1%	0%	0%
	Linear	100	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	90%	38%	1%	56%	10%	0%	6%	0%	0%
-.35	Quad	100	70%	9%	0%	23%	0%	0%	0%	0%	0%
		200	41%	1%	0%	3%	0%	0%	0%	0%	0%
		50	17%	0%	0%	16%	0%	0%	16%	0%	0%
	Pwr	100	2%	0%	0%	2%	0%	0%	1%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.4 (cont'd) Non-significance rates for Vuong's Test for the True Exponential Decay Model

			ICC								
			0.5			0.7			0.86		
			L1SS								
Slope	Alt Model	L2SS	5	9	13	5	9	13	5	9	13
		50	35%	0%	0%	20%	0%	0%	3%	0%	0%
	Linear	100	10%	0%	0%	3%	0%	0%	0%	0%	0%
		200	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	88%	54%	11%	54%	12%	1%	11%	0%	0%
-.25	Quad	100	70%	18%	1%	24%	0%	0%	1%	0%	0%
		200	40%	1%	0%	3%	0%	0%	0%	0%	0%
		50	18%	0%	0%	18%	0%	0%	17%	0%	0%
	Pwr	100	3%	0%	0%	2%	0%	0%	2%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	64%	4%	0%	46%	1%	0%	15%	0%	0%
	Linear	100	46%	0%	0%	20%	0%	0%	2%	0%	0%
		200	22%	0%	0%	4%	0%	0%	0%	0%	0%
		50	88%	58%	28%	63%	18%	2%	22%	1%	0%
-.15	Quad	100	74%	27%	3%	38%	1%	0%	4%	0%	0%
		200	53%	4%	0%	13%	0%	0%	0%	0%	0%
		50	32%	0%	0%	31%	0%	0%	31%	0%	0%
	Pwr	100	10%	0%	0%	10%	0%	0%	9%	0%	0%
		200	1%	0%	0%	0%	0%	0%	1%	0%	0%
		50	81%	29%	1%	65%	10%	0%	34%	1%	0%
	Linear	100	66%	6%	0%	46%	0%	0%	11%	0%	0%
		200	51%	0%	0%	19%	0%	0%	1%	0%	0%
		50	91%	68%	37%	77%	29%	5%	43%	2%	0%
-.10	Quad	100	83%	39%	8%	59%	7%	0%	15%	0%	0%
		200	69%	12%	0%	27%	0%	0%	1%	0%	0%
		50	50%	0%	0%	48%	0%	0%	47%	0%	0%
	Pwr	100	27%	0%	0%	26%	0%	0%	20%	0%	0%
		200	6%	0%	0%	5%	0%	0%	4%	0%	0%

Power. Results of a logistic regression model examining the main effects of study factors on the power of Vuong's test to detect the true model indicated significant main effects for all study factors. As Level 1 sample size (χ^2 (2)=30759.14, p < .0001), Level 2 sample size (χ^2 (2)= 16490.86, p < .0001), and ICC (χ^2 (2)=12958.02, p < .0001) increased, so did the power of Vuong's test. The manipulation of slope coefficient magnitude (χ^2 (5)= 5696.25, p < .0001) exhibited the expected complex relationship discussed above: as slope became stronger, power increased up to a point and then began to decrease as model fit became more similar.

A logistic regression model exploring the interactions among study factors indicated a significant four-way interaction between Level 1 sample size, Level 2 sample size, slope magnitude, and ICC (χ^2 (40)= 130.42, p < .0001). As would be expected, the effects of the higher order interaction were generally observed as moderate sample sizes, ICCs, and slopes grew to larger values as effects would tend to slow as power increased. In general, the effects of study factors on the power of Vuong's test can be explained by the main effects described throughout the previous section.

True Power Model.

Selecting the Correct Model. Correct classification rates of Vuong's test and information criteria when the true data generating process was a power model can be found on Tables 5.5 (ICC = .5) and 5.6 (ICC = .86). Generally the trends observed across ICCs were the same, increases in any study factor improved power, however, power tended to be greater when ICC was large. When Level 1 sample size had 13 observations, almost every comparison had a correct classification rate of at least .8. The only exceptions across both ICC conditions occurred when slope was weakest; comparing the quadratic model to the power model did not select the correct model in at least 80% of replications unless there were 200 observations at Level 2 when

ICC was .5. When ICC reached .86 the only underpowered condition when Level 1 sample size was large was when Level 2 sample size was small and the alternative model was quadratic.

The correct classification rate decreased slightly as the Level 1 sample size decreased to 9 observations. Again, correct classification rates for Vuong's test were inadequate only when the slope was weak and were mainly an issue for the curvilinear models. In the weakest slope condition, the correct classification rate only exceeded .8 when the alternative model was an exponential form and Level 2 sample size was large or the alternative was a linear model with Level 2 sample sizes of 100 or 200. At medium Level 1 sample sizes, comparing the true power model to the quadratic form never achieved correct classification of .8 in the weak slope condition.

Finally, when there were only 5 observations at Level 1 adequate rates of correct classification were achieved when the comparison models were linear or exponential in almost every case when the true slope was stronger than -.10. Comparisons to the exponential model were underpowered at the small Level 2 sample size when the slope was at least -.25. When comparing the true power model to the quadratic model, however, Vuong's test did not reach adequate correct classification rates until slope was at least -.45. Even when the slope was relatively strong, Vuong's test failed to select the correct model unless Level 2 sample size was at least 100.

Information criteria generally performed well when the true data generating model was a power function. Only when slope was particularly weak and Level 1 sample size was small did correct classification rates from information criteria drop below 90%. When the true slope was stronger than -.10 information criteria almost perfectly selected the correct model. However, when the slope was weak, information criteria selected the incorrect model rather frequently.

Table 5.5 Correct Model Selection Rates for True Power Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	71%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.65	Quad	100	94%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	43%	97%	97%	99%	96%	100%	100%	100%	100%	100%	100%	100%
-.45	Quad	100	74%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	89%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	80%	99%	99%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	18%	89%	89%	94%	65%	98%	98%	99%	92%	100%	100%	100%
-.25	Quad	100	35%	95%	95%	98%	92%	100%	100%	100%	100%	100%	100%	100%
		200	64%	98%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	59%	97%	97%	97%	97%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	82%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.5 (cont'd) Correct Model Selection Rates for True Power Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	25%	82%	82%	82%	59%	98%	98%	98%	85%	100%	100%	100%
	Linear	100	41%	91%	91%	91%	86%	99%	99%	99%	98%	100%	100%	100%
		200	62%	98%	98%	98%	98%	100%	100%	100%	100%	100%	100%	100%
		50	5%	80%	81%	94%	13%	85%	85%	93%	29%	93%	93%	96%
-.10	Quad	100	8%	81%	82%	96%	28%	92%	92%	97%	52%	97%	97%	99%
		200	13%	85%	85%	97%	52%	97%	97%	99%	84%	100%	100%	100%
		50	20%	80%	80%	80%	51%	96%	96%	96%	79%	99%	99%	99%
	Exp	100	31%	88%	88%	88%	78%	99%	99%	99%	96%	100%	100%	100%
		200	50%	96%	96%	96%	95%	100%	100%	100%	100%	100%	100%	100%

Table 5.6 Correct Model Selection Rates for True Power Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.65	Quad	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	97%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.45	Quad	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	89%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	94%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Linear	100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	72%	99%	99%	99%	98%	100%	100%	100%	100%	100%	100%	100%
-.25	Quad	100	92%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		50	62%	97%	97%	97%	98%	100%	100%	100%	100%	100%	100%	100%
	Exp	100	85%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.6 (cont'd) Correct Model Selection Rates for True Power Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
-.10	Linear	50	38%	90%	90%	90%	79%	99%	99%	99%	95%	100%	100%	100%
		100	62%	97%	97%	97%	98%	100%	100%	100%	100%	100%	100%	100%
		200	86%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Quad	50	17%	88%	89%	95%	45%	96%	97%	98%	69%	99%	99%	100%
		100	34%	94%	94%	97%	76%	99%	99%	100%	93%	100%	100%	100%
		200	62%	99%	99%	100%	96%	100%	100%	100%	100%	100%	100%	100%
	Exp	50	22%	80%	80%	80%	53%	95%	95%	95%	80%	99%	99%	99%
		100	32%	88%	88%	88%	80%	100%	100%	100%	98%	100%	100%	100%
		200	52%	94%	94%	94%	97%	100%	100%	100%	100%	100%	100%	100%

Misclassification. Due to high correct model selection rates, there was not much misclassification except when the true slope was weakest (-.10). Tables 5.7 and 5.8 display the misclassification rates of Vuong's test and information criteria in the small and large ICC conditions. Vuong's test never selected the incorrect model in more than 2% of replications. Using 2% as a base rate, the results of the sensitivity analysis above indicated that any misclassification greater than 4% was significantly worse than Vuong's test. When the slope was stronger than -.10, information criteria only selected the incorrect model at a significant rate when the alternative candidate was quadratic, Level 2 sample size was small or medium and Level 1 sample size was small. With larger Level 1 sample sizes or larger slopes, there was never significant misclassification with a slope stronger than -.10.

In the weak slope condition, however, information criteria performed significantly worse than Vuong's test whenever Level 1 sample size was small with the exception of the comparison to a linear model with a large Level 2 sample size. As Level 1 sample size increased, misclassification rates dropped to non-significant rates when comparing the power model to linear and exponential models, however comparisons to the quadratic model continued to misclassify at a significant rate. In these cases BIC outperformed AIC and AICc because of its preference for more parsimonious models.

Non-significance rates of Vuong's test are presented on Table 5.9. Shaded cells denote conditions in which information criteria performed significantly worse than Vuong's test. As would be expected, information criteria performed at their worst when Vuong's test exhibited high rates of non-significance.

Table 5.7 Incorrect Model Selection Rates for True Power Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
-.65	Quad	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	3%	3%	1%	0%	0%	0%	0%	0%	0%	0%	0%
-.45	Quad	100	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	11%	11%	6%	0%	2%	2%	1%	0%	0%	0%	0%
-.25	Quad	100	0%	5%	5%	2%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.7 (cont'd) Incorrect Model Selection Rates for True Power Models when Residual ICC = .5

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
-.10	Linear	50	0%	18%	18%	18%	0%	3%	3%	3%	0%	0%	0%	0%
		100	0%	9%	9%	9%	0%	1%	1%	1%	0%	0%	0%	0%
		200	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	Quad	50	2%	20%	19%	6%	1%	15%	15%	7%	0%	7%	7%	4%
		100	1%	19%	18%	4%	1%	8%	8%	4%	0%	3%	3%	1%
		200	1%	15%	15%	3%	0%	3%	3%	1%	0%	0%	0%	0%
	Exp	50	1%	21%	21%	21%	0%	4%	4%	4%	0%	1%	1%	1%
		100	0%	12%	12%	12%	0%	1%	1%	1%	0%	0%	0%	0%
		200	0%	5%	5%	5%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.8 Incorrect Model Selection Rates for True Power Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
-.65	Quad	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
-.45	Quad	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
-.25	Quad	100	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	Exp	100	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.8 (cont'd) Incorrect Model Selection Rates for True Power Models when Residual ICC = .86

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC	Vuong's Test	AIC	AICc	BIC
		50	0%	10%	10%	10%	0%	1%	1%	1%	0%	0%	0%	0%
	Linear	100	0%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	12%	12%	5%	0%	4%	3%	2%	0%	1%	1%	0%
-.10	Quad	100	0%	6%	6%	3%	0%	1%	1%	0%	0%	0%	0%	0%
		200	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	1%	20%	20%	20%	0%	5%	5%	5%	0%	1%	1%	1%
	Exp	100	0%	12%	12%	12%	0%	1%	1%	1%	0%	0%	0%	0%
		200	0%	6%	6%	6%	0%	0%	0%	0%	0%	0%	0%	0%

Table 5.9 Non-significance rates of Vuong's Test for True Power Models.

ICC

			0.5			0.7			0.86		
			L1SS								
Slope	Alt Model	L2SS	5	9	13	5	9	13	5	9	13
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	29%	0%	0%	9%	0%	0%	0%	0%	0%
-.65	Quad	100	6%	0%	0%	1%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	4%	0%	0%	4%	0%	0%	3%	0%	0%
	Exp	100	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Linear	100	0%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	57%	4%	0%	32%	1%	0%	4%	0%	0%
-.45	Quad	100	26%	0%	0%	6%	0%	0%	0%	0%	0%
		200	4%	0%	0%	0%	0%	0%	0%	0%	0%
		50	12%	0%	0%	11%	0%	0%	11%	0%	0%
	Exp	100	2%	0%	0%	1%	0%	0%	1%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	20%	0%	0%	16%	0%	0%	6%	0%	0%
	Linear	100	3%	0%	0%	2%	0%	0%	0%	0%	0%
		200	0%	0%	0%	0%	0%	0%	0%	0%	0%
		50	81%	35%	8%	64%	18%	3%	28%	3%	0%
-.25	Quad	100	65%	9%	0%	44%	3%	0%	8%	0%	0%
		200	36%	1%	0%	12%	0%	0%	0%	0%	0%
		50	41%	3%	0%	42%	2%	0%	39%	3%	0%
	Exp	100	18%	0%	0%	18%	0%	0%	15%	0%	0%
		200	2%	0%	0%	2%	0%	0%	2%	0%	0%

Table 5.9 (cont'd) Non-significance rates of Vuong's Test for True Power Models.

			ICC								
			0.5			0.7			0.86		
			L1SS								
			5	9	13	5	9	13	5	9	13
		50	75%	41%	15%	71%	34%	9%	62%	22%	5%
	Linear	100	58%	14%	2%	55%	10%	1%	38%	3%	0%
		200	38%	2%	0%	31%	1%	0%	14%	0%	0%
		50	93%	87%	71%	90%	77%	57%	82%	55%	32%
-.10	Quad	100	91%	72%	48%	87%	55%	29%	66%	25%	7%
		200	86%	49%	16%	74%	26%	7%	38%	4%	0%
		50	80%	49%	21%	80%	49%	19%	78%	47%	20%
	Pwr	100	69%	22%	4%	65%	21%	3%	68%	21%	3%
		200	50%	5%	0%	50%	3%	0%	48%	3%	0%

Power. The same approach using logistic regression was employed when the true data generating process was a power model as when it was an exponential model; a main effects logistic regression model was first used to understand the behavior of Vuong's test over the study factors. Main effects were found for Level 1 sample size (χ^2 (2)= 18978.06, p < .0001), Level 2 sample size (χ^2 (2)= 10323.70, p < .0001), slope (χ^2 (3)= 30242.22, p < .0001), and ICC (χ^2 (2)= 4010.31, p < .0001). As any study factor increased, so did the power of Vuong's test to detect the correct model. It would be expected that the same non-linear relationship for effect size that occurred when the true data generating process was exponential would occur when the data generating process was a power model. However, as noted above it was not possible to fir certain models to data generated with large enough slopes to observe the nonlinearity. Therefore in the observed range of effect sizes power increased monotonically with slope.

Results from a full logistic regression model including interactions among all study factors indicated a four-way interaction among Level 1 sample size, Level 2 sample size, slope and ICC, χ^2 (24)= 39.76, p =.023. Similar to the true exponential models, this interaction was expected, improving study factors had diminishing returns as power increased and approached its asymptote. Thus, there tended to be larger differences between study factors when the two conditions being compared had lower power.

Random Intercept and Slope Model.

Method

Data Generation

The data generation program used to generate data for the random intercept only models was altered to generate data for the random intercept and slope models. Specifically, instead of a

scalar for the residual Level 2 intercept variance, a matrix was specified. Values for the residual Level 2 variances were adopted from Cudeck and Harring (2007):

$$\boldsymbol{\tau} \sim N\begin{pmatrix} 0 \\ 0' \end{pmatrix}, \begin{matrix} .63 & -.18 \\ -.18 & .04 \end{matrix}.$$

Because of considerable convergence issues when a random slope was included in the model, only the two largest slopes from the study of non-linear models with a random intercept were able to be generated for each model. Data were generated using the same equations as above except with an additional residual term associated with the random slope. Start values were identified as either the true values from the data generation program or values found through trial and error.

After generating outcomes, four models were fit to the data: linear, quadratic, power, and exponential. Again, specification was the same as above except for the addition of the random slope term, its variance, and the random intercept and slope covariance. Every model included a random time slope in addition to the random intercept. A random quadratic term was omitted from the model of quadratic effects so as to maintain the same number of random effects terms across models. Vuong's test was conducted on each pair of models for each replication and results saved for further analysis.

Sample Size. The same sample sizes that were used for the random intercept only models was used for the random intercept and slope models. Conditions were generated for 50, 100, or 200 Level 2 units and 5, 9 or 13 Level 1 units.

ICC. ICC was not manipulated due to the fragility of the estimation with a random slope. Thus the residual ICC was .86, reflecting the value in Cudeck and Harring (2007).

Effect Size. The two largest coefficient magnitudes sizes from the random intercept only models were included here. Effect size (i.e., the difference between models) was manipulated by

varying these slope coefficients. Thus, slopes for the exponential models took values of -.776 and -.55 while the power models took values of -.86 and -.45.

Data Analysis

To understand the performance of Vuong's test relative to the performance of information criteria I contrasted the two methods for model selection in their correct classification rates and misclassification rates. Further, I examined the non-significance rates of Vuong's test relative to misclassification resulting from information criteria. An additional table is provided with non-significance rates of Vuong's test. Following the same methodology as the previous studies, G*Power 3.1.5 was used to determine the difference from a constant proportion that would indicate significantly worse model selection. The largest misclassification rate for Vuong's test when non-linear models with random slopes were compared was 10%. Using this maximum misclassification as a conservative constant, a sensitivity analysis indicated that a difference of 2.8% (from 10%) could be detected with a power of .8 and a total sample size of 998, the minimum sample size for a given condition. Thus, misclassification of at least 13% can be considered significantly worse than Vuong's test in the worst case.

Logistic regression with Firth's penalized likelihood was then used to determine the study factors that significantly impacted the power of Vuong's test to detect the true model. Main effects models were first used to understand the general behavior of Vuong's test when comparing non-nested functional forms with random slopes. Then, logistic regression models were run with all higher order interactions. Non-significant effects were removed one-by-one until only significant effects remained in the model. If lower order effects contributed to a significant higher order interaction they were retained regardless of their significance level.

Results

True Exponential Decay Model.

Selecting the Correct Model. To compare the performance of Vuong's test to the performance of information criteria, correct classification rates were examined for each model selection method (Table 5.10). Again, results of a sensitivity analysis suggested that any misclassification made by the information criteria of more than 13% could be considered significantly worse than Vuong's test's worst case. When the slope of the true exponential model was strong, power was above .8 when comparing the quadratic model to the exponential model and there were 9 or more observations at Level 1.

When the slope coefficient was strong, power increased with either Level 1 or Level 2 sample size for both the quadratic and power comparisons. Vuong's test only reached adequate power when compared to the power model when Level 2 sample size was large and Level 1 sample size was at least 9 observations. In other cases Vuong's test was underpowered.

Table 5.10 Correct Model Selection for True Exponential Decay Model with Random Intercept and Slope

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICC	BIC	Vuong's Test	AIC	AICC	BIC	Vuong's Test	AIC	AICC	BIC
-.776	Quad	50	14%	88%	88%	93%	97%	99%	99%	99%	99%	100%	100%	100%
		100	33%	93%	94%	97%	99%	100%	100%	100%	99%	100%	100%	100%
		200	62%	98%	98%	100%	98%	100%	100%	100%	98%	100%	100%	100%
	Power	50	37%	88%	88%	88%	52%	93%	93%	93%	56%	94%	94%	94%
		100	54%	96%	96%	96%	71%	97%	97%	97%	77%	96%	96%	96%
		200	75%	99%	99%	99%	87%	98%	98%	98%	86%	96%	96%	96%
-.55	Quad	50	7%	72%	73%	80%	61%	87%	88%	88%	80%	90%	90%	90%
		100	13%	74%	74%	82%	65%	85%	85%	86%	73%	85%	85%	86%
		200	20%	77%	77%	82%	59%	79%	79%	80%	63%	79%	79%	79%
	Power	50	36%	88%	88%	88%	50%	85%	85%	85%	48%	79%	79%	79%
		100	55%	96%	96%	96%	52%	85%	85%	85%	45%	72%	72%	72%
		200	74%	99%	99%	99%	55%	87%	87%	87%	38%	66%	66%	66%

In the weaker of the two slope conditions, power increased as Level 2 sample size increased for both model comparisons when Level 1 sample size was small. However, as Level 1 sample size increased there was an unexpected trend: power started to decrease as sample size grew. This counter intuitive effect is explored in more detail in the next section.

Information criteria generally performed well when comparing a true exponential model with random intercepts and slopes to a quadratic and power model with random slopes. When the slope coefficient was large, information criteria tended to select the correct model in over 90% of cases across all conditions with the only exception arising when Level 1 and Level 2 sample size were both small. When the alternative model was quadratic, BIC performed better than AIC or AICc as the penalty term more heavily favored the more parsimonious exponential model. In the weaker slope condition, information criteria performed more poorly. When comparing the true exponential model to the quadratic model, information criteria only selected the correct model in roughly three-quarters of replications when Level 1 sample size was small, however, it still performed reasonably well when the alternative model was the power model. Similar to the results of Vuong's test, correct classification rates decreased with increasing Level 2 sample sizes at larger Level 1 sample sizes.

Misclassification. Misclassification rates of Vuong's test and information criteria can be found on Table 5.11. When the slope coefficient was large, there was only a significant degree of misclassification for AIC when comparing the exponential model to a quadratic model in small Level 1 and Level 2 sample sizes. In no other condition was there a significant degree of misclassification of information criteria when the exponential slope was strong.

In the weaker exponential slope condition, there was a significant degree of misclassification among information criteria whenever Level 1 sample size was medium or large.

Regardless of Level 2 sample size, information criteria had a significant degree of misclassification for both comparison models. When Level 1 sample size was large, misclassification rates increased with increases in Level 2 sample sizes, reflecting the anomalous results found for correct model selection. In the small Level 1 sample size, there was a significant degree of misclassification when comparing the exponential and quadratic models. At this small Level 1 sample size, the effect of Level 2 sample size was as expected; increasing Level 2 sample size reduced the degree of misclassification.

Misclassification of Vuong's test exhibited the same patterns as information criteria. When Level 1 sample size was large, the misclassification rate of Vuong's test increased with Level 2 sample size and when Level 1 sample size was small the effect of Level 2 sample size was as expected. The largest degree of misclassification occurred when both Level 1 and Level 2 sample sizes were large.

Non-significance rates of Vuong's test are presented on Table 5.12. Shaded cells represent conditions in which information criteria selected the incorrect model in a significant proportion of replications. As can be seen from the table, when non-significance rates of Vuong's test were large, misclassification rates of information criteria tended to be large as well. Once again this illustrates the benefit of Vuong's test over that of information criteria. While information criteria were leading to incorrect results in upwards of 30% of replications, Vuong's test tended to report non-significance.

Table 5.11 Incorrect Model Selection for True Exponential Decay Model with Random Intercept and Slope

			L1SS											
			5				9				13			
Slope	Alt Model	L2SS	Vuong‘s Test	AIC	AICC	BIC	Vuong‘s Test	AIC	AICC	BIC	Vuong‘s Test	AIC	AICC	BIC
-.776	Quad	50	1%	13%	12%	7%	0%	1%	1%	1%	0%	0%	0%	0%
		100	0%	7%	7%	3%	0%	0%	0%	0%	0%	0%	0%	0%
		200	0%	2%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%
	Power	50	0%	12%	12%	12%	0%	7%	7%	7%	1%	6%	6%	6%
		100	0%	4%	4%	4%	0%	3%	3%	3%	0%	4%	4%	4%
		200	0%	2%	2%	2%	0%	2%	2%	2%	1%	4%	4%	4%
-.55	Quad	50	3%	28%	27%	20%	4%	13%	13%	12%	5%	10%	10%	10%
		100	2%	26%	26%	18%	4%	15%	15%	14%	5%	15%	15%	14%
		200	2%	24%	24%	18%	4%	21%	21%	20%	8%	21%	21%	21%
	Power	50	0%	12%	12%	12%	2%	15%	15%	15%	6%	21%	21%	21%
		100	0%	4%	4%	4%	2%	15%	15%	15%	6%	28%	28%	28%
		200	0%	1%	1%	1%	2%	13%	13%	13%	10%	34%	34%	34%

Table 5.12 Non-significance Rates of Vuong's Test for the True Exponential Decay Model with Random Intercepts and Slopes

			L1SS		
Slope	Alt Model	L2SS	5	9	13
-.776	Quad	50	85%	3%	1%
		100	67%	1%	1%
		200	38%	2%	2%
	Power	50	63%	48%	43%
		100	46%	29%	23%
		200	25%	13%	13%
-.55	Quad	50	90%	35%	15%
		100	86%	31%	22%
		200	78%	37%	30%
	Power	50	64%	49%	46%
		100	45%	46%	49%
		200	27%	43%	52%

Power. Logistic regressions with Firth's penalized likelihood were used to test main effects models of study factors on the power of Vuong's test to detect the true data generating process when random intercepts and slopes were included in non-linear models. Results indicated significant effects of Level 1 sample size (χ^2 (2)= 3303.76, p < .0001), Level 2 sample size (χ^2 (2) = 630.66, p < .0001), and slope coefficient, χ^2 (1)= 2001.04, p < .0001. Power increased as any study factors increased.

A full logistic regression model resulted in a significant three-way interaction was between Level 1 sample size, Level 2 sample size and effect size, χ^2 (4)= 57.76, p < .0001. Empirical power rates only for Vuong's test can be seen on Table 5.13 to aid in understanding the interaction. The two way interaction between Level 1 and Level 2 sample size for the small effect size, tended to differ in the larger effect size. When effect size was small, the Level 1 by Level 2 sample size interaction was such that power decreased as Level 2 sample size increased at larger Level 1 sample sizes. In the large effect size condition, there were diminishing returns

on increasing Level 2 sample sizes within Level 1 sample sizes. Still, power generally increased as Level 2 sample size increased at specific Level 1 sample sizes, except for very specific and anomalous instances.

Table 5.13 Empirical Power rates of Vuong's Test for Exponential Data Generating Process

			L1SS		
True Model	Candidate	L2SS	5	9	13
-.776	Quadratic	50	14%	97%	99%
		100	33%	99%	99%
		200	62%	98%	98%
	Power	50	37%	52%	56%
		100	54%	71%	77%
		200	75%	87%	86%
-.55	Quadratic	50	7%	61%	80%
		100	13%	65%	73%
		200	20%	59%	63%
	Power	50	36%	50%	48%
		100	55%	52%	45%
		200	74%	55%	38%

To explore what might be contributing to this phenomenon, I examined parameter bias for the estimated exponential model for each condition (Table 5.14). As evidenced by Table 5.14, there was significant bias in fixed effects contributing to increased misfit as Level 1 sample size and/or Level 2 sample size increases. When Level 1 sample size was large, the largest degree of bias was present in the estimated parameters; the intercept estimate was negatively biased whereas the slope estimate was positively biased. Within Level 1 sample sizes, bias also increased as a function of Level 2 sample size. Furthermore, the acceleration of bias also appeared to increase across Level 2 sample sizes as Level 1 sample size increased. The same phenomenon appeared to be true for the large effect size condition, however, the effects of Level 1 and Level 2 sample size, as well as the absolute degree of bias, tended to be lower. It is reasonable to assume that this estimation error is driving the curious power results.

Table 5.14 Absolute and Percent Bias in Fixed Effects Estimates for Exponential Decay Models

		True Values							
		27		-.55		27		-.776	
		γ_{00}		γ_{10}		γ_{00}		γ_{10}	
		Estimation Bias							
L1 SS	L2 SS	Abs	%	Abs	%	Abs	%	Abs	%
	50	-2.439	9%	0.104	19%	-1.751	6%	0.079	10%
5	100	-2.534	9%	0.107	19%	-1.750	6%	0.080	10%
	200	-2.583	10%	0.109	20%	-1.885	7%	0.082	11%
	50	-4.127	15%	0.163	30%	-1.945	7%	0.088	11%
9	100	-4.678	17%	0.181	33%	-2.210	8%	0.096	12%
	200	-5.106	19%	0.196	36%	-2.364	9%	0.101	13%
	50	-5.277	20%	0.196	36%	-2.277	8%	0.098	13%
13	100	-6.235	23%	0.227	41%	-2.359	9%	0.102	13%
	200	-7.290	27%	0.255	46%	-2.508	9%	0.106	14%

True Power Model.

Selecting the Correct Model. Overall, Vuong's test performed well when detecting the correct model when the true data generating process was a power model with random intercepts and slopes. Vuong's test was only underpowered in one condition, when Level 1 and Level 2 sample size were both small, the slope was weak and the alternate model was a quadratic. In this condition the correct classification rate to detect the true power model was 63%. Information criteria performed nearly perfectly across all conditions. Clearly, misclassification was not an issue and so I omit its discussion and the table. I omit the non-significance rate table for the same reason. Correct classification rates can be found on Table 5.15.

Table 5.15 Correct Model Selection Rates for True Power Model with a Random Intercept and Slope

		L1SS												
			5				9				13			
Slope	Alt Model	L2SS	Vuong's Test	AIC	AICC	BIC	Vuong's Test	AIC	AICC	BIC	Vuong's Test	AIC	AICC	BIC
-.86	Quad	50	96%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		100	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	50	88%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		100	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
-.45	Quad	50	59%	96%	96%	96%	98%	100%	100%	100%	100%	100%	100%	100%
		100	81%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
		200	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Exp	50	21%	80%	80%	80%	52%	95%	95%	95%	80%	99%	99%	99%
		100	32%	88%	88%	88%	80%	99%	99%	99%	97%	100%	100%	100%
		200	51%	95%	95%	95%	96%	100%	100%	100%	100%	100%	100%	100%

Power. Logistic regression models with Firth's penalized likelihood were used to explore the main effects of study factors that influence power when Vuong's test was used to distinguish between non-linear models with random intercepts and slopes before exploring complete models with interactions among all study factors. Results of the main effects model indicated significant main effects of Level 1 sample size (χ^2 (2)= 165.98, p < .0001), Level 2 sample size (χ^2 (2)= 308.88, p < .0001), and effect size,χ^2 (1)= 287.33, p < .0001. Increases in any study factor resulted in more power to detect the true power model on average. Analyses of a full model examining the interactions among study factors indicated only a significant interaction between Level 1 and Level 2 sample size, χ^2 (4)= 9.64, p =.047, such that as either sample size increased, the effect of additional units of the other sample size decreased. This interaction was unsurprising considering that power was basically at its asymptote in all conditions.

Table 5.16 shows bias in estimated parameters for the true power model. Similar to the exponential model, the intercept analogue was consistently downwardly biased whereas the slope parameter exhibited consistent upward bias. Despite this bias, it is likely that a ceiling effect was present as a result of the effect sizes generated. Unfortunately, it was not possible to consistently estimate smaller effect sizes under the existing conditions.

Table 5.16 Absolute and Percent Bias in Fixed Effects Estimates for Power Models

		True Values							
		27		-.45		27		-.86	
		γ_{00}		γ_{10}		γ_{00}		γ_{10}	
		Estimation Bias							
L1 SS	L2 SS	Abs	%	Abs	%	Abs	%	Abs	%
5	50	-0.091	0%	0.021	5%	-0.061	0%	0.021	2%
	100	-0.103	0%	0.022	5%	-0.082	0%	0.020	2%
	200	-0.044	0%	0.024	5%	-0.048	0%	0.020	2%
9	50	-0.249	1%	0.035	8%	-0.105	0%	0.028	3%
	100	-0.200	1%	0.038	8%	-0.123	0%	0.029	3%
	200	-0.243	1%	0.037	8%	-0.072	0%	0.029	3%
13	50	-0.360	1%	0.049	11%	-0.147	0%	0.034	4%
	100	-0.388	1%	0.049	11%	-0.187	0%	0.035	4%
	200	-0.414	2%	0.051	11%	-0.156	0%	0.035	4%

Discussion

In this study I examined the effects of study factors on the power of Vuong's test to detect the true model when non-nestedness manifested as different functional forms of growth. I also compared the performance of Vuong's test to the performance of information criteria to select the correct model. This evaluation was conducted in two contexts, random intercept only models and random intercept and slope models.

The factors that contributed to and the performance of Vuong's test in random intercept only non-linear models was generally as expected with the exception of the ICC effect. When the true data generating process was exponential, increases in Level 1 sample size, Level 2 sample size, and ICC tended to increase power. While increases in power would be expected from increases in sample size at either level, ICC is usually expected to decrease power as each individual data point contributes less unique information. When alternative models were other non-linear forms (i.e., exponential or power) there was a curvilinear effect of slope coefficient magnitude on the power to detect the correct model. As the slope coefficient magnitude used to

manipulate the effect size increased, power increased to a point before it decreased again. This effect was not surprising, however, as Figure 5.1 showed that models will tend to fit similarly when slopes are large or small. Overall, Vuong's test tended to be better at distinguishing between quadratic and exponential models than power and exponential models. In comparison to the information criteria, Vuong's test tended to outperform information criteria in that it rarely, if ever, selected the incorrect model.

When the data generating process was a power function, the correct classification rate was generally high across medium and large Level 1 sample sizes unless Level 2 sample size was small and the slope was weak. In general, where power had not yet reached its asymptote increases in Level 1 sample size or Level 2 sample size tended to improve power, however, increases in Level 1 sample size tended to improve power to a greater degree. When comparing Vuong's test to the information criteria, results for power models were very similar to those for exponential models. When information criteria selected the incorrect model, Vuong's test failed to find significance in favor of either model in a large proportion of replications. Vuong's test was beneficial over information criteria in that it rarely if ever suggested the incorrect model entirely.

When random slopes were included, power became significantly worse when the true data generating process was exponential. Power was greater for the large effect size condition than the small effect size condition, however, Level 1 and Level 2 sample size exhibited a strange interaction. When Level 1 sample size was small, increases in Level 2 sample size increased power whereas when Level 1 sample size was large, increases in Level 2 sample size decreased power but only when effect size was small. Exploratory analyses indicated that in the

small effect size condition the intercept parameter was downwardly biased while the slope parameter exhibited upward bias.

When comparing the performance of Vuong's test to the performance of information criteria, once again the biggest benefit of Vuong's test occurred when the alternative model was quadratic. However, when Level 1 sample size was large and effect size was small, the amount of misclassification when the alternative model was a power model was the highest recorded in this particular study. Misclassification rates reached as high as 34%. In the large effect size condition, Vuong's test outperformed information criteria generally in the small Level 1 sample size when the alternative candidate was a quadratic model, and across all Level 1 sample sizes when the alternative was a power model.

For the cases in which the data generating process was a power model, power was almost always adequate regardless of sample size or comparison and there was almost never any misclassification from information criteria. It appears that the estimable slopes resulted in a ceiling effect of power.

While statistical power decreased at either end of the slope spectrum, the decrease in power manifested in different ways. Referring back to Table 5.1**,** when comparing power and exponential functions, statistical power was similar at the largest and smallest slope values. However, when comparing quadratic and exponential functions, power was greater when the slope was stronger compared to when it was weaker. In stronger slope conditions, the true model likely reached the asymptote within the functional space even when there were few Level 1 units and created additional misfit in the quadratic model. With additional Level 1 units, this effect was exaggerated and it became easier to distinguish between the true non-linear forms and the quadratic models.

The parameter bias that was encountered when the slope coefficient was small, the data generating process was exponential, and random intercepts and slopes present was an unexpected finding. It is unclear as to why adding a random slope into the data generation would introduce bias into the fixed effects. It is also unclear as to why bias became worse as sample size became larger. This is an open question for research and should be evaluated before more work on this topic is done with non-linear models.

An additional issue in this study was the difficulty with determining "effect size" in terms of model fit. While the slope parameter is a convenient way to manipulate the degree to which models take a specific shape it is clear that if research into this method should continue, it will be important for researchers to assess and manipulate model fit directly. One potential avenue for such a metric is the root mean square error. If a researcher knows the data generating process, as they would in a simulation study, the difference in predicted values from the estimated models from the true model would provide a closer analogue to K-L divergence than that achieved through manipulating model parameters. While RMSE is a good metric on which to base future research, it is unclear as to how it can be directly manipulated.

In this study I explored the study factors that contribute to power in Vuong's test to distinguish between models with different functional forms and compared it to the performance of information criteria. While Vuong's test outperformed information criteria outright in a number of cases, there were other cases in which its superiority was not so obvious. As was shown in the previous two chapters, those circumstances in which the power of Vuong's test was low (generally with small differences between models) are precisely the circumstances in which using Vuong's test provided benefit over information criteria. While Vuong's test may not find a significant difference between models, it rarely if ever suggests the incorrect model, except when

a random slope was included in the model. Information criteria, on the other hand, will suggest the incorrect model more often that Vuong's test when the power of Vuong's test is low.

While the previous three chapters have examined the performance of Vuong's test in simulation studies, the test has not yet been applied to real data using multilevel models. In the next chapter chapters I use real data to apply Vuong's test to three cases which might be seen in practice: non-nested covariate sets, non-nested Level 1 residual error structures, and non-nested functional forms. I conclude the dissertation with a general discussion highlighting overall findings, limitations of these studies, and future directions.

Chapter 6: Applications

In the previous three chapters I have extended Vuong's test to three scenarios where non-nested models commonly arise in multilevel data: different predictor sets, Level 1 residual covariance structures, and non-linear functional forms. While the previous studies have illustrated the utility of Vuong's test through simulation, they have not applied the test to any real data. In this chapter I illustrate how using Vuong's test might influence results from two previously published studies and a publically available data set. The first application revisits data from Moreno et al. (2016) in which the relationship between positive affect and inflammation in breast cancer survivors was explored. While the original study was focused on individual differences in inflammation related to positive affect, the current application of Vuong's test will address a modeling decision: which of two fatigue indicators to control for in the model. Because the simulation study conducted in Chapter 3 included only a small number of covariates beyond time, the application will be examined in two parts. The first analysis will examine growth curves where only fatigue is controlled for. The second analysis will examine the fully conditional growth models including all of the variables modeled in Moreno et al. (2016). Further, both of these analyses will be applied to the two inflammation markers of interest in the original study, soluble tumor necrosis factor receptor type II (sTNFR-II) and C-reactive protein (CRP).

The second application will explore the case of non-nested Level 1 residual covariance structures. Following the lead of Bollen and Curran (2004), I use household income data in two year increments between 1986 and 1994 from the freely available National Longitudinal Study of Youth (NLSY). In doings so, I compare the autoregressive structure with a lag of one (AR(1)) to the three-banded Toeplitz structure (TOEP(3)) as done in Chapter 4. While it is unclear which

Level 1 residual covariance structure should be preferred in these data, this example reflects the modeling process that would be undertaken in an applied setting.

Finally, the third application will examine competing functional forms of growth. For this illustration, previous data from Weisz et al. (2012) will be examined. These data were used to compare the effect of a modular therapy on children's problem behaviors to standard manualized treatment and a "treatment-as-usual" control group. Overall, children's improvement progressed nonlinearly with rapid improvement early in treatment which slowed over time. In a randomized controlled trial such as Weisz et al., this behavior is exactly what researchers would hope to observe; the treatment is highly effective for clients at the outset of treatment and improvement slows as the target behaviors approach a baseline. To accommodate this nonlinearity, Weisz et al. applied a logarithmic transformation to linearize the time trend.

While there may be substantial practical benefits for linearizing a time trend through transformation, not least of which are specification and estimation, there are also arguments in favor of modeling the time trend on its original scale. Discussing results in terms of "days", for example, is more intuitive than the adjusted time scale of "log days". Additionally, linearizing time does not allow one to incorporate certain parameters that map on to theoretically important aspects of growth (e.g. asymptotes). Excluding such parameters has the potential to limit the understanding of the nature of change and its determinants (Grimm, Ram, & Hamagami, 2011).

Non-nested Covariate Sets

Moreno, Moskowitz, Ganz, and Bower (2016) examined data to explicate the relationship between high and low arousal positive affect and inflammation in 186 women who had recently completed treatment for breast cancer. Although the study was focused on the relationship between positive affect and inflammation, a number of covariates were included in the model to

control for individual factors such as type of treatment, age, and time since last treatment, among others. During peer review, one reviewer had noted that our measure of high arousal positive affect (e.g., “excited”, “active”, and “enthusiastic”) might simply capture the absence of fatigue, rather than positive feelings. Cancer survivors in particular have been shown to experience elevated levels of fatigue and as a result the reviewer thought our study would be strengthened by including a measure to control for it.

Two candidates for fatigue were eligible for inclusion in the model: fatigue severity and fatigue interference, both of which were measured by the Fatigue Symptom Inventory (Hann et al., 1998). Fatigue severity assessed how fatigued a participant had been for the past week, how many days they had felt that way, and the extent of each day on average they felt fatigued (none of the day to the entire day; Hann, Jacobsen, Azzarello et al., 1998). Fatigue interference assessed the degree to which participants felt that fatigue hindered their performance of normal activities. While arguments could be made for including either predictor, only one predictor was found to be related to the outcome despite their moderate to high collinearity (r=.77). Two models were fit using SAS Proc MIXED with default REML estimation and Satterthwaite degrees of freedom. Both models were identical in the measures of positive affect, time, and other covariates (e.g., gender, treatment, etc.) but differed in the inclusion of fatigue: one contained interference and the other severity. Results indicated that fatigue severity was significantly related to sTNF-RII levels whereas fatigue interference was not. Neither measure of fatigue was significantly related to CRP. Therefore, Moreno et al (2016) chose to control for fatigue severity in the models.

Determining which covariate to control for is a prime example of where Vuong's test for non-nested models could be applied. Thus, I revisited the decision of which fatigue covariate to include with Vuong's test. For each of the two inflammation markers, sTNF-RII and CRP, I fit restricted models including time and either fatigue interference or fatigue severity as well as complete models with full predictor sets differing only in their fatigue variables. All models included random intercepts and time slopes as well as an intercept and slope covariance. They were estimated using SAS Proc MIXED with maximum likelihood estimation and Satterthwaite degrees of freedom. The goal of this approach was to first generate results directly comparable to the simulation study in Chapter 3 and then to examine how Vuong's test would be used in a truly naturalistic setting. For all models, data with any missing values in the predictors or outcomes were omitted resulting in a 10% loss of subjects. In its current state, the SAS macro developed to conduct Vuong's test cannot incorporate missing values.

sTNF-RII. Models examining sTNF-RII were estimated on data from 167 individuals with at most 3 time points of data. The residual ICC after accounting for the fixed and random effects of time was .84, indicating that much of the variability in sTNF-RII was between persons. Intercept and slope were highly correlated at -.45.

When comparing the candidates with only time and the fatigue variable, Vuong's test was non-significant at the nominal .05 level, $Z_{VT} = -1.425$, $p_{interference} = .923$, $p_{severity} = .077$. The notation used here is similar to that used by Merkle et al. (2015) to convey the results of Vuong's test. Z_{VT} refers to Vuong's test statistic given by Equation 21. Each "p" refers to the probability of the model defined by the variable in the subscript; $p_{interference}$ refers to the probability associated with preferring the model with fatigue interference whereas $p_{severity}$ refers to the probability associated with preferring the model with fatigue severity. The information criteria

(Table 6.1) uniformly suggested that fatigue severity should be preferred over fatigue interference. In the simulation study reported in Chapter 3, Vuong's test suggested the incorrect model in at most 1% of replications when non-nestedness occurred in Level 2 covariates, Level 2 sample size was 100, and Level 1 sample size was 5. While this study contains fewer Level 1 samples, it contains more Level 2 samples, which, according to Chapter 3, has a greater effect on the power of Vuong's test when non-nestedness occurs at Level 2. Additionally, fatigue severity explained 3.1% more variance in the intercept than fatigue interference. The empirical power rates observed in Chapter 3 suggest that power for this test is somewhere between 9% and 18%. Given that Vuong's test appears underpowered, because it is marginally significant in support of fatigue severity in a decision commensurate with the information criteria it would be reasonable to select the model with fatigue severity over fatigue interference.

Table 6.1 Information Criteria in Growth Models for sTNF-RII Conditional on Only Fatigue

	AIC	AICc	BIC
Fatigue Interference	-244.55	-244.31	-222.72
Fatigue Severity	-250.30	-250.06	-228.48

A more realistic comparison was made using the full models estimated by Moreno et al (2016). Because fatigue was simply a covariate to be controlled for in this analysis and not a variable of interest, it is unrealistic that it would be tested by itself. Using the same methods as the previous comparison, two models were estimated with all of the effects of interest included. That is, models were completely specified as they were in Moreno et al. Results with the full model were less convincing. While the information criteria maintained support for fatigue severity (Table 6.2), Vuong's test trended away from significance, $Z_{VT} = -.892$, $p_{interference} = .814$, $p_{severity} = .186$. Given the study conditions and the reduced difference between models, now 2.3% more variance explained, the power of Vuong's test would now be even lower. While arguments

can be made for a more lenient cutoff for the test statistic in the previous example, the current results require a larger stretch. Still, taken together there appears to be modest support for the preference of fatigue severity, although more evidence should be gathered to make a stronger claim.

Table 6.2. Information Criteria in Fully Conditional Growth Models for sTNF-RII.

	AIC	AICc	BIC
Fatigue Interference	-294.19	-292.7	-238.06
Fatigue Severity	-297.86	-296.35	-241.74

CRP. Models testing the relationship between positive affect and CRP were also reexamined using Vuong's test. Again, these analyses utilized only 167 of the original 189 participants due to missing data. Residual ICC was almost as high as it was for sTNF-RII at .75. A similar analysis with options identical to those specified above was used to test the CRP models.

When comparing candidates with only time and each respective fatigue variable, Vuong's test was non-significant, Z_{VT} = -.381, $p_{interference}$ = .648, $p_{severity}$ = .352. While information criteria uniformly agreed that the model including fatigue severity should be the preferred, differences were exceedingly small (Table 6.3). While the simulation results in Chapter 3 suggest that Vuong's test would be severely underpowered (Level 1 observations: 3, Level 2 observations: 167, variance explained difference = .2%), the thin margin by which fatigue severity fits the data better than fatigue interference indicates that there is too little information on which to make a decision in favor of one model over the other. Again, even if .05 is too stringent a criteria for Vuong's test, the .35 significance level is far too liberal to suggest support.

Table 6.3 Information Criteria in Growth Models for CRP Conditional on Only Fatigue

	AIC	AICc	BIC
Fatigue Interference	1391.09	1391.33	1412.91
Fatigue Severity	1390.65	1390.89	1412.47

A more realistic approach was employed by testing the final model from Moreno et al. (2016) in its entirety. Similar to the results for sTNF-RII, there was less of a difference between fully conditional models when comparing fatigue interference with fatigue severity. Vuong's test and the information criteria all agree that choosing between either of these two variables for inclusion in the model would be trivial. AIC, AICc, and BIC are essentially identical values (Table 6.4). Even the probability of selecting either model from Vuong's test was approaching 50%, $Z_{VT} = -.028$, $p_{interference} = .511$, $p_{severity} = .489$. As a result, it can be concluded that both covariates would function equally well in this context.

Table 6.4. Information Criteria in Fully Conditional Growth Models for CRP.

	AIC	AICc	BIC
Fatigue Interference	1335.7042	1337.21	1391.83
Fatigue Severity	1335.6969	1337.20	1391.82

Taken together, the results above suggest that there is mild evidence supporting the decision to include fatigue severity over fatigue interference, at least when examining sTNF-RII. Both fatigue severity and fatigue interference performed equally well when predicting CRP. There is no evidence clearly favoring one covariate over the other. For continuity it would be appropriate to include fatigue severity in the model when analyzing CRP. While it may be unrealistic to expect substantive researchers to report these tests, especially when they do not concern the focal predictors, the method by which these modeling decisions are made should be acknowledged at some point in a manuscript.

Non-nested Residual Variance Structures.

Kwok, West, and Green (2007) showed that misspecification (i.e., non-nestedness) of the Level 1 residual covariance matrix can lead to a number of issues in parameter estimation, including overestimation of variances of the random effects at Level 2, overestimation of standard errors of growth parameters, and lower statistical power to detect fixed effects. In order to mitigate these negative effects of misspecification, it is important to try to recover the true residual covariance matrix as closely as possible. To illustrate how Vuong's test could be used to compare residual Level 1 covariance structures in multilevel models, I provide an example using the National Longitudinal Study of Youth (NLSY) following the methods of Bollen and Curran (2004).

NLSY data were originally collected to understand, in detail, the life course experiences of young adults in America (Bureau of Labor Statistics). Specifically, data were collected on a variety of topics including labor market behavior, health issues, financial information, etc. to examine the transition of young adults into the work force. Beginning in 1979, the original sample included survey responses from 12,686 individuals between the ages of 14 and 22, half of whom were female. A majority of the original sample was white (59%), however black (25%) and Hispanic or Latino (16%) respondents were represented as well. This sample was intended to be nationally representative of the United States at the time.

Following Bollen and Curran (2004) I extracted respondents' total net family income over the previous calendar year in two year increments from 1986 to 1994 (N = 3995). All respondents reported complete data at all time points. Bollen and Curran report using the same data extracted for this study, however, for an unknown reason their sample consisted of only 3912 individuals. The current sample matched Bollen and Curran in terms of mean age (24.7

years, SD = 2.2) with a minimum age of 21 years and a maximum of 29 years, however, gender and ethnic composition varied slightly. Despite the unexplained differences in samples, I maintain methodological consistency with Bollen and Curran and use the square root transformation of the net household income variable to reduce kurtosis and skewness of the data in its original scale.

In the motivational example, Bollen and Curran (2004) used these data to illustrate the utility of their Autoregressive Latent Trajectory model. Because of their assumption that these data are autoregressive, I take the position that Vuong's test should prefer the autoregressive model to a Toeplitz model with three bands.

Results of Vuong's test comparing the AR(1) model to the TOEP(3) model indicate that evidence the more flexible TOEP(3) model fits the data better than more restricted AR(1) model, $Z_{VT} = -3.052$, $p_{AR(1)} = .999$, $p_{TOEP(3)} = .001$. Differences in information criteria also supported the TOEP(3) model despite the additional complexity. Information criteria can be found on Table 6.5.

Table 6.5 Information Criteria for Models Non-nested in Level 1 Residual Variance Structure

	AIC	AICc	BIC
Autoregressive	227925.42	227925.42	227963.18
Toeplitz(3)	227878.19	227979.19	227922.23

The estimated residual covariances for both the AR(1) and TOEP(3) were

$$AR(1) = \begin{bmatrix} 4292.33 & 433.65 & 43.81 & 4.43 & .45 \\ 433.65 & 4292.33 & 433.65 & 43.81 & 4.43 \\ 43.81 & 433.65 & 4292.33 & 433.65 & 43.81 \\ 4.43 & 43.81 & 433.65 & 4292.33 & 433.65 \\ .45 & 4.43 & 43.81 & 433.65 & 4292.33 \end{bmatrix}$$

$$TOEP(3) = \begin{bmatrix} 4376.45 & 542.59 & 392.91 & 0 & 0 \\ 542.59 & 4376.45 & 542.59 & 392.91 & 0 \\ 392.91 & 542.59 & 4376.45 & 542.59 & 392.91 \\ 0 & 392.91 & 542.59 & 4376.45 & 542.59 \\ 0 & 0 & 392.91 & 542.59 & 4376.45 \end{bmatrix}.$$

Examining the estimated values of these matrices conveys obvious differences in the representation of the residual covariances. The AR(1) was forced to estimate the residual covariances in an exponential form. That is the coefficient scaling the variance for the off-diagonal elements decreases exponentially as it gets further away from the diagonal.

The TOEP(3) structure implies that the residual covariances do not necessarily decrease at an exponential rate. While the relationship between the diagonal and the first off diagonal element of the TOEP(3) matrix resembled that of the AR(1) matrix, the values in the second off diagonal were markedly different. There was considerably more covariance estimated between time points of lag 2 for the TOEP(3) structure than there was for the AR(1) structure. While misspecification likely remains in the third and fourth off diagonal elements of the TOEP(3) matrix, it is likely that the discrepancy between the two matrices in the second off diagonal element provides enough of a difference in fit for Vuong's test to prefer the TOEP(3) model.

Non-nested Functional Forms

To examine how Vuong's test could be applied to real data when comparing different functional forms, I revisited data from Weisz et al (2012). In this study, researchers examined the effect of a modular therapy on children's problem behaviors and compared it to the efficacy of standard manualized treatment and a treatment-as-usual control group. Researchers reported results of their analyses with time transformed to log days, indicating that a long right tail existed in the dataset that may have attenuated the treatment effects had the trend been modeled as linear. The log transformation effectively linearized this curvilinear trend by transforming time values on the x-axis to be less extreme. Functionally, this transformation condensed the larger

time values to give them less leverage. While log transforming the time variable is an accepted method for linearizing the curvilinear relationship that was initially observed (Grimm, Ram, & Hamagami, 2011), doing so may complicate interpretation.

Alternatively, the curvilinear relationship can be modeled directly. While this approach might ameliorate some problems resulting from transformations (i.e., interpretability), others may be introduced. For instance, while a quadratic time effect might be suitable to model a curvilinear relationship within a restricted space, the form of the growth implies a change of direction in the future (or past). Furthermore, the function will decrease (or increase) to infinity after the shift. Another option to address curvilinear relationships is to model a naturally non-linear form such as exponential decay. In cases when values decrease and reach some asymptote without changing direction, exponential decay is an appealing model. Non-linear models are not without their own issues, however, in that they may be exceedingly hard to estimate and model parameters do not necessarily map neatly onto those of the more common linear models. Still, if the more complex model can provide a better fit to the data and help to explain the phenomenon under study with more accuracy, it can be worth the effort.

One-hundred and seventy four children (70% male) were assessed and treated over an average of 221.9 days (SD = 143.7, Mean sessions = 16.8, SD = 11.4) for anxiety, depression, or disruptive behavior (Weisz et al. 2012). However, as the program developed to conduct Vuong's test is unable to incorporate cases with missing values, some data was excluded from the analysis. Complete data from 129 (59% male) children treated over an average of 132.14 days (SD = 55.16, Mean sessions = 15.22, SD = 7.06) were included in these analyses. While this exclusion is not ideal and makes the unrealistic assumption that missing data is 'missing completely at random' (Enders, 2010), the remaining data should suffice for the purposes of this

demonstration. Weisz et al (2012) found a curvilinear relationship in the data to warrant transformation to log days, I compared the fit of the linear model with log days to the naturally non-linear exponential decay model. A logical first step in analyzing these data would be to determine the time trend that best models the observed change.

Examining unconditional growth models where the growth trend was either non-linear or linear on a log transformed time variable resulted in residual ICCs of .75 and .77, respectively. Results of Vuong's test comparing the two models indicated a non-significant difference in fit, but support for the non-linear exponential model trended toward significance, $Z_{VT} = 1.37$, $p_{exponential} = .086$, $p_{linear} = .914$. Information criteria uniformly support this conclusion. AIC, AICc, and BIC all suggest the exponential model is a better fit to the data than the linear model with the log transformation of time. Information criteria can be found in Table 6.6.

Table 6.6 Information Criteria for Unconditional Growth Models of Problem Behaviors

	AIC	AICc	BIC
Exponential Decay	9945.2	9945.2	9962.3
Linear Log Transform	10027.2	10027.3	10044.4

The analysis program was further adapted to incorporate a model testing the effect of modular therapy and standard manualized treatment on the time trend (i.e., the treatment by time interactions). Thus, two new models were fit incorporating dummy variables for modular treatment and standard manualized treatment compared to the treatment-as-usual control group. These Level 2 indicator variables were included as predictors of both the random intercept and random time slope. The time by treatment interaction, specifically the effect for modular treatment (γ_{11}), would typically be the parameter of interest in this type of analysis. The two-level equations for this analysis were specified as:

$Linear$:

$$Level\ 1: BPC_{tot} = \beta_{0j} + \beta_{1j}\ln(time_{ij}) + \epsilon_{ij}$$

$$Level\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01}MMT_{.j} + \gamma_{02}SMT_{.J} + u_{0j} \quad (52)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}MMT_{.j} + \gamma_{12}SMT_{.j} + u_{1j}$$

And,

$Exponential$:

$$Level\ 1: BPC_{tot} = \beta_{0j} * e^{\beta_{1j} * time_{ij}} + \epsilon_{ij}$$

$$Level\ 2: \beta_{0j} = \gamma_{00} + \gamma_{01}MMT_{.j} + \gamma_{02}SMT_{.j} + u_{0j} \quad (53)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}MMT_{.j} + \gamma_{12}SMT_{.J} + u_{1j}$$

where,

$$var\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N(0, \boldsymbol{\tau}) \quad (54)$$

and,

$$var(\epsilon_{ij}) \sim N(0, \sigma^2). \quad (55)$$

Results of Vuong's test for the conditional growth curves reflected the same result as the unconditional models. When treatment variables were included in the model Vuong's test remained marginally significant in favor of the exponential decay model, $Z_{VT} = 1.42$, $p_{exp} = .077$,

p_{lin} = .923. Information criteria (Table 6.7) also uniformly suggested that the exponential decay model be preferred when treatment variables are included in the model.

Table 6.7 Information Criteria for Unconditional Growth Models of Problem Behaviors

	AIC	AICc	BIC
Exponential Decay	9939.5	9939.7	9968.1
Linear Log Transform	10025.1	10025.2	10053.7

Because treatment effects are of central importance to these analyses, it is worth examining them as well. Parameter estimates and significance tests for the exponential decay and linear models can be found on Table 6.8. Beginning with the exponential decay model, there was a significant difference in initial values between the group that received modular treatment and the control group. On average, children in the modular therapy group began with about 3 more reported problem behaviors. Furthermore, the group receiving modular therapy improved at a faster rate than the control condition. For each day in treatment, the exponential slope decreased by .002 units. There was no significant difference in initial values between the standard manualized treatment group and control nor was there a significant difference in their exponential time trends.

Table 6.8 Parameter Estimates of Fixed Effects for the Exponential Decay and Linear Model with Log(Time) Predicting Differences in Problem Behaviors (BPC) Across Treatments.

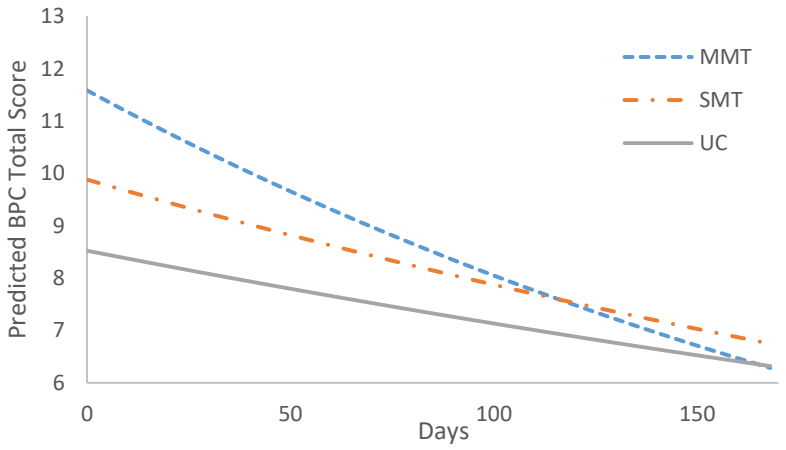
	Exponential Decay	Linear w/ log(time)
Int	8.5223***	9.6024***
Time	-.0018**	-.3568*
MMT	3.0611**	3.2346**
SMT	1.3578	1.5600
MMT*Time	-.0019*	-.6185**
SMT*Time	-.0005	-.4514^{+}

$^{+}$p < .10 * p < .05 ** p < .01 *** p < .001

Results were basically identical for the linear model with the log transformed time variable. Initial values for the modular treatment group were significantly higher than for the

control group. Additionally, the modular treatment group improved at a faster rate than controls; the for each log day unit the difference between the modular and control conditions increased by .62. Once again, there was no significant difference between the manualized treatment group and control at the study outset. However, the difference in slopes between manualized treatment and control groups was marginally significant.

Finally, the implications for the differences in trends can be seen in Figure 6.1. The top panel displays the results from the exponential decay model whereas the bottom panel displays the results of the linear model with the log transformed time variable. Predicted values from the exponential decay model indicate that although initial values for the modular treatment are greatest, by the end of the study (25 weeks indexed by days) the modular therapy group has the lowest number of problem behaviors. Conversely, the linear model suggests that the standard manualized treatment has the best outcome at the end of the study. Although significance tests would lead to similar conclusions, modular treatment performs better than usual care, plots of the effects over time paint a different picture. This example illustrates precisely why it is important to consider non-linear models when selecting among candidates to fit curvilinear data.



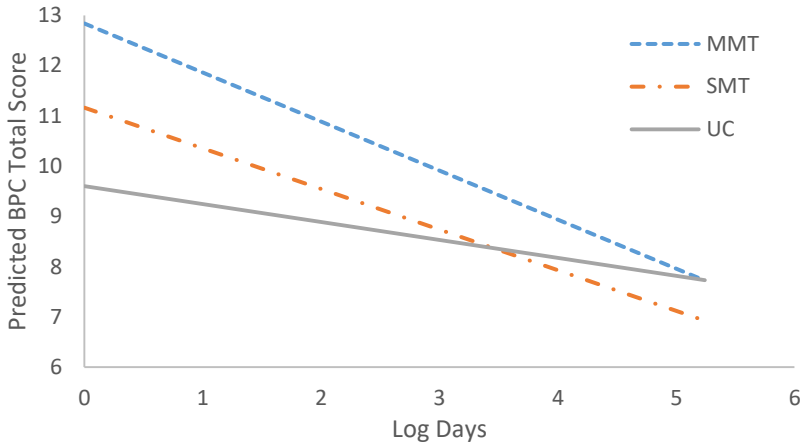


Figure 6.1 Predicted Values from the Exponential Decay (Top) and Linear Model with log transformed time variables.
Note: The y-axis was adjusted to highlight non-linearity in the exponential decay model. The adjustment was carried over to the linear model for consistency.

Results from these analyses suggest exponential decay model should be preferred over the linear model with a log transformation of time. While both the unconditional and conditional growth curves fail to reach significance at the traditional .05 level, they both fall in the range of marginal significance in favor of the exponential model. Comparisons of information criteria overwhelmingly support the exponential decay model. Still, because the results of Vuong's test were inconclusive, preference for one model over another should be discussed with caution. However, with the evidence provided in this example it would be reasonable to conclude that the exponential decay model best represented the non-linear trends in the data.

Discussion

In this chapter I have applied Vuong's test to three real data sets in which candidate models were non-nested in either their covariate sets, Level 1 residual covariance structures, or functional forms. The first illustration applied Vuong's test to a case where researchers needed to decide between two covariates to include in the model. For the first outcome, sTNF-RII, a model conditional only on time and the covariates suggested weak support in favor of fatigue severity,

whereas for a fully conditional model Vuong's test was not able to distinguish between the two covariates. Furthermore, information criteria always preferred the model including fatigue severity, however, Vuong's test only trended toward fatigue severity in the unconditional model and had almost equal probabilities of preferring both models when the fully conditional model was specified. In the second illustration, Vuong's test tended to be inconclusive as were the information criteria for both the restricted and full models. Because neither fatigue severity nor fatigue interference provided a better fit to the data, it was determined that fatigue severity be included in models of CRP for consistency.

Next, an illustration was provided exploring Vuong's test when models were non-nested in their Level 1 covariance structures. Results suggested that the TOEP(3) structure be preferred over the AR(1) structure when examining net household income in the NLSY data. When comparing the TOEP(3) model to AR(1), information criteria indicated that the increased flexibility afforded by the extra parameter in the Toeplitz model significantly improved model fit. By examining the residual covariance matrices directly, it was clear that the autoregressive model was underestimating covariability among observations with lags greater than one.

Finally, an illustration was provided comparing a truly non-linear functional form and a linear model with a log transformed time variable. This illustration served two purposes. The first was to mimic the initial step in fitting a growth curve model by testing different functional forms in an unconditional model. The second was to provide an example in which Vuong's test with non-linear models was adapted to include predictors or covariates. Including additional fixed effects required modifications to the partial derivatives used to create the random effects design matrix used in calculating the case-wise log likelihoods whereas additional random effects would require additional partial derivatives.

Results indicated that information criteria preferred the non-linear form for both the unconditional and conditional models, with Vuong's test also trending toward significance in this direction. When examining the significance of the effects of interest, differences in time trend across treatment groups, conclusions based on statistical tests would be the same across models. However, when examining the predicted values of each model differences started to emerge. In the better fitting, exponential decay model, the group who received modular therapy had the best predicted outcomes. This group also displayed the most non-linear slope. Conversely, predicted values from the linear model indicated that the standard manualized treatment group had the best outcomes at the study's conclusion. Not only did the exponential decay model fit the data better, but it provided a more nuanced and accurate insight into the data.

While Vuong's test provided consistent results in these examples it may not always do so. Given that both candidates are in line with theory, it would stand to reason that Vuong's test could be used to assess models at different stages in the fitting process until one distinguishes itself from the other. For instance, while it might be ideal to attempt to determine the functional form of growth prior to including treatment variables, if at first models fit equally well, it would be wise to compare them at each step until one is preferred over the other.

Perhaps the biggest drawback of Vuong's test is the ambiguity that remains when two models fit the data equally well. As a non-significant difference between two models does not imply that candidates are not good representations of the data, but rather fit the data no better than one another, a researcher would need to make a decision as to which model to report. To researchers in this quandary, I would suggest first choosing the model with the most theoretical justification. If there is ample theory for both candidates according to Vuong's test, I would suggest the model with fewer parameters. Using the same logic that is typically applied to

likelihood ratio tests of nested models, if the added complexity of a model does not significantly improve its fit, the more parsimonious model should be preferred. It is an open question as to whether preferring the more parsimonious model in the case of a non-significant result from Vuong's test would introduce substantial bias if the true model was in fact more complex. In this case where candidate models contain the same number of parameters, when Vuong's test is non-significant I would suggest using the model supported by the information criteria. Although not ideal, following information criteria in light of a non-significant result from Vuong's test is not technically wrong in that both models fit the data equally well. Under these circumstances, alternative models should be acknowledged and statements about model preferences tempered until more definitive conclusions can be reached.

Chapter 7: General Discussion

In this dissertation I have made a case for Vuong's Likelihood Ratio Test to be used as an alternative to information criteria when comparing non-nested multilevel models. I have shown that while information criteria might be more sensitive to differences in model fit, they are far more error prone than Vuong's test. Whereas Vuong's test rarely suggested the wrong model, information criteria selected the incorrect model in as many as 75% of replications. Additionally, I have evaluated the effects that study factors have on the power of Vuong's test to detect the best model when non-nestedness manifested in several different scenarios. Overall study factors behaved as expected with some deviation in effect sizes and ICCs for non-linear models and Level 2 covariates. Finally, I provided three examples of when Vuong's test could be used in practice, acknowledging the ambiguity that remains in the presence of a null result.

While Vuong's test may not be as sensitive as information criteria to differences in competing models, its potential lies in that Vuong's test rarely selects the incorrect model. From this perspective, Vuong's test outperformed information criteria in every scenario. In fact, when Vuong's test was underpowered and had difficulty determining which candidate fit the data best, information criteria tended to perform at their worst. That is, information criteria always performed well when there was enough power for Vuong's test to detect the best model but encountered considerable issues in other cases.

Across the different manifestations of non-nestedness, information criteria performed differentially. In the case where the true model had more parameters than the alternative model, the larger penalty of the BIC resulted in more misclassification than AIC or AICc. Conversely, when the true model contained fewer parameters than the alternative model BIC outperformed AIC and AICc. While this behavior might be useful when a researcher intends to find the most

parsimonious model, it is problematic if they are attempting to determine the best model from a number of candidates. Without knowing if the best model in the population has more or fewer parameters than alternatives, it is difficult to decide which information criterion to use. In cases where information criteria may disagree, it is likely that Vuong's test will be inconclusive and fail to provide evidence in support of either model.

In these instances where Vuong's test is unable to determine which candidate should be preferred, Merkle et al. (2015) advocate for specifying a larger model that encompasses both candidates as special cases. Thus, the original candidate models would be nested within the newly specified more complex model. Researchers would then either use this more complex model for inference or continue with the modeling process to create an alternative that combines the two original candidates. Alternatively, a researcher could compare the more complex model to each original model with a nested likelihood ratio test to determine if the more complex model provides a significantly better fit than the original models. While this advice may prove helpful in certain situations, researchers may still encounter problems. Should the nested likelihood ratio test produce a null result, the researcher is back to square one: the more complex model does not provide an improvement in fit and a choice must be made between two non-nested candidates. Additionally, should a researcher decide to retain the more complex model regardless of a likelihood ratio test, there is no guidance as to how model fitting should proceed to reduce it. In other cases, it may not be possible or may not make theoretical sense to specify a more complex model in which both candidates are nested. Thus researchers may be left with a difficult decision and inadequate evidence on how to proceed with their study.

Because Vuong's test does not provide sufficient evidence to support either model does not mean that a researcher cannot continue their study. Recall that Vuong's test is a relative fit

statistic. That is, it does not indicate the degree to which a model fits in a population, but rather compares the fit of one model to that of another. Leveraging the relative nature of Vuong's test, a researcher can choose a number of approaches. Ideally, and especially when Vuong's test is non-significant, modeling decisions should be heavily based on theory. With the caveat that there is not enough empirical support for either theory, a researcher can discuss results in the context of the hypothesized theory while simultaneously reporting the results of the model fit. That is to say, because the alternative theory is no better than the proposed theory (and vice versa) it would not be incorrect to discuss one theory's implications. However, acknowledging the possibilities of and differences in both theories would be the best course of action and allow researchers to temper their arguments in support of either claim while setting the stage for future work.

When candidate models have different numbers of parameters and Vuong's test does not indicate which model should be preferred, a researcher should be more skeptical of results from information criteria, as they may mistakenly select the more parsimonious model when there is little difference between candidates. Without a strong theory for the more complex model, a researcher could take the non-significant result to mean that the added complexity of the extra parameter does not significantly improve the fit of the model. This approach would be analogous to that traditionally taken for the nested likelihood ratio test and would likely align with the results implied by comparisons of information criteria. A researcher could also proceed by fitting a simpler alternative model. For example, fitting a model with a TOEP(2) structure instead of a TOEP(3) structure if the more complex model does not significantly improve model fit. If models still fit the data equally well by Vuong's test, researchers should note the non-significant result and differences between the models and proceed with the model best supported by their theory (or most empirically interesting) with appropriate caveats.

The decision to include focal variables in tests of certain non-nested models is also an open question. While results from examples in Chapter 6 were consistent regardless of the inclusion of focal variables they may not always lead to the same conclusions. Initial analyses of the non-linear example in Chapter 6 including only a single vector representing the modular treatment group led to conflicting results when the variable was included or excluded. When the treatment effect was excluded, Vuong's test showed mild support for the non-linear model. However, when the treatment effect was included in the model (predicting both intercept and slope), Vuong's test was highly significant in favor of the model with a linear transformation. It is a matter of debate at which point non-nested models should be tested, however, when the focus of a study is treatment differences in trends over time I take the perspective of including the focal variable. The implications of these differing perspectives are a direction for further research,

Alternatively, researchers still have the option of utilizing to information criteria to guide model selection in the event that Vuong's test is non-significant. While the results of this dissertation indicate that information criteria tend to perform poorly when Vuong's test is non-significant and as such would suggest that this approach is ill-advised, it does still provide *some* empirical basis on which to choose a model. Even when using information criteria for model selection, the results of Vuong's test, albeit non-significant, should be taken into account. Should Vuong's test approach significance in favor of the model suggested by information criteria, a researcher could be more confident that they are selecting the best model. Caution should remain when using information criteria, however, if Vuong's test does not lean in favor of either model. In these cases where Vuong's test does not support either model, information criteria tend to perform at their worst.

The power of Vuong's test to detect the best model behaved as expected. Generally, as sample size increased at either level, power to detect the best model increased. As power approached the asymptote of 100%, the relative effects of increases in sample size were diminished. This result came as no surprise, however, as diminishing returns of increases in sample size as power approaches 100% are widely known.

Increases in effect size (i.e., differences in model fit) also tended to increase power, however manipulating the slope coefficients did not always produce the expected result. In Chapter 5, I showed that as the slope coefficient increased power increased up to a point and then started to decline. While this result was unexpected, it was easily explained when examining plots of the predicted effects. When the non-linear slope was moderate, the alternative models were unable to adequately approximate the slope of the exponential model, especially when Level 1 sample size was small. By definition, this poor approximation led to larger differences in fit between the true and alternate models. As coefficients deviated from the median, alternative models were able to better approximate the true exponential model and power decreased.

This curvilinear effect of the effect size manipulation underscores the need to develop a well understood and easily manipulated effect size for likelihood based tests in multilevel models. In this dissertation I manipulated effect size by altering the magnitude of parameters. While this approach served its purpose of creating differences between models, it was not entirely clear as to how the magnitude of each effect, especially when non-nestedness occurred in Level 1 covariance structures and non-linear models, was to manifest in model fit or differences in model fit. Creating a well understood and easily manipulated measure of likelihood difference would help to advance the understanding of this, and other, likelihood based model fitting approaches.

The effects of ICC were mixed between models that included non-nestedness at Level 1 or at Level 2. When non-nestedness occurred at Level 1 (e.g., Level 1 covariate sets, Level 1 residual covariance structures) power tended to decrease when ICC was large. Conversely, when non-nestedness occurred at Level 2 (i.e., Level 2 covariate sets) power was greater at the larger ICC. In non-linear growth models, power was greater in the large ICC condition than in the small ICC condition. This result was surprising because increased variability at Level 1, and more unique information, should provide more power. Although this test was statistically significant, there was no qualitative difference in power between ICC conditions.

The applications explored in Chapter 6 illustrate how Vuong's test can be used to test non-nested models in real data. Importantly, these examples highlight the difficulties that arise when using the test to guide decision making and the ambiguity that remains from a null result. As was seen when Vuong's test was used to compare different functional forms, even when results are significant, conclusions are not necessarily straightforward. Through these examples, I attempted to illustrate how a researcher would proceed with their study given the results of their model comparisons. Taken together the results of these examples illustrate that above all, it is paramount that model building be an iterative process that continues to acknowledge alternative conclusions.

In addition to the scenarios discussed in this dissertation, there are many other areas of research which might benefit from Vuong's test. Cross-classified models are one instance in which adopting Vuong's test might prove especially useful. On occasion, researchers are tasked with deciding between a cross-classified and a three-level model. To facilitate this comparison, Vuong's test would need to be adopted to compare data with unequal numbers of cases. That is, the number of cases in a three-level model would not be equivalent to the number of cases in a

cross-classified model due to differences in the structures data are representing. A "case" in the cross-classified sense would be each unique combination of upper level nesting units. For instance, if cross classification existed for schools and neighborhoods cases would refer to each unique pairing of schools and neighborhoods. This scenario poses both a philosophical and a computational issue for Vuong's test. Namely, would Vuong's test be *valid* when comparing two different nesting structures, and by extension different cases, and can it be adapted to handle such a structure? Understanding Vuong's test in this context would have more general implications for comparing non-nested random effects given that their differences would represent different cases.

Another potential application for Vuong's test involves alternative growth models. Curvilinear forms and non-linear models are not the only two methods by which to measure growth. For instance, continuous and discontinuous piecewise models are often used to measure growth as well. As mentioned in the discussion in Chapter 3, a piecewise model compared with a quadratic model would be a simple case of non-nested fixed effects. Comparing a quadratic model with a piecewise model would require little extra effort computationally and require a simple modification to the **X** matrix, assuming the random effects structure remained the same.

Finally, while this dissertation focused exclusively on growth models, it was shown in Chapter 2 that Vuong's test should apply in the same way to data for individuals nested within groups. It would be instructive for a future study to explore smaller Level 2 sample sizes, larger Level 1 sample sizes, and lower ICCs to expand the understanding of Vuong's test to a wider variety of research applications and capitalize on the elegance of multilevel models when individuals are nested within groups.

Over the course of this dissertation I have provided a method by which researchers can test the difference between non-nested multilevel models. I have shown that Vuong's test is an improvement over information criteria in that it rarely, if ever, suggests that the wrong model fits data best. While Vuong's test constitutes an advancement in how researchers can test certain hypotheses, it remains only one piece of information in a broader context of evaluation. Rarely are model comparisons so simple as to be decided by a single result and instead the results of Vuong's test should be considered in a broader context merging theory, other tests of significance, and the ultimate goals of the study to reach a conclusion.

Footnotes

[1]Note: "*I*" is used to represent "information" and was chosen here for consistency with various discussions on the topic. While K-L information and K-L divergence (or discrepancy) are synonymous I employ the latter term throughout this dissertation because divergence or discrepancy apply naturally to the concept of model selection. I prefer the use of *I* as opposed to the logical *D* as *D* commonly refers to deviance when discussing model selection.

Appendix A

SAS Macro for Vuong's test of Fixed Effects

```
proc IML;
start ML_Vuong(data,cluster,dv,m1_fixed,m1_random, m1_rep, m1_rtype,
m2_fixed, m2_random, m2_rep, m2_rtype,
                        ddfm,m1_type,m2_type,method,datname, tnum);

/*************************************************************/
/* This section of code to parse the fixed and random        */
/* effects was adopted from                                  */
/* the code provided by Stephen A. Mistler in his            */
/* MMI IMPUTE and MMI ANALYZE macro                          */
/* Mistler, S. A. (2013) A SAS macro for computing pooled    */
/* likelihood ratio tests with multiply imputed data.        */
/* SAS Global Forum 2013.                                    */
/* Much of the rest of this program was written using        */
/* Mistler's programs as examples.                           */
/*************************************************************/


      pi = arcos(-1);
      method = upcase(method);

      m1_f = m1_fixed;              * Parsing the fixed effects of model 1;
      m1_f_n = countn(m1_f);        * Counting the fixed effects of model 1;

      m1_r = m1_random;             * Parsing the random effects of model 1;
      m1_r_n = countn(m1_r) +1;     * Counting the random effects for model
1. One is added for intercept;

      m2_f = m2_fixed;              * Parsing the fixed effects for model 2;
      m2_f_n = countn(m2_f);        * Counting fixed effects for model 2;

      m2_r = m2_random;             * Parsing the random effects for model 2;
      m2_r_n = countn(m2_r) +1;     * Counting the random effects for model
2. One is added for intercept;

* Submit statement to call proc;
*This line is used to initialize macro variables in submit;

submit data cluster dv m1_fixed m1_random m1_rep m1_rtype ddfm m1_type
m2_fixed m2_random m2_type m2_rep m2_rtype method;

*model 1 for two-Level data;

proc mixed data = &data method = &method noclprint covtest;
class &cluster;
model &dv = &m1_fixed / s ddfm = &ddfm notest;
random int &m1_random /g sub = &cluster type= &m1_type;
repeated &m1_rep /r sub = &cluster type = &m1_rtype;
ods output
R                       = rmat_m1
G                       = psi_m1
covparms          = sigma_m1
```

```
solutionF          = beta_m1
iterhistory        = iter_m1
modelinfo          = model_m1
FITSTATISTICS      = fit_m1;
run;

proc mixed data = &data method = &method noclprint covtest;
class &cluster;
model &dv = &m2_fixed/ s ddfm = &ddfm notest;
random int &m2_random/g sub = &cluster type= &m2_type;
repeated &m2_rep /r sub = &cluster type = &m2_rtype;
ods output
R                        = rmat_m2
G                        = psi_m2
covparms           = sigma_m2
solutionF          = beta_m2
iterhistory        = iter_m2
modelinfo          = model_m2
FITSTATISTICS      = fit_m2;
run;

ods select all;
endsubmit;

*Read in model fitting information and parameters;
use model_m1;
read all var {Value} where (descr = "Degrees of Freedom Method") into ddfm;
read all var {Value} where (descr = "Covariance Structures") into type;
close model_m1;

use sigma_m1 where (covparm = 'Residual');
read all var {estimate} into sigma_m1;
close sigma_m1;

sigma_m1 = sigma_m1[:,];
_names_ = {Residual};
mattrib sigma_m1 colname = _names_;

use sigma_m2 where (covparm = 'Residual');
read all var {estimate} into sigma_m2;
close sigma_m2;

sigma_m2 = sigma_m2[:,];
_names_ = {Residual};
mattrib sigma_m2 colname = _names_;

if countn(cluster) ^= 0 then do;

      *Read in each models L2 covariance matrix;

      use psi_m1;
      read all var _num_ into psi_m1;
      close psi_m1;


      * Parse variance covariance parameters;
      n_psi_m1= ncol(psi_m1);
```

```
      psi_m1 = psi_m1[,3:n_psi_m1];

      use psi_m2;
      read all var _num_ into psi_m2;
      close psi_m2;

      n_psi_m2 = ncol(psi_m2);
      psi_m2 = psi_m2[,3:n_psi_m2];
end;

      *Read in model parameters;
      use beta_m1;
      read all var {effect} into beta_m1_names;
      read all var {estimate} into beta_m1;
      beta_m1 = beta_m1`;
      mattrib beta_m1 colname = beta_m1_names;
      close beta_m1;


      use beta_m2;
      read all var {effect} into beta_m2_names;
      read all var {estimate} into beta_m2;
      beta_m2 = beta_m2`;
      mattrib beta_m2 colname = beta_m2_names;
      close beta_m2;


* Read in data for calculations;
use (data); *use data;
read all var dv into y;             *read in DV for residual calculation;
read all var m1_f into x_m1;        *Read in values of predictors;
if m2_f_n ^=0 then read all var m2_f into x_m2; *If model 2 has any fixed
effects read them in;
if m1_r_n ^=1 then read all var m1_r into z_m1; *If model 1 has any random
effects other than an intercept read their values;
if m2_r_n ^=1 then read all var m2_r into z_m2; *If model 2 has any random
effects other than an intercept read their values;
if countn(cluster) > 0 then read all var cluster into id;
close (data); *close data;


n = nrow(x_m1);                     *Count rows of design matrix for model 1;

*Add intercept column to design matrix;
if m1_f_n > 0 then x_m1 = j(n,1,1) || x_m1;
else x_m1 = j(n,1,1);
if m2_f_n > 0 then x_m2 = j(n,1,1) || x_m2;
else x_m2 = j(n,1,1);

if countn(cluster) ^=0 then
do;
*Create random effects design matrices;

      id_l1 = unique(id[,1])`;
      m = nrow(id_l1);

      if m1_r_n ^= 1 then z_m1 = j(n,1,1) || z_m1;
```

```
      else z_m1 = j(n,1,1);

      if m2_r_n ^= 1 then z_m2 = j(n,1,1) || z_m2;
      else z_m2 = j(n,1,1);
end;

if countn(cluster) ^=0 then
do;

*Name columns of random effects design matricies;

      mattrib cluster colname = cluster;
      mattrib id_l1 colname = cluster;

      _z_m1_ = {Intercept} || m1_r;
      mattrib z_m1 colname = _z_m1_;

      _z_m2_ = {Intercept} || m2_r;
      mattrib z_m2 colname = _z_m2_;
end;

*     Name Fixed effects design matrices;
_x_m1_ = {Intercept} || m1_f;
mattrib x_m1 colname = _x_m1_;

_x_m2_ = {Intercept} || m2_f;
mattrib x_m2 colname = _x_m2_;


/*************************************************************/
/* Calculating the Likelihood for each cluster               */
/* for model 1 and model 2. Currently, this is only          */
/* supported for two-Level data. Support for n               */
/* levels of clusters will be added in a future              */
/* version.                                                  */
/* This section of code to extract the individual            */
/* Specific log likelihoods was based largely on             */
/* the code provided by Stephen A. Mistler in his            */
/* MMI IMPUTE and MMI ANALYZE macro                          */
/* Mistler, S. A. (2013) A SAS macro for computing pooled    */
/* likelihood ratio tests with multiply imputed data.        */
/* SAS Global Forum 2013.                                    */
/*************************************************************/



use rmat_m1;
read all into rmat_m1;
close rmat_m1;

rmat_m1 = rmat_m1[,3:ncol(rmat_m1)];

use rmat_m2;
read all into rmat_m2;
close rmat_m2;

rmat_m2 = rmat_m2[,3:ncol(rmat_m2)];
```

```
*ML estimation;
if method = 'ML' then do;


      *Multilevel Data;
      if countn(cluster) ^= 0 then
      do;
      ind_ll_m1 = j(m,1,0);
            do j = 1 to m;
                  ***Model 1***;
                  v_log_det_m1 = 0; *Reset log det v to 0;
                  rvr_m1 = 0;       *reset rvr to 0;
                  temp_x_m1 = x_m1[loc(id[,1] = id_l1[j,1]),];
            * create fixed eff design matrix;
                  temp_y = y[loc(id[,1]=id_l1[j,1]),];
                  temp_z_m1 = z_m1[loc(id[,1]=id_l1[j,1]),];
                  * create random eff design matrix;
                  temp_n_m1 = nrow(temp_z_m1);
                  * count how big the cluster is;
                  temp_v_m1 = temp_z_m1 * psi_m1 * temp_z_m1` + rmat_m1;
            *creating total variance matrix V;
                  temp_r_m1 = temp_y - temp_x_m1 * beta_m1`;
                  v_log_det_m1 = v_log_det_m1 + log(det(temp_v_m1));
                  rvr_m1 = rvr_m1 + temp_r_m1` * inv(temp_v_m1) * temp_r_m1;
                  ind_ll_m1[j,1] = (-1/2) * v_log_det_m1 - (1/2) * rvr_m1 -
(temp_n_m1/2) *log(2*pi);

end;
end;
***Model 2***;
      ind_ll_m2 = j(m, 1, 0);
            do j = 1 to m;
                  v_log_det_m2 = 0;
                  rvr_m2 = 0;
                  temp_x_m2 = x_m2[loc(id[,1]=id_l1[j,1]),];
                  * create fixed eff design matrix;
                  temp_y = y[loc(id[,1]=id_l1[j,1]),];
                  * create Y vector;
                  temp_z_m2 = z_m2[loc(id[,1]=id_l1[j,1]),];
                  * create fixed eff design matrix;
                  temp_n_m2 = nrow(temp_z_m2);
                  temp_v_m2 = temp_z_m2 * psi_m2 * temp_z_m2` + rmat_m2;
            *creating total variance matrix V;
                  temp_r_m2 = temp_y - temp_x_m2 * beta_m2`;
                  v_log_det_m2 = v_log_det_m2 + log(det(temp_v_m2));
                  rvr_m2 = rvr_m2 + temp_r_m2` * inv(temp_v_m2) * temp_r_m2;
                  ind_ll_m2[j,1] = (-1/2) * v_log_det_m2 - (1/2) * rvr_m2 -
(temp_n_m2/2) *log(2*pi);

end;
end;


use sigma_m1;
read all var {Estimate} into sig_m1;
```

```
close sigma_m1;


use sigma_m2;
read all var {Estimate} into sig_m2;
close sigma_m2;

sig_m1 = t(sig_m1);
sig_m2 = t(sig_m2);

***Computing Omega squared;

omega2 = (m-1)/m * var(ind_ll_m2- ind_ll_m1); * taken from nonnest2 package;

lr = sum(ind_ll_m1 - ind_ll_m2);

Vuong_LR = (1/sqrt(m))*(lr/sqrt(omega2));
tot_parm_m1 = m1_f_n + ncol(sig_m1) ; *Number of fixed effects +num random
effects;
tot_parm_m2 = m2_f_n + ncol(sig_m2) ;
*adjustments to test statistics. Information Criteria;
V_AIC = Vuong_LR - (tot_parm_m1 - tot_parm_m2); *Difference in Length of
Coefficients;
V_BIC_diffAB =  Vuong_LR - (tot_parm_m1 - tot_parm_m2) * log(m)/2;

pLRTA = 1 - cdf('NORMAL', Vuong_LR,0,1);
pLRTB = cdf('NORMAL', Vuong_LR,0,1);


use fit_m1;
      read all var {Value} into value_m1;
      value_m1 = t(value_m1);
      close fit_m1;


      use fit_m2;
      read all var {Value} into value_m2;
      value_m2 = t(value_m2);
      close fit_m2;

testdat = nrow(ind_ll_m1)|| nrow(temp_x_m1)|| lr || Vuong_LR || omega2 ||
pLRTA || pLRTB || V_AIC || V_BIC_DIFFAB || value_m1 || value_m2;

create (datname) from testdat[colname ={"n" "m" "lr" "Vuong_LR" "omega2"
"pLRTA" "pLRTB" "V_AIC" "V_BIC" "-2LL_m1" "AIC_m1" "AICC_m1" "BIC_m1" "-
2LL_m2" "AIC_m2" "AICC_m2" "BIC_m2"}];
append from testdat;
close (datname);


finish ML_vuong;
store module=(ML_vuong);

quit;
```

```
/*************************************************************/
/* Example Test Code with data generation program            */
/* 09/07/16 Currently works for nonnested fixed effects      */
/* 11/07/16 Changed Data generation for growth models in ch2*/
/*************************************************************/
/*
data a;
do i = 1 to 200;
      b0j = 10 + rannor(0)*2;
      b1j = 3 + rannor(0);
      w1 = ranbin(0,1,.5);
      w2 = ranbin(0,1,.5);
      w3 = ranbin(0,1,.5);
      do j = 1 to 25;
            eij = rannor(0) *1.73205080756;
            yij = b0j + b1j*j + 1.6*w1 + 1*w2 + .4*w3 + eij;
            output;
      end;
end;
run;

      proc iml;
      load module=(ML_vuong);

      data                    = "work.a";
      cluster                 = "i";
      dv                      = "yij";
      m1_fixed          = {j};            *no quotes necessary;
      m1_random         = {j};            *only quotes for an empty model;
      m1_rep                        = {""};
      m1_rtype                = {ar(1)};
      m2_fixed                = {j};
      m2_random               = {"j"};
      m2_rep                        = {""};
      m2_rtype                = {toep(2)};
      ddfm                    = "SATTERTH";
      m1_type                 = "UN";
      m2_type                       = "UN";
      method                  = "ML";
      run ml_vuong(data,cluster,dv,m1_fixed,m1_random, m1_rep, m1_rtype,
m2_fixed, m2_random, m2_rep, m2_rtype,
                        ddfm,m1_type,m2_type,method);
      quit;
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*(2), 135-167. doi:10.3102/10769986028002135

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods & Research, 32*(3), 336-383. doi:10.1177/0049124103260222

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective* (Vol. 467): John Wiley & Sons.

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205-229. doi:10.1177/0049124192021002004

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370. doi:10.1007/BF02294361

Burnham, K. P., & Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information Theoretic Approach*: Springer, Berlin, Heidelberg, New York.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioural Sciences* (Rev ed.). New York: Academic Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multple Regression/Correlation Analysis for the Behavioral Sciences* (Third ed.): Routledge.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58*(1), 615-637. doi:10.1146/annurev.psych.58.110405.085520

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*(4), 529-569. doi:10.1207/s15327906mbr3804_5

Delattre, M., Lavielle, M., & Poursat, M.-A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics, 8*(1), 456-475. doi:10.1214/14-EJS890

Dimova, R. B., Markatou, M., & Talal, A. H. (2011). Information methods for model selection in linear mixed effects models with application to HCV data. *Computational Statistics & Data Analysis, 55*(9), 2677-2697. doi:10.1016/j.csda.2010.10.031

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*(3), 379-403. doi:10.1207/S15327906MBR3703_4

Gagné, P., & Dayton, C. M. (2002). Best regression model using information criteria. *Journal of Modern Applied Statistical Methods, 1*(2), 57. doi:10.22237/jmasm/1036110180

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology, 44*(1), 153-170. doi:10.22237/jmasm/1036110180

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163-171). New York: Routledge.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika, 97*(4), 773-789. doi:10.1093/biomet/asq042

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development, 82*(5), 1357-1371. doi:10.1111/j.1467-8624.2011.01630.x

Hann, D. M., Jacobsen, P. B., Azzarello, L. M., Martin, S. C., Curran, S. L., Fields, K. K., . . . Lyman, G. (1998). Measurement of fatigue in cancer patients: Development and validation of the Fatigue Symptom Inventory. *Quality of Life Research, 7*(4), 301-310. doi:10.1023/A:1024929829627

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 190-195.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine, 21*(16), 2409-2419. doi:10.1002/sim.1047

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (Second Edition ed.): Routledge.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307. doi:10.2307/2336663

Huttenlocher, J., Haight, W., Bryk, A. S., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236 - 248. doi:10.1037/0012-1649.27.2.236

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in medicine, 30*(25), 3050-3056. doi:10.1002/sim.4323

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. (1998). A comparison of to approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics - Simulation and Computation, 27*(3), 591-604. doi:10.1080/03610919808813497

Kitagawa, G., & Konishi, S. (2010). Bias and variance reduction techniques for bootstrap information criteria. *Annals of the Institute of Statistical Mathematics, 62*(1), 209-234. doi:10.1007/s10463-009-0237-1

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Kullback, S. (1959). *Statistics and Information Theory*: J. Wiley and Sons, New York.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics, 22*(1), 79-86.

Kwok, O.-m., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research, 42*(3), 557-592. doi:10.1080/00273170701540537

Levy, R., & Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research, 42*(1), 33-66. doi:10.1080/00273170701329112

Levy, R., & Hancock, G. R. (2011). An extended model comparison framework for covariance and mean structure models, accommodating multiple groups and latent mixtures. *Sociological Methods & Research, 40*(2), 256-278. doi:10.1177/0049124111404819

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 673-687. doi:10.2307/2532087

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

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

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259. doi:10.1037/1082-989X.10.3.259

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1), 23. doi:10.1037/1082-989X.5.1.23

Merkle, E. C., & You, D. (2014). Package 'nonnest2'.

Merkle, E. C., You, D., & Preacher, K. J. (2015). Testing nonnested structural equation models. *Psychological Methods*. doi:10.1037/met0000038

Millsap, R. E. (2010). *A simulation paradigm for evaluating "approximate fit" in latent variable modeling.* Paper presented at the "Current topics in the Theory and Application of Latent Variable Models: A Conference Honoring the Scientific Contributions of Michael W. Browne", Ohio State University, Columbus, OH.

Mistler, S. A. (2013). *A SAS macro for computing pooled likelihood ratio tests with multiply imputed data.* Retrieved from

Moreno, P. I., Moskowitz, A. L., Ganz, P. A., & Bower, J. E. (2016). Positive affect and inflammatory activity in breast cancer survivors: Examining the role of affective arousal. *Psychosomatic medicine, 78*(5), 532-541. doi:10.1097/PSY.0000000000000300

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28*(2), 135-167. doi:10.1214/12-STS410

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). Taking into account sampling variability of model selection indices: A parametric bootstrap approach. *Multivariate Behavioral Research, 48,* 168-169. doi:10.1080/00273171.2013.752266

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17*(1), 1-14. doi:10.1037/a0026804

Pu, W., & Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of multivariate analysis, 97*(3), 733-758. doi:10.1016/j.jmva.2005.05.009

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology, 25*, 111-164. doi:10.2307/271063

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Second ed.): Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. doi:10.1037/1082-989X.5.2.199

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*(4), 387-401. doi:10.1037/1082-989X.6.4.387

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34*(2), 199-244. doi:10.1207/S15327906Mb340204

Rivers, D., & Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal, 5*(1), 1-39. doi:10.1111/1368-423X.t01-1-00071

. SAS Institute. SAS OnlineDoc 9.1.3. (2002-2005). Cary, N. C.: SAS Institute Inc.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347-367. doi:10.1177/1094428107308906

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Shang, J., & Cavanaugh, J. E. (2008a). An assumption for the development of bootstrap variants of the Akaike information criterion in mixed models. *Statistics & Probability Letters, 78*(12), 1422-1429. doi:10.1016/j.spl.2007.12.015

Shang, J., & Cavanaugh, J. E. (2008b). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics & Data Analysis, 52*(4), 2004-2021. doi:10.1016/j.csda.2007.06.019

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica,* 375-394.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling,* (Second ed.). London Sage.

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods, 17*(4), 582-599. doi: 10.1037/a0029253

Timmons, A. C., & Preacher, K. J. (2015). The importance of temporal design: how do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research? *Multivariate Behavioral Research, 50*(1), 41-55. doi:10.1080/00273171.2014.961056

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(1), 10-21. doi:10.1027/1016-9040.12.1.10

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior Research Methods, 43*(1), 18-36. doi:10.3758/s13428-010-0040-1

Vallejo, G., Tuero-Herrero, E., Núñez, J. C., & Rosário, P. (2014). Performance evaluation of recent information criteria for selecting multilevel models in behavioral and social sciences. *International Journal of Clinical and Health Psychology, 14*(1), 48-57. doi:10.1016/S1697-2600(14)70036-5

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica, 57*(2), 307-333. doi:10.2307/1912557

Wang, J., & Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in statistics-Simulation and computation, 38*(4), 788-801. doi:10.1080/03610910802645362

Weakliem, D. L. (2016). *Hypothesis Testing and Model Selection in the Social Sciences*. New York: Gulliford Press.

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., . . . Research Network on Youth Mental, H. (2012). Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: a randomized effectiveness trial. *Archives of General Psychiatry, 69*(3), 274-282. doi:10.1001/archgenpsychiatry.2011.147