Title
Nonlinear Multilevel Model Selection Using Information Criteria

Author
Christensen, Wendy

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Nonlinear Multilevel Model Selection Using Information Criteria

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Wendy Christensen

2019

ABSTRACT OF THE DISSERTATION

Nonlinear Multilevel Model Selection Using Information Criteria

by

Wendy Christensen

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2019

Professor Jennifer Lynn Krull, Chair

Multilevel modeling is a common approach to modeling longitudinal change in behavioral sciences. While many researchers use linear functional forms to model change across time, researchers sometimes anticipate nonlinear change. In such cases, researchers often fit polynomial functional forms, such as quadratic or cubic forms. Polynomial functional forms are suitable in many situations, but there are other functional forms that could potentially better match the researcher's theory about the nature of the longitudinal change. "Truly" nonlinear models, such as exponential and logistic models, have been used to model biological phenomena and may also be useful for psychological research. Such models, however, are non-nested, meaning that likelihood ratio tests cannot be used to select among models if one or more truly nonlinear models are in the candidate model set. Information criteria offer a flexible framework for model selection that can accommodate truly nonlinear models, but there currently is no research directly exploring the ability of information criteria to select truly nonlinear multilevel

models. In this dissertation, two Monte Carlo simulation studies were conducted to examine the performance of two frequently used information criteria: AIC and BIC. The goal of the first study was to examine their ability to select unconditional models with correctly specified nonlinear functional forms. Higher L1 and L2 sample sizes, a higher ICC, and greater distinction between nonlinear functional forms generally improved correct model selection rates, but BIC appeared to be better than AIC when identifying more distinct nonlinear functional forms and AIC appeared to be better when the forms were less distinct. The goal of the second study was to examine the ability of AIC and BIC to select a model with a "more correct" predictor set when the underlying functional form was truly nonlinear. In many cases, information criteria were able to identify models determined to be more correct, but no clear pattern emerged between AIC and BIC. Finally, the utility of truly nonlinear functional forms was demonstrated in two behavioral health applications, both of which contained substantively interesting nonlinear trends that would have been missed if analysis had been limited to the linear functional form.

The dissertation of Wendy Christensen is approved.

Craig Kyle Enders

Andrew J. Fuligni

Ayako Janet Tomiyama

Jennifer Lynn Krull, Committee Chair

University of California, Los Angeles

2019

DEDICATION

I dedicate this dissertation to my loving parents and my loving partner – without you, none of this would have been possible.

I also would like to express my deepest gratitude to each of my committee members for their invaluable contributions of knowledge, experience, and generous feedback throughout the dissertation process.

Finally, many thanks to the faculty, staff, and graduate students of the UCLA Psychology Department for fostering a research and teaching environment abounding with collegiality and intellectual rigor.

TABLE OF CONTENTS

1. Introduction 1

2. Study 1: Model Selection using Information Criteria to Identify Nonlinear Functional Forms 20

3. Study 2: Model Selection to Identify Non-Nested Predictor Sets in Nonlinear Models 38

4. Interlude: Fitting Covariance Structures to Truly Nonlinear Multilevel Models 64

5. Application: Discrepancy in Body Image Across Childhood and Adolescence 73

6. Application: The Effects of Sleep, Interpersonal Interactions, and Demands on Daily Distress in Adolescents 91

7. Discussion 105

8. Tables 108

9. Figures 132

10. References 137

VITA

EDUCATION

M.A. in Psychology — Dec 2015
University of California Los Angeles, Los Angeles, CA
Thesis title: The effect of group sizes, number of individuals, and true group-level covariance structure type on estimates and fit statistics of dynamic group models

M.S. in Applied Statistics and Research Methods — May 2013
University of Northern Colorado, Greeley, CO

B.S. in Psychology — May 2010
Colorado State University, Fort Collins, CO

PAPERS AND PRESENTATIONS

Christensen, W., & Krull, J. L (2019, July). *Using information criteria to select among polynomial and "truly" nonlinear multilevel models*. Poster presented at the Joint Statistical Meeting, Denver, CO.

Christensen, W. (2018). *Agreeing to disagree: Using SAS® to make reasoned decisions when information criteria select different models*. Paper presented at the Western Users of SAS Software conference, Sacramento, CA.

Christensen, W. (2018). *Model selection using information criteria (Made easy in SAS®)*. Paper presented at SAS Global Forum, Denver, CO.

Christensen, W., & Krull, J. L. (2015, May). *Determining necessary sample size for dynamic group models*. Poster presented at the 2015 Modern Modeling Methods conference, Storrs, CT.

Christensen, W., & Gibbons, A. M. (2010, May). *Development of a learning environment scale for postsecondary classrooms.* Poster presented at the 2010 Annual Meeting of the Association for Psychological Science, Boston, MA.

AWARDS AND SCHOLARSHIPS

Shepard Ivory Franz Distinguished Teaching Assistant Award — 2019

Western Users of SAS Software Academic Scholarship — 2018

SAS Student Ambassador — 2018

Graduate Student of the Year — 2013

1. Introduction

Nonlinear Multilevel Model Selection Using Information Criteria

It is a truth, universally acknowledged, that a researcher in possession of good data must be in want of a statistical model. Per the logic of the scientific method, the selection of a good statistical model should be made based on theory. If a researcher wants to test the plausibility of a single theory, the choice of statistical model is clear: the fitted model should be the model that best matches the theory. However, it is not uncommon for researchers to have multiple plausible theories that they want to test simultaneously. For example, a researcher may have competing theories, which could involve different predictor sets or even different conceptualizations of the nature of a change trend. In addition, complex statistical techniques require researchers to make complex model specification choices, some of which may be outside the conceptual space of the theory being tested. For example, multilevel models permit a researcher to specify a variety of covariance structures to model error structure at different levels, but it is often difficult to choose a structure solely on the basis of theory (Bauer, Gottfredson, Dean, & Zucker, 2013). For these reasons, statistical model selection methods are popular in quantitative behavioral research.

Model Selection for Nested Models

Null hypothesis statistical testing (NHST) is widespread in quantitative behavioral research, including in the context of model selection. This could be done through testing a series of nested models. Models are considered nested if a more specific model can be derived by removing parameters from a more general model (Hox, 2010). For example, a researcher could begin with a multiple regression model containing two predictors, which is shown below as Model 1. Model 2 contains both of the predictors in Model 1, along with two additional new predictors. In this case, the two-predictor Model 1 is said to be nested within the four-predictor Model 2 because

Model 1 could be fully specified solely by removing the third and fourth predictor in Model 2. One could create as many additional models as there are predictors available to be added, which could be added one at a time or as sets.

$$\text{Model 1: } Y_i = b_o + b_1X_{1i} + b_2X_{2i} + e_i$$

$$\text{Model 2: } Y_i = b_o + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + e_i$$

Models with interactions between predictors may also be thought of as being more general models compared to models that contain the lower-order terms only. Continuing the example, Model 3 contains the two predictors in Model 1, as well as an interaction between those two predictors. Because the interaction includes both lower-order predictors—that is, the interaction in Model 3 can be removed to fully express Model 1—Model 1 is nested within Model 3. Like before, one could specify an even more general model with additional predictors and higher-order interactions, but that model would need to include the two predictors and their interaction for Model 3 to be nested within that more general model.

$$\text{Model 3: } Y_i = b_o + b_1X_{1i} + b_2X_{2i} + b_3X_{1i} * X_{2i} + e_i$$

It is possible to use NHST to test the difference between any two nested models. In the context of model selection, the model with fewer predictors is referred to as the reduced model, and the model with more predictors is referred to as the full model. Nested model testing in multiple regression is conducted using R-square change tests, which involves an F test that uses the difference in the number of predictors between the full and reduced models as the numerator degrees of freedom (Cohen, Cohen, West, & Aiken, 2003). If this test is significant, the researcher would conclude that the full model accounted for a significantly greater proportion of the variance in the outcome than the reduced model. While not usually done for the purpose of model selection, a test of the significance of a single model's R-square is implicitly a model

selection procedure, as it is testing the fitted model against a mean-only model.

For multilevel models, a common method of model selection is the likelihood ratio test, sometimes also called the deviance change test or deviance difference test (Cohen et al, 2003; Hox, 2010). Although some details of the test are different, the likelihood ratio test is akin to the R-square change test. To conduct a likelihood ratio test, a researcher would begin by fitting a multilevel model with some number of parameters. These parameters could be things like fixed effects and random effects. The researcher would then fit a new model with one or more new parameters in addition to the parameters in the preceding model. The maximum likelihood estimation process for both the reduced and full models produces both a likelihood and a deviance. A model's deviance is a transformation of the model's likelihood; specifically, the deviance is computed by taking the natural log of the likelihood and multiplying that value by negative two. As long as the reduced model can be specified entirely by removing parameters from the full model, a chi-square test may be used to determine if the decrease in model deviance in the full model compared to the reduced model is statistically significant, with the difference in the number of parameters being the degrees of freedom for the test.

The likelihood ratio test is a common procedure in psychological research, having the benefits of being relatively intuitive and simple to conduct once the deviances have been computed using statistical software. There are, however, some disadvantages to this approach. Hamaker, van Huttum, Kuiper, and Hoijtink (2010) discussed three specific limitations of likelihood ratio tests. The first of these, and perhaps the most immediately noticeable of the three, is that likelihood ratio tests require models to be nested. Continuing the previous example, one could write a different four-predictor regression model, denoted Model 4, for which Model 1 would be considered a nested model.

$$\text{Model 4: } Y_i = b_o + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{5i} + e_i$$

While the differences between Models 1 and 3 and Models 1 and 4 could be tested with likelihood ratio tests, the difference between Models 3 and 4 could not be tested in this manner because they are not nested models relative to each other.

The second disadvantage discussed by Hamaker et al (2010) is that likelihood ratio tests only permit the testing two models at a time, meaning that a set of three or more models cannot be compared simultaneously. They note that this could cause results that seem to be in conflict depending on the choice of which two models to compare. Continuing the ongoing example to illustrate this limitation, if a researcher decided to conduct a likelihood ratio test between Models 1 and 4, which are nested, the researcher could potentially conclude that the difference between the two models is significant. However, this finding may not apply to the likelihood ratio tests of intermediate models; even if the difference between Models 1 and 4 is significant, there is no guarantee that difference between Models 1 and 2 and between Models 2 and 4 are significant as well. This is particularly problematic when a series of nested models is being tested incrementally and the researcher stops when the likelihood ratio tests cease being statistically significant – in this situation, the significant difference between Models 1 and 4 would be missed entirely if any of the intermediate comparisons have non-significant results. The third and final disadvantage discussed by Hamaker and colleagues (2010) is that the nature of NHST means that reduced models can only fail to be rejected; that is, the logic of hypothesis testing does not permit one to conclude that the reduced model is a better model even when one may have substantive interest in doing so. For these reasons, they suggest the use of information criteria to select among multilevel models when these limitations are undesirable.

Information Criteria

Information criteria developed for statistical model selection purposes have an intellectual root in information theory, which emerged from Shannon's (1948) groundbreaking work in signal processing and data compression. Akaike (1973/1992) developed the first information criterion, which is now called the Akaike Information Criterion (AIC). In particular, AIC is based on Kullback-Leibler (K-L) distance (Kullback, 1959), which connects information theory to random variable distributions.

The brief overview of K-L distance provided here closely follows the conceptual summary provided in Burnham and Anderson's (2002) widely-cited applied book on model selection. K-L distance is named as such because it refers to the distance between different models. However, this is not a simple distance; rather, it is a directed distance, or discrepancy, based on which model is acting as the reference point. For example, if one wished to compare model *f(x)* and model *g(x)*, one could compute the K-L distance between *f(x)* when *g(x)* is the reference model, as well as the K-L distance between *g(x)* when *f(x)* is the reference model. Because K-L distances are directed distances, or discrepancies, these two seemingly similar distances do not have to be equal. If *f(x)* and *g(x)* are identical models, both of these distances would be equal to zero. Otherwise, the K-L distance will always be greater than zero due to how it is computed, which is shown below.

$$\text{Kullback-Leibler Information: } I(f,g) = \int f(x) * \log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

The formula as shown represents the distance from *g(x)* to *f(x)*. Because of the information theory framework, this is understood as the information lost when *g(x)* is used as an approximation of *f(x)*. As this distance increases, so does the information lost, which would provide greater basis to conclude that *g(x)* is a poor approximation of *f(x)*. In the context of model selection, *f(x)* is the "true" model or "full reality" from which the data are drawn, and *g(x)*

is an approximating model specified to represent it. In practice, the underlying model *f(x)* is rarely known; in fact, Burnham and Anderson (2002) argued that the very notion of a "true model" is nonsensical outside of Monte Carlo simulation studies. Regardless of why *f(x)* is unknown, not knowing *f(x)* means that this distance cannot be computed directly. If, however, one were to compare any number of approximations of *f(x)*, these approximations would all have *f(x)* in common. In other words, *f(x)* is a constant across all possible approximating models. As such, while absolute distances between approximating models cannot be computed when *f(x)* is unknown, relative distances can be. This means that the relative distances can be used as a basis to compare different approximating models on the basis of ranking their relative distances.

One of Akaike's (1973) key insights was the relationship between K-L distance and maximum likelihood estimation, the result of which he called "an information criterion". The first half of the AIC computation includes the deviance produced by maximum likelihood estimation for a given model, a property shared with likelihood ratio tests. The second half incorporates the number of parameters, denoted *K*, in the model. The latter part is often thought of as a penalty for model complexity. It has this useful property, but the ability to interpret it as such is incidental; rather, the purpose of its inclusion is to account for the asymptotic bias in the likelihood (Bozdogan, 1987; Burnham & Anderson, 2002).

$$\text{AIC: } -2LL + 2K$$

To compare a set of candidate models, one would fit each candidate model to the data, compute AIC values for each model, and then compare all of the AIC values. The model with the smallest AIC value, or a subset of candidate models with the smallest AIC values if selecting more than one model is acceptable, would be selected. If two candidate models have the same number of parameters, then the model with the higher likelihood (and thus smaller deviance) will

have the smaller AIC value. In this case, the models are equally parsimonious because they have the same number of parameters being estimated, but the selected model has stronger explanatory power. On the other hand, if two candidate models have the same likelihood, then the model with fewer parameters will have the smallest AIC value. This is because the models have the same explanatory power, but the selected model is more parsimonious because it used fewer parameters to obtain the same likelihood. In practice, both the likelihood and the number of parameters tend to change across candidate models as they are fit to data, so the model with the smallest AIC value may be thought of as the model that best balances complexity and parsimony among the candidate models.

Since Akaike's (1973) initial contribution, numerous alternative information criteria have been developed based upon it or in reaction to it. Many statistical software packages used to fit multilevel models will produce one or more information criteria for a model, whether by default or upon request. Even if a statistical package does not compute any particular information criterion, most are easily computed by hand after obtaining the model's deviance from the output. Along with AIC, four additional information criteria will be discussed in this paper: AICC, BIC, CAIC, and HQIC. These specific information criteria are more likely to be used by applied researchers than other choices, such as the Deviance Information Criterion (Spiegelhalter, Best, Carlin, & van der Linde, 2002) or conditional-AIC (Vaida & Blanchard, 2005). Part of the reason for this may be because of which criteria are included in statistical software packages. Whittaker and Furlow (2009) chose to examine AIC, BIC, CAIC, and HQIC because those information criteria were available as part of the MIXED procedure in SAS 9.2, the most current version of SAS at the time. In SAS/STAT 14.3, which was the most current version of SAS at the time of this writing, at least 33 procedures provided at least one maximum

likelihood-based information criterion as part of their output, the most common of which were AIC and BIC (for more information about how this was determined, see Christensen, 2018). Regardless of the reason, the relative popularity of these particular information criteria makes them of interest.

AIC does not take sample size into account, and later research found that it can perform poorly when sample size is small relative to the number of parameters in the model; when this is the case, AIC is a biased estimator of the K-L distance, making it more likely to select over-parameterized models (Hurvich & Tsai, 1989). A corrected form of AIC, called AICC, was developed to adjust for this bias. AICC was first developed by Sugiura (1978) for regression models, and later generalized by Hurvich and Tsai (1989). The generalized form is the version that tends to be referenced in the multilevel model selection literature. Burnham and Anderson (2002) recommended that AICC be used instead of AIC when the ratio of the size of the sample and the number of parameters in the most highly parameterized model (n/K) is small, giving less than 40 as an example of a small ratio. If the sample size is adequately large relative to the number of parameters, AICC and AIC become very similar.

$$\text{AICC: } -2LL + 2K\left(\frac{n}{n-K-1}\right)$$

AIC and AICC are examples of information criteria that are *efficient*, which means that they will select the best model when the true model is of infinite dimension. The dimensionality of a model refers to the number of estimated parameters in the model. As such, a model of infinite dimension is a model with an infinite number of parameters, making it impossible for a researcher to specify a "true model". Burnham and Anderson (2002) viewed this property of efficient criteria as matching the reality of the scientific process for complex phenomena. Researchers formulate hypotheses and theories that are inherently simplifications of reality, and

the best of these are distillations containing only the most important facets of reality. To paraphrase Box's (1976) famous aphorism, all models a researcher could conceivably fit are wrong, but the best ones are useful.

Information criteria can also be *consistent*, which represents a different point of view about what should be optimized when selecting among models. When the true model actually is in the set of candidate models, a good selection criterion should select that true model, as opposed to selecting a good approximating model. The terminology comes from the idea of an information criterion being dimension consistent. It is possible for AIC to overestimate the dimension of a true model even in the asymptotic case; that is, as sample size approaches infinity, the probability of selecting a model that is overly complex is non-zero (Bozdogan, 1987). Because of this, AIC is considered dimension inconsistent. Consistent information criteria, however, will select the true model approaching 100% of the time when the model is of finite dimension and in the set of candidate models. Burnham and Anderson (2002) expressed skepticism about the widespread use of consistent criteria because they believed that these conditions are very strong assumptions that can only be met in Monte Carlo simulation studies or in the simplest of scientific contexts. Whether consistent criteria represent a reasonable approach to truth and the finding thereof or not, efficient and consistent information criteria are usually presented alongside each other in software and are often used jointly for interpretation in applied research.

The first of these consistent criteria to be developed was the Bayesian Information Criterion (BIC), sometimes called the Schwarz-Bayesian Criterion (Schwarz, 1978). As suggested by the name, BIC was derived using a Bayesian framework, which seeks the most probable model given the data. Specifically, each candidate model has the same prior probability, and posterior probabilities are computed based on the actual data collected.

$$\text{BIC: } -2LL + \ln(N)K$$

The next consistent criteria to be developed, HQIC (Hannan & Quinn, 1979) and CAIC (Bozdogan, 1987), were developed to reformulate AIC to have properties of a consistent criterion like BIC. The formulae for these are shown below. Similar to AICC, the difference in the computation between each of these and AIC is the penalty term, which is adjusted for sample size.

$$\text{CAIC: } -2LL + [\ln(n) + 1]K$$

$$\text{HQIC: } -2LL + 2Kln(\ln(n))$$

Multilevel Model Selection Using Information Criteria

The performance of different information criteria has been evaluated for different types of models that are of interest to psychological researchers, such as latent variable modeling (e.g., Vrieze, 2012) and item response theory (e.g. Whittaker, Chang, & Dodd, 2012). Because of the popularity of multilevel modeling in psychological research, quantitative researchers have examined the properties of information criteria in these models. As noted by Burnham and Anderson (2002), the methodology of Monte Carlo simulation studies implicitly requires the perspective that reality is knowable and truth can be fully specified. This is because the goal of such research is making inferences about the performance of models based on how well these models recover parameters from data that have been generated by the researcher to have particular properties of interest. The literature on model selection in multilevel models using information criteria appears to universally follow this practice; no simulation study in this literature engaged in model selection without the generating (true) model included in the set of candidate models being tested. In the presence of correctly-specified models, there is evidence that certain factors influence the performance of information criteria, three of which will be

discussed here: the effect of different kinds of "wrong-ness" (misspecification) in candidate models, the effect of choosing different likelihood estimation methods, and the effect of the unique complexity of sample size in multilevel modeling.

Nature of misspecification

Generally speaking, AIC tends to select overly-complex multilevel models compared to other criteria, which is in line with its dimension inconsistency (Bozdogan, 1987). This can be advantageous, however, as AIC tends to select the correct model more often than consistent information criteria when the true model is complex, such as when there are cross-level interactions (Whittaker & Furlow, 2009), correlated random effects (Vallejo, Tuero-Herrero, Nunez, & Rosario, 2014), or complex covariance structures (Vallejo, Fernandez, Livacic-Rojas, & Tuero-Herrero, 2011). When the true model does not have these kinds of complexities, consistent criteria tend to perform better than efficient criteria (Gurka, 2006; Whittaker & Furlow, 2009; Vallejo et al., 2011; Vallejo et al., 2014). Typically, simulation studies on this topic create these different kinds of misspecification through different sets of fixed and random effects. The performance of specific information criteria for selecting the correct set of fixed and random effect structures is highly related to the specific conditions of the simulation study, particularly fixed effect and random effect parameter magnitude, intraclass correlation, and covariance structure type and magnitude (Gurka, 2006; Wang & Schaalje, 2009; Whittaker & Furlow, 2009; Vallejo, Ato, & Valdés, 2008; Vallejo et al., 2014). In general, higher ICCs and greater fixed and random effect magnitudes (i.e., effect sizes) makes the selection of the correct set of parameters more likely across all of the information criteria, but this was not uniformly true across all conditions in all studies.

Estimation methods

All of the information criteria discussed in the previous section require an estimate of the model likelihood. The choice of likelihood estimation method does not change how the information criteria are computed, but it does determine what value is used as part of the computation and should be held constant across models when computing the information criteria for a set of candidate models. Full maximum likelihood (ML) and restricted maximum likelihood (REML) are two common choices, both of which are included in most statistical software packages. In practice, researchers tend to use whichever estimation method is the default in their statistical software package of choice, but both methods have specific properties that may make one more appropriate than the other in certain circumstances (Hox, 2010). ML includes both fixed and random effects in the likelihood function to be maximized, making it computationally less intensive and creating a logical basis for comparing models with different sets of fixed effects. REML estimation includes only the random effects in the likelihood function, with the fixed effects estimated in a second estimation step (Hox, 2010). REML produces less biased estimates of the likelihood than ML when sample size is small, and also provides less biased estimates of the random effects of the model (Hox, 2010). However, because of the removal of fixed effects during estimation, using likelihoods produced by REML estimation for model comparison may be more tenuous than using ML when the fixed effects are not the same across all of the candidate models (Hox, 2010; Verbeke & Molenberghs, 2000).

Gurka (2006) argued that the differences between ML and REML are not substantial enough to justify a blanket recommendation against using REML for model selection purposes and used simulation to demonstrate this assertion empirically. He compared the performance of four information criteria (AIC, AICC, BIC, and CAIC) with a set of candidate models, estimated using both REML and ML estimation, that had different fixed effects. Specifically, three

situations were tested: when the random effects of a model were known (i.e. correctly specified) and the fixed effects were unknown (i.e. potentially misspecified), when the fixed effects were known and the random effects were unknown, and when both the fixed and random effects were unknown. In most circumstances, the performance of the information criteria was comparable under both REML and ML estimation; in fact, REML-based information criteria performed better than their ML counterparts in some conditions. This finding has been replicated in later multilevel modeling-specific information criteria literature (Vallejo et al., 2011; Vallejo et al., 2014). In light of these findings, the use of REML to estimate the likelihood of a model for selection purposes appears to be a viable choice.

Sample size

All information criteria except AIC incorporate some direct adjustment to the penalty term based on the sample size. In single-level regression, this is always the total number of observations. Sample size is made more complex in the multilevel modeling framework because there are effectively two potential sources of sample size within the same data set (Hox, 2010). One potential source is the total number of observations, often denoted as *N* in the multilevel information criteria literature. If the data are balanced (i.e., the number of L1 units is the same across all L2 units), the total number of observations can be found by multiplying the number of L2 units by the number of L1 units. Another potential source is the number of L2 units only, usually denoted as *m*. The differences between these two different ways of conceptualizing sample size can potentially lead to dramatically different penalty terms, even if all else is equitable across computations. For example, if a balanced data set has 50 groups of 20 individuals, using *m* would give a sample size of 50, while using *N* would give a sample size of 1000. The potential magnitude of this difference could influence the relative performance of

information criteria that incorporate sample size because each one incorporates it differently (e.g. Keselman, Algina, Kowalchuk, & Wolfinger, 1998).

As with estimation methods, the decision about which sample size source to use is often made implicitly through the default settings in statistical software. In PROC MIXED, the default sample size source is *m* (number of L2 units). Previous research has shown that the exact effect of sample size source is influenced by other factors in the model (e.g., ICC, degrees of difference in covariance structures), but some trends are apparent. First, whether *N* or *m* is used for the computation of the information criteria, the performance of sample size-dependent information criteria improves when the sample size source used is larger than when it is smaller (Whittaker & Furlow, 2009; Vallejo et al., 2011). Second, sample size-dependent information criteria are differently affected by the choice of *N* or *m*, particularly depending on whether a given criterion is consistent or efficient. Some research suggests that consistent criteria may perform slightly better when *m* is used (Gurka, 2006; Vallejo et al., 2014; Wang & Schaalje, 2009), while other research suggests that consistent criteria perform better when *N* is used (Vallejo et al., 2011). Other research suggests that the difference in performance is akin to that of model misspecification, in that the effect is not necessarily uniform across circumstances and may be reliant on the specific properties of candidate models (Whittaker & Furlow, 2009; Vallejo et al., 2014).

Nonlinear Multilevel Models

Whether longitudinal or cross-sectional, the vast majority of multilevel models used in psychological literature are specified with linear trajectories in mind. For example, if a researcher uses multilevel modeling to model repeated measures data, equations can be written to represent the participants (Level 2) and the repeated measures taken from the participants at

different time points (Level 1). If the researcher believes that the outcome of interest will change steadily as time passes, then specifying a linear trend in the model is reasonable. The Level 1 equation for this type of change is shown below.

$$\text{Linear Multilevel Model: } Y_{ij} = \pi_{0i} + \pi_{1i}TIME + \varepsilon_{ij}$$

There are occasions, however, when the linear trend is not a sensible model for the phenomenon of interest or does not align with the theory the researcher wants to test. The flexibility of the multilevel framework allows for alternative specification methods to model trajectories that are not straight lines. Singer and Willett (2003) showcased three methods of modeling nonlinear change. The first method was transforming variables based on the "rule of the bulge" (Mosteller & Tukey, 1977). If successful, this method will produce transformations that linearize a nonlinear trend, allowing the researcher to use the standard linear multilevel model shown above. This method effectively treats nonlinearity as a nuisance to be corrected; however, sometimes the nonlinearity is itself of theoretical interest. Also, the rule of the bulge depends on the nonlinear trend resembling a monotonic sloping curve, which is not always the case.

The second method described by Singer and Willett (2003) was incorporating polynomial time terms into the model. Polynomial terms produce nonlinear trajectories by introducing powered vectors that create "bends" in the line. Each additional polynomial term added to the model changes the trajectory of the trend. Strictly speaking, all multilevel models are polynomial models, with a mean-only model being a zero order polynomial model and a linear model being a first order polynomial model. In practice, this phrasing is typically reserved for models that include second order or higher polynomial terms. The Level 1 equations associated with the quadratic (second order polynomial) and cubic (third order polynomial) models are shown

below. The quadratic model produces a trend with no inflection point (i.e., no change in concavity), with the sign of the quadratic term determining if the trend is convex or concave. The cubic model produces a trend with one inflection point, such that one part of the function is convex and the other part of the function is concave. The Level 1 equations for these models is as follows:

$$\text{Quadratic model: } Y_{ij} = \pi_{0i} + \pi_{1i}TIME + \pi_{1i}TIME^2 + \varepsilon_{ij}$$

$$\text{Cubic model: } Y_{ij} = \pi_{0i} + \pi_{1i}TIME + \pi_{1i}TIME^2 + \pi_{3i}TIME^3 + \varepsilon_{ij}.$$

Polynomial functions allow the nonlinear trend to be explicitly modeled, producing estimates for the coefficients associated with each polynomial term. For all polynomial models, the intercept is the expected value of Y_{ij} when TIME = 0. Singer and Willett (2003) note that the addition of polynomial terms means that the interpretations of the coefficients used in the standard linear multilevel do not apply when higher order polynomials are present in the model. In the case of the quadratic model, the linear term in it can no longer be interpreted in terms of continuous change. Instead, the linear term in this model becomes the instantaneous rate of change when TIME = 0. The quadratic term is the acceleration or curvature of the entire trend. Models that include cubic or higher trends have growth parameters that are difficult to interpret individually (Singer & Willett, 2003).

Pairs of polynomial models are always nested models, meaning that a researcher could use likelihood ratio tests for model selection, and can easily be specified in any statistical software that can fit linear multilevel models. This is because the nonlinearity in the polynomial models emerges from exponentiation of the predictors, not the coefficients of those predictors; that is, the model is linear in the parameters. Singer and Willett (2003) identified these models as being *dynamically consistent* and discussed two important properties of dynamically consistent models.

The first property is that models that are linear in the parameters will obtain equivalent results if either of two methods of computing an "average" trajectory are used. In the curve of averages method, one would estimate the average outcome for all individuals at each time point, then plot the result. In the average of curves method, one would estimate the growth parameters for each individual, then find the average of these growth parameters, plotting these. The second property is that the shape, or *functional form*, of the average trajectory will be the same as that of the individual trajectories used to create it.

Because polynomial models are linear in their parameters, Singer and Willett (2003) considered them to be not "truly" nonlinear, reserving that terminology for models that are nonlinear in their coefficients, the third and final method they described. The truly nonlinear models discussed included hyperbolic models, exponential models, and inverse polynomial models, but many choices are available. Because the nonlinearity in these models is expressed in the parameters, truly nonlinear models are more difficult to specify and not all statistical software packages support such specification. SAS supports this type of specification in the NLMIXED procedure, which can be used to fit many kinds of functional form for a multilevel model. A search of the currently-existing applied literature showed that truly multilevel models have greater representation in the research literature in biological sciences (Davidian & Giltinan, 2003), such as forestry (e.g., Hall & Bailey, 2001) and pharmacology (e.g., Karlsson & Sheiner, 1993) than in psychology. Because of the large number of choices available to researchers, only a few possible nonlinear models will be described here, with a focus on functional forms that are readily applicable to psychological concepts.

Exponential models are models that allow the modeling of "explosive" growth from a starting point. The Level 1 specification for the exponential model, which follows Singer and

Willett's (2003) notation, is as follows:

$$\text{Exponential model: } Y_{ij} = \pi_{0i} e^{\pi_{1i} TIME_{ij}} + \varepsilon_{ij} .$$

The first estimated parameter is the intercept of the model and is the expected value of the outcome at the initial time point. It retains this simple interpretation because e^0 is equal to 1. The second estimated parameter is contained in the exponential portion of the model. While not a slope in the sense of a linear model, it is similar to a slope in the sense that increases its value indicate more rapid increases in exponential growth. Regardless of relative size, exponential models will always start at some starting point and grow without leveling off; the variability between people occurs based on where their values are at the start of the study and the "explosiveness" of their exponential growth.

Logistic models are part of the exponential family of models (Singer & Willett, 2003) because their functional form includes *e*. The Level 1 equation for a four parameter logistic model, as shown in Singer and Willett (2003), is as follows:

$$\text{Logistic model: } Y_{ij} = \alpha_{1i} + \frac{(\alpha_{2i} - \alpha_{1i})}{\left(1 + \pi_{0i} e^{-\pi_{1i} TIME_{ij}}\right)} + \varepsilon_{ij}.$$

These models have additional parameters to model an S-shaped functional form. What distinguishes the logistic model from the exponential model is two additional parameters that represent two asymptotes. The first parameter in the model, α_{1i},is the lower asymptote, which acts as a "floor" for the trajectory. The quantity in the numerator of the fraction is the difference between the lower asymptote and the upper asymptote, α_{2i}, which acts as a "ceiling" for the trajectory. The quantity in the denominator generates the smooth S-shaped curve that connects the two asymptotes.

Unlike polynomial models, nested model tests like the likelihood ratio test cannot be used to compare truly nonlinear models and linear models. Because of this, information criteria are the

only readily available options for model selection for researchers who want to include truly nonlinear models in the set of candidate models. As discussed in the preceding section, attention has been paid to the performance of different information criteria when selecting among linear multilevel models. However, there is not yet literature that addresses how information criteria perform when comparing polynomial or truly nonlinear multilevel models.

The focus of this dissertation is two-fold. First, I use Monte Carlo simulation to empirically examine the performance of information criteria for selecting the proper functional form and selecting among non-nested predictor sets of models fit in PROC NLMIXED, along with a short discussion about fitting covariance structures using NLMIXED. Second, I apply these methods to two existing data sets to demonstrate their utility for using information criteria to compare competing theories about psychological phenomena, expressed in the form of nonlinear functional forms, non-nested predictor sets, and covariance structures.

2. Study 1: Model Selection using Information Criteria to Identify Nonlinear Functional Forms

When fitting multilevel models with linear functional forms, it is reasonable to begin by fitting an unconditional model – that is, a model that incorporates no predictors – because it serves as a natural baseline for further model adjustments (Hox, 2010). In the context of longitudinal modeling, Singer and Willett (2003) referred to models that only included a temporal variable at Level 1 as unconditional *growth* models. If a researcher has decided in advance which functional form to fit, then no model selection is needed to determine which unconditional growth model to fit. For example, if previous literature has established that the change trajectory of a particular phenomenon is quadratic in nature, then the researcher would start the model building process by fitting an unconditional quadratic growth model. If, however, the researcher has competing theories about the shape of the trajectory, then a decision about the functional form will need to be made before moving on to adding predictors or fitting covariance structures. Information criteria are a useful method of model selection among different functional forms when nonlinear functional forms are among the candidate models because information criteria do not require the models to be nested relative to each other.

The number and form of the temporal predictors used in a nonlinear unconditional growth model depends on the functional form. For example, unconditional linear growth models have two fixed effects: the intercept and the slope associated with the time variable. The unconditional growth model can have as many random effects as parameters that can be allowed to vary. The choice of which random effects to specify depend on the structure of the data (i.e., the nesting structure), the research question, and practical considerations such as sample size and whether a model can be estimated without convergence issues or errors. Because truly nonlinear models

can be difficult to estimate, the models that were fitted in this and the succeeding chapters had one functional form parameter specified as a random effect. The approach used in the simulation studies was to choose the parameter in the functional form that was most related to the intercept.

Method

Functional forms for data generation

Data were generated according to one of three functional forms: quadratic, exponential, and logistic. Each of the data generating models included the time-related predictors necessary to create the functional form as fixed effects, as well as a random intercept (b_{0i} for quadratic and exponential) or random intercept-related (b_{2i} for logistic) parameter. Originally, the logistic model was to be specified to match the four-parameter model discussed in Singer and Willett (2003), but this model either did not converge or produced estimates with errors in the larger of the two application studies (over 2000 individuals across 10 time points). To ensure that logistic models would be able to be estimated without error most of the time, the logistic functional form was re-parameterized into a three-parameter model by setting the lower asymptote to zero. Wolfinger (1999) and Zheng (2010), the former of whom adapted the same specification and example used by Pinheiro and Bates (1995), described fitting three-parameter logistic functional forms to applied agronomy examples using the NLMIXED procedure in SAS. The equations used for data generation were as follows:

$$\text{Quadratic model: } Y_{ij} = b_{0i} + b_1 TIME + b_2 TIME^2 + u_{0j} + \varepsilon_{ij}$$

$$\text{Exponential model: } Y_{ij} = (b_{0i} + u_{0j}) e^{b_1 TIME_{ij}} + \varepsilon_{ij}$$

$$\text{Logistic model: } Y_{ij} = \frac{(b_1)}{\left(1 + e^{\frac{-(TIME_{ij} - (b_{2i} + u_{2j}))}{b_3}}\right)} + \varepsilon_{ij}$$

Intraclass coefficients for data generation

The intraclass correlation coefficient (ICC) is a common index of within-group similarity,

which in the longitudinal case refers to the similarity of the repeated measures within individuals. Because of this, ICCs tend to be higher in repeated measures data than in cross-sectional data (Singer & Willett, 2003). In this study, the ICCs were 0.4 and 0.6. To obtain the desired ICCs, the ICC formula can be rearranged to solve for the necessary L2 intercept variance and L1 error variance. For this study, the L1 error variance was set to 1 and the L2 intercept variance was set to 0.67 (ICC = 0.4) and 1.5 (ICC = 0.6). Error terms were generated from a normal distribution with a mean of zero and a variance in accordance with the intended ICCs. All generating models had a homogeneous L1 error covariance structure.

Effect sizes of coefficients for data generation

There is no pre-existing method of determining the effect size for the difference between nonlinear functional forms. It is likely, however, that the differences in the shapes of the nonlinear functional forms contribute to the ability of information criteria to distinguish among such forms. In practice, when fitting nonlinear functional forms, truly nonlinear or otherwise, the range over which the shape of each trajectory can be influenced by the data is limited by the number of waves of data collection. The extent to which a given functional form can express the characteristics that are quintessential to its form within the observed range is related to the differences in the trajectories of the curves and could be thought of a type of effect size. Thus, effect size in this study was operationalized as the distinctiveness of each functional form within the range of the largest L1 sample size used (13 time points, numbered zero to twelve).

I began this process by iteratively graphing quadratic, exponential, and logistic functions. The values for one set of these functions were chosen such that the predicted values from those functions produced curves that were very similar to each other within the 13-time point range. The values for the other set of functions, which represented a higher effect size, were chosen

such that the distinctive features of each function were expressed within the 13 time points, enhancing the differences between the curves. The quadratic function is visually distinguished by its single optimum and parabolic shape, the exponential function by its explosive growth, and the logistic function by explosive growth from a lower asymptote that tapers as it approaches an upper limit. In both cases, the coefficients produced curves that increased monotonically within the observed range. The two sets of curves are shown in Figures 2.1 and 2.2, respectively, and the coefficient values used to make these curves is shown in Table 2.1. The curves representing a lower effect size have trajectories that are difficult to distinguish visually within the observed range. The curves representing a higher effect size have each functional form's distinct characteristics while maintain the overall trend of positive monotonic growth.

The next step of this process was to confirm that the distinctiveness of these two sets of coefficient values would be retained across ICCs and across misspecified functional forms. To do this, 12 single-replication simulations with a large number of L2 units were conducted. The data generated in each simulation had one of three underlying functional forms (quadratic, exponential, and logistic), used one of two sets of coefficients for that true functional form (lower- or higher-distinctiveness), and was generated to have one of two ICCs (0.4 or 0.6). Three models – quadratic, exponential, and logistic – were fitted to each of the generated data sets. In all cases, two of these models were misspecified and one was correctly specified. To increase the number of L2 units that could be used, the upper asymptote and the variance components of the logistic functional form were rescaled by taking the square root of each parameter (Kiernan, Tao, and Gibbs, 2009). This adjustment allowed the misspecified logistic models to successfully estimate when there were as many as 100 L2 units. The number of L1 units was held constant at 13.

Although the primary operationalization of effect size was the visual distinction among functional forms across the observed range, several numeric metrics were also computed to quantify the differences and similarities of the estimated functional form models. One of these was the mean of simple differences in the predicted values between functional form models. The differences were obtained by solving for the predicted value at each time point in the observed range for all three models (one correctly specified, the other two misspecified). These values were then compared across fitted functional forms in a pairwise fashion at each time point. For example, the predicted value when TIME = 1 of the estimated exponential models was subtracted from the predicted value of the estimated logistic models at the same time point. Finally, the mean of these across all time points was taken. Higher means implied greater distinction between forms, which can be seen in many of the higher-effect size curves in Figure 2.2.

Another of these metrics was an area under the curve metric. It was not possible to find definite integrals of all of the model equations, so the area under each of these fitted curves was approximated using the midpoint rule (Rogawski, 2008). These values are shown on Table 2.2, along with the differences in areas under the curves between estimated functional form models. When the underlying functional form was more distinct, the areas under the curves were higher than when the functional forms were less distinct, even across misspecified models. In addition, the differences between areas under the curves were greater when the functional forms were more distinct.

Sample sizes

Longitudinal multilevel models have repeated measurement nested within individuals, meaning that the number of participants determines the number of L2 units available. There is a

wide range in number of participants in longitudinal behavioral health research, with some large-scale panel studies having thousands of participants. In the interest of examining the feasibility for researchers who do not have panel data, possibly because they are testing interventions, the number of L2 units were 30, 50, and 100.

In longitudinal studies, the number of measurement occasions determines the amount of complexity in change over time that may be modeled, with a quadratic model requiring no fewer than 4 time points (Singer & Willet, 2003). For this study, the L1 sample sizes were 5, 7, 9, and 13. These L1 sample sizes reflect a baseline measure plus 4, 6, 8, or 12 additional measurements from the same individuals. This range of measurement occasions is representative of many longitudinal studies in behavioral health, where data could be collected with relatively little time between measurement occasions (e.g., from diaries or smartphone applications) or with as much as a year passing between measurement occasions (e.g., large-scale panel studies). Timmons and Preacher (2015) demonstrated that it is possible to fit nonlinear models within this range of L1 units, though increasing the number of time points improves the estimation of nonlinear model parameters.

In this study, the L2 sample size was directly manipulated – that is, data were specifically generated to have 30, 50, or 100 individuals. Each data set was generated to contain 13 time points by increments of 0.5 (0, 0.5, 1, 1.5, etc.). To create the four L1 sample sizes in such a way to maintain an equivalent range of X, the generated data set was "reduced" to the appropriate number of time points. The 13-time point condition was created by dropping all of the non-integer time values (0, 1, 2, etc.). The 9-time point condition was created by retaining each 1.5 increment (0, 1.5, 3, etc.). The 7-time point condition was created by retaining all even-integer time points (0, 2, 4, etc.). Finally, the 5-time point condition was created by retaining every third

integer (0, 3, 6, etc.). Although the spacing is not equal across L1 sample sizes, the spacing between measures within each L1 sample is was always equidistant.

Set of candidate models

In this study, five candidate models - linear, quadratic, cubic, exponential, and logistic - were fit to each generated data set and compared using information criteria. All of these were specified to have a random intercept or intercept-related parameter. For any given data set, one candidate model was correctly specified and the other four were misspecified. For all candidate models, a homogeneous L1 error covariance structure was fitted, so the covariance structure was always correctly specified.

Information criteria

AIC, AICC, BIC, CAIC, and HQIC were computed for all of the candidate models. AIC is not sample size-dependent, so there was only one calculation of AIC. In total, there were nine information criteria computed for each candidate model: AIC, AICC(N), AICC(m), BIC(N), BIC(m), CAIC(N), CAIC(m), HQIC(N), and HQIC(m).

Overview

For each generated data set, there were four candidate models, each of which had nine associated information criteria. The performance of the different information criteria was likely to be affected by the underlying functional form (3 levels), the effect size of those forms relative to each other (2 levels), ICC (2 levels), number of individuals sampled (3 levels), and number of measures from each individual (4 levels), making for a total of 144 conditions in this study. Replications were run until 1000 valid replications were collected for each condition. Data were generated and candidate models were fitted in SAS 9.4, with candidate models fitted using PROC NLMIXED with maximum likelihood estimation by adaptive Gauss-Hermite quadrature

(SAS Institute Inc, 2015).

Analyses and hypotheses

The primary outcome of interest in this study was the frequency with which each of the nine information criteria selected the candidate model with the functional form that matched the underlying functional form in the data when the candidate models were simultaneously compared. The difference in performance between efficient and consistent information criteria was also of interest, as well as the effect of sample size used in the computation of the sample size-dependent information criteria. Of particular interest were the most commonly used information criteria: AIC and BIC (Whittaker & Furlow, 2009).

Data were analyzed in the GLIMMIX procedure in SAS 9.4 (SAS Institute Inc, 2015) using maximum likelihood with Laplace approximation and between-within denominator degrees of freedom used for the F tests of the fixed effects. The outcome was whether or not the correctly specified model was selected. For each generating functional form, analyses began with fitting a logistic multilevel model with a five-way interaction between L1 sample size (reference group was 5), L2 sample size (reference group was 30), ICC (reference group was low), the distinctiveness of the underlying functional form (reference group was low-distinction), and information criteria (initial reference group was AIC). Replication ID was modeled as a random effect, reflecting the fact that all L1 time points within a replication shared the same generating form. If the five-way interaction was not significant, then a model with four-way interactions was tested. The same process was repeated until there was a significant interaction among the highest-order terms or there were no more interaction vectors to test. Because the focus of this study was information criteria, simple effects tests were conducted for a significant interaction only if the interaction involved information criteria. Finally, AIC, BIC(N), and BIC(m) were

tested against each other and the other information criteria by changing the reference category of the main/simple effect information criteria vectors.

Multilevel logistic models are computationally intensive and sometimes difficult to estimate. The most common error that occurred while fitting these models was the coefficients of one or more dummy vectors having an estimated standard error of zero, creating a test statistic equal to infinity. In other cases, the model did not produce any estimates because it did not converge. If any errors occurred in the five-way interaction model, a four-way interaction model was fitted, and so on until a model that estimated without errors was found. At that point, the analysis procedure described in the previous paragraph was conducted.

Based on the linear multilevel model selection literature, I expected the performance of all information criteria to improve as the number of individuals and the number of time points increased (Whittaker & Furlow, 2009; Vallejo et al., 2014). I also expected performance to improve across all information criteria when the ICC was higher (Gurka, 2006; Whittaker & Furlow, 2009; Vallejo et al., 2014) and when the underlying functional forms were more distinct. There is no literature specifically examining the performance of information criteria for nonlinear multilevel models, but I suspected that consistent information criteria would perform better than efficient criteria for selecting the best model. This is because consistent information criteria tend to perform better than efficient criteria for simple models (Gurka, 2006; Whittaker & Furlow, 2009; Vallejo et al., 2011; Vallejo et al., 2014), and the models being tested in this study all had relatively few parameters. Finally, because of the longitudinal nature of the change being examined, I suspected that the sample size-dependent information criteria would better select the underlying functional form when computed using the number of observed units (N) than when using the number of L2 units (m).

Results

Obtaining 1000 valid replications

Before any analyses could be conducted, 1000 valid replications needed to be obtained for each condition. In this case, "valid" meant that all five candidate models were able to be estimated (i.e., all models converged) and that the estimates produced were trustworthy (i.e., no error messages of any kind). Table 2.3 shows the number of replications ultimately needed to obtain 1000 valid replications for each of these 36 simulations. The lowest rate of successful replications was less than 1% and the highest rate was 94%.

Correct selection rate and most common alternatives

Exponential functional forms. The selection rates for the correctly specified exponential models overall and across levels of distinctiveness are shown in Table 2.4. Across conditions, the set of information criteria as a whole identified the correctly specified exponential model between 88% and 97% of the time. When the underlying exponential functional form was of lower distinction, information criteria identified the correctly specified model between 76% and 93% of the time. When the exponential functional form was of higher distinction, information criteria identified the correctly specified exponential model greater than 99% of the time. Among efficient information criteria, the most commonly selected misspecified models were the linear and logistic models. The misspecified models most commonly selected by consistent criteria were linear models.

Logistic functional forms. The selection rates for the correctly specified logistic models overall and across levels of distinctiveness are shown in Table 2.5. Across conditions, the set of information criteria as a whole identified the correctly specified logistic model about 50% of the time. When the underlying logistic functional form was of lower distinction, information criteria

identified the correctly specified model no more than 3% of the time. When the logistic functional form was of higher distinction, information criteria identified the correctly specified logistic model greater than 99% of the time. Overall and when the distinctiveness of the underlying logistic form was low, the most common model selected by both efficient and consistent information criteria was the linear model. When the distinctiveness of the functional form was high, quadratic models were the most common alternative.

Quadratic functional forms. The selection rates for the correctly specified quadratic models overall and across levels of distinctiveness are shown in Table 2.6. Across conditions, the set of information criteria as a whole identified the correctly specified quadratic model between 51% and 58% of the time. When the underlying quadratic functional form was of lower distinction, information criteria identified the correctly specified model between 5% and 30% of the time. When the quadratic functional form was of higher distinction, information criteria identified the correctly specified quadratic model between 84% and 99% of the time. Overall and when the distinctiveness of the underlying quadratic form was low, the most common model selected by both efficient and consistent information criteria was the linear model. When the distinctiveness of the functional form was high, cubic models were the most common alternative.

Inferential tests of correct model selection among information criteria

Correctly specified exponential models. The five-way, four-way, and three-way multilevel logistic models had coefficient estimates with standard errors equal to zero, but all of the coefficients estimated without issue for the two-way interaction model. Seven two-way interaction terms were significant (all $p < 0.001$), three of which included information criteria: L1 sample size, L2 sample size, and ICC. To examine the simple effects of information criteria, data were subset based on L1 sample size, L2 sample size, and ICC.

There was a significant simple effect of information criteria (all $p < 0.001$) across all L1 sample sizes. All other information criteria except AICC(N) were significantly better than AIC across all L1 sample sizes. The relative performance of both BIC(N) and BIC(m) changed across L1 sample sizes. BIC(N) was significantly better than AIC, AICC (N and m), and HQIC (N and m) for all L1 sample sizes. BIC(N) became significantly better than BIC(m) once the number of L1 units was 7, and also became significantly better than CAIC(m) once the number of L1 units was 13. Across all L1 sample sizes, BIC(m) was significantly better than AIC, AICC (N and m), and CAIC(N), but significantly worse than HQIC (m). BIC (m) became significantly worse than CAIC(m) once the L1 sample size was at least nine.

There was a significant simple effect of information criteria (all $p < 0.001$) across all ICCs. All of the information criteria except for AICC(N) were significantly better than AIC across both ICCs. BIC(N) was significantly better than AIC, AICC (N and m), BIC (m), and HQIC (N and m) across ICCs and became better than CAIC(m) when the ICC was high. BIC(m) was better than AIC, AICC (N and m), and HQIC(m) but worse than BIC(N) and CAIC (N and m) across all ICCs. When the ICC was high, BIC(m) became significantly better than HQIC(N).

There was a significant simple effect of information criteria ($p < 0.001$) across all L2 sample sizes. All other information criteria except AICC(N) were significantly better than AIC across all L2 sample sizes. BIC(N) was significantly better than AIC, AICC (N and m), BIC (m), and HQIC (N and m) across ICCs and became better than CAIC(m) when the L2 sample size was 100. BIC(m) was better than AIC, AICC (N and m), and HQIC(m) but worse than BIC(N) and CAIC (N and m) across all L2 sample sizes. When the L2 sample size was 100, BIC(m) became significantly better than HQIC(N).

Correctly specified logistic models. The five-way, four-way, and three-way multilevel logistic models had coefficient estimates with standard errors equal to zero, but all of the coefficients estimated without issue for the two-way interaction model. Three two-way interaction terms were significant (all $p < 0.001$), one of which, L2 sample size, included information criteria. To examine the simple effects of information criteria across L2 sample sizes, data were subset based on L2 sample size. Regardless if which information criterion was used as the reference category, the only subsetted models that were able to produce estimates were those for which L2 sample size was equal to 30, so continued testing of the simple effects across L2 sample sizes was not possible.

Examination of Table 2.5 showed that the correct selection rates across distinctiveness of the logistic functional form were quite different. In light of this, exploration of the simple effect of information criteria though subsetting by distinctiveness was tried. When distinctiveness was higher, there was no significant simple effect of information criteria. When the distinctiveness was low, however, there was a significant simple effect of information criteria. AIC performed significantly better than all of the other information criteria. BIC(N) was significantly worse than all other information criteria except for CAIC(N). BIC(m) performed significantly worse than AICC (N and m) and HQIC (N and m) but significantly better than BIC(N) and CAIC (N and m).

Correctly specified quadratic models. The highest order interaction term in the five-way interaction model was not significant, so the four-way interaction model was examined. There was one significant four-way interaction vector, but it did not include information criteria. The three-way interaction model had four significant three-way interaction terms, two of which included information criteria. Specifically, those two three-way interaction vectors included L1

sample size, L2 sample size, and distinctiveness of the underlying functional form. Data were subsetted based on their underlying distinctiveness, after which two-way interaction models between information criteria and the different sample sizes were fit.

The interaction between L1 sample size and information criteria was significant, so the data were further subsetted by L1 sample size. Across L1 sample sizes, AIC was significantly better than all of the other information criteria when the distinctiveness was low but was significantly worse than all of the other information criteria when distinctiveness was high. BIC(N) was significantly worse than all of the other information criteria when distinctiveness was low but was better than all other information criteria except CAIC(N) when distinctiveness was high. BIC(m) was significantly better than BIC(N) and CAIC (N and m) when the distinctiveness was low, but worse than AIC, AICC (N and m), and HQIC (N and m). When distinctiveness was high, BIC(m) was significantly better than AIC, AICC (N and m) and HQIC (N and m).

The interaction between L2 sample size and information criteria was significant, so the data were further subsetted by L2 sample size. AIC was significantly better than all of the other information criteria across all L2 sample sizes when distinctiveness was low but was worse than all of the other information criteria when distinctiveness was high. When distinctiveness was low, BIC(N) was significantly worse than all other information criteria except CAIC_N. When distinctiveness was high, BIC(N) was significantly better than all other information criteria except CAIC(N). When distinctiveness was low, BIC(m) was significantly better than BIC(N) and CAIC (N and m) and worse than AIC, AICC (N and m), and HQIC (N and m). When distinctiveness was high, this was reversed.

Discussion

It is appropriate to begin the model fitting process of longitudinal data by fitting an unconditional growth model, but the use of model selection at this step depends on whether or not the researcher has settled on which trajectory to fit. If the researcher has competing theories about the nature of the longitudinal change, then it is possible that model selection could occur as early in the process as the fitting of the unconditional growth model. Information criteria offer a flexible model selection framework for comparing unconditional growth models of all kinds, including truly nonlinear functional forms.

For exponential functional forms, the ability of information criteria to select the correctly specified model depended on the ICC and the number of L1 and L2 units, but the trends within these were similar to each other. Across all levels of L1 sample size, L2 sample size, and ICC, consistent criteria selected correctly specified exponential models more often than the efficient criteria. BIC(N) performed better than all of the other information criteria, including BIC(m). This aligned with my expectation that consistent criteria would perform better due to each of the candidate models have relatively few parameters. When the distinctiveness of the exponential trend was low, the second most commonly selected model when BIC (N and m) was used was a linear model. In contrast, the second most commonly selected models when AIC was used for model selection were linear models (when L1 sample size was lower) or logistic models (when L1 sample size was higher). Unlike BIC (N and m), AIC may select overparameterized models even in the asymptotic case (Bozdogan, 1987). Given this, that AIC sometimes selected runner-up models that were more highly parameterized than the underlying functional was not surprising.

For logistic functional forms, the ability of information criteria to select the correctly specified model appeared to depend on the L2 sample size, but tests of simple effects across the

range of L2 sample size were not possible. This was almost certainly due to the stark difference in the correct selection rates across the levels of distinctiveness in the logistic functional form as shown in Table 2.5. AIC and BIC (N and m) performed equally well when the distinctiveness was high, but AIC performed better than BIC (N and m) when the distinctiveness was low. Of the two, BIC(m) performed better than BIC(N). As also shown in Table 2.5, the selection rate of the correctly specified logistic model when the underlying logistic function was less distinct was dismally low, which was likely driven by the fact that the logistic model was among the most highly parameterized model in the set of candidate models. Because of this, the most common runner-up model was also the model that was selected the most frequently by each information criterion. Both AIC and BIC (N and m) favored the linear model in this case.

For quadratic functional forms, the ability of information criteria to select the correctly specified quadratic model depended on the L1 and L2 sample sizes and the distinctiveness of the underlying quadratic form. When the level of distinctiveness was held constant, the performance of information criteria across L1 and L2 sample sizes was similar. When distinctiveness was low, efficient criteria performed better than consistent criteria, with AIC being the best performing. When distinctiveness was high, however, this trend was reversed, and AIC became the worst performing criterion. Both forms of BIC were better across L1 and L2 sample sizes, but BIC(N) was the better of the two. Per Table 2.6, the selection rate of the correctly specified quadratic model was no more than 30% of the time, meaning that the most selected runner-up model was the most commonly selected model among the information criteria. For both AIC and BIC (N and m), the most commonly selected model was the linear model. When the distinctiveness of the underlying quadratic functional form was high, the runner-up model across all information criteria was the cubic model. This was contrary to what I would have expected because the cubic

model was the most highly parameterized of the models in the candidate model set, but this could be due to how the higher-distinction quadratic form was decided upon for this study. The higher-distinction quadratic trend used for data generation was picked such that the parabolic shape of the quadratic form was distinguishable within the observed range of L1 units. Cubic and quadratic models are nested models, meaning that a cubic function with the cubic term equal to zero would be equivalent to a quadratic function. Even though the cubic model was always misspecified and less parsimonious than other models, it is possible that its ability to capture the quadratic functional form's more distinctive parabolic shape lowered the deviances of the misspecified cubic models sufficiently enough to make it the runner-up model across information criteria.

When the underlying functional form was exponential, BIC(N) was better for selecting the correctly specified exponential model than either BIC(m) or AIC. When the underlying functional forms were logistic or quadratic, however, the best information criteria for selecting correctly specified models depended on the distinctiveness of the functional forms. AIC was better than BIC (N and m) for selecting correctly specified logistic functional form models when the logistic functional form was of lower-distinction, with BIC(m) being better than BIC(N) in this case. For quadratic functional forms, AIC performed better than BIC (N and m) when the distinction of the functional form was low, but BIC (N and m) was better when the distinctiveness was high. In this case, BIC(N) was better than BIC(m). These findings suggest that BIC(N) might be better for identifying correctly specified exponential, logistic, and quadratic models when their respective underlying functional forms are more distinct. When the underlying functional forms are less distinct, then AIC might be better than BIC (N or m) and BIC(m) might be better than BIC(N).

As with all simulation research, the findings of this study may not generalize to conditions not explicitly examined in this study. Although the number of L1 and L2 sample sizes were chosen with common behavioral research contexts in mind, the range of the L1 and L2 sample sizes was limited and the applicability to other sample sizes is unknown. This limitation also applies to the ICCs examined in this study. The most substantial limitation of this study was with regard to the distinctiveness of the underlying functional forms. There is no established way to define the difference between any two nonlinear functional forms. In this study, distinctiveness among the three functional forms of interest was primarily established through iterative visual examination, followed by an attempt to quantify these visual distinctions through comparisons of simple differences of predicted values across time points and of areas under the curves when correctly specified and misspecified models were fit to each form. While this was suitable for the purpose of this study, the curves chosen for each functional form are almost certainly too specific to generalize across contexts. Future work should include a greater variety of distinctiveness of functional forms, perhaps as part of gathering evidence for a more systematic way of defining distinctness between functional forms.

3. Study 2: Model Selection to Identify Non-Nested Predictor Sets in Nonlinear Models

The purpose of the second study in this series was to explore the performance of different information criteria when attempting to identify "more correct" (i.e., less misspecified) models among non-nested predictor sets in the presence of nonlinear functional forms. There were many examples in the linear multilevel model selection literature that use information criteria to simultaneously test a large number of nested models that differ in their fixed effects (e.g. Whittaker & Furrow, 2009; Vallejo et al., 2014). In practice, however, it was likely that applied researchers would use likelihood ratio tests to determine which predictors should be included in the model in such cases. In light of this, the focus of this study was on selecting less misspecified models among truly nonlinear models with non-nested predictor sets that cannot be tested in this manner.

A researcher may engage in model selection to obtain a predictor set in addition to the nonlinear temporal trajectory for several reasons. First, many research hypotheses are best tested by including them as fixed predictors in a model, such as the effect of interventions, demographic variables, or other predictors that contribute to a mechanism in addition to the effect of time. Second, even if the longitudinal change is of greater interest, multilevel models retain many of the assumptions of OLS regression (Hox, 2013), which includes the assumption that predictors are uncorrelated with the error term (Berry, 1993). Misspecification of the predictors in a model can cause this assumption to be violated, the effects of which include introducing bias into the fixed effects estimates and artificially increasing or decreasing their variances (Rao, 1971). Finally, including predictors allows the researcher to interpret the effect of the nonlinear trajectory after controlling for predictors that are correlated with the outcome and remove their variability from the error term of the model, allowing for more sensitive tests.

Method

Functional forms for data generation

Two functional forms, exponential and logistic, were used for data generation. Because the focus of this study was to select among non-nested predictor sets, all of the fitted models were correctly specified with regard to the predictors used to create the functional form. The coefficients used to create the functional form matched those of the more distinct exponential and logistic functions in the first simulation study (shown on Table 2.1).

that produced more distinct exponential and logistic trajectories across the observed range.

Predictor set for data generation

Data were generated to have two predictors at L1 (X1 and X2) and two predictors at L2 (W1 and W2). The first predictor at both levels (X1 and W1) had a correlation of 0.1 with the outcome, and the second predictor at both levels (X2 and W2) had a correlation of 0.5 with the outcome. Hereafter, X1 and W1 will be referred to as the minor predictors at their respective levels and X2 and W2 will be referred to as the major predictors at their respective levels. There were no interactions included in the data generation model. The generated predictors were normally distributed, with predictors at the same level correlated with each other at 0.2.

Set of candidate models

A total of nine candidate models, shown in Table 3.1, were fit to each generated data set. One model, shown in the table as Model 0, included both L1 predictors and both L2 predictors, making it the only correctly specified model. The next set of models were misspecified only at L1, with one model omitting the major L1 predictor (Model 1) and one model omitting the minor L1 predictor (Model 2). These two models were both correctly specified with regard to their Level 2 predictors. The next set of models were similarly misspecified, but at L2. One model

omitted the major L2 predictor (Model 3) and one model omitted the minor L2 predictor (Model 4). These models were both correctly specified with regard to their L1 predictors. Finally, the last set of models had misspecification at both levels. To create this set, the misspecification pattern for Models 2 and 3 was fully crossed with the misspecification pattern for Models 3 and 4. In total, there were four models (Models 5 through 8) with misspecification at both levels.

Relative degrees of misspecification

Although all of the candidate models (M1-M8) were misspecified, there were some models that were almost certainly more correct than others. For example, Models 1 and 2 both omitted a single L1 predictor. It seems reasonable to think that, all else equal, the omission of the single major predictor in Model 2 constituted a greater degree of misspecification than the single omission of the minor predictor in Model 1. Continuing with this logic, Model 1 was even more correct than Model 8 because Model 8 omitted both major predictors. Other models in the candidate model set, however, were not so readily comparable. For example, both Model 1 and Model 3 omitted a major predictor, but the omission occurred at L1 in Model 1 and at L2 in Model 3. It is possible that, even though different predictors were omitted, the omission of any one major predictor constituted the same degree of misspecification. It is also possible that the extent of misspecification depended on the level at which a major predictor was omitted, meaning that actually differed in the degree of their misspecification. A summary table of this a priori ranking can be seen in Table 3.2.

If all of the eight misspecified candidate models were compared in a pairwise fashion, there would be 28 possible pairwise comparisons. Eight of these (e.g., Model 1 vs. Model 5) could have been compared using likelihood ratio tests because the model pairs are nested, leaving 20 non-nested pairwise. These 20 non-nested pairwise comparisons are listed in Table 3.3. Based on

the logic described in the previous paragraph, 15 of the 20 non-nested pairwise comparisons have a model that appears a priori to be less misspecified than the other. Five of these, however, are ambiguous. The comparisons between Models 1 and 3 and Models 2 and 4 are ambiguous because the compared models have the same degree of misspecification, but at different levels. The comparisons between Models 2 and 5 and Models 4 and 5 are ambiguous because it is not clear if omitting two minor predictors or omitting a single major predictor is a greater degree of misspecification. The comparison between Models 6 and 7 is ambiguous because both models omit a major and minor predictor, but at different levels.

In addition to the a priori expectations about the relative degree of misspecification, four large single-replication simulations were conducted for each combination of functional form (exponential and logistic) and ICC (low and high) to examine relative misspecification empirically across model comparisons. The degree of misspecification was examined by comparing the deviances of the models in each pairwise comparison and by comparing the percent reduction in both L2 intercept variance and L1 residual variance when each model was compared to a misspecified unconditional model. Because the functional form specification was always correct in this study, the unconditional models for exponential and logistic models with predictors were the functional form-only models used in Study 1. To compute the percent reduction in variance components for all candidate models, all of the candidate models and the unconditional models had to be estimated without error within each single replication, which was difficult with a large number of L2 units because all of the candidate models were misspecified. All four large single-replication simulations were conducted using 2000 L2 units.

The percent reduction in the L2 intercept variance and the L1 residual variance for all models is shown in Table 3.4. As expected, the correctly specified model is either at the top or near the

top of each model list for both functional forms and both ICCs. The ordering of the models was the same for both ICCs. Across functional forms and ICCs, there appeared to be "tiers" of percent reduction (shown in the shading of different cells in Table 3.4) in the variance components. Both low-ICC and high-ICC exponential functional forms appeared to have two tiers of percentage reduction of the L1 residual variance and four tiers of percentage reduction of the L2 intercept variance. Both low-ICC and high-ICC logistic functional forms appeared to have two tiers of percentage reduction of the L1 residual variance and three tiers of percentage reduction of the L2 intercept variance. Collectively, the percent reduction in the variance components across models suggests that the degree of misspecification may be somewhat different across functional forms and ICCs. The pairwise comparison of model deviances is shown in Table 3.5. The pairwise comparison of model deviances was done for all four combinations of functional form and ICC. Although all models that were compared were all misspecified, the model with the lower deviance within each pair presumably was less misspecified than the model with higher deviance. For each combination of functional form and ICC, the model with lower deviance within each pairwise comparison was the same.

A summary of these findings is shown in Table 3.3, which shows a side-by-side comparison of the model within each pairwise comparison would have been expected to be less misspecified based on a priori judgement, the pairwise deviance comparisons, and the tiers identified when examining the percent reduction in the variance components. There were five comparisons, for which the empirical evaluations were in disagreement. Given that the context of this simulation study was that of longitudinal data, it was tentatively decided that the model favored by the percent reduction in the L1 residual variance would be considered correct in these five models for the purposes of evaluating correct model selection by information criteria.

ICCs and variance components

As in Study 1, the ICCs were 0.4 and 0.6 with the parameters in each functional form most related to the intercept having variance equal to either 0.67 or 1.5, respectively. The L1 error variance was set equal to 1. Error terms were generated from a normal distribution with a mean of zero and a variance in accordance with the intended ICCs.

Sample sizes

As in Study 1, the L2 sample sizes were 30, 50, and 100, and the L1 sample sizes were 5, 7, 9, and 13. Like before, the L2 sample sizes were directly manipulated as part of the data generation step, but the different L1 sample sizes were created by dropping time points from a single generated data set within any given replication. For each of the L1 sample sizes, the time points that were retained were different across sample sizes, but the distances between each time point within an L1 sample size were always equal and maintained the range of the time variable.

Relative effect sizes of model comparisons

It was expected that the ability of information criteria to distinguish between the models being compared would differ across comparisons. If multiple information criteria are computed for a model, those information criteria will incorporate the same deviance. If the differences in the deviances between two models is high, then information criteria are likely to select the model with the lower deviance regardless of which one is used. If, however, the differences in the model deviances is low, then the ability of an information criterion to select a correct (or, in this case, a more correct) model may depend on whether it is efficient or consistent. The absolute values of the model deviance differences within each comparison is shown in Table 3.6. The deviance difference between Models 1 and 3 was the lowest among the functional forms and ICCs, meaning that the characteristics specific to each information criterion (the "penalty" term,

sample size for a sample size-dependent criterion) would be more likely to matter. In contrast, Models 3 and 8 had the greatest difference in model deviances, meaning that the specific characteristics of each information criteria were less likely to matter. Similar to the "tiers" identified in the percent variance reduction in the variance components, three "tiers" of expected difficulty were identified through visual examination of the absolute value of the model deviance differences. These tiers are shown in the middle column of Table 3.6. Model comparisons within the first tier (lower deviance differences) were expected to have more variability in information criteria performance than those in other tiers (moderate and higher deviance differences).

Information criteria

As in Study 1, a total of nine information criteria were computed for each candidate model: AIC, AICC (*N* and *m*), BIC (*N* and *m*), CAIC (*N* and *m*), and HQIC (*N* and *m*). When computing the sample size-dependent criteria, *m* was equal to the number of L2 units (30, 50, or 100) and *N* was equal to the total number of observations (e.g. when L1 sample size was 5 and L2 sample size was 30, *N* was equal to 120).

Overview

The performance of the different information criteria was likely to be affected by the ICC (2 levels), number of individuals sampled (3 levels), and number of measures from each individual (4 levels). One thousand valid replications were simulated in each of these 24 conditions in SAS 9.4. For each generated data set, nine models with correctly-specified functional form were fitted. One of these models (Model 0) was correctly specified, having both Major and Minor predictors at both levels. The other eight (Model – Model 8) had different degrees of misspecification. All of these models were estimated using the NLMIXED procedure with maximum likelihood estimation by adaptive Gauss-Hermite quadrature. A replication was

considered valid if all estimated models converged and produced no error messages. Unlike Study 1, the percentage of valid replications was very high: between 92% (logistic functional form, high ICC, and 100 L2 units) and 100% (logistic functional form, low ICC, 30 individuals).

Analyses and hypotheses

There were two outcomes of interest in this study, both of which involved the ability of the different information criteria to select the more correct model within a set of candidate models. The first outcome, the selection rate of the correctly specified model (Model 0) when the eight other misspecified models are also in the candidate model set, was a conventional outcome that directly built on previous work through the examination of the performance of information criteria when models have truly nonlinear functional forms. The second outcome of interest was the ability of information criteria to select a more correct model when both models in a set of two candidate models are misspecified. As in Study 1, I expected the performance of information criteria to improve (i.e., select the more correct model more often than the less correct model) as the number of individuals and the number of time points increased. The potential impact of the magnitude of the ICC was expected to depend on whether the misspecification occurred at L1, L2, or at both levels.

Data were analyzed in the GLIMMIX procedure in SAS 9.4 (SAS Institute Inc, 2015) using maximum likelihood with Laplace approximation and between-within denominator degrees of freedom used for the F tests of the fixed effects. The outcome was whether or not the more correct model was selected. For each generating functional form, analyses began with fitting a logistic multilevel model with a four-way interaction between L1 sample size (reference group was 5), L2 sample size (reference group was 30), ICC (reference group was low), and information criteria (initial reference group was AIC). Replication ID was modeled as a random

effect, reflecting the fact that all L1 time points within a replication shared the same generating form. If the four-way interaction was not significant, then a model with three-way interactions was tested. The same process was repeated until there was a significant interaction among the highest-order terms or there were no more interaction vectors to test. Because the focus of this study was information criteria, simple effects tests were conducted for a significant interaction only if the interaction involved information criteria. This was done by subsetting data according to what made most sense given the nature of the interaction. Finally, AIC, BIC(N), and BIC(m) were tested against each other and the other information criteria by changing the reference category of the main/simple effect information criteria vectors.

Multilevel logistic models are computationally intensive and sometimes difficult to estimate. The most common error that occurred while fitting these models was the coefficients of one or more dummy vectors having an estimated standard error of zero, creating a test statistic equal to infinity. In other cases, the model did not produce any estimates because it did not converge. If any errors occurred in the four-way interaction model, a three-way interaction model was fitted, and so on until a model that estimated without errors was found. At that point, the analysis procedure described in the previous paragraph was conducted. For all of the non-nested pairwise comparisons, the four-way interaction model either had coefficients with standard errors estimated to be zero or estimated without issue but the four-way interaction was non-significant. Because of this, the estimation of the four-way model is not described in the results section.

Results

Selection rates for correctly-specified models

The selection rate for models that were correctly specified – that is, the fitted models included both major predictors and both minor predictors – is shown in Table 3.7. The

information criterion that was most likely to identify the correct model was AIC (9.8%-14.2% for logistic, 11.5%-18.1% for exponential), followed by AICC-N (8.9%-13.6% for logistic, 10.7%-17.5% for exponential), AICC-m (5.5%-8.1% for logistic, 7.0%-11.2% for exponential), then HQIC-m (5.3%-8.0% for logistic, 6.0%-10.5% for exponential). All other information criteria had correct selection rate of less than 10%.

Non-nested pairwise comparisons – lower deviance differences

The process of determining the ability of AIC, BIC(N), and BIC(m) to select the more correct model when the absolute value of the deviance differences between each of the two models was lower is described in detail in this section. A summary of the results can be found at the top of Table 3.8.

Model 1 vs Model 3. Model 3 (omitted minor L2), which was unclear a priori but was identified by the empirical evaluation methods as being more correct than Model 1 (omitted minor L1 predictor), was selected between 27% and 73% of the time across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way interaction model estimated without issue and contained two significant three-way interaction between that included information criteria ($p < 0.001$). Because both of these interactions included L1 sample size, the data were first subsetted by L1 sample size and the two-way interactions between information criteria, L2 sample size, and ICC were modeled.

When the underlying functional form was exponential, the two-way interactions with L2 sample size and ICC that included information criteria were not significant. There was a significant effect of information criteria across L1 sample sizes (all $p < 0.001$), such that AIC

performed significantly better than BIC(N) and BIC(m). There was no significant difference in performance between BIC(N) and BIC(m).

When the underlying functional form was logistic, the two-way interactions with L2 sample size and ICC that included information criteria were not significant when L1 sample size was 5, 7, or 13. For these models, there was no significant effect of information criteria. When the L1 sample size was 9, L2 sample size and ICC shared interaction terms with information criteria. The data were further subsetted by L2 sample size and ICC separately. The effect of information criteria was significant across these models (all $p < 0.001$), such that AIC performed significantly worse than BIC(N) or BIC(m). There was no significant difference in performance between BIC(N) and BIC(m).

Model 1 vs Model 4. Model 1 (omitted minor L1 predictor), which was identified a priori and by the empirical evaluation methods as the more correct than Model 4 (omitted major L2 predictor), was selected at least 85% of the time across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue. For both main effects models, there was no significant effect of information criteria.

Model 3 vs Model 4. Model 3 (omitted minor L2 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 4 (omitted major L2 predictor), was selected at least 92% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models estimated without issue and had no significant

interaction terms that included information criteria. Information criteria was not a significant effect in the main effects model for either exponential or logistic functional forms.

Model 3 vs Model 6. Model 3 (omitted minor L2 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 6 (omitted minor L1 and Major L2 predictors), was selected at least 69% of the time and up to 94% of the time, across all sample sizes, ICCs, information criteria, and functional forms.

When the underlying functional form was exponential, the three-way interaction model estimated without issue and contained a significant three-way interaction between L1 sample size, L2 sample size, and information criteria ($p < 0.001$). The data were first subsetted by L1 sample size and the two-way interactions between information criteria, L2 sample size, and ICC were modeled. When the L1 sample size was 5 or 7, there was no significant interaction that included information criteria, and information criteria was significant ($p < 0.001$) in the subsetted main effects model. When L1 sample size was 7 or 13, there was a significant interaction between information criteria and L2 sample size, so the data were further subsetted by L2 sample size. Information criteria was significant ($p < 0.001$) across all of these models. AIC performed significantly better than BIC(N) or BIC(m) across all subsetted conditions, and BIC(m) performed significantly better than BIC(N).

When the underlying functional form was logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had three significant interactions that included information criteria: L1 sample size, L2 sample size, and ICC. To examine the simple effects of information criteria, the data set subsetted separately on L1 sample size, L2 sample size, and ICC. There were significant simple effects of information criteria across all of these

subsetted models (all $p < 0.001$). Across all sample sizes and ICCs, AIC performed better than BIC(N) and BIC(m), and BIC(m) performed better than BIC(N).

Model 4 vs Model 5. Model 5 (omitted minor L1 and minor L2 predictors), which was unclear a priori but was identified by the empirical evaluation methods as being more correct than Model 4 (omitted major L2 predictor), was selected at least 86% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had two significant interactions that included information criteria: L1 sample size and L2 sample size. To examine the simple effects of information criteria, the data set subsetted separately on L1 and L2 sample sizes. There were significant simple effects of information criteria (all $p < 0.001$) across all L1 and L2 sample sizes. AIC always performed significantly worse than BIC(N) and BIC(m), and BIC(N) performed better than BIC(m).

When the underlying functional form was logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had two significant interactions that included information criteria: L2 sample size and ICC. To examine the simple effects of information criteria, the data set subsetted separately on L2 sample size and ICC. There were significant simple effects of information criteria (all $p < 0.001$) across all L2 sizes and ICCs. AIC performed significantly worse than BIC(N) or BIC(m), and BIC(N) always performed significantly better than BIC(m).

Model 5 vs Model 6. Model 5 (omitted minor L1 and minor L2 predictors), which was identified a priori and by the empirical evaluation methods as being more correct than Model 7

(omitted major L1 and minor L2 predictors), was selected at least 98% of the time, across all sample sizes, ICCs, information criteria, and functional forms.

When the underlying functional form was exponential, the three-way interaction model failed to converge. The two-way interaction model estimated without issue, and no interaction terms that included information criteria were significant. The main effects model failed to converge. Finally, a model with just information criteria as a predictor was fitted, and the effect of information criteria was not significant. When the underlying functional form was logistic, the four-way, three-way, and two-way interaction models estimated without issue but had no significant interaction terms that included information criteria. There was no significant effect of information criteria in the main effects model.

Model 7 vs Model 8. Model 7 (omitted major L1 and minor L2 predictors), which was identified a priori and by the empirical evaluation methods as being more correct than Model 8 (omitted major L1 and major L2 predictors), was selected at least 90% of the time, across all sample sizes, ICCs, information criteria, and functional forms.

When the underlying functional form was exponential, the three-way interaction model had coefficient estimates with standard errors equal to zero. The two-way interaction model estimated without issue and none of the two-way interactions including information criteria were significant, so a main effects model was fitted. Neither the main effects model nor a model with just information criteria as a predictor successfully converged. Finally, a single-level logistic model was fit using just information criteria as a predictor. This model estimated without error, and there was no significant effect of information criteria. When the underlying functional form was logistic, the four-way, three-way, and two-way interaction models had coefficient estimates

with standard errors equal to zero. The main effects model estimated without issue, and there was no significant effect of information criteria.

Non-nested pairwise comparisons – moderate deviance differences

The process of determining the ability of AIC, BIC(N), and BIC(m) to select the more correct model when the absolute value of the deviance differences between each of the two models was moderate is described in detail in this section. A summary of the results can be found in the middle of Table 3.8.

Model 1 vs Model 2. Model 1 (omitted minor L1 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 2 (omitted major L1 predictor), was selected at least 97% of the time across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue. For both main effects models, there was no significant effect of information criteria.

Model 1 vs Model 7. Model 1 (omitted minor L1 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 7 (omitted major L1 and Minor L2 predictors), was selected at least 86% of the time, across all sample sizes, ICCs, information criteria, and functional forms.

When the underlying functional form was exponential, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had one significant interaction with L1 sample size that included information criteria. To examine the simple effects of information

criteria, the data set subsetted on L1 sample size. AIC performed significantly better than BIC(N) and BIC(m), and BIC(m) performed significantly better than BIC(N).

When the underlying functional form was logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had two significant interactions that included information criteria: L1 sample size and L2 sample size. To examine the simple effects of information criteria, the data was subsetted on these separately. AIC performed significantly better than BIC(N) and BIC(m), and BIC(m) performed significantly better than BIC(N).

Model 2 vs Model 4. This comparison was one for which a priori assessment of relative misspecification was unclear and for which the empirical evaluation methods disagreed. Model 4 (omitted major L2 predictor), which was determined by the model deviance comparison and the percent reduced L1 residual variance to be more correct than Model 2 (omitted major L1 predictor), was selected at least 83% of the time across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue. When the underlying functional for was logistic, the three-way and two-way interaction models estimated without issue and had no significant interaction terms that included information criteria. For both main effects models, there was no significant effect of information criteria.

Model 2 vs Model 5. Model 5 (omitted minor L1 and minor L2 predictors), which was unclear a priori but was identified by the empirical evaluation methods as being more correct than Model 2 (omitted major L1 predictor) was selected at least 98% of the time across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional

form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was a significant effect of information criteria ($p < 0.001$) in both models. AIC performed significantly worse than BIC(N) and BIC(m), and BIC(N) performed better than BIC(m).

Model 2 vs Model 6. This comparison was one for which the empirical evaluation methods disagreed. Model 6 (omitted minor L1 and major L2 predictors), which was determined by the model deviance comparison and the percent reduced L1 residual variance to be more correct than Model 2 (omitted major L1 predictor), was selected between 0% and 18% of the time across all sample sizes, ICCs, information criteria, and functional forms. This selection rate was quite low, which suggests either that the information criteria had difficulty selecting the more correct model for this comparison or that Model 6 may not have actually be less misspecified than Model 2. Because this was one of the five comparisons where the empirical indicators disagreed, the latter seemed more likely. Regardless, whether information criteria was a significant factor in this comparison can be assessed in the same way.

When the underlying functional form was exponential or logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had two significant interactions that included information criteria: L1 sample size and L2 sample size. To examine the simple effects of information criteria, the data set subsetted separately on these separately. Across all sample sizes, AIC performed significantly worse than BIC(N) or BIC(m), and BIC(N) performed significantly better than BIC(m).

Model 4 vs Model 7. This comparison was one for which the empirical evaluation methods disagreed. Model 4 (omitted major L2 predictor), which was determined by the model

deviance comparison and the percent reduced L1 residual variance to be more correct than Model 7 (omitted major L1 and minor L2 predictors), was selected at least 66% of the time and up to 98% of the time, across all sample sizes, ICCs, information criteria, and functional forms.

When the underlying functional form was exponential, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were significant. The two-way interaction model, however had two significant interactions that included information criteria: L1 sample size and L2 sample size. To examine the simple effects of information criteria, the data set subsetted separately on L1 and L2 sample sizes. There were significant simple effects of information criteria (all $p < 0.001$) across all L1 and L2 sample sizes. AIC always performed significantly better than BIC(N) across L1 and L2 sample sizes. AIC also performed significantly better than BIC(m) except when the L1 sample size was 13, at which point they were not significantly different. BIC(m) always performed significantly better than BIC(N).

When the underlying functional form was logistic, the three-way interaction model estimated without issue and contained a significant three-way interaction between L1 sample size, L2 sample size, and information criteria ($p < 0.001$). The data were first subsetted by L1 sample size and the two-way interactions between information criteria, L2 sample size, and ICC were modeled. When the L1 sample size was 7, 9, or 13, there were no significant interactions that included information criteria and the main effect of information criteria was significant (all $p < 0.001$). When the L1 sample size was 5, there was a significant interaction between information criteria and L2 sample size, so the data were further subsetted by L2 sample size. Information criteria was significant ($p < 0.001$) across all of these models. In all of these

subsetted analyses, AIC performed significantly better than BIC(N) and BIC(m). BIC(m) always performed significantly better than BIC(N).

Model 5 vs Model 7. Model 5 (omitted minor L1 and minor L2 predictors), which was identified a priori as being more correct than Model 7 (omitted major L1 and minor L2 predictors), was selected at least 94% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was no significant effect of information criteria.

Model 6 vs Model 7. This comparison was one for which the empirical evaluation methods disagreed. Model 6 (omitted minor L1 and major L2 predictors), which was determined by the model deviance comparison and the percent reduced L1 residual variance to be more correct than Model 7 (omitted major L1 and minor L2 predictors), was selected at least 81% of the time, across all sample sizes, ICCs, information criteria, and functional forms

When the underlying functional form was exponential, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was no significant effect of information criteria. When the underlying functional form was logistic, the four-way, three-way, and two-way interaction models estimated without issue but had no significant interaction terms that included information criteria. There was no significant effect of information criteria in the main effects model.

Model 6 vs Model 8. This comparison was one for which the empirical evaluation methods disagreed. Model 6 (omitted minor L1 and major L2 predictors), which was determined by the model deviance comparison and the percent reduced L1 residual variance to be more

correct than Model 8 (omitted major L1 and major L2 predictors), was selected at least 90% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was no significant effect of information criteria. When the underlying functional form was logistic, the four-way, three-way, and two-way interaction models estimated without issue but had no significant interaction terms that included information criteria. There was no significant effect of information criteria in the main effects model.

Non-nested pairwise comparisons – larger deviance differences

The process of determining the ability of AIC, BIC(N), and BIC(m) to select the more correct model when the absolute value of the deviance differences between each of the two models was higher is described in detail in this section. A summary of the results can be found at the bottom of Table 3.8.

Model 1 vs Model 8. Model 1 (omitted minor L1 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 8 (omitted major L1 and major L2 predictors), was selected at least 89% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was a significant main effect of information criteria ($p < 0.001$). AIC performed better than BIC(N) and BIC(m), and BIC(m) performed better than BIC(N).

When the underlying functional form was logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were

significant. The two-way interaction model, however had two significant interactions that included information criteria: L1 sample size and L2 sample size. To examine the simple effects of information criteria, the data set subsetted separately on these separately. AIC performed significantly better than BIC(N) and BIC(m), and BIC(m) performed significantly better than BIC(N) except when the L2 sample size was 100. In that case, BIC(N) and BIC(m) were not significantly different.

Model 2 vs Model 3. Model 3 (omitted minor L2), which was identified a priori and by the empirical evaluation methods as being more correct than Model 2 (omitted major L1 predictor), was selected at least 98% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue. For both main effects models, there was no significant effect of information criteria.

Model 3 vs Model 8. Model 3 (omitted minor L2 predictor), which was identified a priori and by the empirical evaluation methods as being more correct than Model 8 (omitted major L1 and Major L2 predictors), was selected at least 90% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential, the three-way and two-way interaction models estimated without issue and had no significant interaction terms that included information criteria. There was a significant main effect of information criteria, such that AIC always performed significantly better than BIC(N) or BIC(m). BIC(m) always performed better than BIC(N).

When the underlying functional form was logistic, the three-way interaction model estimated without issue, and none of the interaction terms that included information criteria were

significant. The two-way interaction model, however had three significant interactions that included information criteria: L1 sample size, L2 sample size, and ICC. To examine the simple effects of information criteria, the data set subsetted separately on these separately. Across all sample sizes and ICCs, AIC performed significantly worse than BIC(N) or BIC(m), and BIC(m) performed significantly better than BIC(N).

Model 5 vs Model 8. Model 5 (omitted minor L1 and minor L2 predictors), which was identified a priori and by the empirical evaluation methods as being more correct than Model 8 (omitted major L1 and major L2 predictors), was selected at least 97% of the time, across all sample sizes, ICCs, information criteria, and functional forms. When the underlying functional form was exponential or logistic, the three-way and two-way interactions models had coefficient estimates with standard errors equal to zero. The main effect model estimated without issue, and there was no significant effect of information criteria.

Discussion

There were two outcomes of interest in this study. The first was to examine the correct model selection rate of each information criterion when the correctly specified model was among the set of candidate models. The second outcome of interest was the performance of information criteria when the correctly specified model was not among the set of candidate models; good performance in this case was when the less misspecified, or more correct, model was selected. The former outcome is a common outcome of interest in simulation literature examining the performance of information criteria, and this study extends previous work to truly nonlinear functional forms. The latter outcome is not commonly examined but may better reflect the reality of research (Burnham & Anderson, 2003).

Overall, the correctly specified model was not selected most of the time, which was not

expected. In this study, the correctly specified model was the most parameterized model because it included both major and minor predictors at both levels. All of the examined information criteria incorporate the same deviance as part of their computation, an ordered list of which is shown in Table 3.9. However, how additional model complexity is incorporated differs across criteria, and it is likely that some of the misspecified models – particularly that of Model 3, which was missing just a minor L2 predictor – had deviances close enough to that of the correctly specified model that they were selected because the misspecified models had just two or three predictors. In addition, per Table 3.4, the correctly specified model was always in the highest tier of percent reduction in variance components but was never alone in that tier. For example, the correctly specified model for the low-ICC exponential condition reduced the L1 residual variance by the same percentage as Models 1, 3, 4, 5, and 6. Because of this, the consistent information criteria, which I expected do perform better than the efficient information criteria, homed in on these seemingly more parsimonious models. Efficient information criteria, which are dimension inconsistent, are more likely to select overparameterized models (Bozdogan, 1987), which may be why their selection rate was comparatively better than consistent information criteria in this study. In light of this, even though the low selection rate of the correctly specified model was not anticipated, it is not entirely surprising given the similarity of the candidate models to the correctly specified model.

Selection rates of models identified as more correct within their pairwise comparisons were generally good (with an exception, discussed below). Contrary to what I expected, there was variability in which information criteria performed better across all tiers of model comparisons – that is, no tier had either AIC or BIC (N and m) always or almost always being the best-performing information criterion. There were, however, a couple of systematic

differences of note. First, whenever AIC was the best-performing information criterion of the three, BIC(N) was almost always worse than BIC(m). In contrast, whenever AIC was the worst-performing information criterion of the three, BIC(N) was almost always better than BIC(m). This may be because *N* was much larger than *m,* which enhanced the performance of BIC when BIC was better than AIC and detracted from its performance when AIC was better. There was only one case where there was a significant effect of information criteria that BIC(N) and BIC(m) performed similarly, which was the comparison between Models 1 and 3. Both of these models were missing a minor predictor and had the smallest deviance difference between models. Because of this, there may not have been much room for differences between BIC computations to emerge as there were in other comparisons where there was a significant effect of information criteria. The other systematic difference of note was that the effect of information criteria and the ranking of the performance of AIC and BIC (N and m) was largely consistent across functional forms. The only exceptions to this trend were the comparisons between Models 1 and 3 and Models 3 and 8, which respectively had the lowest and highest model deviance differences among the model comparisons. This may have been due to the differences in deviances of these models between functional forms. Per Table 3.9, the deviances of the logistic models across ICCs was lower than that of the exponential models, meaning that the magnitude of the differences was smaller within logistic models even within the same tier.

One of the greatest limitations of this study had to do with determining the relative degree of misspecification and which of the models in a given pair was more correct. Relative misspecification between models in the candidate model set was explored in multiple ways, including model deviance and percent reduction in variance components. These methods, however, were not always in agreement and none of them provided a definitive metric for

determining relative misspecification. This was most evident in the comparison between Models 2 and 6. Model 6 omitted a major predictor at L2 and a minor predictor at L2, while Model 2 omitted only a major predictor at L1. The model deviance comparison (which was among the largest of the deviance differences) and the percent reduction in L1 residual variance suggested that Model 6 was less misspecified, while the a priori judgment and the percent reduction in L2 intercept variance suggested that Model 2 was less misspecified. Given the disagreement and the fact that the information criteria otherwise performed well, the relative misspecification in this comparison is likely more ambiguous than in the other comparisons. While many of the model comparisons had high rates of selection of the more correct model, which suggests that information criteria could be useful for selecting a more correct model when the correctly specified model is absent from the set of candidate models, it cannot be definitively said the models determined to be more correct actually were.

Like most simulation work, the findings of this study may not generalize to conditions not examined in this study. Although the number of L1 and L2 sample sizes were chosen with common behavioral research contexts in mind, the range of the L1 and L2 sample sizes was limited and the applicability to other sample sizes is unknown. This limitation also applies to the ICCs examined in this study. Although no specific recommendations about which information criteria to use when selecting among candidate models that are all misspecified with regard to their predictor sets emerged from this study, the utility of developing such recommendations is considerable because models are used in behavioral research to explain complex phenomena. Future directions include work on refining the assessment of relative model misspecification, either through simulation or analytical approaches. In addition, future work should include more complex generating models (e.g. cross-level interactions and more predictors) and a greater

variety of types of misspecification. For example, while all candidate models in this study were correctly specified with regard to their functional forms, it would be interesting to see if both functional form and predictor sets could be identified within the same model selection step. Finally, while the focus on this study was on AIC and BIC because of their prevalence in psychological literature (Whittaker & Furlow, 2009) and in statistical software packages, further comparisons with additional information criteria could potentially reveal the usefulness of a less-common criterion in certain circumstances.

4. Interlude: Fitting Covariance Structures to Truly Nonlinear Multilevel Models

Once a functional form and a predictor set has been identified, a likely next step for a researcher fitting truly nonlinear functional forms to longitudinal data would be to fit some kind of covariance structure. It is possible to fit covariance structures at multiple levels when there is sufficient data to do so, but the within-person covariance structure would likely be of particular interest in the context of longitudinal data. The within-person error covariance matrix is a square matrix that has the same number of rows and columns as the number of time points. The main diagonal in this matrix contains the L1 error variance at each time point. The other elements in the matrix contain the covariances between errors at different time points. The default L1 covariance structure in many statistical software packages that fit multilevel models is the conditional independence structure. The conditional independence structure for a five-time point study is shown below:

$$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

This structure implies that the L1 residual variance is the same across all five time points and that there is no correlation between error at different time points. The conditional independence structure may be a reasonable choice in a longitudinal context if the measures are spaced such that a long amount of time, such as a year, passes between measurements (Harring & Blozis, 2014). If, however, the within-person L1 errors are correlated between time points, a model fit using the conditional independence structure would be misspecified. Singer and Willett (2003) noted that substantive hypotheses are usually expressed via the fixed effects of the model and that the estimates of the fixed effects usually do not change regardless of the L1 covariance structure used in a model. The precision of the fixed effects estimates, however, can be increased

by fitting an L1 covariance structure that captures within-person error correlation across time points. Specifically, correct specification will increase the power of the hypothesis tests of the fixed effects by decreasing the standard errors associated with the fixed effects (Kwok, West, & Green, 2007; Singer & Willett, 2003).

It is possible to separately estimate all of the correlations across time points, which is usually referred to as an unstructured covariance matrix and has the following structure

$$\sigma^2 \begin{bmatrix} a & f & g & h & i \\ f & b & j & k & l \\ g & j & c & m & n \\ h & k & m & d & o \\ i & l & n & o & e \end{bmatrix}.$$

Of all of the covariance structures that could be fit as part of a given model, this one will always have the lowest deviance because deviance is not adjusted based on the number of parameters used to obtain the likelihood estimate (Singer & Willett, 2003). While the unstructured matrix can be appealing because it imposes no assumptions about the nature of the within-person variability across time, this structure is the most difficult to estimate because of the large number of unique elements that must be estimated and precludes the estimation of L2 random effects (Bauer, Gottfredson, Dean, & Zucker, 2013). For example, the smallest L1 sample size used in the prior two studies was five, meaning that a fitting an unstructured covariance matrix would require the estimation of 15 additional parameters. The amount of data needed to estimate such a model would be substantial and would likely be impossible to attain in many research contexts. Singer and Willett (2003) suggested beginning exploratory analyses by fitting an unstructured covariance structure if possible, but they also noted that more parsimonious structures are likely to be selected when using model selection methods that adjust based on the number of parameters used to estimate a model.

A common choice for longitudinal research is an autoregressive structure, which models between-time point error correlations that trend exponentially toward zero (Bauer et al., 2013; Singer & Willett, 2003). An autoregressive within-person covariance structure for a five-time point study has the following matrix form:

$$\sigma^2 \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 & \rho^4 \\ \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^4 & \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}.$$

This is an appealing structure for longitudinal data because it allows for errors at adjacent time points to correlate and also allows the correlations between prior time points to decay in a systematic fashion as time passes. The autoregressive structure is created by one estimated parameter, commonly denoted rho (ρ), and the L1 residual variance, making for two parameters total. The change in the correlation over time is introduced by raising rho to successive integer powers up to the total number of time points. Elements in each off-diagonal band are raised to the same power; the first band is always rho raised to the first power, the second off-diagonal band is always rho raised to the second power, etc. As rho is exponentiated, there is a rapid decrease in correlation between errors as more time points separate them. The first band contains the correlation between adjacent time points, which models a situation where adjacent measurements are more highly correlated than measurements with more distance between them.

Another common choice in longitudinal research is a Toepliz-banded structure. This structure can be an appealing choice because, unlike the autoregressive structure, it does not impose the assumption of rapid decay or the presence of a lower asymptote at zero (Bauer et al., 2013). The within-band elements must be identical, but there are no other restrictions; each band can have any correlation with the adjacent band, including zero. Because of this, unlike the estimation of

only a single parameter for the autoregressive structure, each individual band must be estimated separately. Because the elements within each band are constrained to be equal, this structure is more parsimonious than the unstructured matrix. The matrix form Toepliz-banded structure for a five-point time study is as follows:

$$\sigma^2 \begin{bmatrix} 1 & a & b & c & d \\ a & 1 & a & b & c \\ b & a & 1 & a & b \\ c & b & a & 1 & a \\ d & c & b & a & 1 \end{bmatrix}.$$

Note that this structure has five parameters (one for each off-diagonal band and one for the L1 residual variance), unlike the single rho parameter in the autoregressive structure and the 15 parameters in the unstructured matrix. A full Toepliz-banded structure estimates all of the bands, but as the number of time points increases, this structure becomes increasingly difficult to estimate because of the steady increase in the number of parameters. If a researcher wants to fit a Toepliz-banded structure but cannot or would prefer not to estimate all possible bands, the structure can be further constrained by setting one or more bands equal to zero. For example, Harring and Blozis (2014) described a Toepliz-banded specification where only the first off-diagonal band (denoted above as a) was estimated. Such a structure, referred to as a symmetric tridiagonal Toepliz-banded structure, implies that adjacent measurements are correlated, but non-adjacent measurements are uncorrelated.

If a researcher had information from prior research suggesting an appropriate covariance structure or wanted to test a specific hypothesis about the nature of the within-person covariance structure, then model selection for a covariance structure would be unnecessary. Because it is often difficult to choose a structure solely on the basis of theory, it is likely model selection would be needed to benefit from the resulting increase in the precision of the estimates of the

fixed effects when a proper covariance structure is specified. In such cases, fitting several covariance structures and comparing them via a model selection method, such as information criteria, is recommended (Bauer et al., 2013; Singer & Willett, 2003). Many statistical software packages used to estimate linear and polynomial multilevel models have built-in options for specifying different covariance structures. For example, the MIXED procedure in SAS 9.4 can fit over 23 covariance structures directly through the TYPE option (SAS Institute Inc., 2015). In addition, one of these options allows the user to create any covariance structure that may not be offered in MIXED; for example, Bauer and colleagues (2013) created a macro program to be used with MIXED that allows users to fit stabilizing banded structures (i.e., Toepliz-banded structures where the bands eventually stabilize to some correlation not equal to zero).

There are not as many user-friendly options available for fitting a variety of L1 covariance structures when the model is a truly nonlinear model. The NLMIXED procedure, which was used to fit the truly nonlinear functional forms in the previous two simulation studies, fits a conditional independence structure by default and has no built-in options for fitting alternative within-person covariance structures. NLMIXED allows the user to construct and specify a customized log likelihood function (SAS Institute Inc., 2015), which means that a user could write such a function that would fit different L1 covariance structures. However, doing so requires the user to be comfortable building a log likelihood function.

Harring and Blozis (2014) demonstrated how the log likelihood function could be built in NLMIXED to incorporate several L1 covariance structures when fitting truly nonlinear multilevel models. They also included NLMIXED code for the autoregressive structure in the paper's appendix and for five structures in a supplemental document, the latter of which included code for setting up the data for use in NLMIXED and for fitting the customized log likelihood

functions. The first step was to read in the data in wide format, such that each person's repeated measurements were contained in multiple variables in one row. NLMIXED usually requires the data to be in long format, where each person's repeated measures are contained in a single variable and are contained in multiple rows, but this not a requirement if using a customized log likelihood function. The second step was to create two new variables, which were used to keep track of valid (i.e., non-missing) observations. One variable denoted the time at which the valid observation occurred, and the second variable was a count of the number of valid within-person observations. It was necessary to record the time point at which the valid observation occurred because all of the "gaps" in the data created by missing observations were also removed in this step. Finally, the third step was to define the customized log function to estimate one of five covariance structures. This required taking the inverse and determinant of the desired within-person covariance structure. The dimensions of the within-person covariance structure were established by using a looping structure based on the value of the temporal variable. Once this was done, the within-person covariance structure was incorporated into the loglikelihood function. At minimum, the user needed to be able to adapt the single-equation form of the model, the starting values, the L2 random effects, and the number of time points within the time-based looping structure to apply a structure to their own data.

I adapted the three-step supplemental code for three of the five covariances structures: conditional independence, autoregressive, and tridiagonal Toepliz-banded. I adapted these three structures to two kinds of data. The first kind was data collected from over 2000 participants over the course of 5 time points and there were missing measurements from some participants. The second kind was simulated data generated to have conditionally independent, autoregressive, or tridiagonal Toepliz-banded covariance at L1 over five time points. None of the data were

missing in the simulated data set. In the case of the data collected from human participants, the estimated model had a logistic functional form and predictors at both L1 and L2. In the case of the simulated data, the estimated models were functional form only- exponential and logistic models. Unfortunately, my adaptation of the Harring and Blozis (2014) code was unsuccessful, meaning that neither an applied study nor a simulation study using information criteria to selection among different L1 covariances structures could be completed.

When fitting the structures to the simulated data, none of the models with either autoregressive of Toepliz-banded structures produced estimates; these models always terminated due to an execution error. Execution errors can be caused by typographical mistakes, calculations that are impossible or very difficult (e.g., dividing by a number very close to zero), or by fitting misspecified models (Kiernan et al., 2009). Because the models fitted in NLMIXED always matched the generating model exactly, the execution errors could not have been caused by model misspecification. The code was carefully reviewed for typographical mistakes and DATA steps were used to investigate information about the model estimation that was not displayed when an execution error occurs (but is stored in memory). Unfortunately, the cause of the execution errors remained unclear.

When fitting the different structures to the applied data, I started by fitting the functional form-only models first then moved on to the model containing the selected predictor set. The conditional independence structures, which matched the default covariance structure that can be specified in NLMIXED, produced estimates without any accompanying error messages for the functional form-only model and the model with predictors. The autoregressive and Toepliz-banded structures, however, either produced no estimates or produced untrustworthy estimates (specifically, the second order optimality condition was violated). A series of actions was

undertaken to try to coax the autoregressive and Toepliz-banded models into estimating in an interpretable way. First, all of the missing data from the applied example was removed by adapting the second step of the code to exclude missing values in all of the predictors in the model, not just the outcome. The next action was based on how the temporal variable was used in Harring and Blozis' (2014) code. In some parts of the code, the time predictor was used was the input of a prediction equation. The same time predictor was also incorporated into the looping structure of the code, such that certain values of time would either continue or end the loop used to create the L1 covariance structure. In light of this, the values taken by the temporal predictor were was changed from being continuous (e.g., 1.6, 2.4, etc.) to being discrete (i.e., 0, 1, 2, etc.). The autoregressive and Toepliz-banded structure models produced estimates when the data were complete and when time was discrete, but an error about violating the second order optimality condition appeared for both of them. This problem can sometimes be solved by rescaling parameters in the model (Kiernan et al., 2009) and was successfully used for fitting the logistic functional form models in the two simulation studies. In this case, however, rescaling different parameters in the models did not resolve the errors. Finally, to see if the Toepliz-banded structure was having difficulty estimating because these data included fewer time points than the example data used in Harring and Blozis' (2014) example, I removed the determinant computation looping structure and replaced it with a hard-coded determinant calculation for a 5x5 matrix. This also failed to resolve the error messages. Although the conditional independence code was successfully adapted, there were no models with alternative covariance structures to which it could be compared and thus no model selection using information criteria could occur.

The inability to fit different within-person covariance structures to truly nonlinear models as part of a simulation study and as part of the applied chapters was disappointing because it would almost certainly be of interest when fitting models to longitudinal data. While the fixed effect estimates of a model are unlikely to be profoundly changed (Singer & Willett, 2003), the conditional independence structure is likely to be underspecified (i.e., overly simple) when data are longitudinal in nature, causing inflation of the standard errors of the fixed effects (Kwok et al., 2007). Researchers often rely on the null hypothesis tests of fixed model coefficients because the theories being tested are often expressed in the fixed effects of the multilevel model. An underspecified L1 covariance matrix, however, can reduce the power of these tests, making effects that might actually exist more difficult to find.

5. Application: Discrepancy in Body Image Across Childhood and Adolescence

Nonlinear multilevel models allow researchers to specify models that may better match their theories about longitudinal change trajectories than that implied by the steady, monotone linear functional form. The exponential and logistic models of growth are widely used across the life sciences, suggesting their potential utility for modeling behavioral phenomena over time. One area of behavioral medicine that has gotten a great deal of attention over the past several years is the relationship between obesity and psychosocial factors over the lifespan, particularly for adolescents. A handful of studies have used nonlinear modeling methods to explore this relationship, allowing for intriguing conclusions that could not otherwise be tested.

Obesity has been identified an important public health concern in the United States (Brown, Fujioka, Wilson, & Woodworth, 2009), as well as globally (Swinburn et al, 2011). Based on data from the 2011-2012 National Health and Nutrition Examination Survey (NHANES), Ogden, Carroll, Kit, and Flegal (2014) found that approximately 35% of adults and 17% of youth in the United States are obese, and an additional 34% of adults and 14% of youth are overweight. One of the main problems of obesity is that of adiposity, meaning that individuals who are obese have a high amount of body fat. High amounts of adipose tissue are associated with negative health outcomes, including cardiovascular disease, metabolic disorders, cancer, pulmonary disease, musculoskeletal disorders, gastrointestinal disorders, reproductive disorders, and dermatologic disorders (Brown et al, 2009). However, there is growing evidence that many of the risks associated with high adiposity are reversible. A recent systematic review determined that there was high-quality evidence that weight loss in overweight or obese adults is associated with improvements in a variety of cardiometabolic indicators, such as insulin sensitivity, blood lipid profile, and blood pressure (National Heart, Lung, & Blood Institute, 2013). Because of this,

there has been considerable research interest in reducing obesity, often through targeting change in health-related behaviors.

Alongside potential negative physical health outcomes, obesity is also highly stigmatized in Western societies, with concomitant low self-esteem, depression, and body dissatisfaction (Puhl & Heuer, 2010). The stigmatization of obese people is also common among healthcare professionals (Schwartz, Chambliss, Browned, Blair, & Billington, 2003; Tomiyama, 2014) and in public health discourse, where competing narratives regarding responsibility for obesity and the resulting stigma remain (Saguy & Riley, 2005). This stigma may also generate health disparities and interfere with obesity intervention efforts (Puhl & Heuer, 2010). For example, there is evidence that obese adults who feel stigmatized based on their weight are more likely to avoid exercise (Vartanian & Novak, 2011; Vartanian & Shaprow, 2008), avoid healthcare settings (Drury & Louis, 2002), and increase eating behaviors (Tomiyama, 2014). Obesity stigmatization also occurs in children, with children as young as three years old endorsing negative attitudes and beliefs toward obese children (Puhl & Latner, 2007). Children who are obese encounter stigmatization from a variety of sources, including parents, siblings, and other family members (Puhl & Brownell, 2006). Additional sources can also include peers (Latner, Stunkard, & Wilson, 2005), non-family adults (e.g. teachers; Neumark-Sztainer, Story, & Harris, 1999), and mass media (Latner, Rosewall, & Simmonds, 2007).

Stigmatization based on weight is especially pernicious in children and teens, as it appears to have negative psychological and physical impact on youth whether they are obese or not. Mustillo, Hendrix, and Schafer (2012) examined the effects of BMI category change on self-concept and found that girls who were obese as children but normal weight as teens experienced greater body image discrepancy and lower self-esteem than girls who had been of normal weight

as children and as teens. Further work suggested that external labeling - that is, being told that one is too fat - from parents and friends contributes psychological distress in female adolescents across time, even when controlling for actual BMI (Mustillo, Budd, & Hendrix, 2013). In addition, girls who are told that they are too fat are at higher risk of being obese as teens, regardless of their BMI as children (Hunger & Tomiyama, 2014).

While most statistical analyses used in research on weight stigma in youth employ linear models, some researchers have employed modeling techniques that permit nonlinear relationships and correlated effects over time. In one case, a data-driven regression tree method was used to find important predictors of BMI change among a set of 41 predictors, with the ability to evaluate nonlinear relationships explicitly noted as an advantage of that method (Rehkopf, Laraia, Segal, Braithwaite, & Epel, 2011). Mustillo, Hendrix, and Schafer (2012) used growth mixture modeling, which incorporated linear and quadratic terms for BMI, race, and residual variance across a series of models. Mustillo, Budd, and Hendrix (2013) used autoregressive cross-lagged mediation modeling (MacKinnon, 2008), which used an autoregressive structure to model outcome observations. In light of these findings, nonlinear multilevel modeling offers a promising way to examine how body image changes across adolescence, particularly as stigmatizing incidents accumulate during this critical period of change.

There were two outcomes of interest in the current study: body image discrepancy and body dissatisfaction. For each of these, the final models were built according to a three-part process. The first part was an exploration of the nature of the change trajectory in the outcome variables across late childhood and adolescence. Conventionally, this trajectory would be tested with a linear model, which implicitly suggests that body image discrepancy and dissatisfaction change

steadily across adolescence. This may be too simplistic, however, given related findings on the longitudinal relationships between adolescent obesity, weight labeling, and psychological distress (Mustillo, Budd, & Hendrix, 2013). Nonlinear multilevel models offer the opportunity to examine alternative trajectories that may better reflect how body image discrepancy and dissatisfaction change over time. A quadratic change trajectory would indicate that body image discrepancy and dissatisfaction accelerate over time, with teens having greater discrepancy and greater dissatisfaction than when they were children. An exponential trajectory would also indicate that body image discrepancy and dissatisfaction accelerate over time, with an "explosive" increase at a particular time; in this context, this time would likely be puberty. Because of the syncretic relationships among childhood weight, starting of puberty, and later obesity (Jasik & Lustig, 2008), a rapid increase around the start of puberty seems possible. This trajectory suggests, however, that growth in body image discrepancy and dissatisfaction do not slow after the explosion of growth. To account for a ceiling effect, an S-shaped trajectory, such as that of a logistic curve, may be the best overall trajectory.

The second part of the model-building process was evaluating different sets of predictors that affect body image discrepancy and dissatisfaction. These predictor sets included weight labeling, valence toward fatness and thinness, the racial composition of social environments, self-criticism, and unhappiness about one's body. Key covariates were also included as part of each predictor set. The third part of the model-building process was selecting an appropriate covariance structure. Because the models under examination were non-nested, information criteria were used for model selection at each step.

Method

The National Heart, Lung, and Blood Institute National Growth and Health Study

(NGHS) was a longitudinal study conducted from 1987 to 1997 to examine the development of obesity in Black and White pre-adolescent and adolescent girls. Girls between the ages of 9 and 10 were recruited in California, Ohio, and Washington D.C. Data were collected from 2379 participants, who were identified by their guardians as being Black (n = 1213) or White (n = 1166), every year for 10 years. Some measures were recorded every year, including medical history, nutrition, and health beliefs, as well as a physical exam. Other measures were taken only during certain years, including demographics, biomarkers, food intake, physical activity, and psychosocial measures. All analyses were conducted using the MIXED and NLMIXED procedures in SAS 9.4.

Measures - outcomes

Body image discrepancy. At each of the ten time points, participants were asked to answer questions based on a set of nine female figures, which were drawn to be racially ambiguous. The figures acted as a visual reference point for each level of a nine-point Likert scale, with figures at the lower end of the scale appearing very underweight and figures at the higher end of the scale appearing very obese. The participants were first asked which figure they believed they currently looked most like, then asked to identify the figure they would like best to look like. To compute body image discrepancy, the difference between the answers to these two questions was computed, similar to the method of Mustillo and colleagues (2012). Negative values indicated a girl who preferred a body that was thinner than the body she endorsed as being most similar to her current shape, while positive values indicated a preference for a fatter body. Next, the difference was turned into an absolute value, creating an index of absolute discrepancy.

Body dissatisfaction. Participants completed a 66-question survey, "My Feelings", at the third, fifth, seventh, ninth, and tenth time point. Each question was answered on a six-point

Likert scale, with each point indicating how frequently ("always", "usually", etc.) the participant experienced a feeling or thought or behaved in a certain way. Nine of these questions pertained to body dissatisfaction, which were summed to create a composite body dissatisfaction variable. Higher values of this composite variable indicated greater dissatisfaction.

Measures – Temporal predictor and covariates

Time/Age. In this data set, the passage of time is equivalent to the age reported for each participant over ten years. Because of this, age was used to model time slopes as part of the functional form of the different change trajectories. Age was centered for each participant by subtracting 9 from the age reported at each time point. For example, a participant who entered the study at exactly 9 years of age would have a centered age of 0 at the first time point.

Age at menarche. Participants were asked each year at what age they experienced menarche, which is an important marker of pubertal status. The within-participant responses sometimes differed across timepoints, so the within-participant median of these reports was used for analysis.

BMI. BMI was computed each year of the study. Each participant was weighed while wearing either a paper hospital gown or a T-shirt and socks. Height was measured using a stadiometer. BMI was included in each model as a continuous time-variant covariate rather than as BMI cutoffs to preserve a larger range of possible values.

Race. In this study, race was a dichotomous variable, with the participant's guardian identifying each girl as either Black or White. The racial identity endorsed at the beginning of the study was used for analysis. This variable was dummy coded, with White acting as the reference category.

Total household income. At the beginning of the study, the participant's guardian was asked

about the total yearly income for the household. These responses were grouped into one of four income categories: less than $9,999, $10,000-29,999, $30,000-$39,999, and $40,000 or more. These categories were dummy coded, with less than $9,999 acting as the reference category.

Parental/guardian education. At the beginning of the study, the participant's guardian reported his or her own level of education and that of their partner. These responses were grouped into one of three education categories: high school or less, some college, and college or more. These categories were dummy coded, with high school or less acting as the reference category.

Measures – Predictors of interest

Weight criticism. At the first time point, participants were asked if they had ever been told that they were too fat or too thin by one of nine people: mother, father, brother, sister, best girlfriend, best-liked boy, any other girl, any other boy, and any teacher. From this set of questions, three predictors were created. The first was a count of the number of sources who had told the girl that she was too fat. The second was a similar count of sources who had told the girl that she was too thin. The third was the summation of the first two.

Valence toward thinness/fatness. At the first time point, participants were asked questions about their agreement with different statements about a hypothetical fat or thin girl of their age. There were five statements that were positively phrased and two statements that were negatively phrased. The positive statements included the fat/thin girl having more friends, feeling better about herself, being prettier, feeling more like a girl, being less likely to get pushed around, or being healthier. The negative statements included the fat/thin girl looking less grown up and feeling less in charge of things. Participants could agree with the statements, disagree with the statements, or indicate that fatness/thinness would not matter with regard to the statement. From

these, four valence variables were created. Two positive valence variables were created by summing the 'yes' responses to the positive statements when the prompt asked about a thin girl and a fat girl. Two negative valence variables were created by summing the 'yes' responses to the negative statements.

Perceived school minority status. At the first time point, participants were asked if the school they attended had a student body that was all or mostly Black, all or mostly White, or about half Black and half White. This was recoded into a dichotomous perceived school minority variable, which was created by comparing the girl's race with her report of the student body of her school. For example, a Black girl who reported her school's racial composition as being either all or mostly White would be coded as being a perceived school minority. If she instead reported her school's racial composition as being half Black or all/mostly Black, she would be coded as not perceiving herself as a minority within her school. This variable was dummy coded, with not perceiving oneself as a minority acting as the reference category.

Same-race friendships. At the first time point, participants were asked if they had any close friends who were White, Black, Hispanic, Asian, or American Indian/Alaskan Native. Two variables were created from these responses. The first was a variable indicating if the participant reported a same-race close friendship. This variable was dummy coded, such that the lack of a same-race close friendship was the reference category. The second was a count of the number of other-race close friendships reported. For example, a White girl who reported having a close friendship with at least one Black person and at least one Hispanic person would have a value of two.

Self-criticism. Participants completed a 24-item survey, "How I Deal With Things", at the second time point. Participants were asked to imagine something happening, such as having

something bad happen or somebody their age being mean or threatening. Reponses to each item were on a five-point Likert scale, with each point indicating how frequently ("very often", "often", etc.) the participant handled a problem in a certain way. Three of these items pertained to self-blame or having brought the problem upon oneself, which were summed to create a composite self-criticism score. Higher values of this composite variable indicated more frequent engagement in self-criticism when facing a problem.

Unhappiness with body. At the first time point, each participant was asked about her unhappiness with her body, height, weight, and skin color. These were evaluated on a four-point Likert scale, such that higher values meant greater unhappiness.

Model specification – functional form

Model selection started with the selection of the functional form of the trajectory of the outcome under examination. The Level 1 equations for each of the functional forms tested are shown in Table 4.1. The linear, quadratic, and cubic models were each specified to have a random effect associated with the intercept, corresponding to b_o in Table 4.1. The exponential model was specified to have a random effect associated with the initial value of the outcome, which corresponds to b_0 in its formulation. The logistic model had three potential specifications at Level 2. The first logistic model was specified to have a random effect associated with the upper asymptote, corresponding to b_1. The second was specified to have a random effect associated with the placement of the logistic form's inflection point, corresponding to b_2. The third was specified to have a random effect associated with the "rapidity" of increase toward the upper asymptote, corresponding to b_3,. In total, seven functional form models were fit to the data and all models that successfully converged and produced estimates with no errors were compared.

Model specification – predictor sets

Once the best functional form was selected, a second round of model selection was conducted to determine the appropriate predictor set. Seven models were compared at this stage. The first model, referred to hereafter as the base model, included the functional form variables and the set of covariates. BMI was a time-variant covariate entered at Level 1. Parent's education category, the parent's income category, the participant's race, and the participant's age at menarche were time-invariant covariates entered at Level 2. All of the succeeding models included these covariates along with the predictors sets of interest.

The second model and third models, the specific criticism and the general criticism models, included sources of weight criticism from others. The specific criticism model added two predictors, one for how many sources told the participant that she was too fat and another for how many sources told her she was too thin. The general criticism model added a single predictor, which was the sum of the total sources of criticism. The fourth model was the valence model, which had four additional predictors (positivity toward fatness/thinness, negativity toward fatness/thinness). The fifth model was the diversity model, which included the presence of a same-race close friend, the number of other-race close friendships, and whether or not the participant perceived herself as a minority in her school. The sixth model was the self-criticism model, which included the self-criticism composite score. Finally, the seventh model was the unhappiness with appearance model, which included the unhappiness with body, height, weight, and skin color as separate predictors.

Model specification – covariance structure

A third round of model selection was conducted to determine the best Level 1 covariance structure for the selected functional form and predictor set for each outcome. Up to 10

covariances structures, consisting of one of three types, were specified and compared. Examples of these three types are shown in Table 4.2. Table 4.2 displays structures that match those fit to the body dissatisfaction outcome, which was measured at five time points; the same structures were also used for the body image outcome, which was measured at ten time points. The first structure type was a conditional independence structure, which estimates a single parameter for within-person variance. The second structure type was a first-order autoregressive structure, which also estimates a single parameter (ρ), but models autocorrelation by exponentiating this parameter across time points. For example, in the case of five time points, the rho parameter estimate for the correlation between the first and fifth time points would be raised to the fourth power. The third structure type was a Toepliz-banded structure. With this structure, one can potentially estimate any number of bands (i.e. off-diagonal elements of the same degree) up to one less than the total number of time points independently while imposing the restriction that the within-band values must be equal. Any bands that the researcher chooses not to estimated are set to zero. If the selected model had a linear or polynomial functional form, these structures were fit using the MIXED procedure. For this type of model, all possible number of bands were estimated. If the selected model was instead a "truly" nonlinear model, these structures were fit using code written by Harring and Blozis (2014) to adapt NLMIXED to model autoregressive and a single-banded ("tridiagonal") Toepliz structure.

Model selection

To evaluate which functional form to proceed with, AIC, AICC, BIC, CAIC, and HQIC values were computed for each model. For the sample size-dependent criteria, both sample size sources were used. At each model selection step, all nine information criteria values were computed for each model. In the case of all nine information criteria unanimously selecting one

model, that model was chosen. If different models were selected by different information criteria, those models were also fitted.

Results – body image discrepancy

Model selection for functional form. Six functional form models – linear, quadratic, cubic, exponential, logistic with a random upper asymptote parameter (b1), and logistic with a random "rapidity" parameter (b2) – were successfully estimated. The information criteria values for this step of model selection are shown at the top of Table 4.3. All nine information criteria selected the cubic model as the "best" model, as shown in Figure 1. The predicted body image discrepancy was about two at the first time point, dropping to its lowest point within the observed range between four or five years into the study (between 14 and 15 years of age). By the final time point, the discrepancy was near what it was at the beginning.

Model selection for predictor set. All seven models comparing predictor sets were successfully estimated when included with the cubic functional form. All nine information criteria selected the diversity model as the "best" model. The information criteria values for this step of model selection are shown in the middle of Table 4.3.

Model selection for covariance structure. All ten models comparing covariance structure specifications were successfully estimated in MIXED when the diversity predictor set was used. All nine information criteria selected a Toepliz-banded covariance structure, but the number of bands estimated differed slightly. The three efficient information criteria (AIC, AICC-N, and AICC-m) selected the model with eight estimated bands. The six consistent criteria (BIC-N, BIC-m, CAIC-N, CAIC-m, HQIC-N, and HQIC-m) selected the model with nine estimated bands. The information criteria values for this step of model selection are shown at the bottom of Table 4.3.

Final model. The three-step model selection process for the functional form, predictor set, and covariance structure for the body image discrepancy outcome resulted in two models. Both models had a cubic functional form and incorporated the diversity predictor set, which is shown in single-equation form below. The parameter estimates and standard errors, which are shown at the top of Table 4.4, were very close for both of these models.

$$Y_{ij} = b_{oj} + b_{1j}(CenAge) + b_{2j}(CenAge^2) + b_{3j}(CenAge^3) + b_4(BMI) + b_5(RACE) + b_5(INC2) + b_6(INC3) + b_7(INC4) + b_8(EDU1) + b_9(EDU2) + b_{10}(MedMenarche) + b_{11}(SameRace) + b_{12}(OtherRace) + b_{13}(Minority) + u_{0j} + e_{ij}$$

For both models, the three functional form predictors – the linear ($p < 0.001$), quadratic ($p < 0.001$), and cubic ($p < 0.001$) terms – were significant. In addition, three covariates were significant. The first was BMI ($b_4 = 0.069$, $SE = 0.002$, $p < 0.001$), where higher BMI was predictive of greater body image discrepancy after controlling for the functional form and all other predictors. The second was race ($b_5 = -0.192$, $SE = 0.022$, $p < 0.001$), where Black girls were predicted to have lower body image discrepancy than White girls after controlling for the functional form and all other predictors. The third was parental educational attainment ($b_8 = -0.053$, $SE = 0.026$, $p = 0.039$; $b_9 = -0.117$, $SE = 0.029$, $p < 0.001$), where girls whose guardians had not completed high school were predicted to have lower body image discrepancy than those whose guardians had higher educational attainment after controlling for the functional form and all other predictors. None of the predictors of interest in these models – the presence of a same-race close friend, the number of other-race close friendships, and whether or not the participant perceived herself as a minority in her school – were significant. To determine the interrelatedness of the diversity variables, correlations were computed. The correlations between

having a friend of the same race and being a school minority (r = -0.1, $p < 0.001$) and being a school minority and having other race friends (r = -.06, $p < 0.001$) were significant.

Results – body dissatisfaction

Model selection for functional form. Six functional form models – linear, quadratic, cubic, exponential, logistic with a random upper asymptote parameter (b1), and logistic with a random intercept-like parameter (b2) – were successfully estimated. All nine information criteria selected the logistic model with the random upper asymptote parameter (b1) as the "best" model. The information criteria values for this step of model selection are shown at the top of Table 4.5. At the beginning of the study, the expected body dissatisfaction score was approximately 4. As time passed during the study, the expected body dissatisfaction increased, with the most acceleration appearing to occur during the first half of the study (approximately between 9 and 15 years of age).

Model selection for predictor set. All seven models comparing predictor sets were successfully estimated when included with the logistic functional form with a random b2. All nine information criteria selected the diversity model as the "best" model. The information criteria values for this step of model selection are shown in the middle of Table 4.5.

Model selection for covariance structure. Because the selected functional form was "truly" nonlinear, I adapted code written by Harring and Blozis (2014) to fit an autoregressive and a tridiagonal Toepliz structure to both the functional form only and the functional form with the diversity predictor set. While some of these models converged and produced estimates, the ones that did so produced a warning that a second order optimality condition had been violated, making the estimates produced by those models untrustworthy. This particular problem can sometimes be resolved by rescaling parameters (Kiernan et al., 2009), which was used

successfully in the simulation studies to fit correctly specified and misspecified logistic models. Unfortunately, rescaling did not resolve this issue in this application. Because of this, the covariance structure remained a conditional independence structure, which is the only covariance structure that can be specified in NLMIXED by default.

Final model. The model selection process for the functional form and the predictor set for the body dissatisfaction outcome resulted in one model, which had a logistic functional form and incorporated the diversity predictor set. Due to its length, this equation is not shown here. The parameter estimates and standard errors for the final model are shown in Table 4.6. Of the three functional form predictors, only b2 ($b_2 = -4.758$, $SE = 1.351$, $p < 0.001$), the intercept-like parameter, was significant. In addition, four covariates were significant. The first was BMI ($b_4 = 0.672$, $SE = 0.017$, $p < 0.001$), where higher BMI was predictive of greater body dissatisfaction after controlling for the functional form and all other predictors. The second was race ($b_5 = -4.280$, $SE = 0.986$, $p < 0.001$), where Black girls were predicted to have lower body dissatisfaction than White girls after controlling for the functional form and all other predictors. The third was parental educational attainment ($b_5 = -1.053$, $SE = 0.470$, $p = 0.025$), where girls whose guardians had not completed high school were predicted to have lower body dissatisfaction than those whose guardians had completed college after controlling for the functional form and all other predictors. The fourth was the age of menarche ($b_4 = -0.437$, $SE = 0.151$, $p < 0.004$), where experience menarche at higher ages was predictive of lower body dissatisfaction after controlling for the functional form and all other predictors. None of the predictors of interest in these models – the presence of a same-race close friend, the number of other-race close friendships, and whether or not the participant perceived herself as a minority in her school – were significant. To determine the interrelatedness of the diversity variables,

correlations were computed. The correlations between having a friend of the same race and being a school minority ($r = -0.1$, $p < 0.001$) and being a school minority and having other race friends ($r = -.06$, $p < 0.001$) were significant.

Discussion

The final models selected for body image discrepancy and body dissatisfaction have several interesting similarities, one of which is that they share several significant predictors. Specifically, the final models all included BMI, race, and the dummy variable indicating whether the participant's guardian had completed college. The direction of the relationship for all of these predictors was also the same across the final models, such that, controlling for all other predictors, a White girl with a higher BMI whose guardian completed college would be expected to have both greater body image discrepancy and express greater body dissatisfaction than a Black girl with a lower BMI whose guardian had not completed college. There were, however, some differences; body dissatisfaction alone had age of menarche as a significant predictor, and both parental education dummy variables were significant only when the outcome was body image discrepancy.

Another interesting similarity is that the diversity predictor set was selected as the "best" model after the functional form was determined. In addition, for both of these outcomes, none of the predictors that made up the diversity predictor set were significant. A result like this can be jarring but is does not inherently mean that something is amiss with the model selection process. It is likely that the set of these three predictors - the presence of a same-race close friend, the number of other-race close friendships, and whether or not the participant perceived herself as a minority in her school - are jointly able to account for more variability in both outcomes than the other sets of predictors but none of the individual predictors account for enough unique (i.e.,

non-overlapping) variability in the outcomes for any of them to be significant when the others are in the model.

Finally, for both outcomes, the best-fitting functional form was some kind of nonlinear model. Body image discrepancy was modeled using a cubic trajectory, which is not "truly" nonlinear but does allow for the temporal change in body image discrepancy to fluctuate in a nonlinear fashion across late childhood through adolescence. Body dissatisfaction was modeled using a logistic trajectory, which is a "truly" nonlinear trajectory. Among the functional-form only models using both outcomes, the linear functional form models were among the worst fitting of all the forms tested. That is, even though cubic and logistic functional forms were less parsimonious than the linear functional form, the increase in predictive power was such that these more complex models were unanimously selected across both efficient and consistent criteria.

The longitudinal nature of this application meant that fitting and testing different covariance structures could potentially be of interest to a substantive researcher. In addition, previous simulation work has demonstrated that misspecification of the time-specific residual covariance structure can result in both the estimates of the variance components and the standard errors of the functional form estimates to be too high (Kwok et al., 2007). Although not all possible covariance structures were fit to model with the body image discrepancy as the outcome, the ones most likely to be fit in practice were well-represented. Modeling an eight- or nine-banded Toepliz structure is impractical in many behavioral research applications, but it was possible because thousands of participants had data over many time points; it might not be replicable unless another study was similarly large. The conditional independence structure used for the body dissatisfaction outcome, which was chosen due to software difficulties and not as a result of

theory or a model selection method, is almost certainly underspecified (i.e. too simple). Per Kwok, West, and Green (2007), it is likely that statistical tests of the logistic functional form parameters were underpowered compared to a model with the correct (or more correct) covariance structure specification.

As the results of this application demonstrate, nonlinear longitudinal trends exist in behavioral data and can be explicitly modeled using a variety of nonlinear functions. If a researcher had only modeled these outcomes using linear models, the interesting nonlinearity that exist in these data would have been entirely missed. Information criteria are easy to use and allow researchers to engage in model selection when the candidate models are non-nested, which is always true when comparing linear or polynomial functional forms to "truly" nonlinear functional forms, such as those used in this application.

6. Application: The Effects of Sleep, Interpersonal Interactions, and Demands on Daily Distress in Adolescents

The sleep needs of adolescents have gotten an increasing amount of attention from researchers and public policy makers. The American Academy of Sleep Medicine (Paruthi et al., 2016) recommended that children between the ages 13 and 18 get eight to ten hours of sleep each day on a regular basis for optimal health. Later research has shown that there may not be a single "best" amount of sleep for adolescents. Fuligni, Arruda, Krull, and Gonzales (2018) found that the amount of sleep that for which symptomatology (internalizing and externalizing behaviors) was lowest and academic achievement was highest differed by more than an hour. In addition, they found that greater *variability* in adolescent an adolescent's sleep duration was predictive of greater symptomatology. Despite recommendations about the importance of sufficient sleep, sleep deprivation in adolescents is widespread. A nationally-representative sample of over 12,000 high school students found that 68.9% slept less than 8 hours on the average school night, with 38% sleeping 6 hours or less (Eaton, McKnight-Eily, Lowry, Perry, Presley-Cantrell, & Croft, 2010). Previous research has found that insufficient sleep has several negative consequences for adolescents, including obesity, depression, and higher rates of drowsy driving accidents (Owens, Adolescent Sleep Working Group, & Committee on Adolescence, 2014).

A systematic review of literature concluded that there is evidence for a bidirectional relationship between sleep and depression and anxiety, with insomnia potentially being a better predictor of depression than vice versa (Alvaro, Roberts, & Harris, 2013). This idea was further supported by an experimental study (Baum, Desai, Field, Miller, Rausch, & Beebe, 2014), in which adolescents were in bed for 10 hours for one week and for 6.5 hours in another week. The adolescents reported feeling more irritable, anxious, angry, confused, and fatigued during the

week they slept less than the week they slept more. Fuligni, Bai, Krull, and Gonzales (2019) found that adolescents with higher internalizing and externalizing symptoms (as measured at the beginning of each wave) needed more sleep than those with lower symptoms to minimize next-day distress. They were able to come to this intriguing conclusion by fitting a three-level model (repeated measures nested within waves nested within individuals) that allowed sleep to relate to next-day distress quadratically. The normative curve, which was quadratic, was representative of the majority of participants' individual trends, with the minimum of the function being the person-specific number of hours of sleep at which next-day distress was minimized.

In addition to sleep, there is evidence that there are other factors that influence psychological functioning in adolescents. One factor is that of day-to-day stressful events. Higher amounts of daily hassles, or frustrations encountered in daily living, have been linked to greater daily emotional distress in young adults (D'Angelo & Wierzbicki, 2003). Experiencing more daily hassles during adolescence was also predictive of future diagnosed psychopathology up to ten years later (Asselman, Wittchen, Lieb, & Beesdo-Baum, 2017). Another factor that affects adolescent psychological functioning is the quality of interpersonal relationships. Previous research has shown that positive (e.g., companionship) and negative (e.g., conflict) qualities of relationships with friends, romantic partners, and parents affect depression, social anxiety, and emotional distress in adolescents (Kenny, Dooley, & Fitzgerald, 2013; La Greca & Harrison, 2005). Due to the presence of these other factors, it is probable that a model using just sleep to predict psychological functioning in adolescents would be underspecified. Because information criteria permit model selection among non-nested sets of predictors, it is possible to directly compare the predictive utility of daily stressors and daily interpersonal interactions by including models with different predictors in the set of candidate models.

In the two simulation studies and in the previous application, the analytic context was that of a longitudinal study where the change trajectory of interest occurred across time. As shown by Fuligni and colleagues (2019), a nonlinear functional form may be more suited for modeling the relationship between the time spent sleeping and distress. Because of the presence of nonlinearity and because these variables had sufficient range to potentially fit a variety of functional forms, this application used a data set similar to that of Fuligni and colleagues to demonstrate the utility of using information criteria to select among nonlinear functional forms and predictor sets. To ensure that several functional form models would converge and produce interpretable estimates, the models used in this application were simpler than those fit by Fuligni and colleagues. For the same reason, this application used one wave of data and had a two-level multilevel structure, while the study by Fuligni and colleagues used two waves and had a three-level structure.

Method

The UCLA Study of Adolescents Daily Lives' was a large-scale diary study of the psychosocial experiences of ethnically and socio-economically diverse youth in California. The data collected as part of this study were especially rich, including demographic, psychosocial, and interpersonal information collected from adolescents for between one and four years of high school. Students in three high schools were invited to participate, and all students who returned completed parental and personal consent forms were allowed to participate in the study. In 9^{th}, 10^{th}, 11^{th} , and 12^{th} grade, each participant completed a 45-minute questionnaire while at school. In 9^{th}, 10^{th}, and 12^{th} grade, each participant was also given instructions and materials for a two-week diary study. Students who were not included in previous data collection waves were allowed to enter the study in later grades. The data used for this study came from participants in

the 12th grade. Consent from students and their parents was obtained either in an earlier grade (for a returning participant) or upon entry to the study in the current year. At the end of the two-week data collection period, the diaries were collected at school and students who completed at least one diary received $30 in cash and students who completed no diaries received $10 in cash. Grades and course enrollment information was also collected at that point. Students who completed most of their diaries were sent two movie tickets, and one student from each school who completed all of their diaries was randomly selected to win a $100 Borders gift card.

Participants

A total of 681 participants participated in the 12th grade study. Participants were instructed to complete the daily diary before going to bed each night, fold it in half, and seal it with a provided sticker. Participants who had been given time stampers were then to use the stamp on the seal. Those who were not given time stampers instead wrote the date and time of completion on the seal. Diary completion rates across the 14 days were high, with the highest diary completion rate (>99%) occurring on the first day and the lowest diary completion rate (91.6%) occurring on the fourteenth day.

Measures – outcome and functional form

Daily distress. As part of the daily checklist, adolescents and parents completed the anxiety and depression sub-scales of the Profile of Mood States (POMS; Lorr & McNair, 1971). Anxious feelings included feeling on edge, nervous, uneasy, and unable to concentrate. Depressive feelings included feeling discouraged, hopeless, and sad. Participants answered each question on a 5-point scale, where a 1 indicated not having experienced a particular depressive feeling and a 5 indicated having experienced a particular depressive feeling in an extreme way. For data analysis purposes, a value of one was subtracted from each participants' response to

these questions so that a participant who reported having none of a feeling that day would have a score of zero. The subscales were then summed to create an index of daily psychological distress.

Previous night's sleep. For each night of the study, participants were asked to report the number of hours and minutes they slept the night before. Responses were open-ended, such that participants wrote in the amount directly.

Measures – covariates

Age. Participants reported their birthdays, which were converted to age in years at the beginning of this wave of the study. For data analysis purposes, this variable was centered around the grand mean of ages reported by the participants (M = 17.79).

Gender. Participants reported their gender (boy or girl) at the beginning of the study. This variable was dummy coded with boy as the reference category. There were 337 boys and 404 girls included in the sample.

Ethnicity. At the beginning of the study, each participant was presented with a list of 44 ethnic labels and asked to endorse any of them for which the participant identified. In addition, participants could list additional ethnic labels that were not on the provided list. For the purposes of this study, a participant who endorsed any of the following labels was considered Latino: Brazilian, Central American, Chicano/a, El Salvadoran, Guatemalan, Hispanic, Hispanic-American, Honduran, Latino/a, Latino/a-American, Mexican, Mexican-American, Nicaraguan, Nicaraguan-American, Spanish, Spanish-American, and any label written in by a participant that was akin to a label on the list (e.g., Puerto Rican). All others were considered to be non-Latino. This variable was dummy coded such that non-Latino was the reference category. There were 451 students who did not endorse a Latino ethnic category and 293 who did.

Depressive symptoms. At the beginning of the study, each participant completed the Center for Epidemiologic Studies Depression (CES-D) scale, a 20-question survey which asks participants about the frequency of various feelings and experiences over the past month. Responses could range from 1 (rarely) to 4 (most or all of the time). Four of the items were phrased such that a higher response indicated less depressive symptomatology, so those items were reverse coded. Each participant's responses to the CES-D were summed to create an index of depressive symptoms such that higher scores meant more depressive symptoms. The mean summed CES-D score was 1.93 (SD = .53).

School/Weekend night. For each daily diary, the participants were asked to circle the day of the week it was when they filled out the checklist. If the participant circled either Saturday or Sunday, the diary was considered to have been filled out on a weekend night. The diary was considered to have been filled out on a school night if the participant circled any other day of the week. This variable was dummy coded such that weekend nights were the reference category.

Measures – predictors of interest

Daily family assistance behaviors. Participants were asked whether or not they engaged in any of eight family assistance behaviors that day. These included helping clean the house or apartment, taking care of siblings, running an errand for the family, helping siblings with schoolwork, helping parents with official business, helping cook a meal for the family, helping parents at their workplace, and anything else done to help the family. These items were summed to create an index of daily family assistance, ranging from 0 (reported providing no assistance to their families that day) to 8 (reported doing all seven of the acts for which a specific question was asked and an additional act of family assistance). Participants were also asked how much time in total they spent on all of the family assistance activities they reported.

Daily family leisure activities. Participants were asked whether or not they engaged in any of three family leisure activities that day. These included eating a meal with family, spending leisure time with family, and spending time with aunts, uncles, cousins, or grandparents. Participants were also asked how much time in total they spent on all of the family leisure activities they reported.

Daily school activities. Participants were asked whether or not they engaged in any of three school-related activities that day. These included doing homework while at school, doing homework while not at school, participating in extracurricular activities after school. Participants were also asked how much time in total they spent on these activities.

Daily friend activities. Participants were asked whether or not they engaged in any of three activities with friends that day. These included spending time with friends outside of school, talking on the phone with friends, and emailing or instant messaging friends. Participants were also asked how much time in total they spent on these activities.

Daily job activity. Participants were asked if they had worked at a job that day. They were also asked how much time in total they worked.

Daily perceived demand. Each day, participants were asked if they had a lot of work to do at home, at school, or a job. They were also asked if they had a lot of demands made by family or friends on that day. These five items were summed to create an index of the perceived demandingness experienced by the participant.

Daily negative interpersonal events. Each day, participants indicated whether they had experienced negative interactions with family members, friends, adults at school, other students at school, or adults outside of school. Examples included arguing with mother/father/other family member about something, arguing with a close friend/boyfriend/girlfriend, having an argument

or being punished by an adult at school, and being harassed/picked on/teased by a student at school or someone outside of school. In addition, participants were asked if a student at school, an adult at school, or someone outside of school treated the participant poorly because of the participant's race. The number of affirmative responses to these items were summed to create an index of negative interpersonal interactions, ranging from 0 to 12, such that higher scores indicated the occurrence of more negative interpersonal events.

Daily positive interpersonal events. Each day, participants indicated whether they had experienced positive interactions with family members, friends, and adults at school. Examples included getting along with parents, getting along with friends, and getting along with adults at school. In addition, participants were asked if a student at school, an adult at school, or someone outside of school treated the participant well because of the participant's race. The number of affirmative responses to these items were summed to create an index of positive interpersonal interactions, ranging from 0 to 8, such that higher scores indicated the occurrence of more positive interpersonal events.

Model specification – functional form

Model selection started with the selection of the functional form of the trajectory of the outcome under examination. The Level 1 equations for each of the functional forms match those shown in Table 4.1, except that the functional form predictor was hours of sleep instead of time. The linear, quadratic, and cubic models were each specified to have a random effect associated with the intercept, corresponding to b_o in Table 4.1. The exponential model was specified to have a random effect associated with the initial value of the outcome, which corresponds to b_0 in its formulation. Three logistic models were also fit, each with a single random effect. The first logistic model was specified to have a random upper asymptote, corresponding to b_1. The second

was specified to have a random parameter loosely related to the intercept, corresponding to b_2. The third, corresponding to b_3, was specified to have a random "rapidity" parameter. In total, seven functional form models were fit to the data. All models that converged and produced interpretable parameter estimates were compared using information criteria.

Model specification – predictor sets

Once the best functional form was selected, a second round of model selection was conducted to determine the appropriate predictor set. Seven models were compared at this stage. The first model, referred to hereafter as the base model, included the functional form variables and the set of covariates. CES-D score, gender, centered age, and ethnicity were time-invariant covariates entered at Level 2. Whether the diary was completed on a school day or a school night was entered as a time-varying covariate at Level 1. All of the succeeding models included these covariates along with the predictors sets of interest.

A summary of the six models with the predictor sets of interest is shown in Table 5.1. The first three models in the series were the demand models. The first of these was the perceived demand model, which included predictors indicating if the participants felt that they had a lot of work at school, work, or home that day, as well as if they felt a lot of demands from family or friends. The second was the event demand model, which included predictors indicating if the participant had engaged in tasks within the home, leisure activities with their families, activities at school, activities with friends, and whether they worked a job that day. The third was the time demand model, which included predictors of the time spent doing the things in the event demand model. The difference between the event demand model and the time demand model was that focus of the event demand model was on the number of tasks in which the participant engaged, and the focus of the time demand model was on the amount of time spent doing different tasks.

The second series of predictors sets were the interpersonal interactions models. The first was the negative interpersonal interactions model. This model included predictors indicating if participants had negative interactions with family members, friends, adults at school, other students, or others outside of school. In addition, they were asked if an adult at school, a student at school, or someone outside of school had treated them badly because of their race. The second of these was the positive interpersonal interactions model. This model included predictors indicating if participants had negative interactions with family members, friends, and adults at school. Participants were also asked if an adult at school, a student at school, or someone outside of school had treated them well that day because of their race. The combination model included all of the predictors of the negative and positive interaction models.

Model specification – covariance structure

Fitting models with nonlinear functional forms is largely the same when using either time or another variable of interest to create the nonlinear change trajectory. One important difference, however, is that the values of the input variable change from being discrete to being continuous. When time is used as the functional form variable, the within-person covariance structure represents the correlation of within-person observations across time points. When a non-temporal variable is used, the within-person covariance structure represents the correlation of the values of the observations across the range of values of that variable. For example, if time were the variable used to construct the functional form, the off-diagonal elements of the L1 covariance matrix would represent the correlation between measures at different time points. If a variable like hours of sleep were used, then the elements of the L1 covariance matrix would represent the correlation between each discrete value of reported sleep. In the latter case, the conditional independence structure was the L1 covariance structure that was most sensible in

context, so the conditional independence structure was used for the final model.

Model selection

To evaluate which functional form and predictor set to proceed with, AIC, AICC, BIC, CAIC, and HQIC values were computed for each model. For the sample size-dependent criteria, both sample size sources were used. At each model selection step, all nine information criteria values were computed for each model.

Results

Model selection for functional form.

Of the seven functional form models that were fit, five converged and produced interpretable estimates. Those five were the linear, quadratic, cubic, exponential, and the random upper asymptote logistic models. The predicted daily distress for each of these models across the range of hours of sleep can be found in Figure 5.1. All of the models predicted a downward trend in distress between zero and ten hours of sleep. Once the number of hours of sleep was higher than ten, the curves began to noticeably diverge. The linear model had a negative slope, so the predicted distress values continued to steadily decrease. The cubic and quadratic models predicted an upswing in distress when sleep was greater than 10 hours. The predicted distress values from the logistic and exponential models were almost entirely identical within the observed range of sleep, showing a downward trend that began leveling off when sleep was greater than 10 hours. The nine information criteria values for the functional form-only models are shown at the top of Table 5.2. All information criteria selected the exponential model.

Model selection for predictor set.

Because the exponential functional form was selected as the best model in the previous step, all seven models described previously and shown in Table 4.1 were fitted with an

exponential functional form. The nine information criteria values are shown at the bottom of Table 5.2. All of the information criteria selected the time demand model.

Final model.

The final model was the time demand model with an exponential functional form. The L1 equation for this model is as follows:

$$Y_{ij} = b_{0i}e^{b_{1i}Sleep_{ij}} + b_3(age) + b_4(gender) + b_5(CESD) + b_6(Latino) + b_7(schoolday) + b_8(family\ task\ time) + b_9(family\ leisure\ time) + b_{10}(school\ activity\ time) + b_{11}(friend\ time) + b_{12}(job\ time) + \varepsilon_{ij} \ .$$

A table of estimated coefficients and their associated p-values is shown in Table 5.3. Both the intercept (b_{0i}) and the exponential growth (b_{1i}) parameters of the exponential functional form variables were significant ($p < 0.001$). Of the covariates, ethnicity ($b_6 = 0.51, p = 0.02$) and whether the day was a school day or not ($b_7 = -.88, p < 0.001$) were significant. All else held constant, students were predicted to experience less distress on weekend days and Latino students were predicted to have less distress than non-Latino students. Two predictors in the time demand model were significant. The first was the amount of leisure time spent with family ($b_9 = -0.08, p = 0.004$) and time spent on school activities ($b_{10} = -0.08, p < 0.001$). All else being equal, more leisure time spent with family and more time spent on school activities were predictive of less distress.

Discussion

The results of this application demonstrate that nonlinear functional forms in multilevel models have utility that extends beyond modeling change trajectories based on the passage of time and that information criteria can be used to identify such forms. Although likelihood ratio tests could have been used to compare some of the predictor sets (e.g., the positive and negative

interactions models could have been compared to the all interactions model), only by using information criteria could the six predictor sets of have been tested simultaneously.

These findings suggested that getting more sleep reduced daily distress in adolescents, but also that this reduction in distress slowed once the amount of sleep was greater than 10 hours, a value greater than the higher end of the amount of sleep recommended for adolescents by the American Academy of Sleep Medicine (Paruthi et al., 2016). The logistic and exponential models produced almost identical deviances, but the exponential model was selected because it was more parsimonious. In contrast, the selection of the predictor set was mostly driven by larger differences in the deviances. All three demand models were the most parsimonious of the predictors sets of interest, so the differences in their deviances was what distinguished them during model selection. The predictor set findings suggested that the amount of time spent on various activities was more predictive of daily distress than the number of activities, the perceived demandingness of those activities, and positive or negative interactions with others in daily life. Specifically, adolescents experienced less distress on days when they spend more leisure time with family members or spend more time on extracurricular activities.

Although the simplicity of these models was intentional to maximize the number of estimable and interpretable truly nonlinear functional forms, the fact that these models were almost certainly underspecified (i.e., too simple) is a major limitation of the findings from this application. First, there were neither interactions among the predictors in the predictor sets nor interactions between the predictors and the functional form variables. Second, across the different functional form models, only one of the functional form parameters in each model had an associated random effect. For example, the quadratic analysis only had a random intercept parameter; such a specification allows for individual-specific intercept values but assumes that

the linear and quadratic trends are the same across individuals. Given these limitations, the fact that the conclusions regarding the functional form in this application and that of Fuligni and colleagues (2019) differ slightly – this application concluded that an exponential functional form was the best, while Fuligni and colleagues employed a quadratic functional form in order to be able to detect minima – should not be taken as a conflict. Rather, the findings from this application further confirm the existence of a nonlinear relationship between sleep and daily distress in these data by ruling out the linear model through an alternative model selection method. Whether examining the relationship by comparing the linear model to a nonlinear model using a likelihood ratio test or by using information criteria, the linear model was the worst-fitting model among those tested here.

7. Discussion

The scientific method is often conceptualized as a cycle in which researchers develop hypotheses based on theory derived from prior observations and evaluate those hypotheses based their utility in predicting new observations. Researchers in the behavioral sciences frequently express hypotheses by building representative statistical models and test those hypotheses by fitting models to data collected for that purpose. When a researcher wants to compare the plausibility of two or more competing hypotheses, the models representing those hypotheses can be compared using a model selection method. Information criteria offer a flexible framework for selecting among multilevel models because the models being compared do not have to be nested. This is especially useful when the models being compared are truly nonlinear models, such as exponential and logistic models, because such models cannot be nested relative to each other or to polynomial models. To date, there has been little research on the performance of information criteria when selecting models that contain truly nonlinear functional forms, so no empirically-based guidelines for the use in this context currently exist for applied researchers. The work presented here represents some first steps in the creation of such guidelines.

The goal of the first study in this series was to empirically examine the ability of different information criteria to select the model with the correct functional form specification when the candidate model set included linear, polynomial, and truly nonlinear models. The ability of information criteria to detect correctly specified nonlinear models was affected by the L1 sample size, the L2 sample size, ICC, and the distinctiveness of the underlying functional form. The correct selection rate was generally higher as the number of L2 units and the number of L1 units increased. Correctly specified models were also more likely to be selected when the ICC was higher than when it was lower. BIC(N) may be better for identifying correctly specified

exponential, logistic, and quadratic models when their respective functional forms are more distinct, while AIC may be better when these forms are less distinct. Future work in this area may include the exploration of other truly nonlinear functional forms. For example, Singer and Willett's (2003) discussion about truly nonlinear functional forms also included hyperbolic and inverse polynomial functions. In addition, an examination of a broader range of distinctiveness among functional forms would help generalize the current findings.

The goal of the second study in this series was to empirically examine the performance of different information criteria when selecting among different predictor sets when the underlying functional forms were either exponential or logistic. Two outcomes were of interest. The first outcome of interest, the ability of different information criteria to select a correctly specified model in a candidate model set, was a standard outcome of interest in research about the performance of information criteria (e.g., Gurka, 2006; Whittaker & Furlow, 2009; Vallejo et al., 2011). The findings for this conventional outcome suggested that information criteria may have difficulty identifying a model that is correctly specified with regard to both its truly nonlinear functional form and its predictor set if alternative misspecified models are "close enough". In this case, efficient criteria (AIC and AICC) performed better than their consistent counterparts, likely due to their property of dimension inconsistency (Bozdogan, 1987). The second outcome of interest, the ability of different information criteria to select a "more correct" (less misspecified) model when the correctly specified model was not in the candidate model set, was a less conventional outcome but may better match the circumstances of behavioral research because the fitting of models that are fully correctly specified may be unrealistic. In most cases, both efficient and consistent information criteria were able to identify the model determined to be more correct between two models with misspecification in their predictor sets. Future work in

this area could include greater degrees of misspecification or include misspecification with regard to both the predictor sets and the functional forms simultaneously.

Finally, the utility of these models was demonstrated though their application to data collected from participants in two longitudinal behavioral health studies. Nonlinear trajectories are often applied when the trajectory of interest is change over time, such as the case of the changes in within-person body image discrepancy and body dissatisfaction over the course of late childhood and adolescence in girls. Such models can also be applied when the trajectory of interest is not temporal in nature, such as the case of changes in daily distress across within-person ranges of sleep. In both of these cases, model selection using information criteria revealed nonlinear trends. These substantively interesting trends would have been missed if the fitted models were limited to the standard linear multilevel model. Across behavioral research studies, it seems likely that some substantively interesting nonlinear trends are missed because researchers do not realize that these nonlinear trends exist in their phenomena of interest (perhaps because prior research had been limited to the exploration of linear trends) or because researchers may have difficulty modeling a particular nonlinear hypothesis statistically. In light of the technical difficulties that occurred throughout the studies in this dissertation, the latter situation seems particularly likely in the case of a hypothesis involving a truly nonlinear functional form. To encourage the inclusion of truly nonlinear models in behavioral research, quantitative researchers may want to consider providing code, such as Harring and Blozis' (2014) demonstration of how to fit different L1 covariance structures in the NLMIXED procedure in SAS, or detailed didactic examples to help bridge this gap.

8. Tables

Lower effect size	Exponential	Logistic	Quadratic
	b0: 5.35	b1: 20.06	b0: 5.35
	b1: 0.035	b2: 20	b1: 0.19
		b3: 20	b2: 0.0035
Higher effect size	Exponential	Logistic	Quadratic
	b0: 0.5	b1: 22	b0: 1
	b1: 0.28	b2: 10	b1: 0.05
		b3: 2.5	b2: 0.1

Table 2.1. Set of data generation coefficients used for lower and higher distinctiveness in functional forms

	Simple differences (between-models)			Area under the curve			AUC differences (between-models)		
	Log vs Expo	Log vs Quad	Expo vs Quad						
	Mean	Mean	Mean	Log	Expo	Quad	Log vs Expo	Log vs Quad	Expo vs Quad
Exponential form, lower effect size, ICC = 0.4	-0.07	-0.07	0.00	79.31	80.14	80.12	-0.83	-0.82	0.02
Exponential form, higher effect size, ICC = 0.4	-2.83	-2.92	-0.09	24.50	63.44	63.61	-38.94	-39.11	-0.17
Exponential form, lower effect size, ICC = 0.6	-0.15	-0.15	0.00	78.45	80.28	80.26	-1.83	-1.81	0.02
Exponential form, higher effect size, ICC = 0.6	-2.81	-2.87	-0.06	12.79	51.40	51.46	-38.61	-38.67	-0.06
Logistic form, lower effect size, ICC = 0.4	0.01	0.01	0.00	80.10	79.96	80.03	0.14	0.07	-0.07
Logistic form, higher effect size, ICC = 0.4	-0.17	-0.03	0.13	63.89	65.81	64.23	-1.92	-0.35	1.58
Logistic form, lower effect size, ICC = 0.6	0.01	0.01	0.00	80.13	79.94	80.01	0.18	0.12	-0.07
Logistic form, higher effect size, ICC = 0.6	-0.27	-0.10	0.17	64.06	67.06	65.15	-3.01	-1.09	1.91
Quadratic form, lower effect size, ICC = 0.4	-0.03	-0.02	0.00	79.36	79.65	79.65	-0.29	-0.29	0.00
Quadratic form, higher effect size, ICC = 0.4	-0.03	0.05	0.08	106.05	106.41	105.70	-0.36	0.36	0.71
Quadratic form, lower effect size, ICC = 0.6	-0.05	-0.05	0.00	79.12	79.69	79.66	-0.57	-0.54	0.03
Quadratic form, higher effect size, ICC = 0.6	-0.03	0.06	0.10	105.66	106.08	105.22	-0.42	0.44	0.86

Table 2.2 Curve similarity metrics from large-L2 single-replication simulations.

Data generation	Effect size	ICC	L2 units	Total replications
Exponential	1	1	30	2606
Exponential	1	1	50	2402
Exponential	1	1	100	2232
Exponential	1	2	30	2075
Exponential	1	2	50	1861
Exponential	1	2	100	1694
Exponential	2	1	30	1110
Exponential	2	1	50	1165
Exponential	2	1	100	1421
Exponential	2	2	30	2392
Exponential	2	2	50	7043
Exponential	2	2	100	18575
Logistic	1	1	30	9804
Logistic	1	1	50	7255
Logistic	1	1	100	5001
Logistic	1	2	30	4066
Logistic	1	2	50	3182
Logistic	1	2	100	2719
Logistic	2	1	30	1583
Logistic	2	1	50	1489
Logistic	2	1	100	1767
Logistic	2	2	30	1056
Logistic	2	2	50	1142
Logistic	2	2	100	1214
Quadratic	1	1	30	1345
Quadratic	1	1	50	1214
Quadratic	1	1	100	1246
Quadratic	1	2	30	1095
Quadratic	1	2	50	1076
Quadratic	1	2	100	1130
Quadratic	2	1	30	1960
Quadratic	2	1	50	1909
Quadratic	2	1	100	2305
Quadratic	2	2	30	1060
Quadratic	2	2	50	1078
Quadratic	2	2	100	1248

Table 2.3. Number of simulation replications needed to obtain 1000 valid replications by condition

	EXPO FORM		EXPO - LOW DISTINCTION		EXPO - HIGHER DISTINCTION	
	Percent correct	Runner-up	Percent correct	Runner-up	Percent correct	Runner-up
AIC, 13	0.917	Log	0.8343	Log	0.9997	Log
AIC, 9	0.9065	Log	0.813	Log	1	
AIC, 7	0.8966	Lin	0.7933	Lin	1	
AIC, 5	0.8782	Lin	0.7563	Lin	1	
AICC-N, 13	0.9178	Log	0.836	Log	0.9997	Log
AICC-N, 9	0.9081	Log	0.8162	Log	1	
AICC-N, 7	0.8983	Lin	0.7967	Lin	0.9998	Log
AICC-N, 5	0.8802	Lin	0.7603	Lin	1	
AICC-m, 13	0.9307	log	0.8617	Log	0.9997	Log
AICC-m, 9	0.9207	Lin	0.8413	Lin	1	
AICC-m, 7	0.9092	Lin	0.8185	Lin	0.9998	Log
AICC-m, 5	0.8902	Lin	0.7803	Lin	1	
BIC-N, 13	0.964	Lin	0.9283	Lin	0.9997	Log
BIC-N, 9	0.9499	Lin	0.8998	Lin	1	
BIC-N, 7	0.9354	Lin	0.871	Lin	0.9998	Log
BIC-N, 5	0.9123	Lin	0.8245	Lin	1	
BIC-m, 13	0.9503	Lin	0.901	Lin	0.9997	Log
BIC-m, 9	0.9402	Lin	0.8803	Lin	1	
BIC-m, 7	0.9274	Lin	0.855	Lin	1	
BIC-m, 5	0.9058	Lin	0.8115	Lin	1	
CAIC-N, 13	0.9653	Lin	0.9308	Lin	0.9997	Log
CAIC-N, 9	0.9513	Lin	0.9027	Lin	1	
CAIC-N, 7	0.9368	Lin	0.8737	Lin	0.9998	Log
CAIC-N, 5	0.914	Lin	0.828	Lin	1	
CAIC-m, 13	0.9588	Lin	0.918	Lin	0.9997	Log
CAIC-m, 9	0.9468	Lin	0.8937	Lin	1	
CAIC-m, 7	0.9333	Lin	0.8667	Lin	0.9998	Log
CAIC-m, 5	0.9107	Lin	0.8213	Lin	1	
HQIC-N, 13	0.949	Lin	0.8983	Lin	0.997	Log
HQIC-N, 9	0.9381	Lin	0.8762	Lin	1	
HQIC-N, 7	0.9237	Lin	0.8475	Lin	0.9998	Log
HQIC-N, 5	0.9016	Lin	0.8032	Lin	1	
HQIC-m, 13	0.9331	Log	0.8665	Log	0.9997	Log
HQIC-m, 9	0.9254	Lin	0.8508	Lin	1	
HQIC-m, 7	0.9125	Lin	0.8252	Lin	0.9998	Log
HQIC-m, 5	0.8931	Lin	0.7862	Lin	1	

Table 2.4. Proportion of correctly specified exponential models selected and most common alternative

	LOG FORM		LOG - LOW DISTINCTION		LOG - HIGHER DISTINCTION	
	Percent correct	Runner-up	Percent correct	Runner-up	Percent correct	Runner-up
AIC, 13	0.5099	Lin	0.0198	Lin	1	
AIC, 9	0.5123	Lin	0.0245	Lin	1	
AIC, 7	0.5125	Lin	0.025	Lin	1	
AIC, 5	0.5148	Lin	0.0302	Lin	0.9993	Quad
AICC-N, 13	0.5096	Lin	0.0192	Lin	1	
AICC-N, 9	0.5114	Lin	0.0228	Lin	1	
AICC-N, 7	0.5119	Lin	0.0238	Lin	1	
AICC-N, 5	0.5138	Lin	0.0282	Lin	0.9993	Quad
AICC-m, 13	0.5062	Lin	0.0123	Lin	1	
AICC-m, 9	0.508	Lin	0.016	Lin	1	
AICC-m, 7	0.5084	Lin	0.0168	Lin	1	
AICC-m, 5	0.509	Lin	0.0187	Lin	0.9993	Quad
BIC-N, 13	0.5006	Lin	0.0012	Lin	1	
BIC-N, 9	0.5008	Lin	0.0015	Lin	1	
BIC-N, 7	0.5008	Lin	0.0015	Lin	1	
BIC-N, 5	0.5013	Lin	0.0032	Lin	0.9993	Quad
BIC-m, 13	0.503	Lin	0.006	Lin	1	
BIC-m, 9	0.5043	Lin	0.0085	Lin	1	
BIC-m, 7	0.5034	Lin	0.0068	Lin	1	
BIC-m, 5	0.5038	Lin	0.0083	Lin	0.9993	Quad
CAIC-N, 13	0.5003	Lin	0.007	Lin	1	
CAIC-N, 9	0.5004	Lin	0.008	Lin	1	
CAIC-N, 7	0.5005	Lin	0.001	Lin	1	
CAIC-N, 5	0.5004	Lin	0.0015	Lin	0.9993	Quad
CAIC-m, 13	0.5015	Lin	0.003	Lin	1	
CAIC-m, 9	0.5022	Lin	0.0043	Lin	1	
CAIC-m, 7	0.5023	Lin	0.0045	Lin	1	
CAIC-m, 5	0.5021	Lin	0.0048	Lin	0.9993	Quad
HQIC-N, 13	0.5032	Lin	0.0063	Lin	1	
HQIC-N, 9	0.5046	Lin	0.0092	Lin	1	
HQIC-N, 7	0.5041	Lin	0.0082	Lin	1	
HQIC-N, 5	0.5052	Lin	0.011	Lin	0.9993	Quad
HQIC-m, 13	0.5064	Lin	0.0128	Lin	1	
HQIC-m, 9	0.5083	Lin	0.0167	Lin	1	
HQIC-m, 7	0.5073	Lin	0.0147	Lin	1	
HQIC-m, 5	0.5091	Lin	0.0188	Lin	0.9993	Quad

Table 2.5. Proportion of correctly specified logistic models selected and most common alternative

	QUAD FORM		QUAD - LOW DISTINCTION		QUAD - HIGHER DISTINCTION	
	Percent correct	Runner-up	Percent correct	Runner-up	Percent correct	Runner-up
AIC, 13	0.5687	Lin	0.2978	Lin	0.8395	Cubic
AIC, 9	0.5543	Lin	0.2652	Lin	0.8435	Cubic
AIC, 7	0.5378	Lin	0.238	Lin	0.8368	Cubic
AIC, 5	0.5269	Lin	0.2172	Lin	0.8367	Cubic
AICC-N, 13	0.5708	Lin	0.2973	Lin	0.8443	Cubic
AICC-N, 9	0.5551	Lin	0.261	Lin	0.8492	Cubic
AICC-N, 7	0.5388	Lin	0.2333	Lin	0.8443	Cubic
AICC-N, 5	0.5299	Lin	0.2113	Lin	0.8485	Cubic
AICC-m, 13	0.5779	Lin	0.2655	Lin	0.8903	Cubic
AICC-m, 9	0.5628	Lin	0.2302	Lin	0.8955	Cubic
AICC-m, 7	0.5461	Lin	0.2042	Lin	0.888	Cubic
AICC-m, 5	0.5385	Lin	0.187	Lin	0.89	Cubic
BIC-N, 13	0.5283	Lin	0.0673	Lin	0.9893	Cubic
BIC-N, 9	0.5193	Lin	0.0523	Lin	0.9863	Cubic
BIC-N, 7	0.5175	Lin	0.0513	Lin	0.9837	Cubic
BIC-N, 5	0.5145	Lin	0.05	Lin	0.979	Cubic
BIC-m, 13	0.5595	Lin	0.1712	Lin	0.9478	Cubic
BIC-m, 9	0.5429	Lin	0.1307	Lin	0.9552	Cubic
BIC-m, 7	0.531	Lin	0.113	Lin	0.949	Cubic
BIC-m, 5	0.5231	Lin	0.098	Lin	0.9482	Cubic
CAIC-N, 13	0.521	Lin	0.0473	Lin	0.9947	Cubic
CAIC-N, 9	0.5133	Lin	0.0345	Lin	0.9922	Cubic
CAIC-N, 7	0.5118	Lin	0.0335	Lin	0.9902	Cubic
CAIC-N, 5	0.5104	Lin	0.0323	Lin	0.9885	Cubic
CAIC-m, 13	0.5471	Lin	0.1212	Lin	0.973	Cubic
CAIC-m, 9	0.5315	Lin	0.0903	Lin	0.9727	Cubic
CAIC-m, 7	0.5247	Lin	0.0777	Lin	0.9717	Cubic
CAIC-m, 5	0.5176	Lin	0.0643	Lin	0.9708	Cubic
HQIC-N, 13	0.5659	Lin	0.1878	Lin	0.944	Cubic
HQIC-N, 9	0.5496	Lin	0.152	Lin	0.9472	Cubic
HQIC-N, 7	0.5363	Lin	0.136	Lin	0.9365	Cubic
HQIC-N, 5	0.53	Lin	0.128	Lin	0.932	Cubic
HQIC-m, 13	0.5743	Lin	0.2517	Lin	0.897	Cubic
HQIC-m, 9	0.5553	Lin	0.2093	Lin	0.9013	Cubic
HQIC-m, 7	0.5379	Lin	0.1815	Lin	0.8943	Cubic
HQIC-m, 5	0.5295	Lin	0.1652	Lin	0.8938	Cubic

Table 2.6. Proportion of correctly specified quadratic models selected and most common alternative

Models	Included (~~omitted~~)
Model 0 (matched data generation)	X1, X2, W1, W2
Model 1 (missing X1 only)	~~X1,~~ X2, W1, W2
Model 2 (missing X2 only)	X1, ~~X2~~, W1, W2
Model 3 (missing W1 only)	X1, X2, ~~W1~~, W2,
Model 4 (missing W2 only)	X1, X2, W1, ~~W2~~
Model 5 (missing X1 and W1)	~~X1~~, X2, ~~W1~~, W2
Model 6 (missing X1 and W2)	~~X1,~~ X2, W1, ~~W2~~
Model 7 (missing X2 and W1)	X1, ~~X2~~, ~~W1~~, W2
Model 8 (missing X2 and W2)	X1, ~~X2~~, W1, ~~W2~~

Table 3.1: Set of candidate models with non-nested predictor sets.

"Correctness" (descending order)	Candidate model	Included (~~omitted~~)
Fully correct	Data generating model (not included in the candidate set)	X1, X2, W1, W2
Missing one minor predictor	Model 1 (missing X1 only)	~~X1,~~ X2, W1, W2
	Model 3 (missing W1 only)	X1, X2, ~~W1~~, W2
Missing both minor predictors	Model 5 (missing X1 and W1)	~~X1~~, X2, ~~W1~~, W2
Missing one major predictor	Model 2 (missing X2 only)	X1, ~~X2~~, W1, W2
	Model 4 (missing W2 only)	X1, X2, W1, ~~W2~~
Missing one major predictor and one minor predictor	Model 6 (missing X1 and W2)	~~X1,~~ X2, W1, ~~W2~~
	Model 7 (missing X2 and W1)	X1, ~~X2~~, ~~W1~~, W2
Missing both major predictors	Model 8 (missing X2 and W2)	X1, ~~X2~~, W1, ~~W2~~

Table 3.2. Presumed "correctness" of candidate models.

Model pairs for comparison		A priori	Lower deviance	Higher % reduction in L2 variance	Higher % reduction in L1 variance
Model 1 (omitted X1)	Model 2 (omitted X2)	Model 1	Model 1	Expo: Model 1 Log: Similar	Model 1
	Model 3 (omitted W1)	Unclear	Model 3	Similar	Similar
	Model 4 (omitted W2)	Model 1	Model 1	Model 1	Similar
	Model 7 (omitted X2, W1)	Model 1	Model 1	Expo: Model 1 Log: Similar	Model 1
	Model 8 (omitted X2, W2)	Model 1	Model 1	Model 1	Model 1
Model 2 (omitted X2)	Model 3 (omitted W1)	Model 3	Model 3	Expo: Model 3 Log: Similar	Model 3
	Model 4 (omitted W2)	Unclear	Model 4	Model 2	Model 4
	Model 5 (omitted X1, W1)	Unclear	Model 5	Expo: Model 5 Log: Similar	Model 5
	Model 6 (omitted X1, W2)	Model 2	Model 6	Model 2	Model 6
Model 3 (omitted W1)	Model 4 (omitted W2)	Model 3	Model 3	Model 3	Similar
	Model 6 (omitted X1, W2)	Model 3	Model 3	Model 3	Similar
	Model 8 (omitted X2, W2)	Model 3	Model 3	Model 3	Model 3
Model 4 (omitted W2)	Model 5 (omitted X1, W1)	Unclear	Model 5	Model 5	Similar
	Model 7 (omitted X2, W1)	Model 4	Model 4	Model 7	Model 4
Model 5 (omitted X1, W1)	Model 6 (omitted X1, W2)	Model 5	Model 5	Model 5	Similar
	Model 7 (omitted X2, W1)	Model 5	Model 5	Expo: Model 5 Log: Similar	Model 5
	Model 8 (omitted X2, W2)	Model 5	Model 5	Model 5	Model 5
Model 6 (omitted X1, W2)	Model 7 (omitted X2, W1)	Unclear	Model 6	Model 7	Model 6
	Model 8 (omitted X2, W2)	Model 6	Model 6	Expo: Model 6 Log: Model 8	Model 6
Model 7 (omitted X2, W1)	Model 8 (omitted X2, W2)	Model 7	Model 7	Model 7	Similar

Table 3.3. All non-nested pairwise comparisons and the model determined to be more correct of the two by evaluation method.

Models sorted by proportion reduction in variance of random effects									
	Expo, low ICC					Log, low ICC			
	% L1 var			% L2 var			% L1 var			% L2 var
Model 0	10%		Model 0	37%		Model 0	8%		Model 2	9%
Model 3	10%		Model 1	37%		Model 3	8%		Model 1	9%
Model 4	10%		Model 3	37%		Model 4	8%		Model 0	9%
Model 1	10%		Model 5	36%		Model 1	8%		Model 5	8%
Model 5	10%		Model 2	29%		Model 5	8%		Model 3	7%
Model 6	10%		Model 7	29%		Model 6	8%		Model 7	6%
Model 2	1%		Model 4	19%		Model 2	1%		Model 8	1%
Model 7	1%		Model 6	18%		Model 8	1%		Model 4	-25%
Model 8	1%		Model 8	5%		Model 7	1%		Model 6	-27%
	Expo, high ICC					Log, high ICC			
	% L1 var			% L2 var			% L1 var			% L2 var
Model 0	11%		Model 0	26%		Model 0	8%		Model 2	4%
Model 3	11%		Model 1	25%		Model 3	8%		Model 0	4%
Model 4	11%		Model 3	24%		Model 4	8%		Model 1	4%
Model 1	11%		Model 5	24%		Model 1	8%		Model 5	4%
Model 5	11%		Model 2	20%		Model 5	8%		Model 3	3%
Model 6	11%		Model 7	19%		Model 6	8%		Model 7	3%
Model 2	1%		Model 4	14%		Model 2	1%		Model 8	0%
Model 7	1%		Model 6	13%		Model 8	1%		Model 4	-14%
Model 8	1%		Model 8	5%		Model 7	1%		Model 6	-15%

Table 3.4. Percent reduction in variance components compared to the unconditional model.

Deviance comparison - Expo, Low ICC				Deviance comparison - Expo, High ICC			
Model 1	Model 2	**Model 3**	Model 6	**Model 1**	Model 2	**Model 3**	Model 6
88232	90413	**88128**	88756	**89689**	92146	**89618**	89989
Model 1	**Model 3**	**Model 3**	Model 8	Model 1	**Model 3**	**Model 3**	Model 8
88232	**88128**	**88128**	90999	89689	**89618**	**89618**	92482
Model 1	Model 4	Model 4	**Model 5**	**Model 1**	Model 4	Model 4	**Model 5**
88232	88638	88638	**88251**	**89689**	89887	89887	**89723**
Model 1	Model 7	**Model 4**	Model 7	**Model 1**	Model 7	**Model 4**	Model 7
88232	90427	**88638**	90427	**89689**	92176	**89887**	92176
Model 1	Model 8	**Model 5**	Model 6	**Model 1**	Model 8	**Model 5**	Model 6
88232	90999	**88251**	88756	**89689**	92482	**89723**	89989
Model 2	**Model 3**	**Model 5**	Model 7	Model 2	**Model 3**	**Model 5**	Model 7
90413	**88128**	**88251**	90427	92146	**89618**	**89723**	92176
Model 2	**Model 4**	**Model 5**	Model 8	Model 2	**Model 4**	**Model 5**	Model 8
90413	**88638**	**88251**	90999	92146	**89887**	**89723**	92482
Model 2	**Model 5**	**Model 6**	Model 7	Model 2	**Model 5**	**Model 6**	Model 7
90413	**88251**	**88756**	90427	92146	**89723**	**89989**	92176
Model 2	**Model 6**	**Model 6**	Model 8	Model 2	**Model 6**	**Model 6**	Model 8
90413	**88756**	**88756**	90999	92146	**89989**	**89989**	92482
Model 3	Model 4	**Model 7**	Model 8	**Model 3**	Model 4	**Model 7**	Model 8
88128	88638	**90427**	90999	**89618**	89887	**92176**	92482
Deviance comparison - Log, Low ICC				**Deviance comparison - Log, High ICC**			
Model 1	Model 2	**Model 3**	Model 6	**Model 1**	Model 2	**Model 3**	Model 6
79613	81289	**79552**	80292	**81354**	83148	**81317**	81716
Model 1	**Model 3**	**Model 3**	Model 8	Model 1	**Model 3**	**Model 3**	Model 8
79613	**79552**	**79552**	81450	81354	**81317**	**81317**	83221
Model 1	Model 4	Model 4	**Model 5**	**Model 1**	Model 4	Model 4	**Model 5**
79613	80169	80169	**79620**	**81354**	81660	81660	**81356**
Model 1	Model 7	**Model 4**	Model 7	**Model 1**	Model 7	**Model 4**	Model 7
79613	81363	**80169**	81363	**81354**	83172	**81660**	83172
Model 1	Model 8	**Model 5**	Model 6	**Model 1**	Model 8	**Model 5**	Model 6
79613	81450	**79620**	80292	**81354**	83221	**81356**	81716
Model 2	**Model 3**	**Model 5**	Model 7	Model 2	**Model 3**	**Model 5**	Model 7
81289	**79552**	**79620**	81363	83148	**81317**	**81356**	83172
Model 2	**Model 4**	**Model 5**	Model 8	Model 2	**Model 4**	**Model 5**	Model 8
81289	**80169**	**79620**	81450	83148	**81660**	**81356**	83221
Model 2	**Model 5**	**Model 6**	Model 7	Model 2	**Model 5**	**Model 6**	Model 7
81289	**79620**	**80292**	81363	83148	**81356**	**81716**	83172
Model 2	**Model 6**	**Model 6**	Model 8	Model 2	**Model 6**	**Model 6**	Model 8
81289	**80292**	**80292**	81450	83148	**81716**	**81716**	83221
Model 3	Model 4	**Model 7**	Model 8	**Model 3**	Model 4	**Model 7**	Model 8
79552	80169	**81363**	81450	**81317**	81660	**83172**	83221

Table 3.5. Comparison of pairwise model deviances. Selected model is in bold typeface.

Expo, low ICC			Difference "tier"	Expo, high ICC		
Comparison		|Deviance diff|		Comparison		|Deviance diff|
Model 1	Model 3	104.39	Lower	Model 1	Model 3	71.07
Model 4	Model 5	387.34	Lower	Model 4	Model 5	164.34
Model 1	Model 4	405.98	Lower	Model 1	Model 4	198.53
Model 5	Model 6	505.51	Lower	Model 5	Model 6	266.34
Model 3	Model 4	510.37	Lower	Model 3	Model 4	269.6
Model 7	Model 8	571.95	Lower	Model 7	Model 8	305.15
Model 3	Model 6	628.54	Lower	Model 3	Model 6	371.6
Model 2	Model 6	1656.89	Moderate	Model 2	Model 6	2156.96
Model 6	Model 7	1670.98	Moderate	Model 6	Model 7	2187.13
Model 2	Model 4	1775.06	Moderate	Model 2	Model 4	2258.96
Model 4	Model 7	1789.15	Moderate	Model 4	Model 7	2289.13
Model 2	Model 5	2162.4	Moderate	Model 2	Model 5	2423.3
Model 5	Model 7	2176.49	Moderate	Model 5	Model 7	2453.47
Model 1	Model 2	2181.04	Moderate	Model 1	Model 2	2457.49
Model 1	Model 7	2195.13	Moderate	Model 1	Model 7	2487.66
Model 6	Model 8	2242.93	Moderate	Model 6	Model 8	2492.28
Model 2	Model 3	2285.43	Higher	Model 2	Model 3	2528.56
Model 5	Model 8	2748.44	Higher	Model 5	Model 8	2758.62
Model 1	Model 8	2767.08	Higher	Model 1	Model 8	2792.81
Model 3	Model 8	2871.47	Higher	Model 3	Model 8	2863.88
Log, low ICC			**Difference "tier"**	**Log, high ICC**		
Comparison		|Deviance diff|		Comparison		|Deviance diff|
Model 1	Model 3	61.08	Lower	Model 1	Model 3	37.04
Model 7	Model 8	87.16	Lower	Model 7	Model 8	48.78
Model 4	Model 5	548.62	Lower	Model 4	Model 5	304.02
Model 1	Model 4	555.2	Lower	Model 1	Model 4	305.53
Model 3	Model 4	616.28	Lower	Model 3	Model 4	342.57
Model 5	Model 6	671.78	Lower	Model 5	Model 6	360.02
Model 3	Model 6	739.44	Lower	Model 3	Model 6	398.57
Model 2	Model 6	997.48	Moderate	Model 2	Model 6	1432.53
Model 6	Model 7	1070.99	Moderate	Model 6	Model 7	1456.05
Model 2	Model 4	1120.64	Moderate	Model 2	Model 4	1488.53
Model 6	Model 8	1158.15	Moderate	Model 6	Model 8	1504.83
Model 4	Model 7	1194.15	Moderate	Model 4	Model 7	1512.05
Model 2	Model 5	1669.26	Moderate	Model 2	Model 5	1792.55
Model 1	Model 2	1675.84	Moderate	Model 1	Model 2	1794.06
Model 2	Model 3	1736.92	Moderate	Model 5	Model 7	1816.07
Model 5	Model 7	1742.77	Moderate	Model 1	Model 7	1817.58
Model 1	Model 7	1749.35	Higher	Model 2	Model 3	1831.1
Model 5	Model 8	1829.93	Higher	Model 5	Model 8	1864.85
Model 1	Model 8	1836.51	Higher	Model 1	Model 8	1866.36
Model 3	Model 8	1897.59	Higher	Model 3	Model 8	1903.4

Table 3.6. Absolute values of the differences in deviances across non-nested model pairs, divided into three “tiers” of size of the differences.

	AIC	AICC-N	AICC-m	BIC-N	BIC-m	CAIC-N	CAIC-m	HQIC-N	HQIC-m
Expo – 13	18.1%	17.5%	11.2%	0.7%	4.5%	0.3%	2.2%	5.9%	10.5%
Expo – 9	15.1%	14.5%	9.5%	0.8%	3.5%	0.4%	1.9%	4.7%	8.4%
Expo – 7	13.6%	12.7%	8.1%	0.5%	2.7%	0.3%	1.2%	4.3%	7.5%
Expo – 5	11.5%	10.7%	7.0%	0.7%	2.2%	0.3%	1.2%	3.6%	6.0%
Log – 13	14.2%	13.6%	8.1%	0.7%	3.3%	0.3%	1.7%	4.0%	8.0%
Log – 9	12.9%	12.2%	7.5%	0.5%	2.2%	0.2%	1.2%	3.7%	7.3%
Log – 7	11.1%	9.0%	6.1%	0.4%	2.2%	0.2%	1.0%	3.3%	6.0%
Log - 5	9.8%	8.9%	5.5%	0.5%	1.9%	0.2%	0.8%	2.9%	5.3%

Table 3.7. Overall correct selection rate across time points and information criteria.

Deviance difference tier	Comparison		Correct selection range	Criteria ranking for exponential	Criteria ranking for logistic
Lower difference	Model 1	Model 3	27% - 73%	AIC BIC(N and m)	BIC(N and m) AIC
	Model 1	Model 4	At least 85%	(No effect)	(No effect)
	Model 3	Model 4	At least 92%	(No effect)	(No effect)
	Model 3	Model 6	69% - 94%	AIC BIC(m) BIC(N)	AIC BIC(m) BIC(N)
	Model 4	Model 5	At least 86%	BIC(N) BIC(m) AIC	BIC(N) BIC(m) AIC
	Model 5	Model 6	At least 98%	(No effect)	(No effect)
	Model 7	Model 8	At least 90%	(No effect)	(No effect)
Moderate difference	Model 1	Model 2	At least 97%	(No effect)	(No effect)
	Model 1	Model 7	At least 86%	AIC BIC(m) BIC(N)	AIC BIC(m) BIC(N)
	Model 2	Model 4	At least 83%	(No effect)	(No effect)
	Model 2	Model 5	At least 98%	BIC(N) BIC(m) AIC	BIC(N) BIC(m) AIC
	Model 2	Model 6	0% - 18%	BIC(N) BIC(m) AIC	BIC(N) BIC(m) AIC
	Model 4	Model 7	66% - 98%	AIC BIC(m) BIC(N)	AIC BIC(m) BIC(N)
	Model 5	Model 7	At least 94%	(No effect)	(No effect)
	Model 6	Model 7	At least 81%	(No effect)	(No effect)
	Model 6	Model 8	At least 90%	(No effect)	(No effect)
Higher difference	Model 1	Model 8	At least 89%	AIC BIC(m) BIC(N)	AIC BIC(m) BIC(N)
	Model 2	Model 3	At least 98%	(No effect)	(No effect)
	Model 3	Model 8	At least 90%	AIC BIC(m) BIC(N)	BIC(N) BIC(m) AIC
	Model 5	Model 8	At least 97%	(No effect)	(No effect)

Table 3.8. Summary of selection rate ranges of the more correct model across non-nested pairwise comparisons, with performance ranking of AIC, BIC(N), and BIC(m).

Expo, low ICC			Log, low ICC		
Model	Deviance	Difference	Model	Deviance	Difference
Model 0	88111.9		Model 0	79518.3	
Model 3	88128	16.01	Model 3	795`52.2	33.93
Model 1	88232.3	104.39	Model 1	79613.3	61.08
Model 5	88251	18.64	Model 5	79619.9	6.58
Model 4	88638.3	387.34	Model 4	80168.5	548.62
Model 6	88756.5	118.17	Model 6	80291.7	123.16
Model 2	90413.4	1656.89	Model 2	81289.2	997.48
Model 7	90427.5	14.09	Model 7	81362.7	73.51
Model 8	90999.4	571.95	Model 8	81449.8	87.16
Expo, high ICC			Log, high ICC		
Model	Deviance	Difference	Model	Deviance	Difference
Model 0	89585.8		Model 0	81308.8	
Model 3	89617.7	31.93	Model 3	81317.2	8.42
Model 1	89688.8	71.07	Model 1	81354.3	37.04
Model 5	89723	34.19	Model 5	81355.8	1.51
Model 4	89887.3	164.34	Model 4	81659.8	304.02
Model 6	89989.3	102	Model 6	81715.8	56
Model 2	92146.3	2156.96	Model 2	83148.3	1432.53
Model 7	92176.4	30.17	Model 7	83171.9	23.52
Model 8	92481.6	305.15	Model 8	83220.6	48.78

Table 3.9. Ordered list of deviance differences among models, including correctly specified models.

Functional form	Level 1 specification
Linear – 4 parameters	$Y_{ij} = b_o + b_1 CenAge + e_{ij}$
Quadratic – 5 parameters	$Y_{ij} = b_o + b_1 CenAge + b_2 CenAge^2 + e_{ij}$
Cubic – 6 parameters	$Y_{ij} = b_o + b_1 CenAge + b_2 CenAge^2 + b_3 CenAge^3 + e_{ij}$
Exponential - 4 parameters	$Y_{ij} = b_0(e^{b_2 CenAge}) + e_{ij}$
Logistic - 5 parameters	$Y_{ij} = \frac{b_1}{1 + e^{\frac{(-(CenAge-(b_2))}{b_3}}} + e_{ij}$

Table 4.1. Functional form specifications at Level 1.

Conditional independence	First-order autoregressive	Toepliz-banded (2 bands)
$\sigma^2\begin{bmatrix}1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1\end{bmatrix}$	$\sigma^2\begin{bmatrix}1 & & & & \\ \rho & 1 & & & \\ \rho^2 & \rho & 1 & & \\ \rho^3 & \rho^2 & \rho & 1 & \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1\end{bmatrix}$	$\sigma^2\begin{bmatrix}1 & & & & \\ a & 1 & & & \\ b & a & 1 & & \\ 0 & b & a & 1 & \\ 0 & 0 & b & a & 1\end{bmatrix}$

Table 4.2. Examples of covariance structure types for a five-time point model.

Models	*Efficient criteria*			*Consistent criteria*					
Functional form	**AIC**	**AICC-N**	**AICC-m**	**BIC-N**	**BIC-m**	**CAIC-N**	**CAIC-m**	**HQIC-N**	**HQIC-m**
Linear (b0)	45638	45638	45638	45670	45661	45674	45665	45649	45647
Quadratic (b0)	44911	44911	44911	44951	44940	44956	44945	44924	44922
Cubic (b0)	44874	44874	44874	44921	44908	44927	44914	44889	44886
Expo (b0)	45483	45483	45483	45515	45506	45519	45510	45494	45492
Logistic (b1)	45893	45893	45893	45933	45922	45938	45926.9	45906	45904
Logistic (b2)	45123	45123	45123	45162	45151	45167	45156	45135	45133
Logistic (b3)	.	.	.	.	.	.	.	.	.
Predictors									
Cubic/Base	41182	41182	41182	41293	41262	41307	41276	41218	41211
Cubic/specific criticism	40832	40832	40832	40958	40924	40974	40940	40873	40865
Cubic/general criticism	40885	40885	40885	41003	40971	41018	40986	40924	40916
Cubic/valence	40745	40745	40745	40883	40844	40901	40862	40791	40781
Cubic/diversity	38451	38451	38451	38584	38548	38601	38565	38495	38486
Cubic/self-criticism	40293	40293	40293	40411	40378	40426	40393	40332	40324
Cubic/appearance	40424	40424	40424	40566	40527	40584	40545	40470	40462
Covariance structure									
Cubic/diversity/VC	38451	38451	38451	38584	38548	38601	38565	38495	38486
Cubic/diversity/AR(1)	37465	37465	37465	37606	37567	37624	37585	37511	37502
Cubic/diversity/TO(9)	37115	37115	37116	37311	37257	37336	37282	37179	37167
Cubic/diversity/TO(8)	37116	37116	37117	37304	37253	37328	37277	37178	37166
Cubic/diversity/TO(7)	37128	37128	37128	37308	37258	37331	37281	37187	37175
Cubic/diversity/TO(6)	37174	37174	37175	37347	37299	37369	37321	37231	37220
Cubic/diversity/TO(5)	37208	37208	37208	37372	37327	37393	37348	37262	37252
Cubic/diversity/TO(4)	37262	37262	37263	37419	37376	37439	37396	37314	37304
Cubic/diversity/TO(3)	37396	37396	37396	37545	37504	37564	37523	37445	37435
Cubic/diversity/TO(2)	37709	37709	37710	37850	37811	37868	37829	37755	37747

Table 4.3. Information criteria values for the three-step model selection process.

	TO(9) Model			TO(8) Model		
Parameters	Estimate	SE	P	Estimate	SE	P
Fixed effects						
Intercept (b0)	-0.187	0.138	0.1752	-0.193	0.138	0.16
Linear (b1)	-0.186	0.015	< 0.001	-0.186	0.015	< 0.001
Quadratic (b2)	0.025	0.003	< 0.001	0.024	0.003	< 0.001
Cubic (b3)	-0.001	0.000	< 0.001	-0.001	0.000	< 0.001
BMI (b4)	0.069	0.002	< 0.001	0.069	0.002	< 0.001
Race (b5)	-0.192	0.022	< 0.001	-0.192	0.022	< 0.001
Income2 (b6)	-0.037	0.034	0.27	-0.037	0.034	< 0.001
Income3 (b7)	-0.008	0.029	0.79	-0.008	0.029	0.27
Income4 (b8)	-0.008	0.030	0.8	-0.007	0.030	0.79
Education (b9)	-0.053	0.026	0.04	-0.053	0.026	0.81
Education (b10)	-0.117	0.029	< 0.001	-0.117	0.029	0.04
Menarche (b11)	0.003	0.008	0.72	0.003	0.008	< 0.001
Friend race (b12)	-0.105	0.073	0.15	-0.104	0.073	0.16
Other friend race (b13)	0.004	0.008	0.58	0.005	0.008	0.57
School minority (b14)	0.049	0.036	0.17	0.049	0.036	0.17
Random effects						
Intercept variance	0.030			0.041		
TO(2)	0.225			0.214		
TO(3)	0.186			0.175		
TO(4)	0.143			0.132		
TO(5)	0.111			0.100		
TO(6)	0.086			0.074		
TO(7)	0.067			0.055		
TO(8)	0.035			0.023		
TO(9)	0.014			.	.	.
Residual variance	0.521			0.510		

Table 4.4. Parameter estimates and standard errors for body image discrepancy final models.

Models	*Efficient criteria*			*Consistent criteria*					
Functional form	**AIC**	**AICC-N**	**AICC-m**	**BIC-N**	**BIC-m**	**CAIC-N**	**CAIC-m**	**HQIC-N**	**HQIC-m**
Linear	65176	65176	65176	65205	65199	65209	65203	65186	65184
Quadratic	65146	65146	65146	65182	65175	65187	65180	65158	65156
Cubic	65146	65146	65146	65189	65181	65195	65187	65161	65159
Expo	65003	65003	65003	65032	65026	65036	65030	65013	65011
Logistic (b1)	64670	64670	64670	64706	64699	64711	64704	64682	64680
Logistic (b2)	64912	64912	64912	64948	64941	64953	64946	64924	64922
Logistic (b3)	.	.	.	.	.	.	.	.	.
Predictors									
Logistic/Base	61452	61452	61452	61545	61526	61558	61539	61484	61479
Logistic/specific criticism	60939	60939	60939	61047	61025	61062	61040	60976	60970
Logistic/general criticism	61010	61010	61010	61111	61090	61125	61104	61044	61039
Logistic/valence	60697	60697	60697	60819	60794	60836	60811	60738	60732
Logistic/diversity	56503	56503	56503	56617	56594	56633	56610	56542	56536
Logistic/self-criticism	60151	60151	60151	60251	60231	60265	60245	60185	60180
Logistic/appearance	60470	60470	60470	60592	60567	60609	60584	60511	60505

Table 4.5. Model selection for body dissatisfaction outcome.

	Diversity Model		
Parameters	Estimate	SE	P
Fixed effects			
Log (b1)	-0.189	1.963	0.92
Log (b2)	-4.758	1.351	<0.001
Log (b3)	9.387	6.904	0.17
BMI (b4)	0.672	0.017	<0.001
Race (b5)	-4.280	0.986	<0.001
Income2 (b6)	-0.680	0.502	0.18
Income3 (b7)	0.323	0.415	0.78
Income4 (b8)	0.272	0.435	0.62
Education (b9)	-0.522	0.379	0.17
Education (b10)	-1.053	0.470	0.03
Menarche (b11)	-0.437	0.151	<0.001
Friend race (b12)	-2.023	1.127	0.07
Other friend race (b13)	0.104	0.116	0.37
School minority (b14)	0.411	0.519	0.79
Random effects			
Intercept variance	31.426		
Residual variance	21.313		

Table 4.6. Parameter estimates and standard errors for body dissatisfaction final model.

Model	# Parameters	Unique L1 Predictors
Perceived demand	14	A lot of work at home (0-1), at job (0-1), at school (0-1), demands from family (0-1), demands by friends (0-1)
Event demand	14	Family tasks (0-8), leisure with family (0-3), school events (0-3), friend events (0-3), worked a job (0-1)
Time demand	14	Time spent on family tasks, time spent on leisure with family, time spent on school events, time spent on events with friends, and time spent working a job
Negative interactions	17	Negative interactions with family (0-4), friends (0-1), adults at school (0-2), with other students (0-1), outside of school (0-1), race-specific with adults at school (0-1), race-specific with students at school (0-1), race-specific with someone outside of school (0-1)
Positive interactions	15	Positive interactions with family (0-1), friends (0-3), adults at school (0-1), race-specific with adults at school (0-1), race-specific with students at school (0-1), race-specific with someone outside of school (0-1)
All interactions	23	Combination of all negative and positive interactions

Table 5.1: Six sets of predictors and the number of parameters in each model, including the parameters needed for the exponential functional form and the covariates.

Models	*Efficient criteria*			*Consistent criteria*					
Functional form	**AIC**	**AICC-N**	**AICC-m**	**BIC-N**	**BIC-m**	**CAIC-N**	**CAIC-m**	**HQIC-N**	**HQIC-m**
Linear	46052.0	46052.0	46052.1	46080.2	46070.0	46084.2	46074.0	46061.6	46059.0
Quadratic	46046.0	46046.0	46046.1	46081.2	46068.5	46086.2	46073.5	46058.0	46054.7
Cubic	46041.0	46041.0	46041.1	46083.3	46068.0	46089.3	46074.0	46055.4	46051.5
Expo	46018.0	46018.0	46018.1	46046.2	46036.0	46050.2	46040.0	46027.6	46025.0
Logistic (b1)	46020.0	46020.0	46020.1	46055.2	46042.5	46060.2	46047.5	46032.0	46028.7
Logistic (b2)	.	.	.	.	.	.	.	.	.
Logistic (b3)	.	.	.	.	.	.	.	.	.
Predictors									
Expo/Base	39452.0	39452.0	39452.3	39514.1	39492.0	39523.1	39501.0	39473.3	39467.5
Expo/Negative interaction	39079.0	39079.1	39080.0	39196.2	39154.5	39213.2	39171.5	39119.3	39108.3
Expo/Positive interaction	39346.0	39346.1	39346.8	39449.4	39412.6	39464.4	39427.6	39381.6	39371.9
Expo/All interactions	39019.0	39019.2	39020.8	39177.6	39121.2	39200.6	39144.2	39073.5	39058.7
Expo/Perceived demand	39393.0	39393.1	39393.7	39489.6	39455.2	39503.6	39469.2	39426.2	39417.2
Expo/Event demand	39429.0	39429.1	39429.7	39525.6	39491.2	39539.6	39505.2	39462.2	39453.2
Expo/Time demand	32049.0	32049.1	32049.7	32142.7	32110.8	32156.7	32124.8	32081.6	32073.0

Table 5.2. Information criteria values for the functional form-only models (top) and for the seven models with predictors (bottom).

Parameter	Coefficient	P value
Fixed effects		
B0	5.0716	<0.001
B1	-0.03995	<0.001
B3 – Age	-0.4001	.4425
B4 – Gender	.05406	.8873
B5 – CESD	-0.07601	.8322
B6 - Latino	-.8779	.0238
B7 – School day	0.5069	<0.001
B8 – Family task time	-0.04118	.2441
B9 – Family leisure time	-0.08123	.0043
B10 – School time	-0.08092	<0.001
B11 – Friend time	-0.02694	.1033
B12 – Job time	-0.04375	.1455
Random effects		
Residual variance	9.7797	<0.001
Intercept variance	19.0785	<0.001

Table 5.3. Coefficient estimates and significance tests of the final model (exponential functional form and the time demand predictor set).

9. Figures

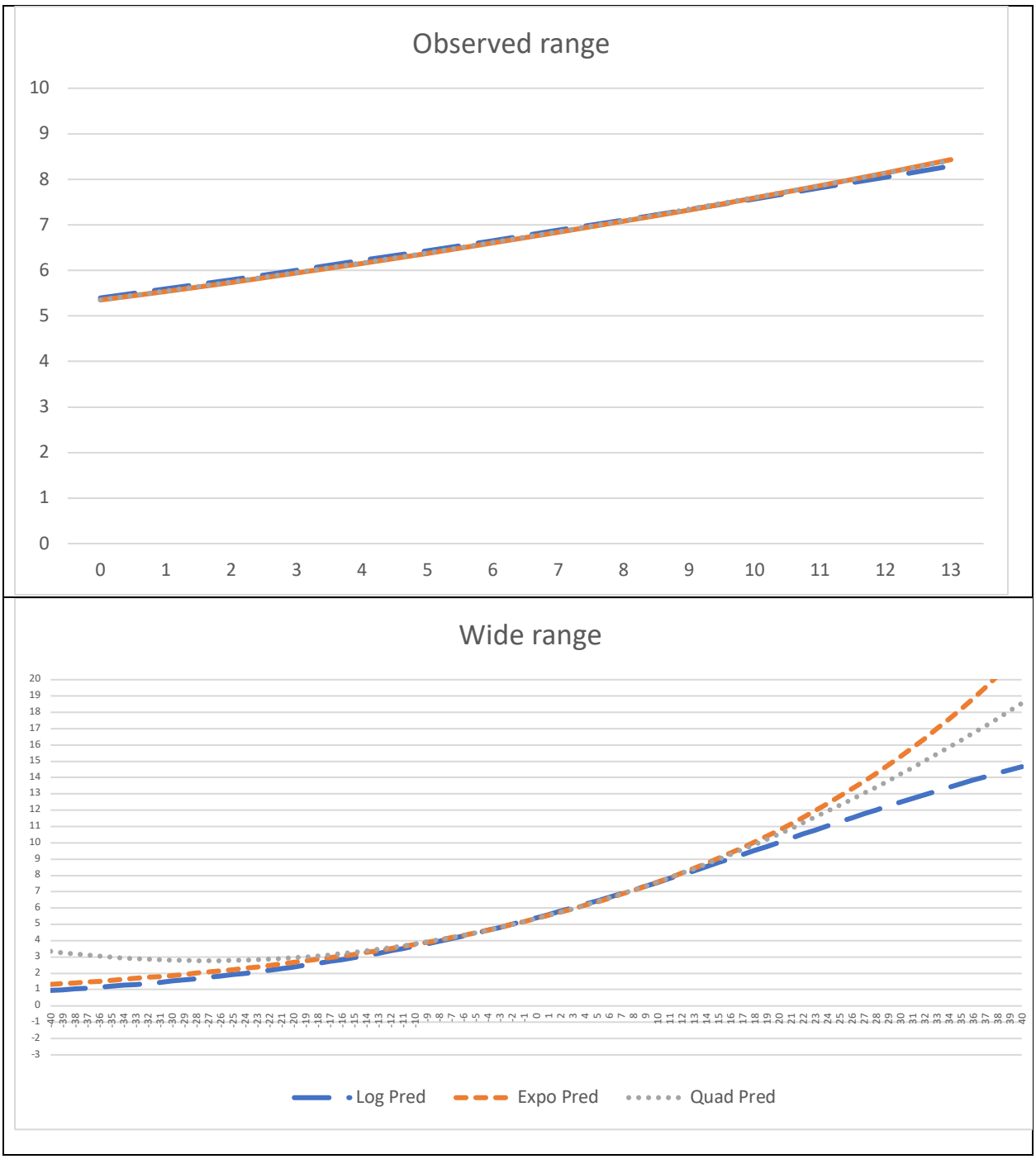

Figure 2.1. Functional form curves that are less distinct within the observed range, with the observed range (0-13) on the top and a wider range (-40 to 40) on the bottom.

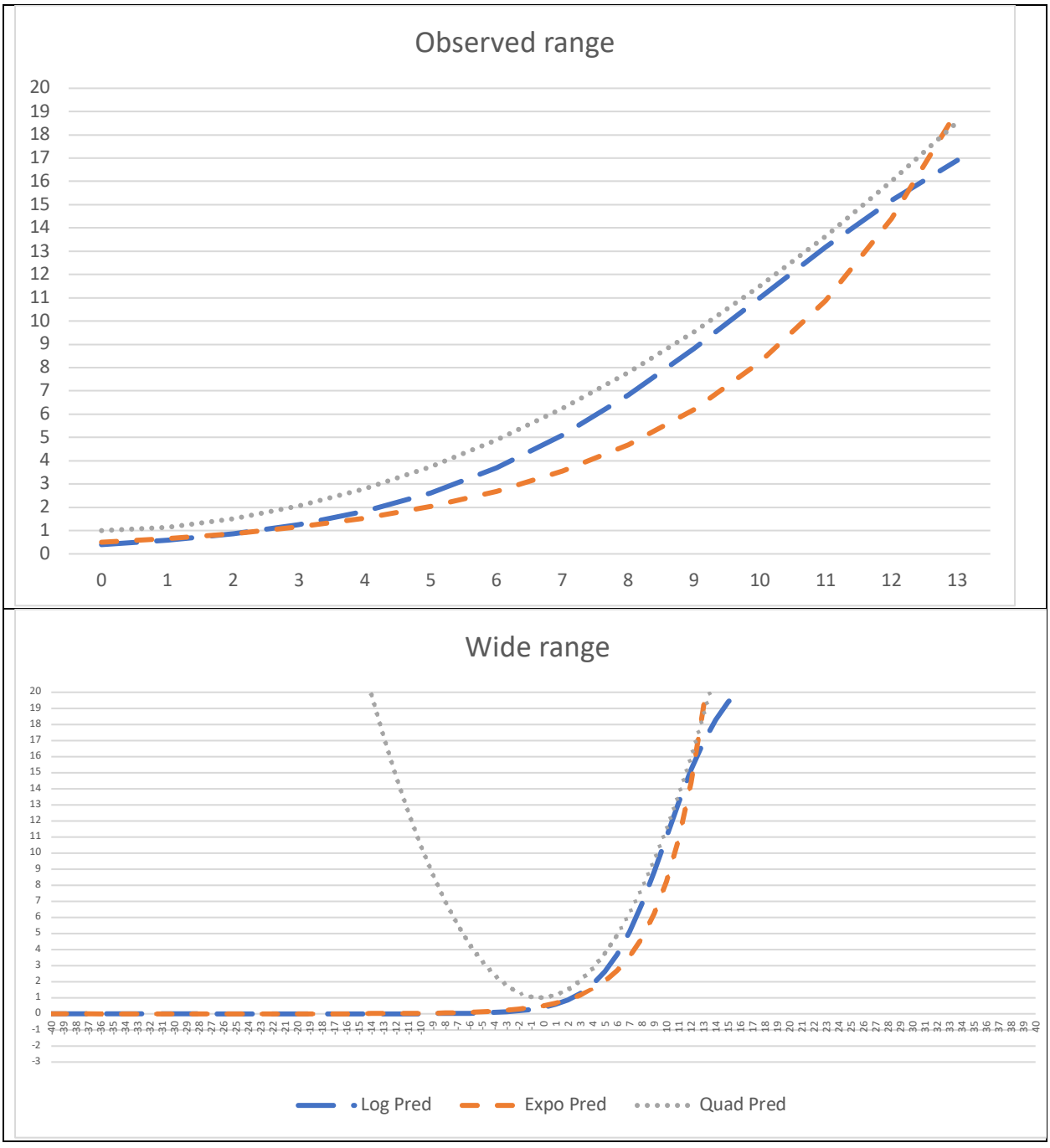

Figure 2.2. Functional form curves that are more distinct within the observed range, with the observed range (0-13) on the top and a wider range (-40 to 40) on the bottom.

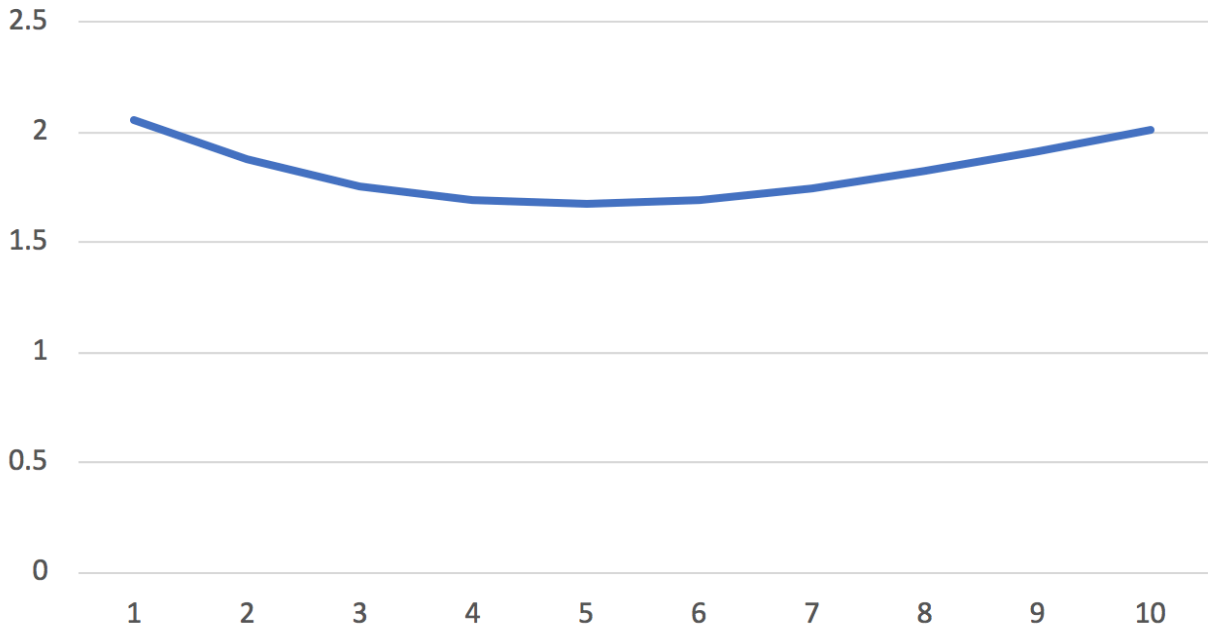

Figure 4.1. The estimated cubic functional form model, which was the best-fitting model, for the body image discrepancy outcome across 10 time points.

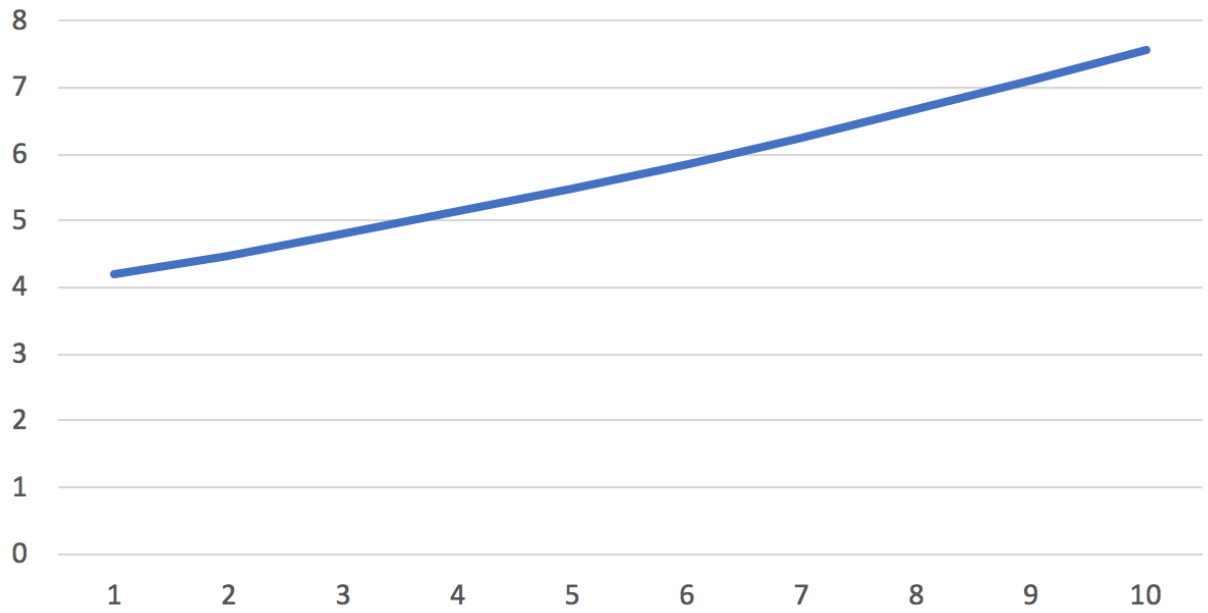

Figure 4.2. The estimated logistic functional form model, which was the best-fitting model, for the body dissatisfaction outcome across 10 time points.

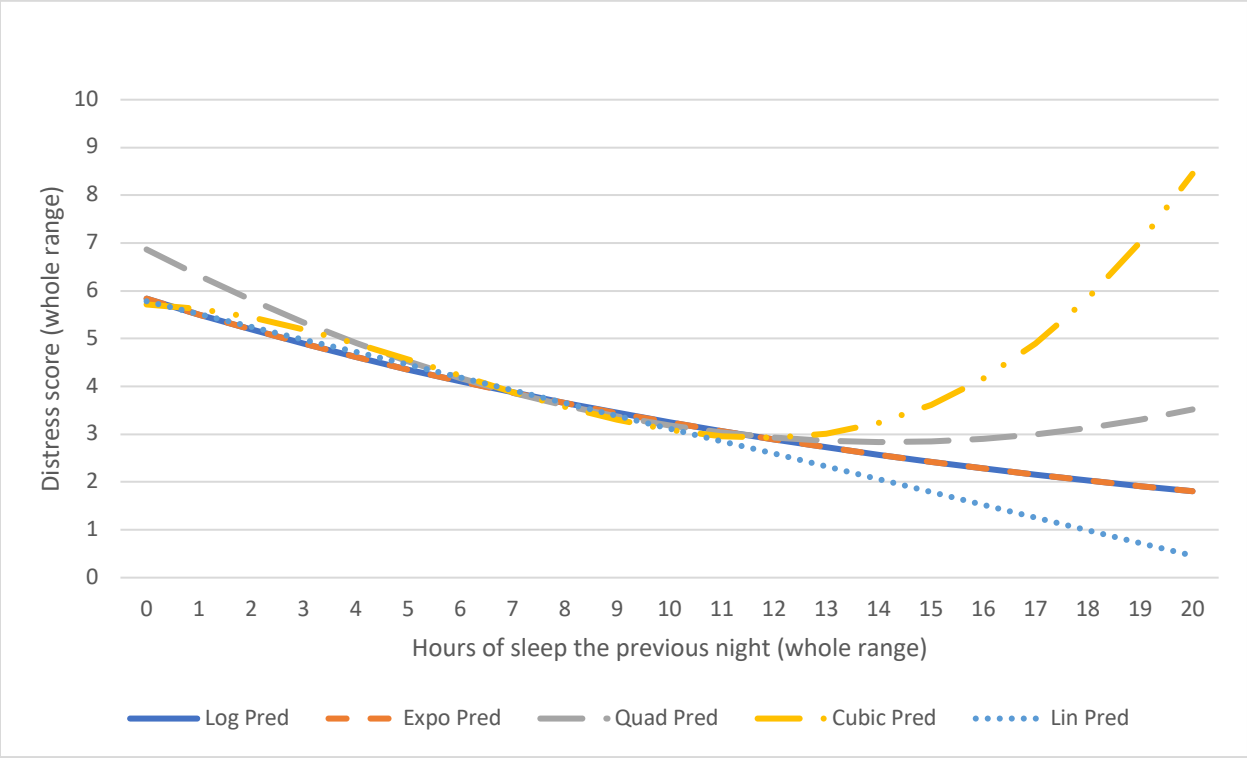

Figure 5.1. Curves representing the predicted daily distress values for each of five functional forms over the observed range (0-20) of prior night's sleep.

10. References

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Foundations and basic theory* (pp. 610-624). New York, NY: Spinger-Verlag. (Reprinted from *2nd international symposium on information theory*, by B. N. Petrov & F. Csàki (Eds.), 1973, Akademai Kiàdo, Budapest.

Alvaro, P. K., Roberts, R. M., & Harris, J. K. (2013). A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. *Sleep*, *36*(7), 1059-1068.

Asselmann, E., Wittchen, H. U., Lieb, R., & Beesdo-Baum, K. (2017). A 10-year prospective-longitudinal study of daily hassles and incident psychopathology among adolescents and young adults: interactions with gender, perceived coping efficacy, and negative life events. *Social Psychiatry and Psychiatric Epidemiology*, *52*(11), 1353-1362.

Bauer, D. J., Gottfredson, N. C., Dean, D., & Zucker, R. A. (2013). Analyzing repeated measures data on individuals nested within groups: accounting for dynamic group effects. *Psychological Methods*, *18*(1), 1-14.

Baum, K. T., Desai, A., Field, J., Miller, L. E., Rausch, J., & Beebe, D. W. (2014). Sleep restriction worsens mood and emotion regulation in adolescents. *Journal of Child Psychology and Psychiatry*, *55*(2), 180-190.

Berry, W. D. (1993). *Quantitative applications in the social sciences: Understanding regression assumptions*. Thousand Oaks, CA: SAGE Publications.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

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

Brown, W. V., Fujioka, K., Wilson, P. W., & Woodworth, K. A. (2009). Obesity: Why be concerned?. *The American Journal of Medicine*, *122*(4), S4-S9.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Christensen, W. (2018). *Model selection using information criteria (made easy in SAS®)*. Paper presented at SAS Global Forum, Denver, CO.

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

D'Angelo, B., & Wierzbicki, M. (2003). Relations of daily hassles with both anxious and depressed mood in students. *Psychological Reports*, *92*(2), 416-418.

Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(4), 387-419.

Drury, A., Aramburu, C., & Louis, M. (2002). Exploring the association between body weight, stigma of obesity, and health care avoidance. *Journal of the American Academy of Nurse Practitioners*, *14*(12), 554-561.

Eaton, D. K., McKnight-Eily, L. R., Lowry, R., Perry, G. S., Presley-Cantrell, L., & Croft, J. B. (2010). Prevalence of insufficient, borderline, and optimal hours of sleep among

high school students – United States, 2007. *Journal of Adolescent Health*, *46*(4), 399-401.

Fuligni, A. J., Arruda, E. H., Krull, J. L., & Gonzales, N. A. (2018). Adolescent sleep duration, variability, and peak levels of achievement and mental health. *Child Development*, *89*(2), e18-e28.

Fuligni, A. J., Bai, S., Krull, J. L., & Gonzales, N. A. (2019). Individual differences in optimum sleep for daily mood during adolescence. *Journal of Clinical Child and Adolescent Psychology*, *48*(3), 469-479.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician*, *60*(1), 19-26.

Hall, D. B., & Bailey, R. L. (2001). Modeling and prediction of forest growth variables based on multilevel nonlinear mixed models. *Forest Science*, *47*(3), 311-321.

Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In Hox, J. J. & Roberts, J. K (Eds.), *Handbook of advanced multilevel analysis* (pp. 231-255). New York, NY: Routledge.

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.

Harring, J. R., & Blozis, S. A. (2014). Fitting correlated residual error structure in nonlinear mixed-effects models using SAS PROC NLMIXED. *Behavior Research Methods*, *46*(2), 372-384.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In Little, T. D. (Ed.), *The Oxford handbook of quantitative methods, vol. 2: Statistical analysis* (pp. 281-294). New York, NY: Oxford University Press.

Hunger, J. M., & Tomiyama, A. J. (2014). Weight labeling and obesity: A longitudinal study of girls aged 10 to 19 years. *Journal of the American Medical Association Pediatrics*, *168*(6), 579-580.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 297-307.

Jasik, C. B., & Lustig, R. H. (2008). Adolescent obesity and puberty: The "perfect storm". *Annals of the New York Academy of Sciences*, *1135*(1), 265-279.

Karlsson, M. O., & Sheiner, L. B. (1993). The importance of modeling interoccasion variability in population pharmacokinetic analyses. *Journal of Pharmacokinetics and Pharmacodynamics*, *21*(6), 735-750.

Kenny, R., Dooley, B., & Fitzgerald, A. (2013). Interpersonal relationships and emotional distress in adolescence. *Journal of Adolescence*, *36*(2), 351-360.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics-Simulation and computation*, *27*(3), 591-604.

Kiernan, K., Tao, J., & Gibbs, P. (2009). Tips and strategies for mixed modeling with SAS/STAT® procedures. Paper presented at SAS Global Forum, Orlando, FL.

Kullback, S. (1959). *Information theory and statistics*. New York, NY: John Wiley and Sons.

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.

La Greca, A. M., & Harrison, H. M. (2005). Adolescent peer relations, friendships, and romantic relationships: Do they predict social anxiety and depression?. *Journal of Clinical Child and Adolescent Psychology*, *34*(1), 49-61.

Latner, J. D., Rosewall, J. K., & Simmonds, M. B. (2007). Childhood obesity stigma: Association with television, videogame, and magazine exposure. *Body Image*, *4*(2), 147-155.

Latner, J. D., Stunkard, A. J., & Wilson, G. T. (2005). Stigmatized students: age, sex, and ethnicity effects in the stigmatization of obesity. *Obesity Research*, *13*(7), 1226-1231.

Lorr, M., & McNair, D. M. (1971). The profile of mood states manual. San Francisco, CA: Educational and Institutional Testing Service.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Taylor & Francis Group.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Mustillo, S. A., Budd, K., & Hendrix, K. (2013). Obesity, labeling, and psychological distress in late-childhood and adolescent black and white girls: The distal effects of stigma. *Social Psychology Quarterly*, *76*(3), 268-289.

Mustillo, S. A., Hendrix, K. L., & Schafer, M. H. (2012). Trajectories of body mass and self-concept in black and white girls: The lingering effects of stigma. *Journal of Health and Social Behavior*, *53*(1), 2-16.

National Heart, Lung, & Blood Institute. (2013). Managing overweight and obesity in adults: Systematic evidence review from the Obesity Expert Panel, 2013.

Neumark-Sztainer, D., Story, M., & Harris, T. (1999). Beliefs and attitudes about obesity among teachers and school health care providers working with adolescents. *Journal of Nutrition Education*, *31*(1), 3-9.

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2014). Prevalence of childhood and adult obesity in the United States, 2011-2012. *Journal of the American Medical Association*, *311*(8), 806-814.

Owens, J., Adolescent Sleep Working Group, & Committee on Adolescence. (2014). Insufficient sleep in adolescents and young adults: An update on causes and consequences. *Pediatrics*, *134*(3), e921-e932.

Paruthi, S., Brooks, L. J., D'Ambrosio, C., Hall, W. A., Kotagal, S., Lloyd, R. M, . . . Wise, M. S. (2016). Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine*, *12*(6), 785-786.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12-35.

Puhl, R. M., & Brownell, K. D. (2006). Confronting and coping with weight stigma: an investigation of overweight and obese adults. *Obesity*, *14*(10), 1802-1815.

Puhl, R. M., & Heuer, C. A. (2010). Obesity stigma: Important considerations for public health. *American Journal of Public Health*, *100*(6), 1019-1028.

Puhl, R. M., & Latner, J. D. (2007). Stigma, obesity, and the health of the nation's children. *Psychological Bulletin*, *133*(4), 557-580.

Rao, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician*, *25*(5), 37-39.

Rehkopf, D. H., Laraia, B. A., Segal, M., Braithwaite, D., & Epel, E. (2011). The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. *Pediatric Obesity*, *6*(2), 233-242.

Rogawski, J. (2008). *Calculus: Early transcendentals*. New York, NW: W. H. Freeman and Company.

Saguy, A. C., & Riley, K. W. (2005). Weighing both sides: Morality, mortality, and framing contests over obesity. *Journal of Health Politics, Policy and Law*, *30*(5), 869-923.

SAS Institute Inc. (2015). SAS/Stat 14.1 User's Guide. Cary, NC: SAS Institute Inc.

Schwartz, M. B., Chambliss, H. O. N., Brownell, K. D., Blair, S. N., & Billington, C. (2003). Weight bias among health professionals specializing in obesity. *Obesity Research*, *11*(9), 1033-1039.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(1), 379-423.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*, *7*(1), 13-26.

Swinburn, B. A., Sacks, G., Hall, K. D., McPherson, K., Finegood, D. T., Moodie, M. L., & Gortmaker, S. L. (2011). The global obesity pandemic: Shaped by global drivers and local environments. *The Lancet*, *378*(9793), 804-814.

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.

Tomiyama, A. J. (2014). Weight stigma is stressful. A review of evidence for the Cyclic Obesity/Weight-Based Stigma model. *Appetite*, *82*, 8-15.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*(2), 351-370.

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology*, *4*(1), 10-21.

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Comparison of modern methods for analyzing repeated measures data with missing values. *Multivariate Behavioral Research*, *46*(6), 900-937.

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.

Vartanian, L. R., & Novak, S. A. (2011). Internalized societal attitudes moderate the impact of weight stigma on avoidance of exercise. *Obesity*, *19*(4), 757-762.

Vartanian, L. R., & Shaprow, J. G. (2008). Effects of weight stigma on exercise motivation and behavior: a preliminary investigation among college-aged females. *Journal of Health Psychology*, *13*(1), 131-138.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer Verlag.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228-243.

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.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement*, *36*(3), 159-180.

Whittaker, T. A., & Furlow, C. F. (2009). The comparison of model selection criteria when selecting among competing hierarchical linear models. *Journal of Modern Applied Statistical Methods*, *8*(1), 173-183.

Wolfinger, R. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. Paper presented at the 24th annual SAS Users Group International Conference, Miami, FL.

Zheng, M. (2010). Fitting linear and nonlinear growth curve models using PROC NLMIXED. Paper presented at SAS Global Forum, Seattle, WA.