Article

Educational and Psychological Measurement
2016, Vol. 76(1) 64–87

DOI: 10.1177/0013164415580432
epm.sagepub.com

Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data

Chelsea Muth[1], Karen L. Bales[1], Katie Hinde[2], Nicole Maninger[1], Sally P. Mendoza[1], and Emilio Ferrer[1]

Abstract

Unavoidable sample size issues beset psychological research that involves scarce populations or costly laboratory procedures. When incorporating longitudinal designs these samples are further reduced by traditional modeling techniques, which perform listwise deletion for any instance of missing data. Moreover, these techniques are limited in their capacity to accommodate alternative correlation structures that are common in repeated measures studies. Researchers require sound quantitative methods to work with limited but valuable measures without degrading their data sets. This article provides a brief tutorial and exploration of two alternative longitudinal modeling techniques, linear mixed effects models and generalized estimating equations, as applied to a repeated measures study (n = 12) of pairmate attachment and social stress in primates. Both techniques provide comparable results, but each model offers unique information that can be helpful when deciding the right analytic tool.

[1]University of California, Davis, CA, USA
[2]Harvard University, Cambridge, MA, USA

Corresponding Author:
Chelsea Muth, Department of Psychology, University of California, 266 Young Hall, One Shields Ave, Davis, CA 95616, USA.
Email: cmuth@ucdavis.edu

Keywords

linear mixed effects models, generalized estimating equations, repeated measures ANOVA, small sample, longitudinal data

In psychological studies involving unique populations, costly laboratory procedures, or animals, sample size issues are unavoidable. Although measurement reliability and statistical power hinge on adequate subject pools, it is not often feasible for researchers to conduct large-sample, population-level studies (Button et al., 2013). Small subject pools may yield reduced variability estimates that may not generalize to the population, but if based on robust research methods and free of measurement error, these estimates are still revealing in context. When conducted longitudinally, small-sample studies can provide very rich information about sample-specific phenomena.

Studies that incorporate longitudinal data gain statistical power with repeated measures. Various quantitative methods may be used to analyze repeated measures and account for the dependency of the data, as well as individual and group variability. Standard longitudinal techniques include traditional models based on ordinary least squares (OLS) estimation, such as repeated measures analysis of variance (RM ANOVA). However, traditional OLS models require complete data and can degrade data sets by performing listwise deletion on cases with missingness. In such models, a subject's entire record is dropped from analysis when any one measure is missing. This is a significant drawback for small data sets (Rubin, Witkiewitz, Andre, & Reilly, 2007).

Furthermore, RM ANOVA's inability to flexibly model within-person correlations may result in biased and less-precise estimates when correlations are truly unequal across repeated measures. Theory often indicates that time-based effects cause subject scores to deviate differently across repeated measures, negating assumptions of independent residuals. For example, residuals at different occasions may exhibit autoregressive covariance (covariance changes as a function of the lag time between measures), or unstructured covariance (covariance differs uniquely across measures), in addition to independence (no covariance across repeated measures). If a model does not account for such variation, or its correlation structure is too simple, we underestimate standard errors of our estimates, which may inflate Type 1 error. Thus, RM ANOVA's assumption of equal variance is a challenge for longitudinal designs and increasingly so for smaller samples where standard errors are inherently larger (Howell, 2007).

To avoid these RM ANOVA issues for small longitudinal samples, various techniques offer modeling alternatives. One choice is to use imputations to replace missing data and calculate replacement values based on intact observations. Although imputation helps preserve sample size, this technique may be risky if missing values are outcome measures, the central focus in our model. Imputations add uncertainty and may misrepresent the sample. In a small data set, where individual values have more influence on variability estimates, this added uncertainty risks causing the very

problems that imputations were meant to avert: biased estimates and diminished precision (Cohen, Cohen, West, & Aiken, 2003).

Bayesian estimation methods also provide a sound alternative to RM ANOVA because of their proven advantages in working with limited samples and missing data. These techniques incorporate data-based evidence from previous studies to construct informative priors and offer additional information to model the study at hand. Research has shown great success for Bayesian models applied to small, unbalanced samples, and interested readers should pursue the literature further (Hsieh & Maier, 2009). Bayesian inference is beyond the scope of this article, which is within the frequentist framework.

Last, researchers may consider simply extending the RM ANOVA technique with the variance component approach to handling unbalanced data. Variance component analysis surmounts the issue of listwise deletion by adjusting sums of squares for each individual, independently estimating each subject's variance. Researchers may choose this option if confident that the correlation structure of their variables is equal across repeated measures; however, the variance component technique remains limited by the inability of RM ANOVA to handle alternative correlation structures (Graybill & Wortham, 1956).

In light of the above alternatives, this article intends to explore the advantages of two models that more flexibly handle correlation structures and unbalanced longitudinal data. We will focus on linear mixed effects (LME) models and generalized estimating equations (GEEs) and aim to advance the literature on their use for longitudinal research, in the context of limited sample sizes.

In this article, we apply LME and GEE models to a study of social stress and partnership in monogamous male titi monkeys (*Callicebus cupreus*) and examine whether they provide comparable measures of significance and precision in this small, unbalanced, repeated measures data set. Our goal with this article is to offer multifaceted considerations for modeling small longitudinal psychology data sets and to open the door for future exploration with LME and GEE models.

Broader Goals

This study intends to advance the longitudinal dialogue about LME and GEE models. Among others, we extend from the works of Burton, Gurrin, and Sly (1998) and Krueger and Tian (2004).

In Burton et al.'s (1998) tutorial, empirical application to a large sample longitudinal study (629 measures taken from 12 subjects) demonstrated comparable LME and GEE performance. Results showed that equivalent parameter value estimates across GEE and LME models (LME: $\beta = 0.247$, GEE: $\beta = 0.247$) had substantially larger standard errors in GEEs (LME: $SE = 0.033$, GEE: $SE = 0.065$; Burton et al., 1998). This study highlighted advantages, disadvantages, and deciding factors between LME and GEE, but only as applied to large samples. We aim to step forward from Burton et al.'s tutorial to examine whether GEE and LME performance remains

consistent with small samples. We hypothesize that our results will mirror Burton's—that GEE standard errors will remain larger than LMEs because of their inflated robust estimation.

In light of our small-sample focus, we draw from Krueger and Tian's (2004) research, which similarly examines LME's facility for modeling around small, unbalanced, longitudinal data. This study applied LME and RM ANOVA models to a study of biology and behavior with over 50% of observations missing (105 possible observations decreased to 43). Results showed clear challenges for RM ANOVA (which limited observations from 43 to 22 using listwise deletion), while LME models incorporated the total observations (Krueger & Tian, 2004). We aim to advance Krueger and Tian's findings by examining GEEs in addition to LMEs, with a small longitudinal data set with missing values.

Overall we intend to provide insight into the efficiency of LME and GEEs, in application, to help inform practical model choice. Readers should thus note throughout, our focus on measures of precision across LME and GEE models.

Description of LME and GEE

In repeated measures data, an individual's outcome scores are related to their scores at each subsequent time point. Statistically, these repeated outcomes exhibit within-cluster correlations, or patterns of variation corresponding to an individual, and thus contain dependency. Longitudinal data may contain within-cluster dependency associated with random predictor variables as well as residual scores. Residual correlation signifies that an individual's random deviations from model predictions vary in structured patterns across measures. Such forms of dependency must not be ignored. If treated independent, repeated measures can lead to biased parameter estimates, underestimated standard errors, and untrustworthy estimates (Diggle, Heagerty, Liang, & Zeger, 2002). Longitudinal dependency presents problems for traditional *t* tests, ANOVA, and simple regression models, which assume independent outcomes and residuals. LME and GEE techniques, however, flexibly model within-cluster correlations across measurement occasions and do not assume equal correlations across repeated measures, unlike the RM ANOVA technique (discussed below; Burton et al., 1998). Both LME and GEE efficiently account for dependency in outcome scores without inflating sample size, distorting the true structure of the data set (a consequence of ignoring correlation structure; Burton et al., 1998), or handling unbalanced designs with listwise deletion (Rubin et al., 2007). Likewise, LME and GEE handle correlated and heteroscedastic residuals (Ghisletta & Spini, 2004).

LME and GEE models partition repeated measures dependencies by estimating a general pooled prediction model (variation across all individuals) and accounting for each subject's correlation structure (variation within one individual). LME and GEE differ in how they model this correlation structure—LME estimates the structure simultaneously in a multilevel model and GEE does so orthogonally (Burton et al., 1998), as we will discuss in detail. As a result of these differences, as well as

differences in calculating standard errors, we expect LME and GEE models to yield equivalent estimates under certain conditions and divergent estimates under others. This hypothesis will be explored further in our results and discussion. The next sections provide a brief overview of the RM ANOVA approach, followed by descriptions of and comparisons between LME and GEE techniques.

Repeated Measures Analysis of Variance

Repeated measures ANOVA extends the ANOVA technique for comparing mean differences across repeated measures by accounting for groups with correlated data. This method partials out the dependency in repeated measures by subtracting within-person variability from the sum of squares error (Krueger & Tian, 2004). Although RM ANOVA partials dependency in repeated measures, the technique assumes equal correlations within subject and does not accommodate alternative structures. The RM ANOVA approach is represented by the following structural model (Howell, 2007):

$$m_{ij} = \alpha + s_i + t_j + e_{ij}. \quad (1)$$

In the RM ANOVA model, expected individual outcome scores m_{ij} for individual i in measurement assessment j are represented as a function of a grand mean α, a fixed subject effect s_i for the ith individual, a fixed treatment effect t_j for the jth measurement assessment, and a random residual error for each individual in each assessment e_{ij}.

RM ANOVA models are calculated with OLS estimation, which traditionally requires balanced designs. For repeated measures data, this means a wide-format data set with one row per subject, and equal measures across all rows. In order to avert introducing bias in estimates, the estimation procedure performs listwise deletion when encountering any cell with missing data (Howell, 2007).

Linear Mixed Effects Models

To account for correlated outcome measures, LME models estimate a pooled multilevel equation by simultaneously incorporating fixed and random effects. This model can be represented as (Singer, 1998)

$$m_{ij} = \alpha_{0j} + \beta_{1j} t_{ij} + e_{ij}. \quad (2)$$

In this Level 1 equation, outcome scores m_{ij} for individual i in repeated assessment j are represented as a linear combination of a random mean α_{0j} for each assessment, a random slope β_{1j} for each individual in each assessment (t_{ij}, measurement occasion, or any other underlying metric), and a random residual error for each individual in each assessment e_{ij}.

Simultaneous to this pooled equation, LME models estimate random effects that account for variation across individuals (Singer, 1998). In a basic model, these random effects include an intercept and a slope, as follows:

$$\alpha_{0j} = \alpha_{00} + u_i, \tag{3}$$

$$\beta_{0j} = \beta_{10} + v_i. \tag{4}$$

In these Level 2 equations, random intercepts α_{0j} are represented as a function of a grand mean α_{00} and conditional deviations μ_i from it. Random slopes β_{0j} are represented as a function of a group-specific slope β_{10}, and conditional slope deviations v_i.

As illustrated, the LME model extends the generalized linear model by allowing varying intercepts and slopes across individuals. When calculating parameters, LME's multiple levels allow for a fluctuating structure of correlation between repeated outcome scores and between residuals, thereby accounting for dependency in scores nested within an individual (Burton et al., 1998). Various correlation matrices may be specified to account for differing structures of covariance in random variables as well as residuals. Most statistical programming software assumes default correlation structures. This study used the default structure in R's *lme* package (R Development Core Team, 2011), an independent matrix, corresponding to zero within-group correlations and zero residual correlations.

Generalized Estimating Equations

Like LME models, GEEs estimate pooled regression equations. However, instead of multilevel estimation, they use solely this population-level equation and account for dependency in repeated measures through the residuals and their correlation structure. GEE equations are calculated by first removing residuals from the general model. This marginal model is represented as (Burton et al., 1998) follows:

$$E(m_{ij}) = \alpha + \beta t_{ij}. \tag{5}$$

In the marginal model, expected individual outcome scores $E(m_{ij})$ for individual i in measurement assessment j are represented as a function of a grand mean α and a common slope β for all individuals at each assessment t.

Residuals are then orthogonally and iteratively estimated as

$$r_{ij} = m_{ij} - (\alpha + \beta t_{ij}). \tag{6}$$

Residuals r_{ij} for each individual i are estimated as a function of the individual's observed outcome m_{ij}, minus the combined grand mean α and common slope for all individuals in all assessments βt_{ij}. These residual estimates are incorporated in a predetermined correlation matrix, which iteratively estimates the marginal model parameters. The iterative process continues until marginal estimates stabilize and converge (Burton et al., 1998).

To handle dependency in GEE models, one must explicitly specify the residual correlation matrix for GEE estimation (Halekoh, Hojsgaard & Yan, 2006). This matrix is referred to as a working correlation matrix, because of its estimation of robust standard errors that provide consistent and unbiased estimates regardless of misspecification (Ghisletta & Spini 2004). These robust standard errors are inflated measures, calculated using the matrix of squared residuals (Zorn, 2006). GEEs use various types of working correlation matrices, which include several key structures (Gosho, 2014):

1. Independent: zero correlation over time (i.e., all off-diagonal elements of the correlation matrix are zero)
2. Exchangeable: constant correlation over time (i.e., all off-diagonal elements are equal)
3. Autoregressive: diminishing correlation over time
4. Unstructured: freely estimated correlation (i.e., no equality constraints within correlation matrix, model is saturated)
5. Specified or fixed: fixed correlation, uniquely determined by user (i.e., unique matrix is designed by analyst based on theory) (Ghisletta, & Spini, 2004)

These structures may also be used for LME models, but as mentioned above, LME, unlike GEE, assumes independent matrices by default unless otherwise specified.

Of these working correlation matrices, the more complex structures that estimate a large number of parameters and use a high number of degrees of freedom should not be chosen for small-sample data sets. This includes unstructured matrices (see above), which may be theoretically sound, but freely estimate all correlation parameters and require more degrees of freedom than available in small data (Grace-Martin, n.d.).

Correlation specification should be approached with forethought, preferably using substantive theory of how subject scores relate at each repeated measure. For example, an exchangeable structure would be reasonable if an individual's test scores were expected to correlate equally over time, and earlier tests were not expected to influence later scores. Accuracy of parameter estimates depends on choosing the correct structure. Misspecified structures may result in inefficient and inconsistent parameter estimates (Gosho, 2014). That said, as mentioned above, GEE accounts for potential misspecification with robust standard errors. Although inflated measures are less precise, they remain unbiased, regardless of correlation matrix misspecification (Ghisletta & Spini, 2004).

LME and GEE in Comparison

For many studies, including those that hypothesize unequal within-subject correlations or significant random effects, or that contain high proportions of missing data, these two longitudinal approaches offer clear advantages over traditional OLS-based estimation methods for repeated measures. The LME and GEE techniques are both

sound modeling choices for repeated measures data, capable of handling dependency and missingness efficiently. Both approaches present a general, fixed-level equation, whereas LME models offer multilevel equations that incorporate random effects, and GEE models offer one marginal equation tailored to prespecified correlation matrices. The two models have distinct approaches to partialing dependency in repeated measures, and as a result may provide different estimates for the same data.

We hypothesize that the extent to which LME and GEE estimates differ relates to factors that most influence their unique features: correlation specification and standard error estimation. Differences in the efficacy between LME and GEEs may be most influenced by correlation structure and sample size, for at least two reasons. First, estimates depend on the complexity of the correlation structure and the extent to which LME or GEE capture it accurately. For instances of misspecification, LME and GEE models may diverge, in which case GEEs may be preferable due to robust standard errors. Second, robust standard errors and normal standard errors vary according to sample size, power, and precision. In large samples that more closely represent population behavior, LME and GEE estimates will likely appear more similar, whereas in smaller, less reliable samples, differences between LME and GEE estimates may increase. In small samples with lower power, GEEs may be preferable for their robust standard errors. We explore these hypotheses in the following empirical application by examining differences and similarities between LME and GEEs in light of correlation structure and sample size conditions.

Given the assumptions of LME and GEEs, we aim to examine the relevance of these models for small-sample, longitudinal studies with missing data. The next section provides our empirical application. We first outline the theoretical rationale behind the study, and then apply the proposed methods to the data and describe results. We conclude with comparisons and recommendations for the use of LME and GEE.

Empirical Illustration

Biological Effects of Stress and Separation in Longitudinal Study of Titi Monkeys

Titi monkeys (*Callicebus cupreus*) are a monogamous species that demonstrate strong heterosexual pair bonds in adulthood through such behavioral attributes as proximity seeking and separation distress (increased vocalization, heart rate, and cortisol levels; Mason & Mendoza, 1998). Previous studies indicate that pair bonding has measurable effects on physiological processes and systems, including the hypothalamic–pituitary–adrenal (HPA) axis, the central hub of the neuroendocrine system for physiological stress response regulation and homeostasis restoration. Pair bonding has been linked to at least two change processes involving the HPA axis.

First, titi monkey pairmates exhibit stress buffering. This social mechanism has been observed in a variety of species with relationships characterized by attachment

or emotional bonds. In stress buffering, mates may protect each other from experiencing the full array of physiological responses to stress, which has broad implications for affecting HPA system functionality and reactivity (Hennessy, 1997).

Second, research suggests that pair bonding leads to a regulatory shift in the HPA system. This regulatory shift negatively alters the normal physiological restoration of homeostasis after stress-induced HPA activation—a process known as the negative feedback loop. When separated from their pairmates, titi monkeys exhibit an impaired negative feedback loop, or failure to restore homeostasis after stress (Mendoza, Capitanio, & Mason, 2000). On pairing, subjects also exhibit a change in regulation of negative feedback, leading to lower baseline levels of cortisol and chronically lower basal HPA activity (Mendoza et al., 2000). Moreover, subjects in long-term partnerships have shown globally higher brain activity than unpartnered peers, possibly due to changes in cortisol, a primary metabolic hormone (Bales, Mason, Catana, Cherry, & Mendoza, 2007). It should be noted that the regulatory shift accompanying pair bonding does not alter subjects' ability to respond to stress, demonstrated by the fact that lone and paired males exhibit cortisol elevation in response to acute stressors (e.g., capture and handling; Rothwell, Mendoza, Mason, Ragen, & Bales, 2013).

In sum, pair bonding is predicted to elicit these two HPA change processes, thus suggesting a link between pair bonds and biological social bonding as well as biological stress response (Bales et al., 2007). To date, however, attachment studies have not measured the pair bonding effects of social stress buffering or regulatory shifts across extended levels of partnership (other than early relationship bonds) and separation (other than 1-month separation).

Method

Participants. This study examined data from titi monkeys housed at the California National Primate Research Center, in Davis, California. Subjects were 12 captive-born adult males, with their cohabitating female partners. The mean age of subjects was 5.8 years (range = 2.9-8.7), and their average length of cohabitated partnership with female mates was 0.97 years. Subjects participated in five experimental conditions of separation and partnership, designed to compare whether the magnitude (either long term or short term) of partnership and/or separation affected hormone levels. These conditions included baseline, short-term separation, long-term separation, partnering with strange female, and reunion with long-term mate. All but three subjects were measured at all five conditions (two measured only at baseline and short-term separation, and one measured only at baseline). In addition, cases of missing hormone data varied for individuals at each measurement occasion, because of sensitivity of hormone assays.

Measures. Subjects, pairmates, and offspring less than 1 year old were relocated to a metabolism room 48 hours prior to blood draws. This relocation period was undertaken to reduce possible effects of novel housing on subject metabolism, as described

in Bales et al. (2007). Blood and cerebrospinal fluid (CSF) samples were collected from all subjects, and plasma samples were assayed for plasma cortisol, vasopressin (AVP), oxytocin (OT), cerebrospinal fluid vasopressin (CSF AVP), cerebrospinal fluid oxytocin (CSF OT), plasma glucose, and plasma insulin. These outcome variables are all implicated in pair bonding and stress response (Bales et al., 2007). Details of hormone assays can be found in substantive analyses.

All procedures in this study were approved by the corresponding university's Animal Care and Use Committee and complied with National Institutes of Health ethical guidelines as set forth in the Guide for Lab Animal Care. Blood draw protocol can be found in detail, identical to previous research on pair bonding (Bales et al., 2007).

Design. Male subjects underwent separation and partnership conditions with concurrent hormone measurement. The first two conditions were counterbalanced: baseline (control) and 48 hours of separation from pairmate (short-term separation). All subjects were then separated from their pairmates for approximately 2 weeks and measured for long-term separation (i.e., all subjects had the same third condition). After this 2-week separation, the last two conditions were also counterbalanced: reunion with pairmate and encounter with a strange female. The same stimulus ''stranger'' female was used for all stranger-partnership conditions. This female had been previously hysterectomized and was not ovulating while in the presence of subjects. She was observed closely during repeated exposures to strange males; in all cases she appeared unstressed and interacted normally with subjects. All animals were fed twice daily; details of husbandry, training, and caging are identical to those described in Tardif et al. (2006).

In addition to counterbalancing, it should be noted that measures were not equally spaced for all subjects—some males were separated from their mates longer than others by approximately 1 to 2 weeks.

Experimental conditions were parameterized in three ways such that, including two interaction models, we tested five different categories of models (R Development Core Team, 2011). The first two model types used parameters based on condition, dummy coding either condition by group (partnered vs. separated) or each condition separately. The third model used a predictor based on time to model measurement order, ignoring condition. The resulting models were the following:

1. Partnership: partnership (reference) versus separation conditions
2. Condition: baseline (reference) versus four experimental conditions
3. Time: measurement order, counterbalanced and unique to each subject
4. Partnership × Time: interaction model
5. Condition × Time: interaction model

For each model type, we tested seven hormones as separate outcomes. Figure 1 provides a descriptive summary of mean changes in hormone levels across conditions

(based on the sum of individual *z* scores, standardized at the baseline condition mean for each hormone).

Our analyses aimed to determine whether differences in stress-induced responses exist across long-term versus short-term separation, and long-term versus stranger-partnership conditions. LME and GEE models were applied to address these questions and to examine the unique contributions of each approach. For methodological purposes, the following analyses strive to illustrate the benefits and limitations of GEE and LME as applied to a small sample, repeated measures data.

Results

A Comparative Starting Point: RM ANOVA

In the first set of analyses, we use RM ANOVA as a starting point for comparative assessment. Table 1 summarizes significant parameter estimates from these RM ANOVA tests of two of our five model types. The two model types, (a) partnership or (b) condition, were applied to each hormone outcome. Across all models (non-significant results are not included), RM ANOVA captures multiple measures of significance for plasma cortisol models and separation predictors. For example, consider the significant estimates in the Plasma Cortisol–Partnership and Plasma Cortisol–Condition models (separation: $\beta = 0.577$, $SE = 0.213$; short-term separation: $\beta = 1.059$, $SE = 0.331$). However, RM ANOVA performs listwise deletion for cases of missingness, reducing both the sample size and the analytic power. CSF AVP and CSF OT models, for example, reflect reduced sample sizes of $n = 4$ and $n = 5$. In an already small sample, this reduction in power is quite drastic.

To briefly compare LME and RM ANOVA results, consider the first half of Table 2, which corresponds with Table 1. Like RM ANOVA estimates, LME parameters capture similar measures of significance (magnitude and direction of effects) in plasma cortisol models, also with separation predictors. In contrast, however, LME estimates are more precise due to handling missing data without listwise deletion. For example, consider the estimates of the predictor short-term separation in Plasma Cortisol–Condition models (RM ANOVA: $\beta = 1.059$, $SE = 0.331$; LME: $\beta = 1.012$, $SE = 0.266$). Similarly, in the first half of Table 3, GEE's handling of missing data reflects an advantage over RM ANOVA (short-term separation in Plasma Cortisol–Condition models: RM ANOVA: $\beta = 1.059$, $SE = 0.331$; GEE: $\beta = 0.944$, $SE = 0.258$). The key distinction here is the model estimation method. While LME and GEE work around missing data with maximum likelihood and iteratively reweighted least squares estimation, RM ANOVA performs listwise deletion for cases of missingness.

We do not extend use of RM ANOVA beyond this preliminary overview, but now offer more thorough analyses of LME and GEE models only. These models are a clear choice over RM ANOVA for these data because of their ability to handle missingness without listwise deletion.

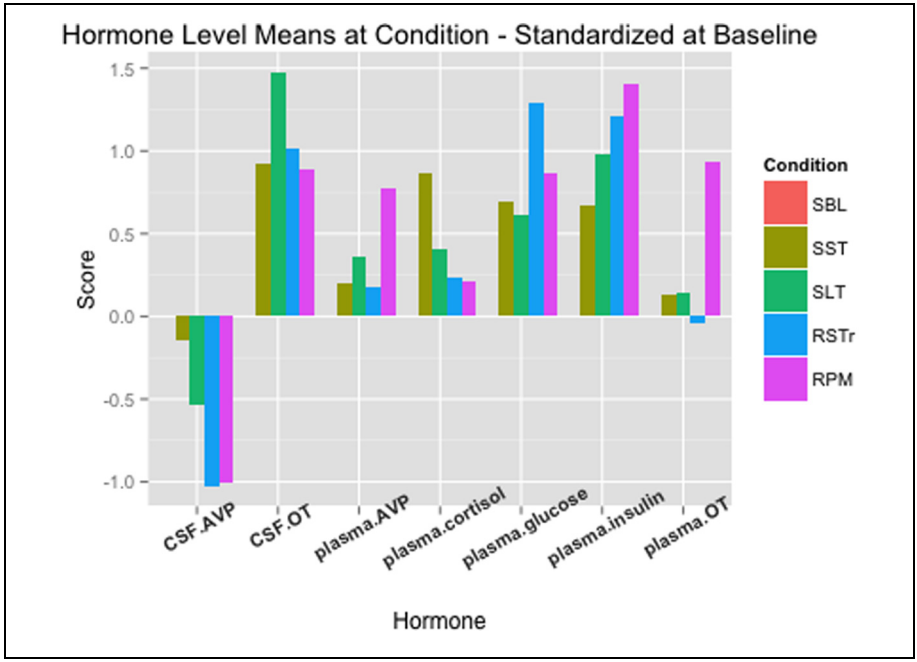

Figure 1. Mean hormone levels at each measurement occasion.
Note. SBL = baseline; SST = short-term separation; SLT = long-term separation; RSTr = partnership with stranger; RPM = reunion with pairmate.

Hormone Measures: LME Versus GEE Models

The following sections summarize results from comparable LME and GEE tests of our five model types. As outlined earlier, these models tested planned comparisons for (a) partnership, (b) condition, (c) time (ignoring condition), (d) time by partnership interaction, and (e) time by condition interaction. These models were chosen to assess parameters that best explain the variance in hormone scores.

Precision and Efficacy. Table 4 illustrates a full comparison of significant and nonsignificant results, across all models tested for plasma cortisol. This is a snapshot (10 models) from 91 total models: 35 total LME models (5 contrast models for each of 7 dependent variables), 35 GEE models with an initial correlation structure (Structure 1), and 21 follow-up GEE models with a comparison correlation structure (Structure 2). Structure 1 GEE models incorporate two types of correlation structures: (a) exchangeable, for models that did not include time as a predictor, and (b) autoregressive, for time-based models. This specification reflects a joint hypothesis: (a) observations within a subject are equally correlated across counterbalanced conditions (i.e., when ignoring measurement order); and (b) when accounting for order, correlations diminish at each subsequent measurement (i.e., scores from conditions that were

Table 1. Parameter Estimates From RM ANOVA, Within-Subjects Effects.

Hormone[b]	Source	Contrast value	*SE*	*df*	*t* contrast	*p* value
Partnership						
Plasma cortisol						
	Separation	0.577	0.213	32	2.706	.011
CSF OT[a]						
	Separation	0.625	0.299	16	2.088	.053
Condition						
Plasma cortisol						
	Short-term separation	1.059	0.331	32	3.204	.003
Plasma glucose						
	Partner stranger	1.167	0.582	32	2.006	.053
Plasma insulin						
	Partner stranger	1.234	0.471	28	2.620	.014
	Reunion pairmate	1.310	0.471	28	2.782	.010
CSF OT[a]						
	Long-term separation	1.580	0.464	16	3.407	.004

Note. RM ANOVA = repeated measures analysis of variance; CSF AVP = cerebrospinal fluid vasopressin; CSF OT = cerebrospinal fluid oxytocin. Only significant effects are reported.
[a]CSF AVP $n = 4$, due to missingness, CSF OT $n = 5$, due to missingness.
[b]Scores standardized at mean of baseline condition.

measured further apart are less correlated than those measured closer together). To explore the possibility of misspecification for the autoregressive component in Structure 1, Structure 2 GEEs test time-based equations with exchangeable correlation. Theoretically, these exchangeable models reflect a second correlation hypothesis: observations within a subject are equally correlated across conditions, regardless of when the measures were taken.

Table 4 gives a closer view of parameter comparisons across LMEs and GEEs (Structure 1 only), and their varying levels of uncertainty in standard error estimates. These estimates show GEE robust standard errors are more precise overall than LME estimates. Structure 2 GEE estimates (not detailed here) are also more precise overall than LMEs. This outcome disproves our initial hypothesis that LME models would be more precise than GEEs, as observed with large sample studies (Burton et al., 1998).

Table 5 gives a full comparison across all models tested for all hormones. In summary, GEE model estimates for both sets of correlation specifications are more precise than LME estimates in over 75% of our results. Although, as seen in Table 4, the differences between LME and GEE standard errors estimates are sometimes slight, the frequency (75.8% to 83.6%) with which we observe GEE's superior precision is notable (Table 5). Moreover, we assess GEE precision in light of the fact that these standard errors are robust estimates, inflated to buffer against correlation structure misspecification. LME standard errors are not inflated estimates; thus, if these

Table 2. Parameter Estimates From LME Models.[b]

Hormone	Source	Value	*SE*	*df*	*t* value	*p* value
Partnership						
Plasma cortisol						
	Separation	0.566	0.207	37	2.730	.010
Condition						
Plasma cortisol						
	Short-term separation	1.012	0.266	34	3.810	.001
CSF AVP[a]						
	Partner stranger	−1.023	0.419	24	−2.440	.022
	Reunion pairmate	−1.000	0.433	24	−2.310	.030
CSF OT[a]						
	Short-term separation	0.786	0.382	27	2.060	.050
	Long-term separation	1.295	0.405	27	3.200	.004
	Partner stranger	0.830	0.397	27	2.090	.046
Plasma insulin						
	Short-term separation	0.670	0.319	33	2.100	.044
	Long-term separation	1.000	0.339	33	2.950	.006
	Partner stranger	1.190	0.444	33	2.690	.011
	Reunion pairmate	1.550	0.328	33	4.720	0
Time point						
Plasma insulin[a]	Time	0.325	0.093	36	3.500	.001
Time Point × Partnership						
Plasma cortisol	Separation	1.156	0.470	35	2.457	.019
Plasma AVP						
	Separation	0.789	0.383	32	2.058	.048
	Time × separation	−0.296	0.135	32	–2.197	.035
Plasma insulin[a]						
	Time	0.408	0.117	34	3.500	.001
Time point × condition						
Plasma cortisol	Short-term separation	1.269	0.532	29	2.390	.024

Note. LME = linear mixed effects; CSF AVP = cerebrospinal fluid vasopressin; CSF OT = cerebrospinal fluid oxytocin. Only significant effects are reported.

[a]Original model did not converge with maximum likelihood estimation; model refit using restricted maximum likelihood, with random effects fixed at 1.

[b]No significant random effects.

models were equal in performance, we would expect LME standard errors to be smaller than GEEs, as observed in large sample studies (Burton et al., 1998).

This distinction suggests that, for this data set, we may trust GEE model measures of significance (and nonsignificance) more than LME model estimates. Because of their superior precision, GEE models may be more accurate and reliable for sample-specific inference.

Estimates. Given the notable differences in standard errors, we now compare measures of significance and corresponding parameters captured by LME and GEE models. Refer again to Table 2 for statistically significant LME results. In 35 LME

hormone models, 17 out of 126 predictors produce p values significant to reject a null hypothesis of parameters equal to zero in the population. No LME models account for significant random effects. Several LME models of CSF AVP, CSF OT, plasma AVP, and plasma insulin scores did not converge with maximum likelihood, and were reestimated with restricted maximum likelihood and random effects set equal to zero. By stabilizing random effects, these models converged, and CSF AVP, CSF OT, and plasma insulin models contained significant fixed effects.

Likewise for GEE models, 24 out of 126 predictors produce p values significant to reject the corresponding null hypotheses (Table 3). Unlike LME, no convergence issues were encountered with GEE models.

Comparison of parameter estimates from Tables 2 and 3 illustrates that LME and GEE models provide similar measures of significance for separation and partnership predictors—most commonly with plasma cortisol models. For example, consider estimates for the predictor Separation, in Plasma Cortisol–Partnership (LME: $\beta = 0.566$; $SE = 0.207$; GEE exchangeable: $\beta = 0.564$; $SE = 0.113$). As discussed, GEE estimates are more precise than the LMEs, with smaller standard errors and p values, excluding several less precise estimates in time-based GEE models that used an autoregressive correlation structure. This may be resultant from a misspecified correlation structure. Perhaps, measurements taken further apart are still equally correlated, not diminishing.

Structure 2 GEE models were implemented to test time-based equations with exchangeable correlation structures. Across all time-based GEE models, 11 out of 91 autoregressive (Structure 1) versus 12 out of 91 exchangeable predictors (Structure 2) produce significant p values. Table 6 provides all significant follow-up measures with exchangeable correlations, to be compared with autoregressive time-based measures in Table 3. The autoregressive versus exchangeable time models share eight significant predictors (all with effect sizes of same direction and relatively equivalent magnitude), and seven of eight are now more precise in models with exchangeable correlations. For example, compare estimates of the predictor Separation, in Plasma Cortisol–Time $\times$ Partnership (LME: $\beta = 1.156$; $SE = 0.470$; GEE autoregressive: $\beta = 1.371$; $SE = 0.639$; GEE exchangeable: $\beta = 1.119$; $SE = 0.527$).

Discussion

Methodological Considerations

The preceding results provide distinct comparisons between LME and GEE models, as applied to this small-sample, longitudinal data. On the whole, LME and GEE models handle the data set well. LME and GEE both preserve sample size despite missingness, an advantage over RM ANOVA, and provide similar measures of significance. However, several indicators suggest the LME approach is less efficient for these measures than GEE. First, in this empirical study, we are only interested in population-level trajectories, not in the estimation of variability in individual trajectories of change, as modeled by LME. Moreover, the small sample does not in fact

Table 3. Parameter Estimates From GEE Models.

Hormone	Source	Value	Robust *SE*	*df* (*df*-resid)	Wald	Pr (> |*W*|)
Partnership (exchangeable)						
Plasma cortisol						
	Separation	0.564	0.113	50 (48)	25.104	.000
CSF OT						
	Separation	0.460	0.196	42 (40)	5.480	.019
Condition (exchangeable)						
Plasma cortisol						
	Short-term separation	0.944	0.258	50 (48)	13.400	.000
CSF AVP						
	Partner stranger	−1.132	0.409	39 (34)	7.680	.006
	Reunion pairmate	−1.096	0.474	39 (34)	5.341	.021
Plasma OT						
	Reunion pairmate	0.872	0.342	49 (44)	6.500	.011
CSF OT						
	Short-term separation	0.772	0.348	42 (37)	4.933	.026
	Long-term separation	1.276	0.507	42 (37)	6.336	.012
	Partner stranger	0.810	0.494	42 (37)	2.688	.101
	Reunion pairmate	0.632	0.225	42 (37)	7.914	.005
Plasma insulin						
	Short-term separation	0.670	0.306	49 (44)	4.790	.029
	Long-term separation	0.950	0.325	49 (44)	8.550	.003
	Partner stranger	1.220	0.431	49 (44)	8.000	.005
	Reunion pairmate	1.380	0.353	49 (44)	15.300	.0000926

(continued)

Table 3. (continued)

Hormone	Source	Value	Robust *SE*	*df* (*df*-resid)	Wald	Pr (> |*W*|)
Time point (autoregressive)						
Plasma insulin						
	Time	0.350	0.127	49 (47)	7.565	.006
Time point × partnership (autoregressive)						
Plasma cortisol						
	Separation	1.371	0.639	50 (46)	4.600	.032
Plasma insulin						
	Time	0.421	0.155	49 (45)	7.338	.007
Time point × condition (autoregressive)						
Plasma cortisol						
	Short-term separation	1.253	0.505	50 (40)	6.156	.013
	Long-term separation	2.535	1.064	50 (40)	5.681	.017
	Time × long-term separation	−0.902	0.405	50 (40)	4.966	.026
Plasma AVP						
	Time	−0.342	0.163	47 (37)	4.432	.035
	Partner stranger	−2.709	1.142	47 (37)	5.624	.018
	Reunion pairmate	−2.508	1.284	47 (37)	3.817	.051
	Time × partner stranger	0.852	0.322	47 (37)	7.008	.008
	Time × reunion pairmate	0.830	0.368	47 (37)	5.078	.024

Note. GEE = generalized estimating equations; CSF AVP = cerebrospinal fluid vasopressin; CSF OT = cerebrospinal fluid oxytocin. Only significant effects are reported.

support tests for random effects. This reinforces previous simulation studies that indicate random effects function poorly with small samples, as a result of making too many estimates from too few pieces of information (Bell, Ferron, & Kromrey, 2008). Last, as mentioned, LME models face convergence issues when handling the most reduced samples in this data set. This is not an issue for GEE models. GEE models, on the other hand, facilitate exploration of the correlation structure over time as well as the effect of time as a predictor of outcome scores. Interestingly, predictors in GEE models are generally more precise than LME predictors, a notable advantage and dissimilarity.

Consistent with our hypothesis that small samples may yield greater differences in LME and GEE estimates, because of reduced reliability, we indeed found divergence of standard error estimates across LME and GEE models. However, we were surprised to find consistently superior precision with robust GEE standard errors. This may be due to multiple factors, namely, those influencing the correlation structure and the standard error estimation. These models provide different estimates, corresponding with different correlation specifications, which may result from each model's ability to accurately model the true correlation. In this study, GEE models appear more reasonable for their theory-driven correlation structure and standard errors that are buffered for misspecification. Moreover, these results raise the possibility that GEEs may be more efficient than LME models for small samples.

Substantive Considerations

First, as mentioned above, inferences from this small-sample study should be cautious. Tables 4 and 5 demonstrate the range of standard errors across models, which make us question the extent to which we can generalize beyond our sample. Yet, as discussed below, these longitudinal measures provide valuable considerations for the behavior of our sample subjects.

From a substantive standpoint, patterns in significant and nonsignificant effects are distinguishable across models. Several key patterns warrant follow-up measures and inferential exploration. The most frequently significant hormone outcome is plasma cortisol, by a far margin. This pattern suggests that separation and partnership may have a stronger impact on this particular hormone, as compared with others.

Condition-based models capture more significant variance than partnership-based models. In other words, more variance in hormone measures is explained by accounting for each condition separately, than by grouping them as partnership versus separation. This suggests that each condition—short-term separation, long-term separation, partnership with stranger, and reunion with pairmate—may have a unique effect on baseline measures of biological attachment and social stress markers in male titi monkey, and provides further cause to investigate the two separate processes linked to pair ponding, stress-buffering effects, and shifts in negative feedback regulation.

Last, time has nonsignificant effects on hormone measures overall and reduces significance of condition predictors in interaction models. In other words, time may be a

Table 4. Plasma Cortisol Model Estimates: Linear Mixed Effects (LME) Models and Generalized Estimating Equations (GEE)[a].

	LME[b] (fixed effects estimates)				GEE[c] (Structure 1)			
	Partnership				Partnership			
Source	Value	*SE*	*t* value	*p* value	Value	Robust *SE*	Wald	*Pr*(>|*W*|)
(Intercept)	0.125	0.257	0.488	.628	0.124	0.241	0.264	.608
Separation	0.566	0.207	2.730	.010	0.564	0.113	25.104	.000
	Condition				Condition			
	Value	*SE*	*t* value	*p* value	Value	Robust *SE*	Wald	*Pr*(>|*W*|)
(Intercept)	0.000	0.291	0.000	1.000	0.000	0.276	0.000	1.000
Short-term separation	1.012	0.266	3.810	.001	0.944	0.258	13.400	.000
Long-term separation	0.421	0.222	1.900	.066	0.385	0.215	3.210	.073
Partner stranger	0.223	0.311	0.720	.478	0.218	0.307	0.507	.476
Reunion pairmate	0.185	0.316	0.590	.562	0.190	0.320	0.355	.551
	Time point				Time point			
	Value	*SE*	*t* value	*p* value	Value	Robust *SE*	Wald	*Pr*(>|*W*|)
(Intercept)	0.289	0.331	0.874	.388	0.299	0.309	0.936	.333
Time	0.020	0.080	0.252	.802	0.008	0.081	0.010	.922
	Time point × Partnership				Time point × Partnership			
	Value	*SE*	*t* value	*p* value	Value	Robust *SE*	Wald	*Pr*(>|*W*|)
(Intercept)	−0.373	0.410	−0.909	.370	−0.477	0.400	1.420	.233
Time	0.160	0.089	1.784	.083	0.177	0.121	2.130	.144
Separation	1.156	0.470	2.457	.019	1.371	0.639	4.600	.032
Time × Separation	−0.185	0.164	−1.125	.268	−0.298	0.251	1.420	.234

(continued)

Table 4. (continued)

Source	LME[b] (fixed effects estimates)				GEE[c] (Structure 1)			
	Time point × Condition				Time point × Condition			
	Value	*SE*	*t* value	*p* value	Value	Robust *SE*	Wald	*Pr*(>|*W*|)
(Intercept)	−0.667	0.545	−1.230	.230	−0.728	0.488	2.223	.136
Time	0.325	0.200	1.630	.114	0.316	0.171	3.426	.064
Short-term separation	1.269	0.532	2.390	.024	1.253	0.505	6.156	.013
Long-term separation	2.002	1.194	1.680	.104	2.535	1.064	5.681	.017
Partner stranger	2.033	1.379	1.470	.151	2.334	1.279	3.328	.068
Reunion pairmate	−0.298	1.355	−0.220	.828	−0.844	1.093	0.597	.440
Time × Short-term separation	−0.126	0.225	−0.560	.580	−0.124	0.232	0.286	.593
Time × Long-term separation	−0.679	0.462	−1.470	.153	−0.902	0.405	4.966	.026
Time × Partner stranger	−0.601	0.396	−1.520	.140	−0.678	0.380	3.187	.074
Time × Reunion pairmate	−0.034	0.384	−0.090	.929	0.091	0.301	0.092	.762

[a]GEE correlation structures: GEE1 = exchangeable for non-time-based models, autoregressive for time-based models; GEE2 = exchangeable for all models (GEE2 results not listed here).

[b]Linear mixed-effects model fit by maximum likelihood (hormone scores standardized at mean of baseline).

[c]Generalized estimating equation model fit by iteratively reweighted least squares (hormone scores standardized at mean of baseline).

Table 5. LME Versus GEE Precision Comparison: Percentage of Most Precise Standard Errors (Out of Total) by Parameter Groupings.

	Comparison 1		Comparison 2	
Model	LME	GEE1[a]	LME	GEE2[b]
Partnership	0.071	0.929	0.071	0.929
Condition	0.382	0.618	0.382	0.618
Time	0.357	0.643	0.071	0.929
Time × Partnership	0.286	0.714	0.250	0.750
Time × Condition	0.114	0.886	0.043	0.957
Cumulative percentage	0.242	0.758	0.164	0.836

Note. LME = linear mixed effects; GEE = generalized estimating equations.
[a]GEE1 = Exchangeable for non-time-based models, autoregressive for time-based models.
[b]GEE2 = Exchangeable for all models.

poor predictor of partnership and separation effects on hormone levels. Our results indicate several substantive possibilities to influence subsequent research questions: (a) hormone measures react differently to time as a predictor (some hormones are less sensitive to time than others); (b) different hormone measures have different residual correlation structures (some hormones are more robust to time effects than others); (c) time is an unreliable predictor in this counterbalanced study, and models best account for variance in outcome scores when ignoring measurement occasion (or time).

These highlights simply touch on the range of patterns distinguishable from LME and GEE models, both individually and in comparison with each other. This level of comparison, precision, and sophisticated inclusion of time would not have been possible with models based on OLS estimation.

Conclusion

This article aimed to illustrate the application of two modeling techniques to analyze repeated measures data with in a small sample. Inferences are limited when analytic models do not adjust for small sample sizes. However, knowing which models best fit one's data enables us to test hypotheses and explore patterns of variability more efficiently.

Through application to small unbalanced longitudinal data, our analyses suggest that GEE models may be more efficient than LME models under the given conditions. To confirm or counter this possibility and make sufficient recommendations for a broader audience, further research should investigate the reliability of these models across multiple small-sample longitudinal data sets.

We hope that our analyses might help inform modeling of repeated measures in studies of limited sample size. Such studies, especially when driven by strong

Table 6. Parameter Estimates From GEE Time-Based Models.

Hormone	Source	Value	Robust *SE*	*df* (*df*-resid)	Wald	*Pr*(> |*W*|)
Time point (exchangeable)						
Plasma insulin						
	Time	0.325	0.114	49 (47)	8.200	.004
Time point × Partnership (exchangeable)						
Plasma cortisol						
	Separation	1.119	0.527	50 (46)	4.520	.034
CSF OT						
	Time	0.214	0.059	42 (38)	13.190	.000
	Separation	1.152	0.278	42 (38)	17.120	.000
Plasma insulin						
	Time	0.421	0.139	49 (45)	8.660	.003
Time point × Condition (exchangeable)						
Plasma cortisol						
	Time	0.313	0.137	50 (40)	5.220	.022
	Short-term separation	1.184	0.509	50 (40)	5.410	.020
	Long-term separation	1.971	1.007	50 (40)	3.830	.050
Plasma AVP						
	Time	–0.294	0.143	47 (37)	4.210	.040
	Partner stranger	–2.267	0.976	47 (37)	5.390	.020
	Time × Partner stranger	0.717	0.306	47 (37)	5.490	.019
CSF OT	Long-term separation	1.444	0.653	42 (32)	4.890	.027

Note. GEE = generalized estimating equations; CSF OT = cerebrospinal fluid oxytocin; AVP = vasopressin. Only significant effects are reported. All models used exchangeable correlation structure, compared with autoregressive time-based models reported in Table 3.

theories about longitudinal correlations, may benefit from the use of GEE models, with its theory-based working correlation matrix. Even if these theories are not strong, robust standard errors may buffer misspecification. On the other hand, if theory is completely unavailable to inform time-based modeling choices, and a researcher is interested in both inter- and intra-individual change, LME may provide less error-prone—although less precise—estimates. Still, it is important to note that GEE and LME are both adequate approaches and helpful as dual techniques that provide multiple perspectives on small sample data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by NIH HD053555 to K.L.B., Office of Research Infrastructure Programs Grant P51OD011107 to CNPRC, and Good Nature Institute to K.L.B.

References

Bales, K. L., Mason, W. A., Catana, C., Cherry, S. R., & Mendoza, S. P. (2007). Neural correlates of pair-bonding in a monogamous primate. *Brain Research*, *1184*, 245-253.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). *Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models*. Paper presented at the proceedings of the Section on Survey Research Methods, Joint Statistical Meetings. Retrieved from https://www.amstat.org/sections/srms/Proceedings/y2008/Files/300933.pdf

Burton, P., Gurrin, L., & Sly, P. (1998). Tutorial in biostatistics. Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine*, *17*, 1261-1291.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376.

Cohen, P., Cohen, J., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation for the behavioral sciences* (3rd ed., pp. 445-447). Mahwah, NJ: Routledge.

Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford, England: Oxford University Press.

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, *29*, 421-437.

Gosho, M. (2014). Criteria to select a working correlation structure for the generalized estimating equations method in SAS. *Journal of Statistical Software, 57*(CS 1), 1-10.

Grace-Martin, K. (n.d.). *The unstructured covariance matrix: When it does and doesn't work*. Retrieved from http://www.theanalysisfactor.com/unstructured-covariance-matrix-when-it-does-and-doesn%E2%80%99t-work/

Graybill, F. A., & Wortham, A. W. (1956). A note on uniformly best unbiased estimators for variance components. *Journal of the American Statistical Association*, *51*, 266-268.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1-11.

Hennessy, M. B. (1997). Hypothalamic-pituitary-adrenal responses to brief social separation. *Neuroscience & Biobehavioral Reviews*, *21*(1), 11-29.

Howell, D. C. (2007). Repeated measures designs. In D. C. Howell (Ed.), *Statistical methods for psychology* (pp. 439-492). Belmont, CA: Thomson Wadsworth.

Hsieh, C. A., & Maier, K. S. (2009). A preliminary Bayesian analysis of incomplete longitudinal data from a small sample: Methodological advances in an international comparative study of educational inequality. *International Journal of Research & Method in Education*, *32*, 103-125.

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, *6*, 151-157.

Mason, W. A., & Mendoza, S. P. (1998). Generic aspects of primate attachments: parents, offspring and mates. *Psychoneuroendocrinology*, *2*(8), 765–778.

Mendoza, S. P., Capitanio, J. P., & Mason, W. A. (2000). Chronic social stress: Studies in non-human Primates. In G. P. Moberg & J. A. Mench (Eds.), *Biology of animal stress. Basic principles and implications for animal welfare* (pp. 227-247). New York, NY: CABI.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Rothwell, E. S., Mendoza, S. P., Mason, W. A., Ragen, B. J., & Bales, K. L. (2013). The role of dopamine D-1 receptors in pair-bond maintenance in monogamous titi monkeys (*Callicebus cupreus*). *American Journal of Primatology*, *75*, 57.

Rubin, L. H., Witkiewitz, K., Andre, J. S., & Reilly, S. (2007). Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. *Journal of Undergraduate Neuroscience Education*, *5*(2), A71.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323-355.

Tardif, S., Bales, K., Williams, L., Moeller, E., Abbott, D., Schultz-Darken, N., & . . . Ruiz, J. (2006). Preparing new world monkeys for laboratory research. *ILAR Journal*, *47*, 307-315.

Zorn, C. (2006). Comparing GEE and robust standard errors for conditionally dependent data. *Political Research Quarterly*, *59*, 329-341.