Assessment and Robust Analysis of Survey Errors

Sharon L. Christ

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Sociology.

Chapel Hill
2008

Approved by
Kenneth BollenPh.D., Paul BiemerPh.D.,
Guang GuoPh.D., Francois NielsenPh.D.,
and Stephen duToitPh.D.

ABSTRACT

Sharon L. Christ: Assessment and Robust Analysis of Survey Errors
(Under the direction of Kenneth Bollen)

Random measurement error and errors due to complex sampling designs may have deleterious effects on the quality of parameter estimates. This dissertation is comprised of three research papers that provide 1) an assessment of random measurement error through estimation of reliability using longitudinal, latent variable models, 2) an evaluation of the various probability weighting methods as corrections to unequal selection probabilities in multilevel models, and 3) an evaluation of several probability weighting and modeling approaches to unequal inclusion of observations in growth curve models.

A popular structural equation model used to estimate reliability for a single measure observed over time is the quasi-simplex model. The quasi-simplex model (QSM) requires assumptions about the constancy of variance components over time, which may not be valid for a given sample and population. These assumptions are tested using models that extend the QSM by using multiple indicator factors. The extended models include item specific error variance and additional factor variance estimates. Reliability estimates and their standard errors for the models with and without the QSM assumptions are compared in light of model fit and test results for several scales using survey data. Reliability estimates for a general model without the QSM assumptions are generally similar to the models with the assumptions indicating that the particular QSM assumption may not be that critical to the reliability estimates obtained from these models. However, variance components due to additional factor and item specific error have the potential of affecting reliability estimates markedly when they are estimated by the model.

Probability weights have traditionally been designed for single level analysis and not for use in multilevel models. A method for applying probability weights in multilevel models has been developed and has good performance with large sample sizes at each level of the model (Pfeffermann, Skinner, Holmes, Goldstein, and Rasbach, 1998). But, the multilevel weighting method in Pfeffermann, et al. can result in relatively poor estimation due to large amount of variation in the multilevel weights. This chapter includes a simulation analysis to evaluate several alternative methods for analyzing two-level models in the presence of unequal selection probabilities. The primary method of interest is to specify the level two part of the model such that it is robust to unequal selection bias in combination with weighting for unequal selection at level one. This "mixed" method does result in less bias, lower variance, and lower mean squared error for some models. A limitation is that the mixed method requires that the model is correctly specified at level two and the appropriate level one weight is used. This "mixed" method is a new approach in that it combines the Pfeffermann et al. (1998) weighting methodology at level one with the use of sample design variables at level two, rather than use the full Pfeffermann et al. approach of weighting at both levels.

Panel studies often suffer from attrition and intermittent nonresponse. Panel data is also commonly selected using complex sampling techniques that include unequal selection of observations. Unequal inclusion of individuals and of repeated measures will result in biased estimates when the missing mechanism is nonignorable, that is, when missing values are related to outcomes. Probability weighting may be used to correct estimates for nonignorable unequal inclusion due to selection and intermittent nonresponse. However the growth curve models frequently used in analysis of change have not traditionally been estimated using sampling weights. These models are usually estimated using a mixed model where the repeated measures are modeled as a function of both fixed and random parameters. Whereas sampling or probability weights have traditionally been applied to marginal models, which do not include random effects parameters. In this

chapter, several weighting approaches are applied to the mixed and marginal modeling frameworks using simulated and empirical data in linear growth models with continuos outcomes. Probability weighting performs the best in a marginal model when missing data are nonignorable. However, in most real situations including the empirical example provided in this chapter, probability weights may need to be combined with estimation that also utilizes variance weighting such as the GLS estimator with a correctly specified repeated measures correlation matrix as the variance weight matrix. This estimation methods can improve efficiency and decrease bias in estimates when data are missing at random (MAR).

ACKNOWLEDGEMENTS

I thank my dissertation chair, Ken Bollen, for his expertise and guidance. As my academic advisor, Ken has provided me with consistent encouragement and great opportunities to study quantitative methods. Without Ken's support, this dissertation would not have been possible. I wish to acknowledge all of my committee members for providing me with invaluable feedback on this research. I especially thank Paul Biemer who taught me much about these research topics.

I also want to express gratitude to my husband, Robert Langager, for his many hours of single parenthood over the last year. He has been wonderful father to our sons, Russell and Willem.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Sampling and nonsampling error in survey data may result in inconsistent and inefficient estimators. Sampling error due to unequal selection of observations is common in survey data and may result from a complex sampling design and/or nonresponse. Parameter estimates will be inconsistent when unequal selection probabilities are related to variables for descriptive statistics or outcome variables for a model. Random measurement error is also widespread in survey data and can result from interviewer, coder, data entry and response bias. Random measurement error decreases power and will result in parameter inconsistency for correlation and regression coefficients when independent variables contain error. The broad topics addressed by this dissertation research includes sampling and random measurement error. One chapter deals with random measurement error via the topic of estimation of reliability for scales in longitudinal data. Longitudinal, latent variable models are used to estimate reliability and test model assumptions that may affect the accuracy of reliability estimates. Estimation approaches that correct for sampling error due to unequal selection of observations is the topic of two chapters. Probability weighting methods are examined in the context of multilevel and longitudinal modeling as to their quality of estimation in the presence of nonignorable unequal selection. The overarching goal of this research was to make contributions to our knowledge about evaluating and reducing the survey errors that arise in complex samples.

Reliability of scale scores may be estimated using latent variable models, which estimate random measurement error separately from factor variance. The quasi-simplex

model (QSM) is a longitudinal, latent variable model that may be used to estimate reliability for a single indicator measured at three or more time points (Heise, 1969; Wiley and Wiley, 1970). The QSM for a single indicator measured over time requires certain model assumptions in order for the model to be identified. These assumptions are not required when the QSM is extended to include multiple indicators of a trait. The research in Chapter 2 uses models that are extensions of the QSM for the estimation of reliability of scale scores and to test the alternative assumptions of the QSM for those scores. The models use the latent variable structure of the QSM along with multiple indicator factors including the individual scale items as indicators and random split-halves of the scale items as indicators. The proposed models allow for the estimation of specific error variance and additional factor variance within time. The ability to partition the specific error variance can greatly affect reliability estimates because this variance is reliable, but would be erroneously applied to random error in the single indicator QSM. The ability to estimate additional factor variance is also important because this variance can then be treated as reliable. The partitioning of all reliable variance allows for the correct estimation of reliability.

The extension models are also used to test two relevant assumptions about the consistency of variance over time. The models are used to test the assumption of constant error variance and constant true score variance over time, where true score variance includes all reliable variance. One or more of these assumptions are required for identifying the original single indicator QSM. Testing the assumptions of the QSM is important because this model has been used in many applied areas as a method for obtaining reliability and stability estimates for scales and other single measures over time. Applying the QSM with the most appropriate constraints is preferable in these applied settings. The assumptions of the QSM have been evaluated in the methodological research using different over-identified models, but a consensus on the best model assumptions has not developed. The analysis in Chapter 2 also contributes to the existing methodological

literature by adding a new testing situation for the QSM, which includes several scale scores from the National Survey of Child and Adolescent Well-Being (NSCAW).

Chapters 3 and 4 both involve the assessment of estimation methods that are robust to sampling error resulting from unequal inclusion of observations. In Chapter 3, a simulation analysis is used to evaluate and compare several weighting methods for correcting unequal selection bias in two-level (mixed) models. The method of particular interest is to combine both probability weighting in estimation and the inclusion of sample design variables in the model. This method is in contrast to the method of using multilevel weighting (Pfeffermann, Skinner, Holmes, Goldstein, and Rasbach, 1998), which has been shown to be a good approach. The weighting and design variable approach is suggested as a potential superior method in certain situations because it can result in lower MSE.

Multistage, complex sample designs that result in nesting of observations are generally designed to limit the effects of unequal selection for a single level analysis. And multilevel modeling of data from such designs can result in severe effects of weights due to the larger degree of unequal selection at the separate sampling stages, which parallel the levels of a multilevel model. This selection results in extreme weight variation and larger MSE. The method that combines a model-based approach of including sample design variables in the model at level two where these variables are often limited in number and weight variation is large, along with weighting at level one is compared to the method of weighting at both levels. In addition, other methods of applying the traditional single-level weight in multilevel models are considered. These methods may be in use by analysts since the traditional, single level weights are often the only weights available to analysts and it is unknown how badly these weights perform in mixed models.

Alternatives to using multilevel weights, besides ignoring unequal selection probabilities all together, have not been evaluated extensively. This research contributes to our knowledge about practical alternatives to multilevel weighting for a common sample design. It is important to understand the conditions under which the alternative

methods perform well or poorly given that access to the weights or information about the sampling design may not be available. It also reveals the possible pitfalls of using a multilevel model with weighted estimation. Another important contribution is that this simulation study is based on a real sampling design and real finite population making the results more generalizable than a simulation using an arbitrary design.

Chapter 4 addresses the issue of sampling error due to unequal inclusion of observations in longitudinal models, specifically linear growth curve models. The focus is also specific to cohort or panel data where observations are followed over time. For this type of data, unequal selection of observations may occur at the first time point due to sample design, non-participation, and other sampling deficiencies while unequal inclusion of observations in the sample at follow up waves is solely due to intermittent nonresponse or attrition. This data structure commonly arises for panel data. The purpose of this chapter is to compare and contrast a number of weighting and estimation methods for dealing with nonignorable unequal inclusion of observations over time in linear trajectory models. Weighting methods are considered in both the mixed model and the marginal model because estimation of parameters in linear trajectory models with missing data at the level of time differs depending on whether a mixed model or a marginal model are used.

The weighting methods considered include panel weighting with complete data, which is the most common method in the survey sampling tradition; weighting for unequal inclusion into the study with no weighting for time-specific nonresponse; weighting for both unequal inclusion and time-specific nonresponse using multilevel weights; weighting for both unequal inclusion and time-specific nonresponse using time-varying weights; and no weighting at all. The weighting methods perform differently depending on whether the mixed or marginal model are used. The various weighting methods in combination with model types were evaluated and compared using an empirical example with the NSCAW data set and using simulated data.

This research contributes to our knowledge about how best to deal with missing data in longitudinal modeling. While unequal selection probabilities is basically a missing data problem, probability weighting has not been a major topic in the missing data literature. Chapter 4 evaluates weighting as a viable option for dealing with missing data in the longitudinal context. As well, the survey sampling literature does not cover the specific case of trajectory modeling very well where the primary method has been the less powerful panel weighting method. "Time-varying" weighting in the marginal modeling context is one alternative to the panel weighting method as is using a mixed model in conjunction with multilevel weights.

Random measurement error and sampling error due to unequal inclusion of observations are common problems in survey data. Analysis methods for evaluating measurement error are critical for an understanding of the quality of the measures that we use from survey data. Analysis methods that correct for bias due to unequal inclusion probabilities, which is tantamount to bias due to missing data, are necessary for accurate estimation. These very important issues are taken up in the chapters that follow.

CHAPTER 2

Scale Reliability Estimation and Testing using Longitudinal, Latent Variable Models

Abstract A popular structural equation model used to estimate reliability for a single measure observed over time is the quasi-simplex model. The quasi-simplex model (QSM) requires assumptions about the constancy of variance components over time, which may not be valid for a given sample and population. These assumptions are tested using models that extend the QSM by using multiple indicator factors. The extended models include item specific error variance and additional factor variance estimates. Reliability estimates and their standard errors for the models with and without the QSM assumptions are compared in light of model fit and test results for several scales using survey data. Reliability estimates for a general model without the QSM assumptions are generally similar to the models with the assumptions indicating that the particular QSM assumption may not be that critical to the reliability estimates obtained from these models. However, variance components due to additional factor and item specific error have the potential of affecting reliability estimates markedly when they are estimated by the model.

2.1 Introduction

The importance of reliability of measurement cannot be overstated. Low reliability may result in inaccurate parameter and variance estimates. While univariate means and covariances are unaffected by random measurement error, correlations and bivariate

regression parameters are attenuated in proportion to the amount of error variance. However, in multiple regression, when multiple independent variables contain measurement error, regression coefficients can be biased upward or downward (Bollen, 1989, pp 151-176). In addition to potential bias in parameter estimates, standard errors of estimates are inflated when measurement error is present. Therefore, the assessment of random measurement via reliability estimation is of great importance.

Structural equation models (SEMs) with latent variables allow for the estimation of reliability under a myriad of theoretical models. One SEM that provides reliability estimates for repeated measures is the quasi-simplex model (QSM). The QSM presented by Heise (1969) and Wiley & Wiley (1970) is a model with three factors each measured by a single indicator where the indicators are the same variable measured at three time points. The single-indicator factors are related over time in a Markov structure. The QSM provides estimates of reliability and stability of an observed variable measured at three or more time points. Unlike traditional test-retest reliability, the QSM does not require perfect stability in true score over time nor parallel measurement. However, the QSM is limited in that it requires assumptions about the equality of true score variance and/or the equality of error variance over time. It also does not allow for the estimation of specific error variance (Palmquist & Green, 1992; Wiley & Wiley 1974). Specific error variance represents variance due to an item and is separate from the common variance of the item and the random error of the item. Specific item error variance will increase reliability estimates since this type of error variance is replicated upon repeated assessment.

This article has several purposes. The first is to present models for estimating reliability and for testing the assumptions required for estimation of the single-indicator quasi-simplex model (QSM). The models combine an inter-item method of reliability estimation (Jöreskog, 1971) with the test-retest method of reliability estimation given by the QSM. In one version of the model the items are random split-halves (Biemer, Christ,

Wiesen, 2008). The hybrid models allow for the estimation of reliability with minimal assumptions and may include estimation of item-specific error variance and correlated measurement error within time/scale. The correlated measurement error is inter-item shared variance that is parameterized as additional factors representing additional traits within time. This variance is considered systematic and therefore reliable. Second, the hybrid models will be utilized to test the assumptions of constant true score variance over time, constant error variance over time, and both assumptions simultaneously (a form of constant reliability). One or the other of these assumptions is required for identification of the single-indicator QSM. In addition, the constant variance components are of substantive interest in their own right for many traits. Finally, reliability of composites and their standard errors are estimated from the hybrid models with and without the traditional QSM assumptions. The estimates are viewed in light of model assumptions and model fit.

The focus of this chapter is the reliability of scale scores, which are sums or averages of several individual items. Scale scores are commonly used by researchers as measures of constructs or broader concepts. Scale reliability is most often estimated using the inter-item correlation method of Cronbach's Alpha (Cronbach, 1951; Hogan, Benjamin, & Brezinsky, 2000). The hybrid models analyzed in this chapter use an inter-item correlation method of reliability estimation with fewer assumptions than Cronbach's method (Jöreskog, 1971). The data required for the hybrid models includes multiple measures of the same scale (construct) over two or more time points. Scale scores from the National Survey of Adolescent and Child Well-Being (NSCAW) are analyzed.

Testing the assumptions of the QSM is important because variants of this model have been widely applied as a method for obtaining reliability and stability estimates for scales and other single measures over time. The QSM models introduced by Heise and Wiley and Wiley have also activated an abundance of methodological work in the area of reliability, stability, and longitudinal analysis. According to the social sciences

citation index, the 1970 Wiley & Wiley paper is currently cited in 142 articles and the Heise (1969) paper is cited in 228 articles. Applying the QSM with the most appropriate constraints is preferable. The assumptions of the QSM have been evaluated in the methodological research using different over-identified models, but a consensus on the best model assumptions has not developed. This article contributes to the existing methodological literature by testing the QSM assumptions in general models that allow for specific item error variance and additional reliable variance theorized as additional traits or dimension. It also contributes by adding a new testing situation for the QSM to include a large and somewhat varied sample of child development scale scores, a large sample of observations (over 1000) for most scales, and another testing population.

The remainder of this article includes 1) background on the classical test theory definition of reliability and latent variable models for estimating reliability, 2) a description of the single-indicator QSM with a review of the methodological literature on the QSM, 3) a description of the data, models, and methods used for testing of the QSM assumptions, and 4) presentation of results and conclusions.

2.2 Background

We start with the classical test theory definition of reliability. Reliability is the ratio of true score variance to total observed variance (Anastasi, 1988; Lord & Novick, 1968; McDonald, 1999) where true score variance is the variance of any valid or invalid consistency in a measure. True score is therefore the expected value of a score over the population of scores (Lord & Novick 1968, p.173-176). Reliability may equivalently be defined as one minus the ratio of pure error variance to total observed variance.

Consider the observed score, y_i, which is comprised of two components: true score, τ_i, and unique error, ε_i.

$$y_i = \tau_i + \varepsilon_i \tag{2.1}$$

Where true score, τ_i, represents any systematic part of y_i including the trait or factor , t_i, and systematic error (or bias), s_i, such that $\tau_i = t_i + s_i$ where s_i is uncorrelated with t_i. Therefore, equation (2.1) may be written in more detail as follows

$$y_i = t_i + s_i + e_i \tag{2.2}$$

where e_i is pure random error. Random error, e_i, has expectation of zero and is uncorrelated with the s_i and the t_i, and, s_i is uncorrelated with t_i. Using equation (2.1), the reliability of y_i is defined

$$\rho_{yy} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} = \frac{\sigma_\tau^2}{\sigma_y^2} \tag{2.3}$$

where $\sigma_\tau^2 = var\,(t_i) + var\,(s_i)$. Equation (2.2) assumes that y_i measures a single trait. The equation can be extended to include multiple correlated traits, t_{ip}, where p indexes the factors. With multiple traits, true score variance becomes $\sigma_\tau^2 = \sum_p var\,(t_p) + 2\sum_{p<}\sum_{p'} cov\,(t_p, t_{p'}) + var\,(s_i)$. Equation (2.3) illustrates that reliability is the proportion of total variance in the observed scores that is not attributable to pure random error. This proportion ranges from zero, completely unreliable, to one, perfectly reliable. Consider multiple y_i, which are part of a composite score. In this case there may be one or more traits and multiple specific error for each observed item comprising the composite.

Depending on what is intended to be measured, the t_{ip} may be considered valid or invalid. If a single trait is intended to be measured, then all but one latent variable would be considered invalid. However, for many scale scores a broader, multidimensional concept may be the intended measure where the scale score is a composite of multiple traits. In this case the validity of each factor should be considered separately. Item-specific, systematic error is generally considered bias as it represents systematic deviation from all traits due to the specific item. Specific error may be considered a method effect because each indicator of the scale is considered a different method of measurement for the scale score (Saris & Aalberts, 2003; Saris & Andrews, 1991). It is also referred to as indicator specificity (Raykov & Tisak, 2004) where the unique error, ε_i, can be

decomposed into error specific to an indicator and "pure" measurement error.

Note that two different populations with equal random error variance for the same scale may have different reliability due to different degrees of true score variation in a population. The population that is more homogenous on true score would have a lower score reliability even though the random error in the measurement process is the same (Lord and Novick, 1968, 199; Wiley & Wiley, 1970, 113). In this sense, reliability is not just a condition of an instrument, but also of a population.

2.3 Latent Variable Models for Estimating Reliability

Reliability estimation methods include split-halves, test-retest, and inter-item (Brown, 1910; Cronbach, 1951; Hoyt, 1941; Jackson, 1939; Spearman, 1910). Split-halves estimates are based on the correlation of two halves of a scale. Test-retest estimates are based on the correlation between two or more measurements of the same scale at different time points. And inter-item (or internal consistency) estimates are based on the shared correlation of the items that make up a scale. All of these methods of estimating reliability are based on the classical test theory operational definition of reliability. In each case, the consistency of $k \geq 2$ measures is obtained.

Structural equation modeling (SEM) provides many methods for assessing reliability, including inter-item, test-retest, and split-halves methods because SEM allows for the separation of random measurement error variance from true score variance (Bollen, 1989; Saris & Andrews, 1991; Cudeck, du Toit, and Sörbom, 2001). Jöreskog (1971) outlines structural equation models for reliability estimation of sets of congeneric items. Congeneric items measure the same construct or trait and allow different measurement error variances and different linear relationships with the trait across items. Congeneric items are less restrictive than parallel items, which require equal linear relationship with the trait and equal error variance across items. Congeneric items are also less restrictive than tau equivalent items, which require equal linear relationship with the trait across items. In many traditional methods for estimating reliability, the assumption of parallel

measures (e.g., test-retest and split-halves) or tau-equivalent measures (e.g., Cronbach's alpha) is made.

Jöreskog's model for analysis of congeneric items is a factor model where multiple measures of the same trait are loaded on a single factor. The system of equations for this model with k measures is

$$y_{ki} = \lambda_k t_i + \varepsilon_{ki}$$

The y_{ki} are the observed items of the factor scores, t_i, with random error component ε_{ki}. The λ_k are the coefficients representing the relation between true (factor) scores and observed scores. There may be item specific error, s_i, in equation 2.3. However with one factor measured at one time point, it is not possible to estimate s_i. In addition to a unidimensional trait, the model assumes the following:

$$\begin{aligned} E(\varepsilon_{ki}) &= 0 \\ cov(\varepsilon_{ki}, \varepsilon_{k'i}) &= 0; k \neq k' \\ cov(\varepsilon_{ki}, t_i) &= 0 \end{aligned}$$

The second assumption (uncorrelated item errors) may be relaxed, though the definition of reliability will change as a result. This model provides estimates of true score variance, $var(t_i)$, as well as error variance for each indicator, $var(\varepsilon_{ki})$. Reliability of the composite of the k measures is given by

$$\rho_{YY} = \frac{\left(\sum_{k=1}^{K} \lambda_k\right)^2 \sigma_{t_i}^2}{\left(\sum_{k=1}^{K} \lambda_k\right)^2 \sigma_{t_i}^2 + \sum_{k=1}^{K} \left(\sigma_{\varepsilon_{ki}}^2\right)} \tag{2.4}$$

where Y is the composite score.

Cronbach's alpha may be estimated in a factor model by placing further restrictions on Jöreskog's model (Werts & Linn, 1970). These restrictions include equal factor loadings, $\lambda_k = \lambda_{k'}$ with factor variance set to one[1] and equality of error variances across

[1] Alternatively all factor loadings may be set to one and factor variance freely estimated.

items $k \neq k'$: $var(\varepsilon_k) = var(\varepsilon_{k'}) = var(\varepsilon)$ so that $y = \lambda t + \varepsilon$ for all items. These additional assumptions make up the requirement of parallel measurement for all of the items in the scale. With these same constraints, traditional split-half reliability estimation may be obtained using this model with two half scores as factor items. A less restrictive reliability estimate using split halves could be obtained using the congeneric model. In general, using SEM or factor models allows for the assumption of parallel or tau-equivalent measures to be tested and, if necessary, relaxed.

2.3.1 The Quasi-Simplex Model for Estimating Test-Retest Reliability

The quasi-simplex model (QSM) (Heise 1969, Wiley & Wiley, 1970) is a SEM that provides reliability estimates for a single measure that is observed at three or more time points. The QSM model measures test-retest reliability with less restrictions than traditional test-retest reliability. Traditional test-retest reliability is estimated by the correlation of the measures of the same trait observed at two (or more) points in time, but assumes parallel measures and no change in true score between measurement (Bollen, 1989, 209-210). In contrast, the QSM allows for change in true score.

The QSM is composed of a set of measurement equations (factors) and latent variable equations. The measurement equations relate the unobserved true scores to the observed scores.

$$Y_t = \lambda_t t_t + \varepsilon_t \tag{2.5}$$

where Y_t are the observed scale scores at time $t = 1, 2, 3$, t_t are the unobserved true scores at $t = 1, 2, 3$, and the ε_t are the random measurement errors at $t = 1, 2, 3$. (The subject subscript, i, is omitted for clarity.) The λ_t are the coefficients representing the relation between true score and Y_t.

The latent variable equations give the relation among true scores and make up a

non-stationary ARMA(1,1) (duToit, 1979), model.

$$t_1 = t_1 \tag{2.6}$$

$$t_2 = \beta_{21} t_1 + \zeta_2 \tag{2.7}$$

$$t_3 = \beta_{32} t_2 + \zeta_3 \tag{2.8}$$

where β_{21} is the effect of true score at time 1 on true score at time 2 and β_{32} is the effect of true score at time 2 on true score at time 3. The β are the parameters that measure stability/change in true score over time. And ζ_2 and ζ_3 are random errors for the autoregressive equations (2.7) and (2.8) and $var(\zeta_t)$ is a component of true score variance at time t.

Assumptions of the QSM include

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ cov(\varepsilon_t, \varepsilon_{t'}) &= 0; t \neq t' \\ cov(\varepsilon_t, t_t) &= 0 \\ cov(\zeta_{t+1}, t_t) &= 0 \\ cov(\zeta_t, \zeta_{t'}) &= 0; t \neq t' \end{aligned}$$

Additional constraints are required to identify this model. The original Heise QSM model included the assumption of constant reliability over time in that the ratio of true score variance to total variance is constant. Heise points out that the test-retest correlation in this model is not simply the squared λ paths unless the $\beta_{21} = \beta_{32} = 1$, which implies that the true score is completely stable over time. The Heise model allows change in the mean true score estimated by the stability coefficients, β. Consequently, the model provides a reliability coefficient that is "uncontaminated by the temporal instability of a [latent] variable" (Heise, 1969, 96). This is useful particularly for variables that are observed at longer intervals where more change in average true score is expected. For instance, the intervals in National Survey of Child and Adolescent Well-Being (NSCAW)

are 18 months and many of the traits are developmental scales observed on children, so that we expect at least some change in the true scores over time.

Wiley & Wiley (1970) [henceforth W & W] did not feel that the assumption of constant true score variance over time was as plausible as the alternative assumption of equal error variance over time.

"Error variance is best conceived as a property of the measuring instrument itself and not of the population to which it is administered. On the other hand, the true score variance is more realistically considered as a property of the population. Thus the specification of stable reliability will normally require assumptions about populations as well as assumptions about the measuring instrument (112)."

They argue that the assumption of constant reliability in the Heise model also implies constant true score variance over time because he used standardized scores. Heise subsequently states that the W & W formulation is to be preferred because of the weaker set of assumptions. "With standardized scores an assumption of parameter equality has to be elaborated by adding the additional assumption that the true-score variances of some of the different measurements are equal" (Heise, 1970).

W & W do view both assumptions as invalid, but make a case that if stability of error variance over time is faulty, then stable reliability will most likely also be in error. The W & W QSM model therefore has different assumptions for identification. This model assumes unit λ_t's, $\lambda_t = \lambda_{t'} = 1$, $t \neq t'$ and equal error variances over time, $var(\varepsilon_t) = var(\varepsilon_{t'}) = var(\varepsilon)$, $t \neq t'$, but allows for true score variance and hence reliability to change over time. Therefore the W&W QSM estimates the parameters β_{21}, β_{32}, $var(\varepsilon)$, $var(t_1)$, $var(t_2)$, $var(t_3)$, $var(\zeta_2)$, and $var(\zeta_3)$ for a model with three time points. The following three, unique reliability estimates for times 1, 2, and 3 are

obtained.

$$\rho_1^2 = \frac{var\,(t_1)}{var\,(t_1) + var\,(\varepsilon)} \tag{2.9}$$

$$\rho_2^2 = \frac{var\,(t_2)}{var\,(t_2) + var\,(\varepsilon)} = \frac{\beta_{21}^2 var\,(t_1) + var\,(\zeta_2)}{\beta_{21}^2 var\,(t_1) + var\,(\zeta_2) + var\,(\varepsilon)} \tag{2.10}$$

$$\rho_3^2 = \frac{var\,(t_3)}{var\,(t_3) + var\,(\varepsilon)} = \frac{\beta_{32}^2\left[\beta_{21}^2 var\,(t_1) + var\,(\zeta_2)\right] + var\,(\zeta_3)}{\beta_{32}^2\left[\beta_{21}^2 var\,(t_1) + var\,(\zeta_2)\right] + var\,(\zeta_3) + var\,(\varepsilon)} \tag{2.11}$$

The estimate of reliability in the Heise model is equal to equation (2.10), or the reliability at time 2.

The QSM is established in published research. Anderson (1959) developed the observed variable simplex model and addressed model identification. Jöreskog (1970, 1979) has developed the estimation and testing aspects of the model as well as identification.

The QSM Assumptions and Reliability Over Time

The W & W article was the beginning of a debate about appropriate assumptions for the QSM. Soon after the Heise and W & W publications, Blalock (1970a, 1970b) discusses the models focusing on the fact that the just identified QSM model does not allow for testing these assumptions. He suggests using multi-indicator, multi-wave models that allow for the testing of homogeneity in variances over time. For the single-indicator models, Blalock (1970a), points out that regression to the mean effects may be due to a homogenizing effect of the population or to measurement. He suggests that if total variance is constant over time, one may attribute change to a measurement effect. Otherwise, if total variance decreases over time, change may be considered due to population homogenization over time. Blalock points to the W & W QSM assumptions as equally arbitrary as the Heise model assumptions indicating that if the variance of either true score or measurement error changes, then parameters linking the true scores to indicators, $\lambda's$, should also be allowed to change.

Various theories about the best assumption exist. First, is the test-retest effect (Campbell & Cook, 1979) or Socratic effect (McGuire, 1960) in attitude theory, which

contends that responses on a scale will become more consistent or homogenous over time. The theory posits that respondent error decreases over time due to familiarity of the test and therefore reliability increases. This theory was tested by Jagodzinski, Kühnel, and Schmidt (1987) using multi-indicator QSMs like the ones used in this chapter. They define consistency in several ways, but attribute the Socratic effect to a decrease in error variance over time. They find decreasing error variance for an attitude from a short-wave (four week) panel study. Jagodzinski & Kühnel (1987) also look at the single indicator Heise and W & W model for short-wave panel study of political attitudes. Not all relevant parameters are separately identified in these models. They show the conditions for perfect reliability and perfect stability and show that both cannot be simultaneously met. Empirical results include low reliability estimates and stability parameters that are low between the first two waves and over 1 for the second two waves. They expected increases in reliability over time due to the test-retest effects, but find the opposite. They surmise that the single-indicator, just identified QSMs are sensitive to sampling fluctuations and/or model misspecifications (J & K, 1987). Coenders, Saris, Batista-Foguet, and Andreenkova (1999) also found that the standard error estimates were very unstable for the single-indicator QSM; however, their results do reveal that the QSM has better MSE than the test-retest parallel measures model when the QSM is the correct model.

Another theory posits increased differentiation over time (Howard, 1964; Howard & Diesenhaus, 1965) where true score variance (inter-individual differences) increases upon repeated measurement. This theory is based on the idea that there is more stress at first measurement and therefore more safe or stereotyped response. Ferrando (2003) tests this theory against the test-retest theory mentioned above using multiple indicator two-wave panel models allowing for correlated errors for the same item across wave (item-specific error variance). Results include change in true score over time with increased item reliability in one out of three scales and increases in reliability for items is due to

increased true score variance, the differentiation effect. He speculates that results are likely different depending on the specific trait being tested.

Werts, Jöreskog, and Linn (1971) use a four panel, single-indicator QSM to test the assumptions of constant error variance, true score variance, and reliability for a quantitative and a verbal test score. They test these for the inner two waves only, since for the first and last time points the three coefficients are not separately identifiable. They find that the equal error variance and equal reliability models fit the data well and the equal true score variance model does not. However, the component fit for the equal reliability were considered theoretically unreasonable while the component fit in the equal error variance were viewed reasonable. Alanen, Leskinen, and Kuusinen (1998) tested equality of reliability and stability in a three-wave, multiple-indicator model for psycholinguistic abilities and found that models with constant reliability and stability were acceptable. They also tested these assumptions in the context of a seven-wave, single-indicator model for a reading test and again found constant stability and reliability.

To summarize, results vary on which assumptions are more appropriate in the QSM. This may be in part because the best assumptions differ depending on the population and trait of interest as well the length of time between assessments. It also makes a difference depending on whether the single indicator or a multiple indicator model is used. There is more instability in the single indicator models. The limitation of the single-indicator QSM with only three time points is that these assumptions cannot be tested.

2.3.2 Specific Item Error Variance

The QSM is a just identified model and therefore does not permit estimation of consistent item error variance. Wiley & Wiley (1974) present a version of the QSM that measured specific item error variance. This model includes a Markov structure on both the error terms and the true score. This model is intended to measure repeatability in the errors over time and requires an additional constraint. The constraint proposed is that the autoregressive parameters linking true scores over time are equal: $\beta_{21} = \beta_{32}$. The

autoregressive parameters linking errors over time are also held equal. While this model allows for the estimation of a systematic error component, the model has rarely been used and typically results in anomalous estimates and hugely inflated standard errors due to empirical underidentification (Palmquist & Green, 1992). Underidentification of this kind will occur if the change process for the errors is not different from the change process for the true scores and/or if true score variance at time one and time two are not different (Wiley & Wiley, 1974). Palmquist & Green (1992) show that the model requires non-constant observed variances and since the error variance is held equal in the model, the true score variances must be the source of increase or decrease. Therefore, the assumption of constant true score variance can never hold in this model. Palmquist and Green (1992) show that obscure results due to large sampling variability decreases with more than three repeated measurements. Incidentally, the QSM cannot have both perfect stability ($\beta_{21} = \beta_{32} = 1$) in addition to constant reliability (Jagodzinski & Kühnel, 1987).

Systematic error in the form of item-specific error variance may be estimated in multiple-indicator Markov models such as the hybrid models used in this chapter. This may be measured by the covariance of measurement error for the same indicator over time or using method factors that load on the same indicator at multiple time points (Raffalovich & Bohrnstedt, 1987; Raykov and Tisak, 2004; Saris & Andrews, 1991). A few of the articles reviewed in Section 2.1.2 include specific error when testing the assumptions of the QSM (e.g., Ferrando, 2003). These models may also have empirical underidentification issues as will be discussed in subsequent sections of this chapter.

Testing of the assumptions of constant error variance, constant true score variance, and constant reliability has been done in several published articles. The testing has been done in both single indicator QSM models with more than 3 time points as well as in multiple indicator models. The length of time between waves varies across the studies as does the types of constructs being measured. Therefore, it is difficult to arrive at any broad conclusions about the viability of each of the assumptions. However, in most of the

articles reviewed here reliability tends to increase over time and constant error variance seems to be a more frequent occurrence than constant true score variance. It is unknown how variance component assumptions hold up in the context of a model that includes the measurement of item specific error variance and additional construct variance. It is also not clear which assumptions might operate best for a population that is changing a lot on the constructs, such as one would expect of developmental measure on children over the course of 36 months.

2.4 Methods

2.4.1 Data

Data used for the analysis come from the National Survey of Child and Adolescent Well-Being (NSCAW), which is a panel survey of children in the child welfare system in the United States. The target population of the NSCAW Child Protection Services (CPS) sample is "all children in the U.S. who are subjects of child abuse or neglect investigations (or assessments) conducted by CPS and who live in states not requiring agency first contact." (Dowd, et al., 2006). The NSCAW sample design is a complex design that includes stratification, clustering, and unequal selection probabilities. The NSCAW Child Protection Services (CPS) cohort includes 5,501 children, ages birth to 14 (at the time of sampling), who had contact with the child welfare system within a fifteen-month period which began in October, 1999. Face-to-face interviews were administered at three points in time: Wave 1, 18 months post-Wave 1, and 36 months post-Wave 1.

Each of the three waves of data includes many types of scales for measuring the health and development of children. These scales are mostly sum scores and are the variables of primary importance for most NSCAW researchers. The NSCAW scales analyzed in this study are generally child developments scales, but others measure characteristics of caregivers or the home environment. Some of the scales are psychometrically designed and some are ad hoc. Sample sizes for each scale range from around 1,000 to over 5,000

cases. A description of the scales used in this analysis is provided in the appendix.

It is expected that there will be change in the true scores over time since the intervals are 18 months and the scales are mostly measuring development in children. If this is true for the population, the stability coefficients will be estimated as less than or greater than 1. The length of time between assessments should also reduce memory effects thereby resulting in less item specific error variance. Theories about change in the variance components are mixed for this population. It is expected that the children and their caregivers would be subject to test-retest effects, particularly for children who are aging and perhaps becoming more adept at answering survey questions. This would imply decreases in random error variance over time. True score variance may also decrease over time since for this population being in contact with social services is a natural intervention that would remove outliers. For example, children who substantiated abuse may be removed from their home or the perpetrator of such abuse subject to incarceration. Children with the most severe developmental issues may regress to the mean after intervention with social services, a homogenization effect.

2.4.2 Hybrid Models

A model that combines the W & W QSM test-retest model with the congeneric Jöreskog model is used to estimate reliabilities and test model assumptions. The measurement model is made up multiple indicators, which are the items for the NSCAW sum scores. In another version of the combined model, two equivalent halves of the scale were created by randomly selecting half of the scale items and summing the items into two scores. The half scores are subsequently used as indicators of the measurement model. In its least restrictive form the model allows for correlated item error over time for the same indicator and within time for different indicators of the same factor. Correlations of the same items over time may be parameterized as factors s_k representing item specific variance due to unique aspects of the items, for example, similarities due to memory effects. The correlations of different items within time may be theorized as additional

dimensions or method effects c_t. This least restrictive model for a scale with two items (or half scores) is pictured in Figure 2.1.

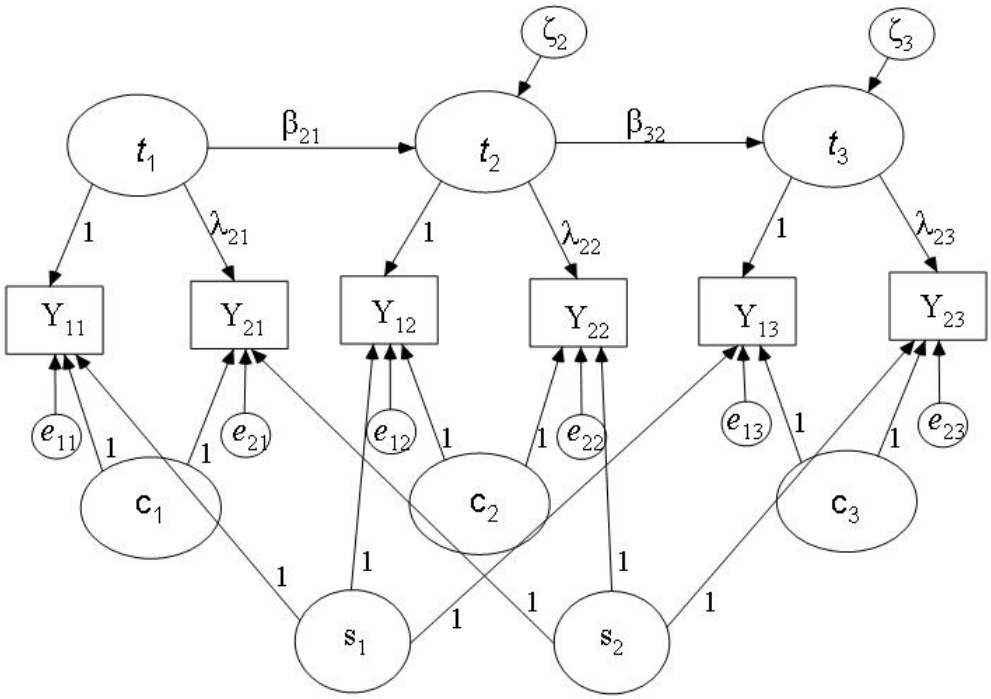

Figure 2.1: Free Model

The latent variable equations for the free model are identical to equations (2.6), (2.7), and (2.8) for the QSM. The measurement model equations for the free model with t time points and k items per scale are

$$y_{tk} = \lambda_{tk} t_t + s_k + c_t + e_{tk} \tag{2.12}$$

Where the t_p from the general model in equation (2.2) include both the t_t and the

c_t. In this model, true score is made up of all factors except for the random error where $\tau = \lambda_{tk} t_t + c_t + s_k$. The number of items per scale may vary across scales in the item-level models and will be two for the models using split-half indicators. One of the λ_{tk} will be constrained to 1 for identification and the others are allowed to be freely estimated. For models with $k > 2$, there are $q = k\,(k-1) \div 2$ factors per time point representing within time correlated items. This means that there is a factor, c, representing each possible correlation between items within time. Therefore, each item is involved in $k-1$ of the additional factors per time point. If one were to parameterize the model using covariances between items within and over time rather than using factors as pictured in Figure 2.1, the variance for both the s_k and c_t would be included in the error variance estimates. That is, when $s_k = cov\,(\varepsilon_{tk}, \varepsilon_{t'k})$ and $t_{t2} = cov\,(\varepsilon_{tk}, \varepsilon_{tk'})$ then $var\,(\varepsilon_{tk}) = var\,(s_k) + var\,(c_t) + var\,(e_{tk})$. But in the factor parameterization used here, the $var\,(s_k)$ and the $var\,(c_t)$ are separately estimated and the estimate of the error variance is an estimate of pure random error variance where $var\,(\varepsilon_{tk}) = var\,(e_{tk})$.

Assumptions of this model include the following

$$
\begin{aligned}
E\,(e_{tk}) &= 0 \\
cov\,(e_{tk}, t_t) &= 0 \\
cov\,(\zeta_{t+1}, t_t) &= 0 \\
cov\,(\zeta_t, \zeta_{t'}) &= 0 \\
cov\,(s_k, s_{k'}) &= cov\,(c_t, c_{t'}) = cov\,(s_k, c_t) = 0 \\
cov\,(t_t, c_t) &= cov\,(t_t, s_k) = 0 \\
var\,(c_t) &= var\,(c_{t'})
\end{aligned}
$$

Basically, all factors are orthogonal except for the three primary factors in the autoregressive system. The last assumption implies that item error associations have constant variance over time. This is a necessary assumption for identification of the model and is tantamount to assuming that the source of additional correlation, e.g., item ordering,

between items operates the same at each time point (Blalock, 1970b).

The model pictured in Figure 2.1 and described here is similar to the multiple indicator model presented in Blalock (1970b). The differences are that Blalock's model assumes equal λ over time for each item, i.e., $\lambda_{tk} = \lambda_{t'k}$, is parameterized using covariances between items rather than factors, and allows the correlation between items for adjacent time points to differ from correlation between items with lag time of 2. Also, the model outlined by Blalock was not generalized to more than two items per scale. The model used in this chapter is also very similar to some of the models used in other papers to test the QSM assumptions. For example, Ferrando (2003) and Jagodzinski, Kühnel, & Schmidt (1987) use models that are similar in that they are longitudinal, multiple indicator factor models with estimates of specific error variance. However, these models do not allow for additional common variance among items within time, i.e., they assume a single dimensional trait with no additional traits or measurement factors.

By allowing the c factors, the assumption that the scales are measuring a single, unidimensional trait is relaxed. Estimating reliability using a single factor model when more than one underlying factor actually exists will distort reliability estimates. Other sources of correlated items vary and may be due to systematic factors such as additional traits, higher order traits, or experimental/method conditions that are common to certain items in the composite (Heise & Bohrnstedt, 1970; Gerbing & Anderson, 1984). The model presented here is quite general in that the c factors allow for all possible item consistent error variation rather than, say, a single c factor that estimates common variance among all the items within time. Also, the models used in this chapter do not attempt to correctly identify all trait and method factors and in this way is not a good model for estimating stability and other specific model parameters unless the single factor in the Markov structure is theoretically meaningful. However, the model is good for estimating variance components that may be applied to one's operational definition of reliability. In this chapter the definition of reliability is broadly defined and includes all

shared variance, both within time and across time, among the items in the composite.

For the NSCAW data, many of the scales assessed are measuring multiple traits (see the appendix). For example, the YSR scale is intended to measure "youth behavior problems", which includes deviance, depression, anxiety, among other traits. Items from these subdomains could be defined as separate trait factors in the same model and the reliability estimates from the separate traits as well as the composite could be obtained (Bentler, 2005; Raykov & Shrout, 2002). However, many analysts are using the composite and may not be interested in the reliability of the separate domains. In the subsequent analysis of NSCAW scales, the c factors (or within time correlated error) are taken to represent repeatable variance from additional traits and possibly methods common among certain items. Alternatively, one could relegate additional shared variance to non-repeatable, pure random error variance depending on the assumptions about the correlated error (e.g., Biemer, Christ, & Wiesen, 2008).

The following equations estimate true score variance and error variance from the free model for each of the three time points for k number of indicators. True score variance is the sum of all reliable variance, therefore, true score variance at time 1 for the sum score is:

$$\left(\sum_{k=1}^{K}\lambda_{1k}\right)^2 (var\,(t_1)) + \sum_{k=1}^{K} var\,(s_k) + \sum_{q=1}^{k(k-1)/2} var\,(c_q) \tag{2.13}$$

True score variance for the sum score at time 2 is:

$$\beta_{21}^2\left(\sum_{k=1}^{K}\lambda_{1k}\right)^2 (var\,(t_1)) + \left(\sum_{k=1}^{K}\lambda_{2k}\right)^2 (var\,(\zeta_2)) + \sum_{k=1}^{K} var\,(s_k) + \sum_{q=1}^{k(k-1)/2} var\,(c_q) \tag{2.14}$$

True score variance for the sum score at time 3 is:

$$\beta_{32}^2\beta_{21}^2\left(\sum_{k=1}^{K}\lambda_{1k}\right)^2 (var\,(t_1)) + \beta_{32}^2\left(\sum_{k=1}^{K}\lambda_{2k}\right)^2 (var\,(\zeta_2)) + \left(\sum_{k=1}^{K}\lambda_{3k}\right)^2 (var\,(\zeta_3)) + \sum_{k=1}^{K} var\,(s_k) + \sum_{q=1}^{k(k-1)/2} var\,(c_q) \tag{2.15}$$

And, error variance for the sum score at time t is:

$$\sum_{k=1}^{K} \left(var \left(e_{tk} \right) \right) \tag{2.16}$$

Reliability for each time point is estimated as the ratio of the true scale score variance over total (true score plus pure error) variance for that time point. Since the model does not include covariances of factors, those are not included in the true score variance.

Identification

Blalock (1970b) shows over-identification for a two-indicator simplex model with correlated errors within and across time similar to the model used here. The model used here differs from the Blalock model because it does not constrain loading equal over time, that is, $\lambda_{21} = \lambda_{22} = \lambda_{23}$ but in the Blalock model error correlations specific to an item over time are allowed to differ for different time lags whereas the above model assumes equal loadings across the items in the s_k factors which is tantamount to constraining the error covariances the same regardless of the lag time. The model in Figure 2.1 is therefore nested within the Blalock model when λ are held equal over time. Most of the models ultimately used for reliability estimates in this chapter include the equal lamda over time. Nevertheless, the additional lambda parameters can be shown to be overidentified because they are estimated by more than one combination of observed covariances. The degrees of freedom for the free model above can be calculated using the following equation:

$$df = \frac{3k \left(3k + 1 \right)}{2} - \left[\frac{k \left(k - 1 \right)}{2} + 7k + 2 \right] \tag{2.17}$$

where the first term represents the number of nonredundant observed covaraiances for k items and three time points and the second term represents the number of parameters being estimated.

Like the single-indicator QSM, empirical underidentification is also a problem for the two-indicator hybrid models. For example, Blalock shows that the estimation of many of the parameters depends on the assumption that the change between waves 1 and 2 and

the change between waves 2 and 3 are different in the population (Blalock, 1970b, p.109). When the two stability parameters are equal in the population, many of the parameters are unstable since the difference between the two stability parameters is involved in the denominator of equations used to identify several parameters. Even if these stability parameters are different in the sample, if the population values are similar then standard error estimates for the estimates involving the ratio of the two stability coefficients may be very volatile. For the models used in this chapter, it can also be shown that when the stability parameters approach zero or one, $var\,(t_1)$ is underidentified. This is due to the fact that the variance at time 1 includes $\beta_{21}\,(\beta_{32}-1)$ and $\beta_{21}\beta_{32}\,(\beta_{21}-1)$ in the denominator. The estimation of time 1 variance parameter is critical to the model since it is involved in the identification of all other parameters.

The hybrid models are very general and should be simplified when there is no specific item error variance or no additional factor variance within time. Therefore the specific error variance factors s_k and the additional c_q factors may be removed from the model for certain scale scores. The models may also be simplified by constraining item loadings over time as in the Blalock (1970b) model. It may also be desirable to limit the number of additional factors particularly for scales with many items where the model may become very large.

2.4.3 Testing

Within the context of the larger, "free" hybrid model specified above, a series of tests were performed for each scale score to determine the best "general model" for testing the QSM assumptions of constant error variance and constant true score variance over time. The general model testing includes: (1) specific error variance (any s_k), (2) additional factors within time (any c_q), and (3) equal loadings for the same item over time ($\lambda_{tk} = \lambda_{t'k}$). Multiple degree of freedom testing using the Wald chi-square (Agresti, 2002) is undertaken for the following hypotheses:

1. Specific error variance: H_o: all $var\,(s_k) = 0$

2. Additional factors: H_o: all $var\,(c_q) = 0$

3. Equal loadings over time: H_o: all $\lambda_{1k} = \lambda_{2k} = \lambda_{3k}$

Testing is in the above order where constraints are retained for QSM assumption testing. The factor structure is tested first in steps 1 and 2 prior to testing for equal loading. Following the general model testing and specification of a general model for each scale, testing of the various QSM assumptions was done where constant error variance over time, constant true score variance over time, and both constant error and true score variance (constant reliability) over time are tested separately. These hypotheses are tested in the context of the general model chosen for each scale, which is the free model with additional constraints as determined by the general model testing. The specific null hypotheses are:

4. Constant error variance over time: H_o: $var\,(e_t) = var\,(e_{t'}) = var\,(e)$, $t \neq t'$

5. Constant true score variance over time: H_o: $var\,(t_t) = var\,(t_{t'}) = var\,(t)$, $t \neq t'$. where $var\,(t_1)$, $var\,(t_2)$, and $var\,(t_3)$ are given in equations (2.13), (2.14), and (2.15), respectively.

6. Constant reliability over time: H_o: $var\,(e_t) = var\,(e_{t'}) = var\,(e)$ and H_o: $var\,(t_t) = var\,(t_{t'}) = var\,(t)$, $t \neq t'$

Notice that testing of constant true score variance and constant error variance is unaffected by how the s_k and c_q variances are treated since both are constant over time in the model. The s_k variance is generally always treated as consistent (part of true score variance) and in this case the c_q variance is treated as consistent. Because it is required for identification of the general model that the variance of additional factors is constant over time, that is, $var\,(c_{tq}) = var\,(c_{t'q})$, there is thereby a partial assumption of equal

true score variance over time since the $var\,(c_{tq})$ is a component of $var\,(\tau)$. The part of true score variance due to the factor in the autoregressive equations may change over time.

2.4.4 Reliability Estimation and Model Assessment

The NSCAW was sampled with a complex design including unequal selection probabilities. However, the reliability estimates of the NSCAW scales are unweighted because we are assuming population homogeneity of reliability for each scale. This assumption means that the expected value of the weighted reliability estimates, ρ_w^2, will equal the expected value of the unweighted reliability estimates, ρ^2. A sandwich estimator was used to estimate standard errors that are robust to clustering. And standard errors will also account for stratification in the NSCAW where estimate variance is taken within stratum and subsequently summed over the strata. Missing data were handled with direct maximum likelihood methods (Arbuckle, 1996). Mplus 5.0 (Muthén & Muthén, 1998-2008) software was used to fit all models, perform nested testing, and estimate reliabilities and their standard errors.

The mean and variance adjusted chi-square statistic (Satorra & Bentler, 1988), CFI, TLI, and the RMSEA (Browne & Cudeck, 1993; Steiger & Lind, 1980) are used to assess overall model fit. The mean and variance adjusted chi-square statistic is used because the nonindependence of observations in the data results in a statistic that is not exactly distributed χ^2 and the adjustments create an approximate χ^2 variate.

2.5 NSCAW Application

There are a variety of scale scores in the NSCAW, many of which have 30 or more items including some scales with around 100 items. For several of these scales, the item-level models with all possible c_q factors have more parameters than may be fit given the sample sizes of the data, particularly because of the nested structure of the data. For these scales, the split-halves (or split-thirds, etc.) model or a model with fewer additional

c_q factors are more appropriate in terms of model size. There are also several scales with fewer than 5 items. These scales are easily estimated using the item-level model and there is no need to perform the random split-halves step. For example, splitting a scale with 3 items into two halves with 1 and 2 items, respectively, is an unnecessary additional step. Therefore, some of the NSCAW scales are evaluated using the split-halves model and some are evaluated using the item-level models based on the number of items comprising the scale. A few of the scales are evaluated using both types of hybrid models for children in the same age group and their results may be compared across the model types. These scales include the CDI, School Engagement, and YSR scales.

2.5.1 General Model Selection Results

The "general" model used for the nested tests of the QSM assumptions is selected for each scale based on a series of the three general model tests, which include tests for specific error variance, additional factor variance, and equal item loadings for the same indicator over time. These tests are multiple degree of freedom tests using a Wald chi-square test (Agresti, 2002) with 0.05 alpha level. The Wald chi-square test is calculated automatically in Mplus and is equal to $P' \left[COV\left(P\right)\right]^{-1} P$ where P is the vector of parameter estimates to be tested and $COV\left(P\right)$ is the variance-covariance matrix of the parameters estimates, which is distributed as chi-square with degrees of freedom equal to the number of parameters tested (see Long, 1997). Tables 2.1 and 2.2 give the results of the this testing.

For both the split half models and the item-level models, all scales have statistically significant item specific error variance (Tables 2.1 and 2.2). While this variance is statistically significant in all cases, the item specific variance is relatively small compared to other variance components such as pure error variance, correlated error within time variance (the c_q), and variance of the Markov factors. For example, in the split-halves models the proportion of total variation due to specific error variance averages 2 percent and for the item level models it averages 3 percent. Specific error variance as a percent of

total variance only exceeds 6 percent for two scales. Because specific item error variance includes memory effects, the magnitude of variation due to specific items over time is probably low here since the time intervals between the three waves are substantial for this population. In addition, for the split half models, the estimation of specific variance due to a half scale (sum of half of the items) may be lower than would be expected for separate specific items since any method (or memory) unique to a given question is combined with methods specific to other questions.

In several instances specific error variance for one indicator is estimated as negative although non-significant value. This is a problem similar to a negative error variance or Heywood case (Bollen, 1989) since the association of the same item over time should not be negative. This problem occurs for both the item-level and split-half level models, but is more prominent in the split-half models. To investigate this problem, the split-half models with specific error variance due to one half only were estimated and in all cases estimates were always positive and significant. Therefore, the simultaneous estimation of specific error variance for more than one indicator is resulting in the negative estimates. For this analysis, non-significant negative specific error variance is ignored. In the case of the split-halves model for the PLS-3 Auditory, PLS-3 Expressive, VABS, and MBA Reading there is one half score with significant negative specific error variance indicating model specification problems or model overfitting when all specific variance factors are estimated. This is particularly true for the MBA and PLS-3 Auditory because the total specific error variance (sum of variance due to each half score) is negative. Nevertheless, this variance generally has very little effect on overall reliability estimates since it is not of substantial size for this application.

Tables 2.1 and 2.2 also present the test results for additional factor variance (item correlations within time). The majority of scales have significant additional factor variance. The test results for scale scores that do not have additional factor variance are highlighted in bold. Variance due to additional factor correlations is fairly substantial for

these scale scores and this population since it is relatively large compared to the other variance components. For example, as a percent of total variation the additional factor variance averages 6 percent and 13 percent for the split-halves and item-level models. For some scales, the addition variance represents up to 21 percent and 34 percent of the total variance for the split-halves and item-level models, respectively. Therefore, additional consistent variance as measured by the c_q factors has a non-trivial impact on the reliability estimates. The test of equal loadings over time indicate that most of the scales have equal loadings for the same indicators over time. Test results are bold in the table for those scales that did not have equal loadings over time.

The general model for each scale score is listed and described under the column "General Model" in Tables 2.3 and 2.4 and the model fit for each general model is given in the columns under "General Model Fit". The general model for a particular scale is the model that fit best given the tests of specific error variance, additional factors, and equal loadings over time. The "free" model is the model described above (see Figure 2.1) with no additional constraints. The "equal loadings" model is a model that includes both specific error variance and additional factor variance but constrains loadings for the same indicator equal over time. This model is the most common for these scale scores and this population. The "no correlated" model is the same as the "equal loadings" model except without additional factor variance. Overall model fit based on the RMSEA, CFI, and TLI indicates acceptable fit for most of the models. However, there are significant chi-square tests in several cases. Given the large sample sizes and the additional model fit statistic results, the significant chi-square tests are likely detecting small departures due to high power of the tests (Bentler & Bonett, 1980). Also, with the number of scales and hypotheses being tested, there is a multiple testing problem where some of the tests will be significant due to chance alone. For example, tests of model fit on 42 models should result in two tests that reach significance at the alpha level of 0.05 by chance. For many of the hypotheses, the p-values for the tests meet an alpha criterion of 0.01 or

smaller.

2.5.2 Simplex Assumption Testing and Reliability Estimates

The general model for each scale, which is given in Tables 2.3 and 2.4 for each scale, is used as the larger model for testing the QSM assumptions of equal true score variance, equal error variance, and equal reliability over time. Tables 2.5 and 2.6 give results of these tests with non-significant tests highlighted in bold. For the majority of scale scores at least one of the assumptions of constant true score variance or constant error variance may be made. In cases where both assumptions are reasonable, the simultaneous assumption of both equal error variance and equal true score variance over time (that is, equal reliability over time) is also acceptable. Neither one of the assumptions seems dominant for the split-halves models where 11 of the 23 scales have equal error variance over time and 9 of the 23 scales have equal true score variance over time (Table 2.5). The assumption of both equal error variance and true score variance over time holds for the four scales where both assumptions hold separately in the split halves models. There are 7 out of the 23 scales where neither QSM assumption holds indicating that reliability estimates obtained for these scales from a model with one of the QSM assumptions may result in poor estimates.

For the item-level models, the assumption of constant true score variance over time is more prevalent than the assumption of constant error variance over time. Constant true score variance holds for 10 of the 17 scales evaluated using the item-level model while only 3 of the 17 scales have constant error variance over time. These results support the idea of a possible test-retest or Socratic effect for this population since the error variances are declining over time while the true score variances are steady over time. The decreases in error variance are expected since the children are developing and thereby increasing their ability to answer the test questions. There are also a fair number (6 out of 17) of scale scores where neither assumption is reasonable. This occurs at higher rate for the item-level models compared to the split-halves models.

It is interesting that for scales that are comparable across the split-halves and item-level models, which include the CDI, School Engagement, and YSR, the QSM assumption test results sometimes differ. For example, for the CDI the split-halves model indicates equal error variance over time while the item-level model indicates neither assumption is acceptable. Also, the school engagement test results are opposite for the two types of models. One difference is that in the split-halves version, the best model for the CDI has equal loadings over time while the item-level model does not. Also, the split-halves general models fit better for the CDI, School Engagement and the YSR compared to the item-level models. This indicates that there are differences in the two model forms, split-halves and item-level. As noted, the equal error variance assumption does not often hold in the item-level model and the equal true score variance holds more often as compared to the split-halves version of the model.

The reliability estimates for the general model, the constant error variance, and constant true score variance models are presented in Tables 2.7 and 2.9. For the split-halves models, 9 out of the 17 constant true score variance models did not converge (Table 2.7). Four of the 9 scales with non-convergence issues were found to have constant true score variance in the testing of that assumption (see Table 2.5). Most of these have issues of non-invertible matrices, which are likely due to empirical underidentification of one or more of the parameters. As discussed in Section 3.1.3, this may occur in these models when both of the stability parameters approach 1 or are equal to each other. The constant true score variance over time constraint may be causing this to occur for many scales. Therefore, no reliability estimates for the constant true score model are available for these 9 scales. For the scales where reliability estimates are available, the estimates do not differ much across the various models. This indicates that the QSM assumption chosen would not affect the reliability estimates dramatically. In some cases applying a QSM constraint can decrease standard errors, but this is not a consistent result. Reliability estimates are efficient for these data including the general model estimates.

For the item-level models, 4 out of the 17 scales did not estimate for any of the models. These were generally scales with more indicators (15 - 32) and were simply too large to estimate. Also, most of the constant true score models did not estimate likely because of empirical underidentification issues. This is interesting in the item-level model case since almost 60% of the tests of constant true score variance were accepted (Table 2.6). Some of the other item-level models produced reliability estimates that are greater than one. In these cases, the pure random error is negative after applying all possible correlations between indicators within time, the c_q. This indicates that the model should be simplified to include fewer c_q elements for these scales. Of the six scales where this occurs, five have statistically significant negative pure error variances indicating that the number of additional factors should be reduced. Like the split-halves models, the reliability estimates for the item-level models do not differ greatly across the QSM assumptions indicating that these assumptions may not have a large impact on estimates for these data.

2.6 Conclusions

In this chapter a longitudinal, latent variable model useful for estimating scale score reliability is outlined. The model is a hybrid of the QSM with a Markov structure and multiple item latent factors with congeneric items. The model differs from others in the literature used to test the assumptions of constant error variance over time and constant true score variance over time because it includes separately identifiable and estimable specific error variances for each item and additional factors for all inter-item correlations within time. The additional factors may represent method factors or additional traits for multi-dimensional scales. The very general model presented can be large and unwieldy when there are larger number of items, particularly due to the additional factor terms within time, the c_q. In this instance, the split-halves (or perhaps multi-splits) model works well or an item level model that includes fewer possible additional factors. The additional factors do not necessarily have to be theoretically defined additional traits

or method effects. The model is general in that it only seeks to estimate true score variance as all possible repeatable variance within the composite. A model could be developed that attempts to specify additional hypothesized methods and traits via the additional factors. Examples of specifying models for the estimation of reliability for a multi-dimensional scale are outlined in Bentler (2005) and Raykov & Shrout (2002). In these cases there are no repeated measurement, therefore, one could not define a general model with all possible additional variance beyond the single factor. Therefore, it is necessary to specify the hypothesized dimensions as separate factors.

Estimation of specific error variance in addition to the Markov structure of the primary trait factors poses a problem for some scales and models with the QSM constraints. For example, several models could not be estimated with the constant true score variance constraint even when tests show that this constraint is appropriate. However, the general model without the QSM assumptions seems to estimate without problem in most cases except where there are a large number of items in the scale resulting in very large models. And, the reliability estimates from the models with QSM constraints do not differ markedly from the general model in cases where they can be compared. Reliability estimates for the NSCAW example were most affected by additional factor variance. In other populations and for other scales, specific item error variance may also impact reliability estimates greatly. The general model, which includes these additional variance components, is therefore a good starting point for estimating reliability.

Scale	# items	n	Specific Error Variance Tests			Additional Factor Variance Tests			Equal Loadings Across Time Tests		
			Wald	**df**	**p-value**	**Wald**	**df**	**p-value**	**Wald**	**df**	**p-value**
YSR (11+ years) Internalizing	31	1825	55.00	2	0.000	9.65	1	0.002	0.91	2	0.635
YSR (11+ years) Externalizing	30	1825	137.39	2	0.000	15.97	1	0.000	1.72	2	0.423
YSR (11+ years) Total Problem Behavior	101	1825	80.17	2	0.000	20.27	1	0.000	4.83	2	0.090
TRF (5+ years) Internalizing	31	2642	6.27	2	0.043	22.53	1	0.000	1.16	2	0.561
TRF (5+ years) Externalizing	28	2642	47.39	2	0.000	19.14	1	0.000	2.36	2	0.308
TRF (5+ years) Total Problem Behavior	95	2643	41.03	2	0.000	23.76	1	0.000	1.13	2	0.568
CDI (7+ years) Total	27	2914	113.99	2	0.000	22.83	1	0.000	4.22	2	0.121
School Engagement (6+ years)	7	3189	98.75	2	0.000	20.99	1	0.000	1.46	2	0.483
SF-12 Physical Health of Caregiver	12	5491	366.12	2	0.000	23.08	1	0.000	4.51	2	0.105
SF-12 Mental Health of Caregiver	12	5491	155.00	2	0.000	75.48	1	0.000	5.83	2	0.054
MBA (6+ years) Reading	73	3210	267.42	2	0.000	42.12	1	0.000	**16.25**	**2**	**0.000**
MBA (6+ years) Math	68	3182	99.73	2	0.000	**0.53**	**1**	**0.467**	**59.27**	**2**	**0.000**
PLS-3 (0-5 years) Auditory Comprehension	48	2739	66.29	2	0.000	15.31	1	0.000	**18.34**	**2**	**0.000**
PLS-3 (0-5 years) Expressive Communication	48	2732	97.08	2	0.000	11.31	1	0.001	**17.34**	**2**	**0.000**
CBCL (2+ years) Internalizing	31	5330	347.17	2	0.000	61.14	1	0.000	0.20	2	0.905
CBCL (2+ years) Externalizing	32	5330	585.26	2	0.000	62.25	1	0.000	5.38	2	0.068
CBCL (2+ years) Total Problem Behavior*	118	5330									
HOME-SF (0-9 years) Total	26	4318	258.64	2	0.000	55.34	1	0.000	2.76	2	0.251
HOME-SF (0-9 years) Cognitive Stimulation	14	4318	72.69	2	0.000	**0.16**	**1**	**0.691**	**17.41**	**2**	**0.000**
HOME-SF (0-9 years) Emotional Support	12	4318	67.71	2	0.000	23.74	1	0.000	0.89	2	0.642
Peer Loneliness and Social Dissatisfaction (5+ years)	16	3445	14.67	2	0.001	8.32	1	0.004	0.73	2	0.694
SSRS (Caregiver report) (3+ years)	40	5015	372.03	2	0.000	73.46	1	0.000	1.02	2	0.600
SSRS (Teacher report) (5+ years)	30	2613	51.18	2	0.000	14.93	1	0.000	1.47	2	0.479
VABS Daily Living Skill (0-10 years)	15	4118	233.44	2	0.000	5.50	1	0.019	**184.42**	**2**	**0.000**

Bold = less common test result within column

* no convergence

Table 2.1: Results of General Model Testing of the Split-Halves Simplex Model

Scale	# items	n	Specific Error Variance Tests			Additional Factor Variance Tests			Equal Loadings Across Time Tests		
			Wald	**df**	**p-value**	**Wald**	**df**	**p-value**	**Wald**	**df**	**p-value**
CBCL (4+ years) Internalizing	31	3937	5771.76	31	0.000	8047.98	465	0.000	**126.12**	**60**	**0.000**
CBCL (4+ years) Externalizing	32	3937	6933.44	32	0.000	9843.33	496	0.000	**116.10**	**62**	**0.000**
CDI (7+ years) Total	27	2914	1446.85	27	0.000	3044.61	351	0.000	**105.48**	**52**	**0.000**
Peer Loneliness and Social Dissatisfaction (8+ years)	16	2623	243.91	16	0.000	2371.91	120	0.000	**62.56**	**30**	**0.000**
RAPS (11+ years) Emotional Security Primary CG	3	1821	44.67	3	0.000	15.42	3	0.002	7.46	4	0.114
RAPS (11+ years) Emotional Security Secondary CG	3	1134	13.48	3	0.004	15.27	3	0.002	3.43	4	0.489
RAPS (11+ years) Involvement Primary CG	4	1821	49.23	4	0.000	456.16	6	0.000	3.77	6	0.707
RAPS (11+ years) Involvement Secondary CG	4	1133	10.49	4	0.033	256.05	6	0.000	7.37	6	0.288
RAPS (11+ years) Autonomy Support Primary CG	2	1817	51.90	2	0.000	**0.29**	**1**	**0.588**	0.90	2	0.639
RAPS (11+ years) Autonomy Support Secondary CG*	2										
RAPS (11+ years) Structure Primary CG	3	1817	20.65	3	0.000	35.32	3	0.000	**10.05**	**4**	**0.040**
RAPS (11+ years) Structure Secondary CG	3	1131	33.86	3	0.000	**3.58**	**3**	**0.311**	3.74	4	0.442
School Engagement (6+ years)	7	3189	413.82	7	0.000	381.48	21	0.000	12.29	12	0.422
SF-12 Physical Health of Caregiver	12	5491	1698.61	12	0.000	9656.43	66	0.000	28.30	22	0.166
SF-12 Mental Health of Caregiver	12	5491	1616.51	12	0.000	9448.19	66	0.000	**45.74**	**22**	**0.002**
YSR (11+ years) Internalizing	31	1825	1403.71	31	0.000	3241.50	465	0.000	63.53	60	0.353
YSR (11+ years) Externalizing	30	1825	2367.32	30	0.000	2971.11	435	0.000	23.18	58	1.000
VABS Daily Living Skill (6-10 years)	15	2121	719.38	15	0.000	967.84	105	0.000	**41.51**	**28**	**0.048**

Bold = less common test result within column

* no convergence

Table 2.2: Results of General Model Testing of the Item-Level Simplex Model

	General Model				General Model Fit						
Scale	Model Label	specific error variance	additional factor variance	fixed loadings	obs	Chi-square	df	p-value	CFI	TLI	RMSEA
YSR (11+ years) Internalizing	equal loadings	X	X	X	1825	3.13	6	0.792	1.000	1.002	0.00
YSR (11+ years) Externalizing	equal loadings	X	X	X	1825	8.75	6	0.188	0.999	0.998	0.02
YSR (11+ years) Total Problem Behavior	equal loadings	X	X	X	1825	14.99	6	0.020	0.999	0.997	0.03
TRF (5+ years) Internalizing	equal loadings	X	X	X	2642	4.70	6	0.582	1.000	1.001	0.00
TRF (5+ years) Externalizing	equal loadings	X	X	X	2642	9.96	6	0.126	0.999	0.998	0.02
TRF (5+ years) Total Problem Behavior	equal loadings	X	X	X	2643	1.34	6	0.969	1.000	1.002	0.00
CDI (7+ years) Total	equal loadings	X	X	X	2914	39.12	6	0.000	0.991	0.979	0.04
School Engagement (6+ years)	equal loadings	X	X	X	3189	16.31	6	0.012	0.996	0.991	0.02
SF-12 Physical Health of Caregiver	equal loadings	X	X	X	5491	14.08	6	0.029	0.998	0.996	0.02
SF-12 Mental Health of Caregiver	equal loadings	X	X	X	5491	6.04	6	0.419	1.000	1.000	0.00
MBA (6+ years) Reading	free	X	X		3210	129.84	4	0.000	0.983	0.936	0.10
MBA (6+ years) Math	no correlated	X		X	3182	10.72	5	0.057	0.997	0.992	0.02
PLS-3 (0-5 years) Auditory Comprehension	free	X	X		2739	25.10	4	0.000	0.995	0.980	0.04
PLS-3 (0-5 years) Expressive Communication	free	X	X		2732	63.90	6	0.000	0.979	0.948	0.06
CBCL (2+ years) Internalizing	equal loadings	X	X	X	5330	5.56	6	0.474	1.000	1.000	0.00
CBCL (2+ years) Externalizing	equal loadings	X	X	X	5330	36.52	6	0.000	0.998	0.996	0.03
CBCL (2+ years) Total Problem Behavior*											
HOME-SF (0-9 years) Total	equal loadings	X	X	X	4318	142.25	6	0.000	0.968	0.921	0.07
HOME-SF (0-9 years) Cognitive Stimulation	no correlated	X		X	4318	131.06	5	0.000	0.969	0.906	0.08
HOME-SF (0-9 years) Emotional Support	equal loadings	X	X	X	4318	45.37	6	0.000	0.988	0.970	0.04
Peer Loneliness and Social Dissatisfaction (5+ years)	equal loadings	X	X	X	3445	3.04	6	0.804	1.000	1.001	0.00
SSRS (Caregiver report) (3+ years)	equal loadings	X	X	X	5015	38.85	6	0.000	0.997	0.993	0.03
SSRS (Teacher report) (5+ years)	equal loadings	X	X	X	2613	39.13	6	0.000	0.995	0.986	0.05
VABS Daily Living Skill (0-10 years)	free	X	X		4118	104.14	4	0.000	0.989	0.959	0.08

* no convergence

Table 2.3: Final "General" Model Resulting from Model Testing of the Split-Halves Simplex Model

	General Model				General Model Fit						
Scale	Model Label	specific error variance	additional factor variance	fixed loadings	obs	Chi-square	df	p-value	CFI	TLI	RMSEA
CBCL (4+ years) Internalizing	free	X	X		3937	8420.86	3687	0.000	0.949	0.941	0.02
CBCL (4+ years) Externalizing	free	X	X		3937	11057.76	3934	0.000	0.947	0.938	0.02
CDI (7+ years) Total	free	X	X		2914	4603.39	2779	0.000	0.937	0.926	0.02
Peer Loneliness and Social Dissatisfaction (8+ years)	free	X	X		2623	1752.85	942	0.000	0.975	0.970	0.02
RAPS (11+ years) Emotional Security Primary CG	equal loading	X	X	X	1821	47.56	23	0.002	0.991	0.985	0.02
RAPS (11+ years) Emotional Security Secondary CG	equal loading	X	X	X	1134	22.64	23	0.482	1.000	1.000	0.00
RAPS (11+ years) Involvement Primary CG	equal loading	X	X	X	1821	85.46	48	0.001	0.986	0.981	0.02
RAPS (11+ years) Involvement Secondary CG	equal loading	X	X	X	1133	77.18	48	0.005	0.984	0.978	0.02
RAPS (11+ years) Autonomy Support Primary CG	no correlated	X		X	1817	4.78	7	0.687	1.000	1.012	0.00
RAPS (11+ years) Autonomy Support Secondary CG*					1130	7.50	6	0.278	0.988	0.971	0.02
RAPS (11+ years) Structure Primary CG	free	X	X		1817	28.96	19	0.067	0.992	0.985	0.02
RAPS (11+ years) Structure Secondary CG	no correlated	X		X	1131	38.70	26	0.052	0.984	0.977	0.02
School Engagement (6+ years)	equal loading	X	X	X	3189	362.20	171	0.000	0.976	0.971	0.02
SF-12 Physical Health of Caregiver	equal loading	X	X	X	5491	2127.48	536	0.000	0.980	0.977	0.02
SF-12 Mental Health of Caregiver	free	X	X		5491	2008.66	514	0.000	0.981	0.977	0.02
YSR (11+ years) Internalizing	equal loading	X	X	X	1825						
YSR (11+ years) Externalizing	equal loading	X	X	X	1825	6428.15	3506	0.000	0.915	0.903	0.02
VABS Daily Living Skill (6-10 years)	free	X	X		2121	1365.36	823	0.000	0.963	0.955	0.02

*no convergence

Table 2.4: Final "General" Model Resulting from Model Testing of the Item-Level Simplex Model

Scale	# items	n	Equal Error Variance Across Time			Equal True Score Variance Across Time			Equal Reliability Across Time		
			Wald	**df**	**p-value**	**Wald**	**df**	**p-value**	**Wald**	**df**	**p-value**
YSR (11+ years) Internalizing	31	1825	**4.01**	**2**	**0.135**	24.07	2	0.000	28.10	4	0.000
YSR (11+ years) Externalizing	30	1825	11.02	2	0.004	17.14	2	0.000	32.50	4	0.000
YSR (11+ years) Total Problem Behavior	101	1825	17.84	2	0.000	22.24	2	0.000	46.00	4	0.000
TRF (5+ years) Internalizing	31	2642	14.74	2	0.001	**5.56**	**2**	**0.062**	20.46	4	0.000
TRF (5+ years) Externalizing	28	2642	**5.56**	**2**	**0.062**	**5.48**	**2**	**0.065**	**9.47**	**4**	**0.050**
TRF (5+ years) Total Problem Behavior	95	2643	**1.94**	**2**	**0.379**	**4.65**	**2**	**0.098**	**6.98**	**4**	**0.137**
CDI (7+ years) Total	27	2914	**5.88**	**2**	**0.053**	7.93	2	0.019	49.56	4	0.000
School Engagement (6+ years)	7	3189	**1.44**	**2**	**0.488**	10.13	2	0.006	26.33	4	0.000
SF-12 Physical Health of Caregiver	12	5491	**5.61**	**2**	**0.060**	13.69	2	0.001	23.74	4	0.000
SF-12 Mental Health of Caregiver	12	5491	8.21	2	0.017	**3.99**	**2**	**0.136**	15.03	4	0.005
MBA (6+ years) Reading	73	3210	10.27	2	0.006	87.45	2	0.000	107.55	4	0.000
MBA (6+ years) Math	68	3182	**3.88**	**2**	**0.144**	42.22	2	0.000	52.91	4	0.000
PLS-3 (0-5 years) Auditory Comprehension	48	2739	39.24	2	0.000	15.57	2	0.000	63.27	4	0.000
PLS-3 (0-5 years) Expressive Communication	48	2732	56.20	2	0.000	18.34	2	0.000	57.39	4	0.000
CBCL (2+ years) Internalizing	31	5330	13.07	2	0.002	15.01	2	0.001	34.58	4	0.000
CBCL (2+ years) Externalizing	32	5330	**2.96**	**2**	**0.227**	20.59	2	0.000	25.77	4	0.000
CBCL (2+ years) Total Problem Behavior*	118	5330									
HOME-SF (0-9 years) Total	26	4318	28.31	2	0.000	**0.51**	**2**	**0.775**	118.11	4	0.000
HOME-SF (0-9 years) Cognitive Stimulation	14	4318	14.50	2	0.001	**5.75**	**2**	**0.057**	35.16	4	0.000
HOME-SF (0-9 years) Emotional Support	12	4318	7.50	2	0.024	**2.87**	**2**	**0.238**	32.46	4	0.000
Peer Loneliness and Social Dissatisfaction (5+ years)	16	3445	**3.78**	**2**	**0.151**	9.61	2	0.008	15.15	4	0.004
SSRS (Caregiver report) (3+ years)	40	5015	**4.12**	**2**	**0.128**	**0.84**	**2**	**0.658**	**4.97**	**4**	**0.290**
SSRS (Teacher report) (5+ years)	30	2613	**0.32**	**2**	**0.854**	**1.53**	**2**	**0.465**	**1.71**	**4**	**0.789**
VABS Daily Living Skill (0-10 years)	15	4118	13.57	2	0.001	108.66	2	0.000	155.08	4	0.000

Bold = nonsignificant test result

* no convergence

Table 2.5: Results of Simplex Model Assumption Testing of the Split-Halves Simplex Model

Scale	# items	n	Equal Error Variance Across Time			Equal True Score Variance Across Time			Equal Reliability Across Time		
			Wald	**df**	**p-value**	**Wald**	**df**	**p-value**	**Wald**	**df**	**p-value**
CBCL (4+ years) Internalizing	31	3937	620.43	2	0.000	**1.66**	**2**	**0.436**	622.16	4	0.000
CBCL (4+ years) Externalizing	32	3937	341.46	2	0.000	**5.40**	**2**	**0.067**	347.11	4	0.000
CDI (7+ years) Total	27	2914	283.55	2	0.000	39.54	2	0.000	327.36	4	0.000
Peer Loneliness and Social Dissatisfaction (8+ years)	16	2623	250.03	2	0.000	8.52	2	0.014	260.11	4	0.000
RAPS (11+ years) Emotional Security Primary CG	3	1821	41.29	2	0.000	**2.65**	**2**	**0.266**	45.99	4	0.000
RAPS (11+ years) Emotional Security Secondary CG	3	1134	23.90	2	0.000	6.85	2	0.033	35.42	4	0.000
RAPS (11+ years) Involvement Primary CG	4	1821	**0.22**	**2**	**0.898**	**1.32**	**2**	**0.518**	**1.41**	**4**	**0.843**
RAPS (11+ years) Involvement Secondary CG	4	1133	**2.62**	**2**	**0.269**	**0.16**	**2**	**0.926**	**2.95**	**4**	**0.566**
RAPS (11+ years) Autonomy Support Primary CG	2	1817	**1.13**	**2**	**0.568**	**0.59**	**2**	**0.746**	**4.30**	**4**	**0.366**
RAPS (11+ years) Autonomy Support Secondary CG*	2	1130									
RAPS (11+ years) Structure Primary CG	3	1817	11.24	2	0.004	15.77	2	0.000	34.71	4	0.000
RAPS (11+ years) Structure Secondary CG	3	1131	13.50	2	0.001	**1.07**	**2**	**0.584**	16.53	4	0.002
School Engagement (6+ years)	7	3189	59.36	2	0.000	**5.89**	**2**	**0.053**	70.10	4	0.000
SF-12 Physical Health of Caregiver	12	5491	29.70	2	0.000	15.90	2	0.000	44.71	4	0.000
SF-12 Mental Health of Caregiver	12	5491	92.49	2	0.000	**5.48**	**2**	**0.065**	98.05	4	0.000
YSR (11+ years) Internalizing	31	1825	264.39	2	0.000						
YSR (11+ years) Externalizing	30	1825	151.41	2	0.000	19.90	2	0.000	173.34	4	0.000
VABS Daily Living Skill (6-10 years)	15	2121	57.03	2	0.000	**6.07**	**2**	**0.048**	63.16	4	0.000

Bold = nonsignificant test result

* no convergence

Table 2.6: Results of Simplex Model Assumption Testing of the Item-Level Simplex Model

scale	model	obs	rho t1		rho t2		rho t3	
YSR (11+ years) Internalizing	general	1825	**0.913**	0.010	**0.898**	0.010	**0.891**	0.010
	constant error variance		**0.920**	0.007	**0.904**	0.008	**0.879**	0.011
YSR (11+ years) Externalizing	general	1825	**0.898**	0.010	**0.885**	0.008	**0.898**	0.008
	constant error variance		**0.915**	0.010	**0.891**	0.012	**0.887**	0.014
YSR (11+ years) Total Problem Behavior	general	1825	**0.958**	0.004	**0.954**	0.004	**0.955**	0.004
	constant error variance		**0.967**	0.003	**0.958**	0.003	**0.953**	0.005
	constant true score variance		**0.948**	0.004	**0.957**	0.003	**0.963**	0.003
TRF (5+ years) Internalizing	general	2642	**0.909**	0.010	**0.893**	0.010	**0.908**	0.012
	constant error variance		**0.899**	0.013	**0.896**	0.009	**0.874**	0.016
TRF (5+ years) Externalizing	general	2642	**0.944**	0.005	**0.946**	0.006	**0.944**	0.006
	constant error variance		**0.949**	0.005	**0.943**	0.004	**0.938**	0.006
TRF (5+ years) Total Problem Behavior	general	2643	**0.969**	0.003	**0.970**	0.003	**0.967**	0.003
	constant error variance		**0.971**	0.003	**0.969**	0.003	**0.966**	0.003
	constant true score variance		**0.966**	0.003	**0.969**	0.003	**0.970**	0.003
CDI (7+ years) Total	general	2914	**0.839**	0.037	**0.826**	0.013	**0.827**	0.013
	constant error variance		**0.875**	0.009	**0.832**	0.011	**0.815**	0.014
School Engagement (6+ years)	general	3189	**0.718**	0.022	**0.684**	0.017	**0.690**	0.019
	constant error variance		**0.736**	0.014	**0.687**	0.013	**0.677**	0.018
	constant true score variance		**0.683**	0.017	**0.701**	0.018	**0.711**	0.019
SF-12 Physical Health of Caregiver	general	5491	**0.677**	0.029	**0.735**	0.014	**0.772**	0.036
	constant error variance		**0.732**	0.017	**0.731**	0.014	**0.753**	0.014
	constant true score variance		**0.723**		**0.738**		**0.715**	
SF-12 Mental Health of Caregiver	general	5491	**0.721**	0.012	**0.729**	0.012	**0.726**	0.015
	constant error variance		**0.748**	0.014	**0.724**	0.011	**0.723**	0.014
	constant true score variance		**0.709**	0.011	**0.734**	0.010	**0.734**	0.011
MBA (6+ years) Reading	general	3210	**0.984**	0.001	**0.986**	0.001	**0.982**	0.001
	constant error variance		**0.986**	0.001	**0.985**	0.001	**0.982**	0.001
MBA (6+ years) Math	general	3182	**0.948**	0.004	**0.932**	0.004	**0.926**	0.005
	constant error variance		**0.949**	0.003	**0.936**	0.003	**0.921**	0.005
	constant true score variance		**0.936**	0.004	**0.939**	0.005	**0.946**	0.005

Table 2.7: Reliability Estimates and Standard Errors from the Split-Halves Simplex Models

scale	model	obs	rho t1		rho t2		rho t3	
PLS-3 (0-5 years) Auditory Comprehension	general	2739	**0.830**	0.051	**0.735**	0.018	**0.758**	0.017
	constant error variance		**0.964**	0.053	**0.965**	0.050	**0.966**	0.047
	constant true score variance		**1.121**	0.143	**0.764**	0.020	**0.719**	0.021
PLS-3 (0-5 years) Expressive Communication	general	2732	**0.897**	0.216	**0.701**	0.022	**0.758**	0.029
	constant error variance		**0.759**	0.040	**0.777**	0.042	**0.828**	0.018
	constant true score variance		**0.767**	0.052	**0.715**	0.020	**0.726**	0.022
CBCL (2+ years) Internalizing	general	5330	**0.912**	0.006	**0.913**	0.004	**0.913**	0.005
	constant error variance		**0.928**	0.003	**0.916**	0.003	**0.909**	0.005
	constant true score variance		**0.899**	0.004	**0.914**	0.004	**0.922**	0.004
CBCL (2+ years) Externalizing	general	5330	**0.946**	0.005	**0.946**	0.003	**0.944**	0.003
	constant error variance		**0.954**	0.003	**0.946**	0.003	**0.946**	0.003
	constant true score variance		**0.938**	0.003	**0.947**	0.002	**0.947**	0.003
HOME-SF (0-9 years) Total	general	4318	**0.784**	0.032	**0.835**	0.028	**0.865**	0.027
	constant error variance		**0.848**	0.016	**0.818**	0.020	**0.809**	0.023
HOME-SF (0-9 years) Cognitive Stimulation	general	4318	**0.694**	0.011	**0.710**	0.012	**0.733**	0.034
	constant error variance		**0.712**	0.009	**0.688**	0.010	**0.727**	0.009
	constant true score variance		**0.697**	0.008	**0.728**	0.009	**0.701**	0.011
HOME-SF (0-9 years) Emotional Support	general	4318	**0.742**	0.030	**0.709**	0.037	**0.750**	0.040
	constant error variance		**0.724**	0.033	**0.712**	0.026	**0.670**	0.039
	constant true score variance		**0.711**	0.033	**0.715**	0.035	**0.760**	0.031
Peer Loneliness and Social Dissatisfaction (5+ years)	general	3445	**0.867**	0.009	**0.870**	0.008	**0.850**	0.012
	constant error variance		**0.874**	0.008	**0.863**	0.008	**0.849**	0.010
SSRS (Caregiver report) (3+ years)	general	5015	**0.894**	0.005	**0.902**	0.004	**0.904**	0.005
	constant error variance		**0.900**	0.004	**0.899**	0.004	**0.901**	0.003
SSRS (Teacher report) (5+ years)	general	2613	**0.932**	0.006	**0.926**	0.006	**0.928**	0.005
	constant error variance		**0.931**	0.004	**0.927**	0.005	**0.928**	0.005
	constant true score variance		**0.929**	0.005	**0.927**	0.005	**0.929**	0.004
VABS Daily Living Skill (0-10 years)	general	4118	**0.920**	0.005	**0.936**	0.005	**0.877**	0.013
	constant error variance		**0.927**	0.003	**0.920**	0.004	**0.884**	0.007
	constant true score variance		**0.898**	0.005	**0.938**	0.004	**0.959**	0.014

Table 2.8: Reliability Estimates and Standard Errors from the Split-Halves Simplex Models Cont.

scale	model	obs	rho t1		rho t2		rho t3	
Peer Loneliness and Social Dissatisfaction (8+ years)	general	2623	**1.141**	0.043	**1.179**	0.052	**1.222**	0.065
RAPS (11+ years) Emotional Security Primary CG	general	1821	**0.666**	0.022	**0.722**	0.024	**0.715**	0.024
	constant error variance		**0.711**	0.022	**0.715**	0.022	**0.683**	0.021
	constant true score variance		**0.669**		**0.703**		**0.721**	
RAPS (11+ years) Emotional Security Secondary CG	general	1134	**0.801**	0.028	**0.805**	0.035	**0.823**	0.032
	constant error variance		**0.842**	0.028	**0.785**	0.033	**0.803**	0.030
	constant true score variance		**0.775**		**0.830**		**0.829**	
RAPS (11+ years) Involvement Primary CG	general	1821	**0.664**	0.025	**0.659**	0.026	**0.639**	0.024
	constant error variance		**0.661**	0.024	**0.659**	0.025	**0.642**	0.023
	constant true score variance		**0.654**		**0.652**		**0.651**	
RAPS (11+ years) Involvement Secondary CG	general	1133	**0.714**	0.040	**0.722**	0.043	**0.716**	0.041
	constant error variance		**0.722**	0.040	**0.709**	0.041	**0.711**	0.039
RAPS (11+ years) Autonomy Support Primary CG	general	1817	**0.434**	0.049	**0.434**	0.037	**0.414**	0.042
	constant error variance		**0.472**	0.038	**0.423**	0.034	**0.428**	0.038
	constant true score variance		**0.410**	0.030	**0.434**	0.031	**0.428**	0.031
RAPS (11+ years) Autonomy Support Secondary CG	general	1130	**1.187**	3.602	**0.772**	1.974	**0.562**	1.169
	constant error variance		**0.580**	0.176	**0.508**	0.196	**0.534**	0.199
	constant true score variance		**0.524**		**0.654**		**0.567**	
RAPS (11+ years) Structure Primary CG	general	1817	**0.536**	0.026	**0.481**	0.045	**0.415**	0.041
	constant error variance		**0.575**	0.021	**0.454**	0.042	**0.387**	0.039
	constant true score variance		**0.502**		**0.539**		**0.546**	
RAPS (11+ years) Structure Secondary CG	general	1131	**0.627**	0.025	**0.628**	0.025	**0.663**	0.020
	constant error variance		**0.667**	0.020	**0.624**	0.022	**0.630**	0.020
	constant true score variance		**0.611**	0.017	**0.645**	0.016	**0.661**	0.015
School Engagement (6+ years)	general	3189	**0.876**	0.031	**0.895**	0.035	**0.907**	0.038
	constant error variance		**0.898**	0.032	**0.891**	0.035	**0.881**	0.036
SF-12 Physical Health of Caregiver	general	5491	**1.039**	0.015	**1.053**	0.015	**1.046**	0.014
	constant error variance		**1.045**	0.015	**1.044**	0.015	**1.041**	0.014
SF-12 Mental Health of Caregiver	general	5491	**1.118**	0.017	**1.158**	0.019	**1.156**	0.019
YSR (11+ years) Internalizing	general	1825	**1.112**	0.048	**1.151**	0.059	**1.187**	0.072

Table 2.9: Reliability Estimates and Standard Errors from the Item-Level Simplex Models

CHAPTER 3

Multilevel Modeling of Samples with Unequal Selection Probabilities

Abstract Probability weights have traditionally been designed for single level analysis and not for use in multilevel models. A multilevel weighting method for applying probability weights in multilevel models has been developed and has good performance with large sample sizes at each level of the model (Pfeffermann, Skinner, Holmes, Goldstein, and Rasbach, 1998). But, the multilevel weighting method can result in relatively poor estimation because the multilevel weights often have more variation than traditional, single level weights. In this chapter, an alternative weighting method called a "mixed" method is proposed that applies the multilevel weight for unequal selection of level one observations at level one of the model along with including the sample design variables used to select clusters in the model at level two. It is hypothesized that using the sample design variable rather than the level two weight will result in estimates with lower MSE and better confidence interval coverage. A simulation analysis is used to evaluate the two alternative weighting methods for two-level models with unequal selection of level one and level two observations. The mixed method does result in less bias, lower variance, and lower mean squared error for models with random intercepts only. However, both the multilevel weighting and the "mixed" method weighting approaches perform less well for models with random intercepts and random slopes.

3.1 Introduction

Many samples for large-scale data sets are selected with complex probability methods that include both clustering and unequal selection probabilities. When data are clustered (or nested) due to the sampling design and hypotheses involve cluster effects, then multilevel (hierarchical, random effects, mixed effects) modeling is an appropriate method. Multilevel modeling has become a prominent method for dealing with clustered data especially because of the advances in software development for this method. Many of the samples selected in multiple stages that result in clustering also involve unequal selection probabilities at one or more sampling stages. That is, either the clusters or the observations within clusters are not selected using simple random sampling and are therefore not directly representative of their counterpart population distributions without accounting for the source of unequal selection probabilities. There are two ways to correct for unequal selection probabilities. The design based method involves the use of probability or sampling weights in estimation to equalize the probabilities of selection. A model based approach involves building a model that includes relevant sample design variables that describe the unequal probabilities of selection. This latter approach results in a model that is "robust" to the unequal selection if the model properly specifies how the sample design variables affect the outcome.

The design-based approach to inference has traditionally been used for descriptive statistics when analyzing complex probability samples (Cochran, 1977; Kish, 1965). This approach to inference assumes the finite population values are fixed and samples are random and generated by the sampling design. Under this philosophy clustering is considered a nuisance due to the sample design and is corrected in estimation by using standard error estimators such as sandwich estimators or replication methods that correct for the non-independence of observations (e.g., Binder, 1983; Hansen, Hurwitz, & Madow, 1953). Under the model-based philosophy, the model generates repeated realizations of random variables and the selected sample is considered fixed (Binder & Roberts, 2003; Skinner,

Holt, & Smith, 1989, section 1.6.4). Therefore, the clusters are considered relevant to the generation of variables and are included in the model, which is supposed to replicate the population generating model. Multilevel modeling is a model-based method for including one or more levels of clustering and should involve hypotheses about cluster effects.

Comparable approaches for handling unequal selection probabilities are taken under the design and model based traditions. Under the model-based philosophy, unequal selection probabilities of observations are ignored since the model is assumed to condition upon all information generating the model outcome(s) including information related to unequal selection probabilities. The analytic model is assumed to replicate the population generating model perfectly, therefore applying the model to any fixed sample results in estimates that are consistent for the infinite population. Under the design-based philosophy, the unequal selection of sample observations is considered a result of the sample design and requires correcting for consistent estimates of fixed finite population parameters. The two approaches to inference have been more fully described and compared in the literature (Hansen, Madow, & Tepping, 1983; Pfeffermann, 1993).

Taking strictly a design-based or model-based approach has become less common as the use of sampling weights in modeling has increased. Many large scale data sets provide sampling weights to users and advise that they be used in analysis. And while analysts are interested in obtaining the best model possible, data and theoretical complexity may limit modeling. In most instances the analytic model is not identical to the population generating model and when this involves the omission of variables related to both sample selection and the outcome being modeled, then this may result in biased and inconsistent estimation (Nathan & Holt, 1980; Nathan & Smith, 1989; Scott & Wild, 1989; Smith & Holmes, 1989). Additionally, it is sometimes not possible to obtain complete information on the sampling design, there are often simply too many variables to condition upon, and/or the sample design variables have no theoretical importance for the modeler. Combining a modeling approach such as multilevel modeling when hypothe-

ses about effects of different levels are of theoretical importance with a sample design approach like applying sampling weights to prevent bias that may result from unequal selection probabilities is a good approach in many cases. Including sampling weights can protect against inadequacies in a model. That is, when the model does not replicate the population generating model, applying weights ensures consistency in model parameters for the finite population from which the sample is drawn even though the parameter may not be correct for the theoretical infinite population. Combining the design and model based approaches is also possible for dealing with clustering. For example, data with multiple levels of nesting may be analyzed using multilevel models for certain nested levels that are of theoretical interest while other nesting is not modeled but accounted for using robust variance estimators. For example, a sample of children nested within school that are in turn nested within school districts may be modeled with a two-level model of children nested within schools where the nesting of schools in districts is treated as a nuisance.

In this chapter, a mixed approach of combining a two-level model with the application of probability weights is evaluated under several different weighting scenarios. The issue at hand is whether some weighting approaches have better estimation properties than others. One approach is to use multilevel weights, which are sampling weights that are specific to the sampling probabilities at each level of the multilevel model. That is, the weight applied at level two correct for unequal selection of clusters and the weights applied at level one correct for the unequal selection of observations within clusters (Pfeffermann, Skinner, Holmes, Goldstein, and Rasbach, 1998). Another approach is to include the sample design variables that define the unequal selection of clusters in the model (a model based approach to unequal selection of observations) along with applying the weight for observations within clusters. Both of these approaches will be detailed in this paper. The motivation for using the latter approach is that the multilevel weights, particularly for the clusters, can be very volatile resulting in larger mean squared error

for estimates using the multilevel weights. A method that includes the sample design variables involved in the selection of clusters can reduce or eliminate inconsistency and be more efficient than applying the cluster weight.

Two parallel approaches are also evaluated where the traditional weight used in single level analysis is used instead of the multilevel weight specific to level one. The single level weight corrects for the unequal selection probabilities of the observations due to both the unequal selection of cluster and the unequal selection of observations within clusters; therefore, it does not correct for the stages of selection separately. The two weighting approaches include applying the single level weight to the multilevel model at level one with and without including the sample design variables that describes unequal selection of clusters. These two methods are included because the multilevel weights prescribed in the literature are often not available to users of survey data and it is unknown how poorly the single level weights perform in multilevel models. Finally, a method that does not correct for unequal selection probabilities is used as a comparison method. Simulated data are used to evaluate the different weighting approaches in two-level models with fixed and random intercepts and slopes. The sampling design used in the simulation is based off a common multistage design called a probabilities proportionate to size (PPS) design using a real national level finite population.

The succeeding sections of the paper include 1) a description of complex samples and sampling weights with applications for dealing with complex sample features in multilevel modeling , 2) a review of the literature that prescribe and evaluates weighting techniques for multilevel models, and 3) a description of the methods, and 4) results of the simulation analysis.

3.2 Complex Samples and Probability (Sampling) Weights

Most large scale surveys do not use simple random sampling (SRS) where observations are selected independently and with equal probability. Instead, data contain such features as nesting and unequal probabilities of selection due to a complex sampling

design. These types of samples pose problems for standard single-level statistical methods, which assume SRS. Data sets that are selected for inference to national populations and other large groups are generally complex samples because they involve clustering, stratification, and/or unequal selection of observations. Clustering typically results when samples are selected in multiple stages where larger sample elements such as geographic regions are selected first followed by selection of multiple elements such as neighborhoods or households within the earlier stage elements. The final stage of selection in survey sampling is usually the individual and the observations of interest. An example is the National Survey of Child and Adolescent Well-Being (NSCAW), which is a survey that has a multiple staged sampling design where selection of a sample that represents the population of children in the Child Protective Services (CPS) system in the United States involves selection of county CPS agencies first, and children within county agencies second. The first stage sampling elements are referred to as the primary sampling units or PSUs in the sampling literature and these are called clusters in the multilevel modeling literature. There may be additional sampling stages and elements and therefore multiple levels of clustering. Not all multiple stage designs result in non-independence of observations. Only when more than one sample element is selected from within formerly selected elements does clustering potentially occur. Even when the latter occurs, the degree of clustering depends on how correlated the observations within clusters are for a particular outcome. To the degree that observations are correlated, bias in the standard errors of parameter estimates and bias in test statistics will occur for single-level analysis with standard, model-based variance estimation. Either design-based standard error estimates or model-based multilevel models may be applied to clustered data to accommodate the non-independence of observations.

A sample feature that may result from a complex sampling process is the unequal probability of inclusion of observations into the sample. Unequal selection of observations is potentially very influential because it will bias parameter estimates when selection is

related to dependent variables. When unequal selection, and therefore the weight, is related to dependent variables then the weights are deemed informative. The degree of informativeness is the degree to which the weights are related to the outcome. Unequal selection probabilities may occur purposely when observations with certain characteristics are over- or under-sampled. Strata are often developed to categorize observations for differential probability of selection usually for ensuring adequate sample sizes for certain subpopulations. Unequal probability of selection may also be a result of multiple stages of selection. For example, when single individuals are selected from sampled households of various sizes. While households may have been selected randomly with equal probability, observations within households were selected unequally depending on the number of eligible individuals living in the household.

One common multiple stage sampling design is a probabilities proportionate to size (PPS) design, which involves selecting clusters with probability proportionate to the number of within cluster observations (Kish, 1965; Levy & Lemeshow, 2008). This is a popular sampling method because it results in smaller standard error estimates when the clusters being sampled vary markedly in size. Also when a second stage of selection within clusters is necessary, PPS sampling can result in equal probability of inclusion of observations when PPS sampling of clusters is combined with random selection of the same number of observations within each cluster. Two-staged PPS sampling designs may also involve unequal selection of observations within clusters. If the *number* of observations selected is equal or near equal across clusters, this will minimize the degree of unequal inclusion due to the PPS selection of the clusters. The NSCAW, which is used in the empirical example in this chapter, is selected using a PPS design where CPS agencies are selected with probabilities proportional to the number of children served by the agency. Children are subsequently selected with unequal probability, but with near equal numbers within selected agencies.

3.2.1 Probability (Sampling) Weights

Probability or sampling weights are designed to correct estimates for the biasing effects of unequal selection probabilities. Weights are the inverse of the probability of selection of observations into the sample. Therefore, observations with a higher probability of selection are weighted down and observations with a lower probability of selection are weighted up. Weights typically account for unequal selection induced by the sampling itself but also correct for unequal inclusion due to nonrandom non-response or non-participation of observations. Sampling weights are the inverse of the inclusion probability. Thus, the weight for observation j is

$$w_j = \frac{1}{\pi_j} \tag{3.1}$$

where π_j is the inclusion probability for observation j. Weights estimate the number of finite population elements, N, that are represented by each sample observation. So the weight for observation j estimates the number of finite population elements represented by observation j and the sum of the weights is an estimate of the total population size, $\sum_{j=1}^{n} w_j = \hat{N}$.

In multistage cluster selection the probability of inclusion for the final sample element is a multiplicative factor of the probability of selection at each stage of selection when each stage is independently selected. One example is a two-stage cluster sample where the probability of selection for a final sample element is

$$\begin{aligned} & P\left(j^{th} \text{ element in the } i^{th} \text{ cluster selected}\right) \\ = \; & P\left(i^{th} \text{ cluster selected}\right) P\left(j^{th} \text{ element selected } \mid \; i^{th} \text{ cluster selected}\right) \end{aligned} \tag{3.2}$$

In this case the single-level weight is

$$w_{ij} = \frac{1}{\pi_i \pi_{j|i}} = \frac{1}{\pi_{ij}} \tag{3.3}$$

where $\pi_{j|i}$ is the conditional probability of individual j selected in cluster i and π_i is the probability that cluster i is selected. For a PPS sampling design π_i is proportional to the

size of the cluster where the size is the number of population elements within the cluster. PPS sampling at the cluster level may be combined with equal number of observations selected within each cluster because it results in an equal or near equal probability of selection at the lowest observation level. In this case the probability of selection for observation j in cluster i is constant across level-one observations and is equal to

$$\pi_j = \frac{N_i}{M} \cdot \frac{n}{N_i} = \frac{n}{M}$$

Where N_i is the number of population members in cluster i, M is the number of clusters in the population of clusters, and n is the number of observations selected within each cluster. Sample designs sometimes do not select observations within clusters with equal probability and therefore the conditional probability, $\pi_{j|i}$, is not equal to $\frac{n}{N_i}$ for every observation within cluster i. This is the case in the NSCAW where children were selected with different probabilities depending on their age, type of purported abuse, and whether or not they received Child Protection Agency (CPA) services. However, the number of children selected within cluster is nearly the same across clusters, which minimizes the effect of the unequal selection of agencies in proportion to their size. The purpose for trying to obtain equal numbers of children within agencies is twofold. First, this creates equal caseloads for interviewers and second it reduces the variance of the weights. Since weight variance will increase the variance of parameter estimates, less variable weights result in more efficient estimates.

In the design-based tradition, weights are used for descriptive statistics and have also been used in single level models. The use of probability weights in models has been studied (Skinner, Holt, & Smith, 1989; Skinner & Holmes, 2003); and, the standard weights available with data sets, the w_j, are appropriate for descriptive statistics and single level models. However, sampling weights that are available in large scale survey data that also include nesting generally do not provide weights appropriate for multilevel modeling. The multilevel weights require that the cluster probabilities and within cluster conditional probabilities of selection are known separately.

Probability weights are used to correct for biases due to unequal selection, but they also add variance to parameter estimates. This results in less efficient parameter estimates that require larger sample sizes for consistency. The effect that weights may have on an estimator is a function of how variable the weights are. For example the estimated design effect for a mean or total due to unequal weighting, the unequal weighting effect (UWE), is calculated as one plus the coefficient of variation of the weights:

$$UWE = 1 + \frac{\sigma_w^2}{\bar{w}^2} = 1 + v_w^2 \tag{3.4}$$

where v_w is the coefficient of variation of the weights. The UWE is an estimate of the increase in the standard error for a mean estimate where the standard error increases by a factor of $\sqrt{UWE}$. The UWE for regression parameters may also increase with weight variation. However, the UWE is lower when weights are informative (Kish, 1965). In most two-staged PPS sampling designs, the separate probabilities for each stage have more variance than the single level probability that is a multiplicative factor of the probabilities at each stage of selection. This has the potential of resulting in more volatile parameter estimates when using the multilevel weights depending on the degree of informativeness of the weights. The potential additional parameter variance due to the multilevel weights is a major motivation for comparing the weighting approach to other model-based approaches that include the cluster size variable in the multilevel model at level two rather than the cluster weight.

3.2.2 Multilevel Models for Nested Data

When PSUs or clusters are randomly selected, they may be treated as random effects in a multilevel model. Multilevel models include specification of the non-independence of observations through random effects (see for example, Goldstein, 2003; Raudenbush & Bryk, 2002). Parameters are weighted in proportion to the precision of parameters for each cluster using a generalized least squares type estimator. Multilevel modeling is a useful approach for hypotheses about contextual (cluster) effects including variance

components due to clusters.

The level one equation for a two-level model may be written

$$y_{ij} = \beta_{0i} + \sum_{k=1}^{K} \beta_{ki} X_{kij} + \varepsilon_{ij} \tag{3.5}$$

where y_{ij} is the outcome for individual j in cluster i, β_{0i} is the unique intercept for cluster i, and β_{ki} are the level one coefficients for cluster i for the k^{th} predictor. X_{kij} is the level one predictor k for individual j in cluster i. The ε_{ij} are the level one errors, which are assumed normally distributed with homogenous variance across clusters, that is $\varepsilon_{ij} \sim N\left(0, \sigma_{\varepsilon}^2\right)$. The level two equations may be written

$$\beta_{ki} = \gamma_{k0} + \sum_{s=1}^{S} \gamma_{ks} W_{si} + \delta_{ki} \tag{3.6}$$

where γ_{ks} are the level two fixed effects coefficients, W_{si} are the level two predictors s for cluster i, and δ_{ki} is the level two random effect. It is typically assumed that $\delta_{ki} \sim N\left(0, \sigma_{\beta}^2\right)$. The components of the level two equations define the type of mixed model in terms of random effects with fixed level one effects. The level two equation for a random intercept model reduces to

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \delta_{0i} \\ \beta_{ki} &= \gamma_{k0} \ , \ k \geq 1 \end{aligned}$$

Inclusion of the δ_{ki} for additional β_{ki} results in a model with random slopes. For example, the level two equations

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \delta_{0i} \\ \beta_{1i} &= \gamma_{10} + \delta_{1i} \end{aligned}$$

result in the mixed model

$$y_{ij} = \gamma_{00} + \delta_{0i} + \gamma_{10} X_{1ij} + \delta_{1i} X_{1ij} + \varepsilon_{ij}$$

Where γ_{00} is the fixed intercept, δ_{0i} is the random intercept, γ_{10} is the fixed slope for variable X_1 and δ_{1i} is the random slope for variable X_1. Inclusion of level two predictors W_{si} results in a model with level two fixed effects and also the interaction of level two predictors with level one predictors when X_{kij} are present for β_{ki} in equation (3.5). All random effects are assumed to have a mean of zero and be uncorrelated with all fixed effects. Additional assumptions are that the ε_{ij} are uncorrelated with all δ_{ki}.

Estimation

Multilevel models may be estimated using maximum likelihood (ML). The ML estimator assumes that the random effects are from a multinormal distribution. With ML estimation, the likelihood is usually maximized using a computational algorithm because there is no closed form solution for the maximizer of the likelihood (Raudenbush & Bryk, 2002). The computational algorithm used in this chapter is the Iterative Generalized Least Squares (IGLS) algorithm (Goldstein, 1986). Similar to other computational methods, IGLS iterates between estimation of the fixed parameters and the random parameters. In the linear model, the estimator for the fixed effects is given by the generalized least squares estimator, which yields the ML estimates when the residuals are normally distributed. The GLS estimator of the fixed effects is

$$\hat{\beta} = \left(X^T V^{-1} X\right) X^T V^{-1} Y \tag{3.7}$$

where X is the matrix of covariates, Y is the vector of outcomes and V is the block diagonal matrix made up of the covariance matrices of random effects for each cluster.

The initial estimate $\hat{\beta}_{(0)}$ is usually the ordinary least squares estimate, which assumes that observations within clusters are independent, i.e., that there are no clustering effects and $V_{(0)}$ are identity matrices. The initial estimates of the random effects $\hat{V}_{(0)}$ are obtained by calculating the cross product of the the raw residuals using (3.7):

$$\tilde{Y} = Y - \hat{\beta} X \tag{3.8}$$

where $\tilde{Y}$ is the vector of raw residuals and the cross product is $\tilde{Y}\tilde{Y}^T$. The $\hat{V}_{(0)}$ is then used in (3.7) to obtain updated estimates of $\hat{\beta}$ and then alternate between the random and fixed effects estimates until the estimates for the fixed and random effects parameters do not change. The IGLS procedures is outlined by Goldstein (1986) and also described in Goldstein (2003) and Raudenbush & Bryk (2002).

The equations for both the fixed and random effects may be written as a combination of sums (Pfeffermann, et al., 1998, section 3; LISREL Manual, 2008, pp.214-223). The sums are taken over the different levels of the model where sums at level one are then aggregated over clusters. Similar to the GLS approach, a full ML estimator would take integrals rather than sums where the integrals are taken over the different levels (Rabe-Hesketh & Skrondal, 2006). Weighted sums or weighted likelihoods (pseudo-likelihoods) may be used with probability weights where the weights may differ for sums and integrals taken over a given level of a multilevel model. The use of probability weights with the IGLS estimation algorithm is called probability weighted iterative generalized least squares (PWIGLS).

3.3 Probability Weighting in Multilevel Models

The traditional single level weight, $w_{ij} = \frac{1}{\pi_i \pi_{j|i}}$, which is typically available on survey data has been found inappropriate for multilevel models. This is because the overall inclusion probabilities, $\frac{1}{\pi_{ij}}$, for the final sample elements do not carry enough information for appropriate bias correction (Pfeffermann, et al., 1998, 24). This is because bias in parameters associated with variables or random effects at a given level is due to the unequal selection at specific stages in a multistage selection. Also, the finite population values are not independent and therefore cannot be estimated using a simple sum (Pfeffermann, et al., 1998, 24). Therefore, the traditional single level weight given in equation (3.3) also requires disaggregation into its component weights specific to the stages of selection that correspond to levels in the model. A method for doing this has been developed by Pfeffermann, Skinner, Holmes, Goldstein, and Rabash (1998) and is described in the

next section. Prior to the Pfeffermann, et al. (1998) paper, probability weighting for either variance estimators (Graubard and Korn, 1996; Longford, 1995) or fixed effects parameters (Pfeffermann & LaVange, 1989) were separately addressed in the literature.

3.3.1 Pfeffermann, Skinner, Holmes, Goldstein, and Rabash (1998)

The authors prescribe using weights at each level of a two-level model to correct for possible informative sampling at one or both stages of selection. Informative sampling occurs when the inclusion probabilities are related to an outcome variable after controlling for other covariates.

The two-level probability weights are outlined per Pfeffermann, et al. (1998) as follows. The basic requirement to compensate for the biasing effects of unequal inclusion probabilities, which may be informative at any level, is to have weights at each level where weights at level two are the inverse of the probability of selection of the level two units (clusters or primary sampling units, PSUs):

$$w_i = \pi_i^{-1} \tag{3.9}$$

where π_i is the probability of selection for cluster i. The weights at level one are the inverse of the inclusion probability of the level one units conditioned on selection of the level two unit:

$$w_{j|i} = \pi_{j|i}^{-1} \tag{3.10}$$

where $\pi_{j|i}$ is the probability of selection of observation j in cluster i given selection of cluster i. These correspond directly to the components of the single level weight (3.3) from a multiple staged sampling design.

Pfeffermann, et al. (1998) provide weighted estimators in a iterative generalized least squares (IGLS) setting, which they call PWIGLS for probability weighted IGLS. The estimators are equivalent to pseudo-maximum likelihood (weighted ML) estimation under the standard case of normality (Rabe-Hesketh & Skrondal, 2006). PWIGLS estimators are derived by replacing sums of population units by weighted sample sums to get to

the finite population model values. The PWIGLS estimators are outlined in the Pfeffermann, et al. (1998) article as well as the LISREL documentation (duToit, 2006). The PWIGLS estimators are consistent when both the number of clusters and the number of observations within cluster increase. While, the consistency of fixed effects is established for a large sample of clusters (level two units), the consistency of random effects requires large cluster sample sizes (level one units per level two units). This is required because the sums over level one units are non-linear due to the weights and the effect of the non-linearity may not vanish for small cluster sample sizes.

Due to the large within cluster sample size requirement for consistent estimation of random effects, several scaling methods for the conditional level one weights in equation (3.10) are proposed to reduce small sample bias in random effects parameters, especially for the individual error variance σ_ε^2 parameter. One scaling correction factor scales the weights so that they sum to the cluster sample sizes. For example, the sum of the conditional weights of students selected in classroom i sum to the number of students in the sample from classroom i. This scaling factor is

$$\lambda_i = \frac{n_i}{\sum_j w_{j|i}}. \tag{3.11}$$

The scaled level one weight becomes

$$w^*_{j|i} = \lambda_i w_{j|i} = \frac{w_{j|i} n_i}{\sum_j w_{j|i}} \tag{3.12}$$

Another scaling factor scales the weights so that they sum to the effective sample size where the effective sample size is the sample size that we would have if the sample were selected with simple random sampling. The effective sample size is typically smaller than the actual sample size when clustering or unequal selection probabilities exist. This weight scaling was originally proposed by Longford (1995, 1996). The Longford method of scaling did not perform as well as the scaling factor in equation (3.11) in several simulation studies (Asparouhov, 2006; Pfeffermann, et al., 1998; .Rabe-Hesketh & Skrondal, 2006)

Pfeffermann, et al. (1998) test the estimators and various scaling methods in a simulation study and also apply them to real data. Results confirm that not weighting results in serious bias when selection is informative at both levels with bias in fixed effects as well as the random intercept. The random intercept is also slightly biased in the non-informative selection case, which the authors suspect is the small sample bias of the ML estimator. The non-scaled weighted estimators remove the bias in all parameters for the larger cluster sample sizes (around 38). For smaller cluster sample sizes, the random effects at both levels remain inconsistent when the weights are not scaled. Scaling to the actual cluster sample sizes using λ_i worked better for informative selection at level one. Standard errors using the Delta method perform very well except for the standard error of the level one random effect (σ_ε^2), which requires further adjustments. In sum, the optimal multilevel weights were found to be w_i (equation 3.9) and $w^*_{j|i}$ (equation 3.12). These weights may be applied properly in LISREL, HLM using the PWIGLS algorithm and in Mplus and GLLAMM using ML estimation with a numerical integration algorithm.

3.3.2 Simulation Studies Evaluating the Pfeffermann, et al. Method

Several subsequent simulation studies evaluate the multilevel weighting method proposed in Pfeffermann, et al. (1998) under various conditions. These articles are reviewed in this section. Simulation studies are required for evaluating the methods because there are no closed form solutions for the estimators and the properties of the method under small finite samples is not known. Asparouhov (2006) does present a closed form solution for the linear random intercept model only and Kovačević & Rai (2003) give census estimating equations for multilevel models generally. Nevertheless, the multi-level software that implements weighting use iterative algorithms to estimate parameters. The weighted (or Pseudo) maximum likelihood (PML) method (Skinner, 1989) is used in GLLAMM (a Stata program) and Mplus software (Muthén & Muthén, 1998-2006) along with some type of iterative numerical integration. LISREL and HLM software use PML with the PWIGLS algorithm for multilevel probability weighted models.

Asparouhov (2006) performed a simulation to assess the PML with multilevel weights under various conditions including scaling method, cluster size, invariance of level one selection across clusters (same selection of level one within all clusters), the degree of informativeness of selection, intraclass correlation, and standardized weight variability. They compare several scaling methods including the method described in equation (3.11). They found that smaller cluster sample sizes and more informativeness increase the bias in unweighted estimates. They recommend the scaling method given in equation (3.11) and also a scaled weight for level two weights when selection is invariant across clusters.

Rabe-Hesketh & Skrondal (2006) evaluate the Pfeffermann weighting methods for generalized linear mixed models, specifically a two level logistic regression model using a simulation study. They show that the scaling of level one weights is necessary for reducing bias in a logistic regression model not only in the random effects coefficients, but also the fixed effects coefficients since the random effects and fixed effects parameters are correlated. They also use estimates from the mixed-model to get back marginal model estimates to address whether the marginal estimates are unbiased. They compared three scaling methods offered in the literature, the Pfeffermann method presented above, the Longford (1995, 1996) method of scaling weights to sum to the effective sample size in each cluster, and the Korn & Graubard (2003) method of using only level two weights that are the product of the level two weight, w_i, and the conditional level one weight, $w_{j|i}$, that is using single-level weights at level two. Results confirm the findings from Pfeffermann, et al. The random intercept is overestimated with raw weights, especially for smaller cluster sizes. In addition, all three weighted methods produce biased estimates for the regression coefficients when the random intercept standard deviation is biased due to the correlation of the fixed and random effects in the logistic case. These biases approximately cancel out when the multilevel coefficients are used to get back to the marginal effects even with small cluster sizes indicating that interpreting marginal effects (design-based analysis) may be best in many situations involving probability weighting.

Grilli and Pratesi (2004) did a simulation study using the Pfeffermann, et al., (1998) weighting method for multilevel models with ordinal and binary outcomes. They use the PML or weighted log-likelihood estimator with adaptive Gaussian quadrature to obtain parameter estimates and use a bootstrap method for obtaining variance estimates. Their results show that the scaled weighting method produces low bias in fixed effects parameters with modest increases in sampling variance. However, sample size affects the random effects estimators and weighting only gives satisfactory results when the within cluster sample sizes are adequate. They largely verify the Pfeffermann, et al., results for the linear model and find similar results for the ordinal and binary models.

Simulation studies have also been performed using LISREL Multilevel software using multilevel weights with scaling of level one weights to sum to the actual cluster sample sizes (du Toit, 2006; LISREL, 2006). Two and three level linear models were simulated. Random intercepts and slopes are included in the population models. PML with the PWIGLS algorithm is used for parameter estimates and Taylor series linearization is used for robust standard error estimation. This study again shows the poor performance of unweighted estimation when informative selection exists. Weighted estimates using the Pfeffermann et al. method and their coverage are much better. In addition, LISREL program results are compared to HLM and MPlus both of which implement the Pfeffermann, et al. method. All three packages perform similarly using this method.

In summary, all simulation analyses using the Pfeffermann, et al. weighting method have shown that approximate unbiasedness, or consistency, of random effects estimates requires a large number of clusters AND a large number of level one units within clusters. This seems to occur even with non-informative weights at level one (Asparouhov, 2006). The fixed effects parameters only require a large number of clusters for consistency (Pfeffermann, et al., 1998). Weighted parameter results for the linear multilevel model are less sensitive to the number of clusters than the number of level one units per cluster. Weight scaling at level one can markedly reduce bias in random effects due to small

cluster sizes. Scaling level one weights to the actual cluster sample sizes seem to perform the best in most cases. The weighted estimation techniques always perform better than unweighted estimation when informative selection exists. The literature described in this section assessed the quality of the multilevel weights under various conditions and scaling techniques and do not consider alternative model-based approaches or evaluate the use of the traditional weight (designed for single level models) in multilevel models. These alternatives are considered in this chapter.

3.4 Model Based Alternative to Weighting

When unequal selection and weights are unrelated to the outcome of a model, then they are not informative for that model. In this case, the model parameters will be unbiased even though the clusters and observations within clusters were not selected with a simple random sample. Sometimes the unequal selection is simply unrelated to the phenomenon being modeled. In other cases, the sample design variables are somehow related to the phenomena but are properly included in the model resulting in noninformative weights conditioned on the model. The approach of trying to include sample design variables as effects in the model is a model-based approach to the unequal selection problem. Generally, this is very hard to accommodate since the design variables may be unknown or there may be too many to add into the model resulting in huge models. Often, the design variables are not of theoretical interest. Also, the design variables need to be properly specified in interactions and other potential non-linear effects in the model.

The special, though common, case of a multistage PPS sample is less problematic for the model-based approach in that the clusters are selected in proportion to their size. In this case, there is one design variable of interest at level two of the model. The size variable could be included in the model instead of using the cluster weights, which are often very volatile in a PPS sample. Therefore, using the size variable rather than the level two weight has potentially very large payoff in terms of decreasing the variance

in parameter estimates. Of course, there may be other design variables involved in the selection of clusters, which would limit the feasibility of this approach. In the analysis that follows a method that includes the design variable at level two along with the within cluster weight, $w^*_{j|i}$, at level one is compared to the multilevel modeling approach described in Section 5. This is really a "mixed method" approach that combines model-based and design-based (weighting) methods for dealing with unequal selection in the same model.

The problem of not having access to multilevel weights may result in having to include design variables into the model, if they are known. Alternatively, the traditional, single-level weight , w_{ij}, could be used at one or both levels of the multilevel model. This is likely already done in practice, however it is unknown how the traditional weight performs under various conditions. Therefore, in this analysis the single-level weight will be applied in the model at level one as a comparison to using the correct level one weight, $w^*_{j|i}$. The single-level weight will be scaled using

$$\lambda_i = \frac{n_i}{\sum_j w_{ij}}. \tag{3.13}$$

so that the sum of the single level weight within cluster is equal to the cluster sample size. Table 3.1 presents the methods that will be compared in the simulation analysis.

All methods use the PWIGLS algorithm and use the estimation described in Pfeffermann, et al. (1998). Therefore, the differences between the methods analyzed in this study are limited to which weights are applied at each level and whether or not the model includes specification of design variables at level two in lieu of weights. Method 1 does not use any weighting or design variables in the model and is included to evaluate the extent of bias due to the unequal selection probabilities. Method 3 is the Pfeffermann, et al. method using the correct multilevel weights. Method 5 is the "mixed method" that employs the correct multilevel weight at level one along with the design variable at level two. Methods 2 and 4 use the traditional, single-level weight at level one of the model. Method 2 does not include a correction at level two while method 4 includes the

design variable at level two. Therefore, methods 4 and 5 are"mixed methods" because they utilize both design-based weighting at level one to account for unequal selection of observations within clusters along with the model-based method of specifying the design variables in the model at level two to counteract unequal selection of clusters. With the exception of the multilevel weights, method 3, the other methods analyzed have not been considered in prior literature.

3.5 Simulation Design

The methods outlined in Table 3.1 are compared using simulated data that are selected with informative unequal probability. Bias, variance, mean squared error (MSE), and coverage error of the parameter estimates were used to assess and compare the quality of estimates across the five methods of dealing with unequal selection probabilities.

This analysis uses a real finite population that includes observations within clusters where the finite population is one realization among many of the infinite population. The number of clusters and cluster sample sizes and counts for the within cluster stratification variables are known for the real, finite population. Values for variables that are not known for the finite population are generated using models. Sample selection from the finite population is based on the real sample selection design of the National Survey of Child and Adolescent Well-Being, which is a two stage, probabilities proportionate to size (PPS) design. The PPS design includes selection of clusters in proportion to their known size (number of level one observations) and unequal selection across known strata of level one units within clusters. Sample replicates are drawn with varying cluster and level one sample sizes. The IGLS and PWIGLS algorithms in LISREL (2008) software are used for estimation of parameters from two-level linear models. Results from LISREL should match with other software that allows weights to be applied to a specific level and scales the level one weight to sum to cluster sample sizes.

3.5.1 Finite population Information

Population information for the sampling frame of NSCAW II is used. This population is made up of children involved in child protection services (CPS) agencies. The clusters in NSCAW are child protection agencies and these approximately correspond to counties, which are a common geographic unit used as primary sampling units. The finite population has 2,962,541 children in the child protective service (CPS) system in the U.S. in 2005. County level population counts are available for all 3,141 U.S. counties or county-equivalent administrative units. These counts represent the size of the clusters (PSUs), which are CPS agencies in this case. The population counts for children in CPS should parallel the general population counts for these U.S. counties to some degree resulting in a very realistic population of clusters and individuals within clusters. Within counties, there are five strata based on the child's age, whether the child resides in foster care or not, and receipt of CPS services. There are 8 combinations of categories for the three design variables: age (2), foster care (2), and service receipt (2). However, some of the categories are collapsed to create the five within county strata. The population counts for each strata within county are known. Population clusters range in size from 1 to 38,053 with the exception of a large county with around 100,000 children.

3.5.2 Population Generating Models

Model outcomes are generated as a function of the known design variables including cluster (county) size and the five child level demographic strata. Outcomes for some models are also a function of simulated variables that are not involved in the sample design, which were generated $\sim N(0,1)$ and are orthogonal to cluster size and strata for one set of models and correlated with design variables for the other set of models. A single value was generated for each cluster for the level two predictors and different values for each observation are generated for level one predictors. Outcome variables are generated with the following models:

- Random intercept model (excluding the sample design variables):

$$y_{ij} = 88 + 13\text{S1} + 0.05\text{S2} + 7.9\text{S4} + 4.7\text{S5} - 7.7\text{size} + \delta_{0i} + \varepsilon_{ij} \tag{3.14}$$

Where S1, S2, S4, and S5 are dummy variables indicating strata 1, 2, 4, and 5 where stratum 3 is the reference stratum and size is equal to log(cluster size). This model has $var(\delta_{0i}) = 11$ and $var(\varepsilon_{ij}) = 310$ to mimic the NSCAW empirical example, and an additional outcome is generated using $var(\delta_{0i}) = 100$ as an alternative degree of nesting.

- Model 2 is a random intercept model with a level one and a level two fixed effect (excluding design variables)

$$y_{ij} = 87 + 14\text{S1} + 0.03\text{S2} + 8.5\text{S4} + 4.4\text{S5} - 6.4\text{size} - 5X + 5W + \delta_{0i} + \varepsilon_{ij} \tag{3.15}$$

where $X \tilde{} N(0,1)$ is the level one predictor and $W \tilde{} N(0,1)$ is the level two predictor. This model has $var(\delta_{0i}) = 9$ and $var(\varepsilon_{ij}) = 306$ to mimic the NSCAW empirical example, and an additional outcome is generated using $var(\delta_{0i}) = 100$ as an alternative degree of nesting.

- Model 3 is a mixed model with random intercept, a level one and a level two fixed effect, and a random slope (excluding design variables)

$$y_{ij} = 94 + 8\text{S1} - 4\text{S2} + 3\text{S4} - 3.5\text{S5} - 5.5\text{size} - 1.2X + 4.7W + \delta_{1i}X + \delta_{0i} + \varepsilon_{ij} \tag{3.16}$$

where $var(\delta_{0i}) = 6$, $var(\delta_{1i}) = 17$, $cov(\delta_{0i}, \delta_{1i}) = 1.6$, and $var(\varepsilon_{ij}) = 120$. And for a second outcome, $var(\delta_{0i}) = 25$, $var(\delta_{1i}) = 36$, $cov(\delta_{0i}, \delta_{1i}) = 4.6$, and $var(\varepsilon_{ij}) = 140$. The parameter values used in equations (3.14, 3.15, and 3.16) were taken from an empirical NSCAW example using the Preschool Language Scale (PLS) as the outcome. The

PLS scale measures precursors of auditory comprehension and expressive communication skills with tasks that focus on attention abilities, social communication, and vocal development. The same five strata and size variables used in sample selection for this simulation were included in the model to obtain parameter values for the effects of the design variables. The parameter values used for the generated variables, X and W, mimic the values for child age and agency poverty level. The random effects values for this empirical example were used along with the alternative levels that result in a higher degree of nesting. The behavior of the y_{ij} in these models would probably be similar to other quasi-continuous scale score assessments that are sums of multiple items. The PLS scale is normally distributed in this example and ranges from 50 to 150 with a mean and median of 90.

In total, there are ten models. Three model types include the random intercept only model in equation (3.14), the random intercept with fixed slopes model in equation (3.15), and the random intercept and random slope with fixed slopes model in equation (3.16). Each model type has two different intraclass correlation coefficient values with one mimicking the NSCAW empirical example (ICC around 5%) and the other with a higher degree of nesting (ICC around 30%). Also, model types with fixed slopes variables X and W have two degrees of correlation with the sample design variables. In one case, X and W are orthogonal to the design variables and in the other case they are correlated around 0.3. Table 3.2 displays the 10 models.

Note that the finite population values for the synthetic variables including X, W, and the outcomes are one realization of the infinite, superpopulation. Quality of sampling results are based on comparisons with the finite population model results, rather than directly to the superpopulation parameters given in equations (3.14), (3.15), and (3.16). SAS software was used to generate population level data.

3.5.3 Sample Selection

Most simulation studies that involve unequal selection induce informative sampling using selection related to model residuals. This study will do this by selecting on the stratum and cluster size variables that were included in the generating models given in equations (3.14), (3.15), and (3.16). This results in informative selection when the design variables are not controlled for in a model. This mimics the way that informative selection arises in real sample selection situations.

Once all variables are generated, clusters were randomly selected using a probability-proportionate-to-size (PPS) without replacement procedure that gives a higher chance of selection to clusters having more level one observations. A composite cluster size measure is actually used for cluster sampling, which allows for near equal selection of the number of level one units within clusters. A description of the composite size measure is provided subsequently in equation (3.17). The composite size measure results in relative county size proportions that are very close to the relative county CPS population size proportions.

The specific method used to select clusters is a sequential sample selection method (Chromy, 1978). Chromy's method selects units sequentially with probability proportional to size and with minimum replacement, which means that the actual number of hits for a unit can equal the expected number of hits for that unit. This method is used because pure without replacement sampling requires that the expected hits for any one county equal one or less. In the NSCAW population there are several large counties that violate this requirement. When a county was selected multiple times, then each selection was treated as a separate PSU where a county that is selected multiple times, say twice, will subsequently have twice the number of children selected from that county. The sample should be more representative using Chromy's method since larger PSUs should be selected more than once in expectation. Three cluster sample sizes were selected: 50, 100, and 200 clusters. The sample with 100 clusters will replicate the NSCAW study.

To counterbalance the propensity to select areas having the largest caseloads, the same number of children within each cluster was selected regardless of cluster size. Three samples sizes of children within clusters were selected without replacement: 35, 55, and 75. The sample with 55 children per clusters will replicate the NSCAW study. Counties with fewer than 45, 65, and 85 children in CPS were dropped from the population for the three sample sizes, respectively. This was done to ensure adequate samples in each stratum. Sample results are compared to finite populations form which they are drawn to avoid any differences due to dropping the smaller counties. It is very common to choose equal numbers of level one units within level two unit because it minimizes the potential UWE for single-level weights and also controls the interviewer case load for each PSU.

Level one observations were selected with unequal probabilities across five strata using the sampling methodology outlined by Folsom, Potter, & Williams (1987) to be described below. Sampling rates within strata were based on the selection design for NSCAW II. The population and sampling rates are given in Table 3.3 and correspond to substantive strata described above. From Table 3.3 it can be seen that strata 1, 4, and 5 are oversampled and stratum 3 is undersampled. The unequal selection within strata was used to ensure adequate samples of children from the strata.

The Folsom, Potter, & Williams (1987) method is used to ensure a near equal number of level one units (children) is selected within each county while simultaneously retaining the strata sampling rates given in Table 3.3. First, the composite size measure is calculated based on the strata sampling fractions and the strata population sizes. The size measure for cluster i is

$$S_i = f_1 N_{i1} + f_2 N_{i2} + ... + f_7 N_{i7} = \sum_{k=1}^{K} f_k N_{ik} \tag{3.17}$$

where f_k is the sampling fraction for stratum k, and N_{ik} is the population count for stratum k in cluster i. Counties are selected PPS with minimal replacement proportional to this composite size measure, S_i. Next, $n_{ik} = f_k N_{ik}$ children were selected using simple random sampling (SRS) from each stratum within each county.

This two stage selection design was selected using SAS Proc Surveyselect. PPS sampling is common for multiple stage selection (Kish, 1965; Levy & Lemeshow, 2008). The use of a real finite population of clusters that generally corresponds to U.S. counties along with the sample design of the NSCAW II makes this simulation study more generalizable than simulations based off of arbitrary selection designs. Five hundred replicate samples for each of the nine combinations of cluster and level one sample sizes are selected.

3.5.4 Analysis

All models were estimated using PML estimation with the PWIGLS algorithm in LISREL v8.8 software. LISREL v8.8 results are comparable with results from Mplus v5.0. All level one weights are automatically scaled using equation (3.11). Each of the 500 replicate samples were fit to the ten generating models ignoring the design variables except for cluster size when it is part of the analysis strategy. For example, Model 1 is $y_{ij} = \gamma_{00} + \delta_{0i} + \varepsilon_{ij}$. Comparisons will be made in the quality of estimates by looking at relative bias, root mean square error (RMSE), and coverage rates. Evaluation of error is with respect to the finite population model parameters. Relative bias was calculated using

$$\text{relative bias}_r = \frac{\hat{\theta}_r - \theta_{fp}}{\theta_{fp}} \tag{3.18}$$

where r is the replicate sample, $\hat{\theta}_r$ is the parameter estimate from sample replicate r, and θ_{fp} is the finite population parameter. The relative bias is averaged over all 500 replicates for each combination of simulation variables. RMSE was calculated using

$$RMSE = \sqrt{\left(\hat{\theta}_r - \theta_{fp}\right)^2 + var_r\left(\hat{\theta}_r\right)} \tag{3.19}$$

where $var_r\left(\hat{\theta}_r\right)$ is the variance in parameter estimates across the 500 replicates. Coverage rates were calculated as the proportion of finite population parameters that fall within the 95% confidence region for each sample. The 95% confidence region for each replicate was calculated using

$$\hat{\theta}_r \pm 1.96\left(se_r\hat{\theta}_r\right) \tag{3.20}$$

where $\left(se_r\hat{\theta}_r\right)$ is the standard error for estimate $\hat{\theta}_r$ in replicate r.

3.6 Results

The UWE for a mean or total is used to compare the relative variation in the multilevel weights (cluster weight and level 1 weight) and the traditional single level weight (Table 3.4). The actual UWE for the model parameters is not known, but the relative variation as measured by the UWE for a mean will give an estimate of the potential added variation to parameter estimates. Cluster weight variance would affect the efficiency of the level 2 parameters and level one weights would affect the efficiency of the level 1 parameters. Comparison of the UWE for a mean of the weights across the three types of weights shows that the multilevel weights have a greater potential for increasing the standard errors of parameter estimates. The UWE for a mean associated with the cluster level weights decrease with increasing within cluster sample sizes. The UWE associated with the level one weights are smaller compared to cluster weights, but still larger than the single level weights. The UWE associated with level one weights decreases with the number of clusters sampled. This comparative look at the degree of variation in multilevel versus the single level weight should be a good representation for a PPS sample using U.S. counties as primary sampling units (clusters) showing that extracting multilevel weights from weights designed for single level analysis using PPS are potentially markedly less efficient.

The informativeness of the weights was measured by the correlation, or partial correlation controlling for other independent variables, of the sampling weights with the dependent variables. Informativeness results from the fact that the design variables that were used in sampling (strata and cluster size) are related to the outcomes and therefore to the random effects in the analytic models, which do not include design variables. Table 3.5 gives the correlations of the multilevel and single level weights with the outcomes averaged over 500 replicates by model type. The standard deviation of the correlations

across replicates is also presented. The cluster level weights are correlated around 0.3 and the level one weights around -0.3 indicating similar degrees of informativeness of the weights at each level. The single level weights are not informative since they are a lot less variable than the multilevel weights. There is some variation across replicate samples in the degree of informativeness, which will result in different degrees of bias.

The nontrivial degree of informativeness of the weights in this simulation ensure that bias will result in parameter estimates when there is no correction for the unequal selection probabilities. The variation in the multilevel weights should affect the efficiency of parameter estimates for the associated levels of the model, though the unequal weighting effect will be smaller for weights that are more informative.

Regression models that include all simulation design variables were used to determine the most important conditions affecting bias, variance, and MSE. Larger sample sizes decreased bias and standard deviations of the parameter estimates as expected, but did not have a very large effect, so results are not broken down by sample size. The degree of nesting seem to have a greater impact on bias, standard deviations, and MSE. And the correlation of the independent variables with sample design variables also affected bias markedly.

3.6.1 Bias

Relative bias (equation 3.18) averaged across the 500 replications was compared across the four estimation methods. Table 3.6 presents these findings. Positive bias in the random intercept is more prominent in the weighting methods without inclusion of the cluster size variable at level two. Using no corrections for the unequal selection probabilities results in the most severe bias relative to the other methods. And applying only the single level weight at level one is comparable to ignoring weighting altogether. This is due to the fact that the single level weights in this design are nearly equal across observations. In data with a sample design where the single level weights are primarily correcting for unequal selection at level one, single level weighting at level one would

perform better. Average relative bias is positive for random intercept estimates and ranges from 2.9 to 5.2 times the population parameter for the weighting only methods. This is unexpected for the multilevel weighting method, which has been shown to produce good results with scaling of the level one weights. In this study, the large positive bias in the random intercept is related largely to the population models with very small ICCs, which are representative of the NSCAW data. Average relative bias in the random intercept for models with low ICC (around 0.05) and higher ICC (approximately 0.30) are presented in Table 3.7. With clustering as low as 0.05 it is questionable whether multilevel modeling should be used at all and should definitely not be used with multilevel weighting based on the findings from Table 3.7. Random slopes and covariances between the random intercepts and random slopes are overestimated with all of the methods and this is not due to the degree of clustering (Table 3.6). The multilevel weighting approach has the least amount of bias in random slopes and covariances of the random effects. Including the size variable at level two along with the level one weight at level one results in the most bias in random slope estimates and covariances of random effect estimates (Table 3.6), empirical expected values of the estimates are over 2 and 3 times larger than the finite population parameters, respectively.

Average relative bias for random error at level one as well as the level one fixed effect is small for all analysis methods for models without random slopes. However, in the model with random slopes, the level one error as well as the level one fixed effect are underestimated across all methods. The level one fixed effect is more biased for the mixed method of including the size variable along with a level one weight. The fixed slopes at level two are underestimated when no weights or just single level weights are applied at level one. The level two fixed slope is also underestimated with the multilevel weighting, but to a smaller degree. Models that include the size variable at level two have the least degree of bias in the level two fixed slope across all models. The bias in fixed effect estimates is partly a function of the correlation of the independent variables

with the design variables. This is particularly true for the model with random slopes. As shown in Table 3.8, the bias is lower when the independent variable are orthogonal to the sample design variables.

In general, models that include the size variable perform better in terms of bias for the models with random intercepts only. However, when random slopes are introduced these methods exhibit more bias in the random slopes, the associated level one fixed slope, and the covariance between random intercepts and slopes.

In summary, methods that include the size variable at level two seem to perform better in terms of relative bias. This is true for the random intercept models, but not for the model with random slopes. All methods overestimate the random slopes and underestimate the associated fixed level one slope.

3.6.2 Variance

The standard deviation of the parameter estimates averaged across the 500 replicate samples was compared by model type and method in Table 3.9. As expected, methods that use the multilevel weights have the highest standard deviations for most estimates. This is likely due to the larger UWE of the multilevel weights. Methods that include the size variable at level two have the lowest standard deviation for the random intercepts estimates and the fixed slopes at level two. Methods that use the single level weight, which has minimal weight variation, have lower standard deviation in the error estimates and level one fixed slopes. Multilevel weighting has the highest standard deviation in all parameter estimates across all models except for the fixed intercept estimates. The random effect estimates are much more volatile compared to the fixed effects estimates for all model types and analysis methods.

The ICC level also affects the variance in random effects estimates, though in the opposite way compared to relative bias. While there was more bias in random effects estimates with low ICC, there is much more variance in random effects estimates with higher ICC levels. Table 3.10 shows the average standard deviations in these estimates

for outcomes with higher ICC and lower ICC levels. The ICC is also important for the variance of the level two fixed slope estimates, which is smaller for the low ICC outcomes (not shown in the table). This is also true for the level one fixed slope, but only in the random slopes model.

3.6.3 RMSE and Coverage Rates

The RMSE (equation 3.19) of the parameter estimates averaged across the 500 replicate samples was compared by model type and method. RMSE results are presented in Tables 3.11 and 3.12. The mixed methods of combining level one design-based weighting with the size variable at level two give the lowest RMSE in most situations. The mixed methods always perform better than the multilevel modeling, with the exception of the random slope estimates. The method of combining single level weighting with the size variable seems to perform the best in terms of RMSE in most situations. Level one effects such as the error and the fixed x slope are estimated well with no correction and just applying the single level weight at level one. This indicates that the variance of the level one weights, i.e., the level one UWE, has more of an impact on estimation than the bias due to unequal selection at this level. This would not be true for cases where the bias due to unequal selection is larger at level one. It is safer to use the weight specific to level one as has been demonstrated in the literature when weight variation at this level is not as large. Additionally, using the single level weight at level one performs poorly in estimation of level two parameters.

Random intercepts, random slopes, and random error have the most total error in estimation. Analysis method does not seem to matter for the total error in the random slopes, covariance of random intercepts and slopes, and the error in the random slopes models. Total loss in these parameters is fairly consistent across methods. This is due both to some bias and larger standard deviations of these estimates. It is unclear why the multilevel weighting method and the mixed method using the level one weight do not improve upon the random slopes, covariances of random effects, and the error estimates.

Random effects are estimated better according to RMSE in models with outcomes that have lower ICC. This is especially true for the mixed methods and also to a lesser degree the multilevel weighting method. Table 3.12 breaks down RMSE for the random effects by ICC level. For the multilevel modeling method, the larger RMSE for outcomes with higher ICC indicates that the larger standard deviation of the random effects estimates with high ICC outcomes overtakes the added loss from the increase in bias for outcomes with low ICC. Random slope estimates are better for outcomes with low ICC for all analysis methods.

Coverage rates for the parameter estimates by model type and analysis method are presented in Table 3.13. Coverage results generally match RMSE results in terms of comparisons across methods and model types. The mixed methods perform well for the random intercept models and for the fixed effect estimates in the random slopes model. Coverage in random intercepts for the multilevel weighting method are not good. Overall results generally favor the mixed methods, except random effects in random slopes models have poor coverage for all methods.

3.7 Conclusions

This study evaluates several methods for analyzing a two-level model using data selected with unequal probabilities. The current design-based gold standard practice is to apply weights at each level that correct for possible informative selection at each level per Pfeffermann, et al. (1998). This method has been implemented in several multilevel modeling software packages (Asparouhov, 2006; Asparouhov & Muthén, 2006a; du Toit, 2006; Rabe-Hesketh & Skrondal, 2006). However, the weights required in this method are not readily available for most data requiring weighting. Furthermore, the sample designs utilized in collecting data are geared toward single level analysis where weights are designed for efficient estimation at the lowest observation level. Multilevel weights extracted from traditional single level weights after the fact are subject to more variability and thereby less efficient estimation, which can potentially overtake biases due to unequal

selection.

The purpose of this study was to compare the design-based method of using multilevel weights with methods that combine weighting at level one with the model-based strategy of including sample design variables at the cluster level. This strategy is practical for probabilities proportionate to size selection (PPS) where clusters are selected in proportion to their size or any other design where the selection of clusters is based on a limited number of design variables. At level two, the cluster size variable used to select clusters (PSUs) is included in the model as a fixed effect rather than applying a weight at level two. The payoff in terms of increased parameter efficiency is high for a PPS sample with clusters sizes mimicking U.S. counties.

Another purpose of this study was to determine how the traditional weights designed for single level analysis perform when used in a multilevel model. In many circumstances the analyst does not have access to the multilevel weights and only has access to the single level weight. Therefore, using the single level weight in a multilevel model may be common in practice. Some have applied the single level weight at level two in a two-level model with good results (Korn & Graubard, 2003). But this approach worked in this case because there was not unequal selection at level one . Rabe-Hesketh and Skrondal (2006) found that this was not a good approach when unequal selection occurs at level one. This study confirms that with unequal selection probabilities for both the clusters and within cluster observations, the multilevel weights are more appropriate.

Results indicate that RMSE and coverage rates can be improved using the mixed methods. However, the analyst must have access to the variable used in selecting clusters with proportion to size and there should be few if any additional variables affecting unequal selection at the cluster level. There are sometimes additional corrections necessary due to cluster non-response and this can affect the adequacy of simply applying the size variable. In this simulation cluster size affected the outcome as a main effect, but in other situations the effect of the size variable could be nonlinear or interact with other

model variables. If the size variable and the single level weights are available, the level one weights can be extracted by multiplying the single level weights by the size variable. In this study, level one estimates were not that different across methods indicating that there was less effect of unequal selection at that level even though selection was informative and the weights had adequate variance. This is mostly due to the fact that there were no interactions between the level one design variables and the level one independent variables and correlations between design and independent variables at that level were modest. In other studies, this is potentially not the case and the level one weights have been shown to perform better than the single level weight.

Of course, one could take a fully model-based approach and include design variables at each level of the multilevel model rather than apply weights at all. However, this is typically not straight forward as selection of individuals, for example, often involves much more complicated selection including the issue of non-response. There are often too many variables to include in a model and they are may not be simply affecting outcomes as main effects like in this limited simulation study. Adding all design variables in the proper way is difficult and may result in models that are too large and atheoretical. There then become an issue of the larger model decreasing efficiency that may offset any gains by not applying the weights.

The multilevel modeling method with proper weight scaling is a good method that is easy to apply when the proper weights are available. This method is ideal when both the number of clusters and the cluster sample sizes are large. But, even with sample sizes as large as 150 level two units and 75 level one units per level two unit, the mixed method can greatly improve RMSE and coverage as was shown in this study. In addition, random effects estimates may be poorly estimated using the multilevel modeling approach when there is not adequate nesting, for example, ICCs are less than 10%. In this instance, multilevel modeling is probably not the best methodology anyway. Using the single level weight in a marginal model that uses standard error estimates that are robust to the

limited clustering is preferred when nesting is small. Finally, if the multilevel weights are not informative, then the model should be estimated without weights. In this situation the weights would not affect fixed effects point estimates, but would increase the standard errors of such estimates. As well, the random effects coefficients would be less efficient and potentially slightly biased if the noninformative weights are applied (Asparouhov, 2006).

	level one weight*	level two weight	level two design variable
method 1: no corrections	none	none	none
method 2: single level weight	w_{ij}	none	none
method 3: multilevel weights	$w_{j\|i}$	w_i	none
method 4: single level weight with size variable	w_{ij}	none	size
method 5: level one weight with size variable	$w_{j\|i}$	none	size

*all weights applied to level one are scaled to sum to the cluster sample sizes

Table 3.1: Methods Used to Address Unequal Selection

	ICC around 0.05	ICC around 0.30
corr(design vars,X's)=0	3 model types: 1) random intercept, 2) random intercept with fixed slopes, and 3) random intercepts slopes and fixed slopes	3 model types: 1) random intercept, 2) random intercept with fixed slopes, and 3) random intercepts slopes and fixed slopes
corr(design vars,X's)=0.3	2 model types: 1) random intercept with fixed slopes, and 2) random intercepts slopes and fixed slopes	2 model types: 1) random intercept with fixed slopes, and 2) random intercepts slopes and fixed slopes

Table 3.2: Population Generating Model Conditions

	stratum 1	stratum 2	stratum 3	stratum 4	stratum 5
population rates	0.0245	0.2780	0.6221	0.0127	0.0627
sampling rates	0.1096	0.2598	0.3723	0.1082	0.1342

Table 3.3: Population and Sampling Rates

number clusters	within cluster size	Cluster UWE	Level One UWE	Single UWE
50	35	4.34 (1.38)	3.78 (0.78)	1.08 (0.01)
50	55	4.20 (1.20)	3.69 (0.77)	1.11 (0.01)
50	75	3.90 (0.97)	3.75 (0.77)	1.10 (0.01)
100	35	4.78 (1.17)	2.69 (0.18)	1.10 (0.01)
100	55	4.43 (1.10)	2.66 (0.18)	1.13 (0.01)
100	75	3.95 (0.63)	2.68 (0.17)	1.12 (0.01)
150	35	4.98 (1.07)	2.34 (0.12)	1.11 (0.01)
150	55	4.48 (0.77)	2.35 (0.12)	1.15 (0.01)
150	75	4.05 (0.53)	2.36 (0.13)	1.14 (0.01)

Table 3.4: Average Unequal Weighting Effects and their Standard Deviations of Cluster Level, Level One, and Single Level Weights by Cluster Sample Size and Level One Sample Size

Model	cluster weight	level one weight	single level weight
Random Intercept	0.365 (0.046)	-0.354 (0.046)	0.109 (0.053)
Random Intercept Fixed Slopes	0.299 (0.053)	-0.230 (0.094)	0.104 (0.040)
Random Intercept Random Slopes	0.335 (0.054)	-0.314 (0.089)	0.000 (0.055)

Table 3.5: Mean and Standard Deviation of Correlations of Weights with Outcome Variables by Model

	no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model					
intercept	-0.644	-0.627	-0.486	-0.026	-0.081
random intercept	5.410	5.162	3.532	0.027	-0.023
error	0.005	0.009	0.008	0.009	0.014
Random Intercepts and Fixed Slopes Model					
intercept	-0.564	-0.550	-0.430	-0.031	-0.073
level one slope (x)	0.012	0.021	0.022	0.024	0.028
level two slope (w)	-0.528	-0.519	-0.143	-0.033	0.061
random intercept	4.115	3.979	2.857	0.012	-0.068
error	0.006	0.010	0.008	0.010	0.014
Random Intercepts, Fixed Slopes, and Random Slope Model					
intercept	-0.446	-0.440	-0.347	-0.024	-0.031
level one slope (x)	-0.162	-0.241	-0.207	-0.247	-0.462
level two slope (w)	-0.484	-0.473	-0.138	-0.026	-0.015
random intercept	4.300	4.131	2.852	-0.079	-0.231
random slope (x)	1.772	1.740	1.515	1.739	2.188
covariance random effects	2.109	1.601	1.595	2.074	3.688
error	-0.191	-0.184	-0.187	-0.184	-0.177

Table 3.6: Average Relative Proportion Bias in Parameters Estimates by Model Type and Analysis Method

		no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model						
random intercept	low ICC	9.817	9.381	6.385	0.062	0.006
	high ICC	1.002	0.942	0.679	-0.008	-0.051
Random Intercepts and Fixed Slopes Model						
random intercept	low ICC	7.611	7.374	5.299	0.043	-0.058
	high ICC	0.620	0.584	0.415	-0.021	-0.078
Random Intercepts, Fixed Slopes, and Random Slope Model						
random intercept	low ICC	6.899	6.642	4.563	-0.078	-0.268
	high ICC	1.701	1.620	1.140	-0.079	-0.194

Table 3.7: Average Relative Proportion Bias in Random Intercept Estimates by Model Type, Analysis Method, and ICC

		no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts and Fixed Slopes Model						
level one slope (x)	no covariance	0.002	0.001	-0.002	0.001	0.006
level one slope (x)	small covariance	0.021	0.042	0.046	0.046	0.051
level two slope (w)	no covariance	0.021	0.037	0.017	0.016	0.160
level two slope (w)	small covariance	-1.077	-1.074	-0.303	-0.082	-0.038
Random Intercepts, Fixed Slopes, and Random Slope Model						
level one slope (x)	no covariance	0.060	0.004	0.008	0.004	-0.197
level one slope (x)	small covariance	-0.384	-0.487	-0.422	-0.498	-0.728
level two slope (w)	no covariance	0.018	0.032	0.010	0.013	0.046
level two slope (w)	small covariance	-0.986	-0.977	-0.286	-0.065	-0.077

Table 3.8: Average Relative Proportion Bias in Fixed Effects Estimates by Model Type, Analysis Method, and Covariance with Design Variables

	no corrections	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model					
intercept	1.112	7.044	2.644	4.029	5.218
random intercept	20.622	25.012	35.858	8.971	8.866
error	6.681	7.174	15.251	7.174	11.854
Random Intercepts and Fixed Slopes Model					
intercept	1.007	6.064	2.616	4.341	6.547
level one slope (x)	0.275	0.293	0.598	0.292	0.470
level two slope (w)	1.054	1.239	3.034	0.850	0.846
random intercept	16.653	19.242	29.098	8.758	8.543
error	6.607	7.115	15.110	7.115	11.746
Random Intercepts, Fixed Slopes, and Random Slope Model					
intercept	0.718	5.170	1.856	2.419	3.378
level one slope (x)	0.818	0.828	1.876	0.827	0.933
level two slope (w)	0.760	0.915	2.065	0.485	0.439
random intercept	8.410	10.557	12.948	2.669	2.544
random slope (x)	10.343	10.423	20.215	10.416	12.658
covariance random effects	6.583	6.663	13.195	3.815	4.177
error	2.230	2.428	5.208	2.428	4.012

Table 3.9: Average Standard Deviation of Parameters Estimates by Model Type and Analysis Method

		no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model						
random intercept	low ICC	14.533	20.128	20.484	2.687	3.580
	high ICC	26.710	29.895	51.233	15.256	14.153
Random Intercepts and Fixed Slopes Model						
random intercept	low ICC	10.656	14.116	15.017	2.376	3.155
	high ICC	22.649	24.367	43.179	15.139	13.932
Random Intercepts, Fixed Slopes, and Random Slope Model						
random intercept	low ICC	7.178	9.626	9.761	1.337	1.430
	high ICC	9.641	11.487	16.134	4.002	3.659
random slope	low ICC	6.938	7.002	13.561	6.993	8.708
	high ICC	13.749	13.843	26.868	13.839	16.608

Table 3.10: Average Standard Deviation in Random Intercept Estimates by Model Type, Analysis Method, and ICC

	no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model					
intercept	57.285	56.407	43.350	5.860	9.743
random intercept	106.317	102.308	82.401	12.199	12.447
error	9.161	10.108	20.749	10.111	16.539
Random Intercepts and Fixed Slopes Model					
intercept	49.599	48.889	37.944	6.385	11.145
level one slope (x)	0.379	0.420	0.822	0.423	0.657
level two slope (w)	3.528	3.574	4.192	1.184	1.330
random intercept	68.309	66.462	59.409	11.943	12.432
error	9.082	10.132	20.577	10.135	16.482
Random Intercepts, Fixed Slopes, and Random Slope Model					
intercept	42.135	42.001	32.819	3.882	5.394
level one slope (x)	1.152	1.189	2.563	1.191	1.453
level two slope (w)	2.911	2.928	2.893	0.683	0.672
random intercept	47.493	46.203	34.881	4.048	4.936
random slope (x)	48.335	47.485	46.365	47.460	59.889
covariance random effects	10.986	10.373	18.178	8.131	13.234
error	25.062	24.230	25.280	24.225	23.660

Table 3.11: Average RMSE of Parameters Estimates by Model Type and Analysis Method

		no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model						
random intercept	low ICC	107.342	103.857	72.527	3.688	4.886
	high ICC	105.292	100.759	92.275	20.709	20.009
Random Intercepts and Fixed Slopes Model						
random intercept	low ICC	68.884	67.609	50.047	3.246	4.370
	high ICC	67.735	65.314	68.772	20.640	20.494
Random Intercepts, Fixed Slopes, and Random Slope Model						
random intercept	low ICC	48.504	47.392	33.874	2.089	2.778
	high ICC	46.482	45.015	35.888	6.006	7.093
random slope	low ICC	31.801	31.309	30.673	31.290	39.527
	high ICC	64.868	63.661	62.056	63.630	80.251

Table 3.12: Average RMSE of Parameters Estimates by Model Type, Analysis Method, and ICC

	no correction	single level weight	multilevel weights	single weight with size variable	level1 weight with size variable
Random Intercepts Model					
intercept	0.00	0.02	0.00	0.93	0.90
random intercept	0.03	0.06	0.48	0.95	0.95
error	0.94	0.99	0.93	0.99	0.99
Random Intercepts and Fixed Slopes Model					
intercept	0.00	0.02	0.00	0.92	0.93
level one slope (x)	0.94	0.90	0.89	0.89	0.89
level two slope (w)	0.50	0.49	0.83	0.95	0.94
random intercept	0.14	0.18	0.50	0.95	0.92
error	0.94	0.99	0.93	0.99	0.99
Random Intercepts, Fixed Slopes, and Random Slope Model					
intercept	0.00	0.02	0.00	0.84	0.92
level one slope (x)	0.96	0.94	0.90	0.94	0.98
level two slope (w)	0.49	0.48	0.80	0.93	0.94
random intercept	0.01	0.03	0.28	0.85	0.74
random slope (x)	0.01	0.01	0.56	0.01	0.26
covariance random effects	0.91	0.95	0.94	0.68	0.46
error	0.00	0.01	0.10	0.01	0.19

Table 3.13: 95% Coverage Rate of Parameters Estimates by Model Type and Analysis Method

CHAPTER 4

Optimal Probability Weighting Methods in Trajectory Models for Data with Nonignorable Unequal Selection and Wave Nonresponse

Abstract Panel studies often suffer from attrition and intermittent nonresponse. Panel data is also commonly selected using complex sampling techniques that include unequal selection of observations. Unequal inclusion of individuals and of repeated measures will result in biased estimates when the missing mechanism is nonignorable, that is, when missing values are related to outcomes. Probability weighting may be used to correct estimates for nonignorable unequal inclusion. However, weights have not traditionally been applied to the growth curve models frequently used in analysis of change. Growth curve models are estimated using a mixed model where the repeated measures are modeled as a function of both fixed and random parameters. Whereas sampling or probability weights have traditionally been applied to marginal models, which do not include random effects parameters. In this chapter, several weighting approaches are applied to the mixed and marginal modeling frameworks using simulated and empirical data in linear growth models with continuos outcomes. Probability weighting that accounts for both the unequal selection of individuals into the study and dropout over time performs the best in a marginal model when missing data are nonignorable. However, probability weighting does not improve estimates when unequal inclusion of observations is ignorable. In this situation, the use of variance weighting such as the generalized least squares (GLS) es-

timator with repeated measures correlation weight matrix can improve the precision of estimates.

4.1 Introduction

A common issue when estimating change over time using survey data is missing observations. Time-specific observations are often missing due to intermittent nonresponse and dropouts. This results in unbalanced data where some of the repeated measures are unobserved for some individuals. In addition, the panel survey data used in trajectory analysis are often obtained through a complex sample design that includes unequal selection of units, e.g., individuals, into the study (Kish, 1965; Levy & Lemeshow, 2008). When the latter occurs in addition to wave specific nonresponse, there exist missing observations at both levels of a two-level growth model. The unequal inclusion of level two units (individuals) or level one units (repeated observations) due to either unequal selection or nonresponse and dropout is a missing data problem for which the Rubin (1976) typologies apply. Unequal inclusion may result in data that is missing completely at random (MCAR), where missing values are unrelated to repeated measures outcomes in the growth curve model. Unequal inclusion may result in data that are missing at random (MAR) where values for missing observations may be related to the repeated outcome values only indirectly through observed variables. In a growth curve model MAR means that missing observations are related to the missing values of the repeated outcomes only through model covariates including time. Finally, if the unequal inclusion results in missing observations that are related to the values of repeated measure outcomes conditioned on the covariates, the missing data are missing not at random (MNAR) or nonignorable. Nonignorable missing data will result in biased estimates of model parameters.

There is quite a wealth of literature addressing missing data (Allison, 2001; Arbuckle, 1996; Little & Rubin, 1987; Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002) and some of this pertains specifically to missing data in longitudinal analyses (e.g., Diggle & Kenward, 1994; Duncan & Duncan, 1994; Hedeker & Gibbons, 1997; Little, 1995).

However, most of the literature addresses methods that may be used under the MAR assumption. These methods often do not include weighting. When unequal inclusion is nonignorable or MNAR, probability weights may be used to correct for the unequal inclusion of both individuals and time-specific observations. An alternative to weighting for nonignorable inclusion is to try and specify a model that renders the missing data mechanism MAR rather than MNAR and use an estimation method that utilizes all observed information.

Growth curve modeling is a powerful method for estimating change over time using longitudinal data. These models may be estimated in a mixed modeling framework or a marginal modeling framework. The more common approach is to used the mixed modeling framework where repeated measures are estimated as a function of both fixed effects parameters and random effects parameters (Bollen & Curran, 2006; Raudenbush & Bryk, 2002). The fixed effects include an intercept and slope of the trajectory as well as other covariates. The random effects typically include random intercepts, slopes, and time-specific residuals. The mixed modeling framework is appealing because the random effects measure the degree of variation across individuals in growth trajectories. In the marginal model framework (Diggle, Heagerty, Liang, & Zeger, 2002), the expected values of the repeated measures are estimated as a function of an intercept and slope as well as other potential covariates. In this framework, random effects parameters are not estimated. For model with continuous or quasi-continuous outcomes, the interpretation of fixed effects parameters for the mixed or marginal model is the same. However, the optimal application of weights in the two models may differ.

It is unclear what the best probability weighting approach is for a growth curve model when missing data are nonignorable or MNAR. Probability weights have traditionally been used in marginal models (Pfeffermann, 1993; Skinner, Holt, & Smith, 1989). A common weighting approach for repeated measures analysis in a marginal model is to use panel weights along with casewise deletion of individuals who were not observed at

every wave. This approach could also be used in a mixed model where individuals (level two observations) that are missing one or more repeated measures are excluded from the analysis and the panel weight is applied at the level of the individual. But panel weighting does not make use of all of the available observations.

Another method that may be used particularly if panel weights are not available is to simply use the weight for the individuals at the first time point used in the analysis. This method would counteract the informative selection of the individuals into the study; however, it would not correct for any time specific dropout or nonresponse. An alternative to both of these approaches, which has not been addressed in the literature, is to use time varying weights. For this method, the weight for each individual at each time point is used where the time specific weights correct for both unequal inclusion into the study and nonresponse at the specific time point. Finally, a weighting method specific to mixed models has been developed by Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998). While the weighting methods for the mixed model have not been tailored to the longitudinal case, they are easily applied there. The purpose of this chapter is to compare these weighting methods in both the mixed and marginal models using a simulation and an empirical example.

The remainder of this section includes 1) a description of the linear growth model specification and estimation in the random effects and marginal modeling traditions, 2) a description of data structures and missing data in linear growth models, 3) various weighting approaches to dealing with missing data/unequal inclusion probabilities in linear growth models, 4) a description of the simulation design and simulation analysis results, and 5) presentation of methods and results from an empirical example using the National Survey of Child and Adolescent Well-Being (NSCAW).

4.2 The Linear Growth Model

Growth curve models are useful for describing and testing change over time. These models may be estimated in a mixed effects model (mixed model) in either the multilevel

(Raudenbush & Bryk, 2002) or structural equation modeling (SEM) traditions (Bollen & Curran, 2006). In mixed effects growth models, each individual has an estimated trajectory. These individual trajectories are averaged for a population level trajectory estimate, known as a fixed effect, and the individual trajectories (intercept and slope) vary across individuals. The estimated variance of the trajectories (intercept and slope) are known as random effects. The unconditional linear growth model may be written

$$y_{it} = \alpha_i + \beta_i T + \varepsilon_{it} \tag{4.1}$$

where y_{it} is the outcome measured at time point t for individual i. In this model, only the time variable, T, which indicates time point, is included as a predictor. The β_i measures the expected change for individual i, and the intercept, α_i, measures the starting value at the reference level of T for individual i, typically the first observation time. The error term, ε_{it} represents the random error for individual i at time t. The ε_{it} are assumed to have zero mean, $E(\varepsilon_{it}) = 0$ and to be uncorrelated with α_i and β_i. (The errors are also often assumed to have no autocorrelation and equal variance for each t, however these assumptions are not required). The coding of T importantly defines the units of time and the origin of time, where for three time points T may be coded 0, 1, and 2 for assessment at waves 1 through 3. T may also be in some other metric such as age (Bollen & Curran, 2006; Mehta & West, 2000).

Equations for α_i and β_i may be written

$$\alpha_i = \mu_\alpha + \delta_{\alpha i} \tag{4.2}$$

$$\beta_i = \mu_\beta + \delta_{\beta i} \tag{4.3}$$

Equations (4.2) and (4.3) are the "level two" equations where μ_α and μ_β are the mean intercept and mean slope, respectively. The random effects are made up of the individual variation around the average intercept, $\delta_{\alpha i}$, and the individual variation around the average trajectory, $\delta_{\beta i}$. In trajectory models, the intercept and slope are typically

allowed to correlate where $cov\left(\delta_{\alpha i}, \delta_{\beta i}\right) = \sigma_{\alpha\beta}$, that is, the matrix of random effects is unstructured. Equation (4.1) may be written in the following form

$$y_{it} = \mu_\alpha + \mu_\beta T + \delta_{\alpha i} + \delta_{\beta i} T + \varepsilon_{it} \tag{4.4}$$

The assumption of equal error variances, $\sigma^2\left(\varepsilon_{it}\right)$, is often made.

In contrast to mixed effects models, marginal or "population average" models (Diggle, Heagerty, Liang, & Zeger, 2002) have the objective of estimating an average trajectory for the population rather than for each individual. In this framework, average trajectories are estimated, but variation in trajectories are not explicitly estimated. The nesting of time within observation is treated as a nuisance and accommodated by correcting standard error estimation (Binder, 1983; Hanson, Hurwitz, & Madow, 1953). Marginal models are a traditional design-based (survey) estimation approach, which easily incorporate weighting for population average estimates. The model is the same as in equation (4.4) except without the random effects parameters, $\delta_{\alpha i}$, $\delta_{\beta i}$, and ε_{it}. The marginal expectation of the response is modeled as a function of explanatory variables. In the unconditional growth model this includes only time, T.

$$E\left(y_{it}\right) = \alpha + \beta T \tag{4.5}$$

For normally distributed, continuous data, the mixed effects and marginal models lead to the same fixed effects parameter estimates (Diggle, Heagerty, Liang, & Zeger, 2002; Zeger, Liang, & Albert, 1988). This is because the average of the intercepts and slopes for individuals is the same as the population average intercept and slope. For an unconditional linear growth model:

$$E\left(y_{it}\right) = E\left[E\left(y_{it}|z_i\right)\right] = \alpha + \beta T + E\left(z_i\right) = \alpha + \beta T. \tag{4.6}$$

where z_i are the random effects. In the mixed effects approach, a trajectory for each child is estimated either explicitly or implicitly. The individual trajectories are then averaged to get an estimated mean population trajectory. For nonlinear outcomes, the random

effects and marginal models lead to different interpretations because the equivalency in expression (4.6) does not hold under transformation (Zeger, Liang, & Albert, 1988). In generalized mixed models, regression coefficients are scaled by a factor related to the amount of random effect variance.

The hypothesis of interest in an unconditional linear growth model typically includes: (1) $H_0 : \mu_\beta = 0$, the mean trajectory has a slope of zero, that is, no change on average over time. And in the mixed model additional hypotheses include (2) $H_0 : var(\beta_i) = var(\delta_{\beta i}) = 0$, the variance of individual slopes is zero, that is, no variation in change over time across individuals, and (3) $H_0 : var(\alpha_i) = var(\delta_{\alpha i}) = 0$, the variance of individual intercepts is zero, that is, no variation in intercept value across individuals.

It is often not just of interest to describe and test for change over time, but to test for predictors of change. A conditional growth model may be fit to longitudinal data where time varying and time invariant predictors affect the trajectory parameters. The conditional growth model equations in a mixed model are

$$y_{it} = \alpha_i + \beta_i T + \varepsilon_{it} \tag{4.7}$$

$$\alpha_i = \mu_\alpha + \Gamma_\alpha X_{it} + \delta_{\alpha i} \tag{4.8}$$

$$\beta_i = \mu_\beta + \Gamma_\beta X_{it} + \delta_{\beta i} \tag{4.9}$$

Again, y_{it} is the outcome measured at time point t for individual i. The X is the matrix of covariates, which may be time varying. The Γ_α is the matrix of slope parameters for the effects of the covariates on the trajectory intercepts and Γ_β is the matrix of slope parameters for the effects of the covariates on the trajectory slopes. The μ_β parameter measures the trajectory slope of an outcome across time as measured by T. This mean slope parameter is conditional given the other covariates in the model. That is, μ_β is conditioned on the X_i. The parameter μ_α measures the mean intercept of the trajectory, which is the conditional mean at $T = 0$. This intercept is also conditioned on the covariates in the model and measures the mean intercept when all $X_i = 0$. In reduced

form equations (4.7), (4.8), and (4.9) become

$$y_{it} = \mu_\alpha + \Gamma_\alpha X_{it} + \mu_\beta T + \Gamma_\beta X_{it} T + \delta_{\alpha i} + \delta_{\beta i} T + \varepsilon_{it} \tag{4.10}$$

The comparable marginal model is the expectation of the response modeled as a function of time, T, and the explanatory variables, X_{it}.

$$E\left(y_{it}\right) = \alpha + \beta T + \Gamma_\alpha X_{it} + \Gamma_\beta X_{it} T \tag{4.11}$$

Conditional growth models allow hypothesis tests of differences in trajectory intercepts and slopes for subgroups of a categorical covariate and for levels of a quantitative covariate. Hypotheses in both the marginal and mixed model include: (1) $H_0 : \gamma_{\alpha k} = 0$, the effect of X_k on the intercept is zero, i.e., there is no difference in the mean of y_{it} at $T = 0$ across levels of X_k, and (2) $H_0 : \gamma_{\beta k} = 0$, the effect of X_k on the slope is zero, there is no difference in the mean of β_i across levels of $X_{k.}$.

The mixed models presented in this section may be estimated using a latent variable (SEM) approach (Bollen & Curran, 2006; McArdle, 1986; Meredith & Tisak, 1990; and Willett & Sayer, 1994). In the SEM parameterization, the intercept and slope as well as the random effects are unobserved latent variables measured by the outcomes, y_{it} and T represents the factor loading on the slope latent variable. However, the theoretical model is the same regardless of whether the mixed or SEM framework is used.

4.2.1 Estimation

In the mixed model with ML estimation, fixed effects estimates and random effects estimates are simultaneously estimated whereby the fixed and random effect influence the estimation of the other through the likelihood function. The likelihood equation for the fixed and random parameters in θ is (Diggle, Heagerty, Liang, & Zeger, 2002)

$$L\left(\theta; y\right) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f\left(y_{ij} | Z_i; \beta\right) f\left(Z_i; G\right) dZ_i \tag{4.12}$$

where Z_i are the random effects variables, β is the matrix of fixed effects parameters, and G is the matrix of random effects parameters. The inclusion of the distribution of

the random effects $f(Z_i; G)$ means that the likelihood equations for the fixed effects are conditioned on the random effects. Estimation requires an iterative process that alternates between estimating the fixed and random parameters given the updated estimate of the other. The fixed effects estimates are therefore precision weighted by the random effects (Raudenbush & Bryk, 2002, pp.38-45; Diggle, Heagerty, Liang, and Zeger, 2002, pp.64-65). The ML estimator in this case is the generalized least squares estimator:

$$\hat{\beta}(V_0) = \left(X'V^{-1}X\right)^{-1} X'V^{-1}y \tag{4.13}$$

where V is a block-diagonal matrix with common non-zero blocks V_0 and the non-zero elements are proportional to the intra- level two correlation. In the longitudinal case with time nested within individual this is $Corr(y_{it}, y_{it'}) = \rho$. The V_0 blocks with four time points , for example, take the form

$$\begin{bmatrix} \sigma^2_{(i)1} & \sigma_{(i)12} & \sigma_{(i)13} & \sigma_{(i)14} \\ \sigma_{(i)21} & \sigma^2_{(i)2} & \sigma_{(i)23} & \sigma_{(i)24} \\ \sigma_{(i)31} & \sigma^2_{(i)32} & \sigma^2_{(i)3} & \sigma_{(i)34} \\ \sigma_{(i)41} & \sigma_{(i)42} & \sigma_{(i)43} & \sigma^2_{(i)4} \end{bmatrix} \tag{4.14}$$

The diagonal elements within blocks are often assumed equal to each other and a uniform correlation structure assumes that every observation within an individual is equally correlated with every other observation from that individual, $\sigma_{(i)tt'} = \sigma_{(i)ss'}$. The off diagonal elements are the intraclass correlation (ICC), ρ. Therefore, the V_0 blocks have the following correlation elements

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} \tag{4.15}$$

Other V_0 may be specified with efficiency gains for correctly specified structures. For example, for longitudinal data it may be more appropriate to specify an autoregressive

structure that includes ρ, ρ^2, and ρ^3 for the correlation between repeated measures that are lag 1, 2 and 3, respectively. In the mixed model, the default V_0 structure is uniform.

In the linear growth model the $Corr\left(y_{it}, y_{it'}\right) = \rho$ is a function of the random intercept, random slope, and covariance between the random intercepts and slope (Diggle, Heagerty, Liang, and Zeger, 2002, p.133). The ML random effects estimates are estimated by plugging in the values from equation (4.13) into the likelihood equation and solving for the random effects in G. Note that the ρ differ across individuals (level two) only in the unbalanced situation when there is no time-specific nonresponse. With no missing time points, all V_0 have equal $\hat{\rho}$ resulting in the standard least squares estimator.

The default estimation in the marginal model does not include precision weighting using ρ. Instead an independent correlation weight matrix is assumed, that is, the V_0 are identity matrices. The independence correlation matrix assumes no correlation between time points within individual so that off diagonal elements are 0. Estimation in the marginal model with correlation weight matrix is called generalized estimation equations (GEE) (Liang & Zeger, 1986; Zeger & Liang, 1986; and Prentice, 1988). GEE with independent correlation weight matrix is also the standard least squares estimator.[1] However, the marginal model may be estimated using a uniform correlation structure[2], which uses the intra individual correlation as a weight for the fixed effects in the same manner as the mixed model where the uniform correlation weight matrix is given in 4.15. As a result, the estimator for the marginal model parameters is also equation (4.13).

When there are no missing data due to time-specific nonresponse, the estimates from the marginal model with any specified correlation weight matrix structure and the mixed model are identical because there are no differences in precision for the estimates across individuals (level two observations) and the covariance weighting becomes equal across

[1]The correlation weight matrix is sometimes refered to as the working correlation matrix in the GEE literature.

[2]The uniform correlation weight matrix is sometimes refered to as an exchangeable matrix due to the exchangeability of the off-diagonal elements in the matrix.

individuals. However, with time-specific nonresponse, estimates from the mixed and marginal model with uniform correlation matrix may differ due to differences in the $\hat{\rho}$ (see, for example, the discussion in Skinner & Vieira, 2007).

The importance of the GLS estimator with covariance weighting for the missing data problem is that it changes the MAR assumption of the growth curve model. Data are missing at random (MAR) if data are missing randomly as a function of the model covariates *including* the repeated outcomes from observed time points. With correctly specified model and covariance weight matrix, the missing data will be MAR and estimates will be consistent. In this situation, probability weights are not necessary for correcting wave-specific nonresponse. However, if the wave-specific nonresponse is not MAR or the correlation weight matrix is miss-specified, probability weights are necessary for adjusting level one missing. The estimator in (4.13) will not result in consistent estimates when missing repeated measures are missing not at random (MNAR) or "nonignorable". A nonignorable missing mechanism occurs when time-specific missing values are related to the missing values of the outcome at those same time points after controlling for other observed variables (i.e., data are not MAR). When wave-specific missing data are MCAR, the generalized least squares estimator for the linear growth model given in equation (4.13) is consistent for any correlation specification and is fully efficient with the correctly specified correlation weight matrix (Liang & Zeger, 1986).

4.2.2 Estimation with Probability (Sampling) Weights

Probability weights are developed to correct for biases due to unequal inclusion of observations into the data resulting from selection or nonresponse. Probability weights are inversely proportional to the inclusion probability of a sample observation. Thus, the weight for observation i is

$$w_i = \frac{1}{\pi_i} \tag{4.16}$$

where π_i is the selection probability for observation i. Weights estimate the number of population elements, N, that are represented by each sample observation. These weights are usually a combination of a base weight that is a function of sampling design variables as well as correction factors, which are functions of auxiliary variables known for each observation (Biemer & Christ, 2007; Lohr, 1999). The correction factors are therefore based on some type of model where the nonparticipation or frame noncoverage is assumed to be MAR given the auxiliary information. The next section discusses weighting in more detail.

For the linear growth model, the weighted ML estimator for fixed effects in the mixed model and the weighted GEE estimator is

$$\hat{\beta}_w = \left(WX'V^{-1}X\right)^{-1} WXV^{-1}y \tag{4.17}$$

where W is the a diagonal matrix of probability weights for each person-by-time observation and V is the block-diagonal correlation matrix with blocks V_0. The variance estimator with probability weights is the sandwich estimator (Binder, 1983; Skinner, 1989, p.78)

$$var\left(\hat{\beta}_w\right) = \left[\sum_{i\in s} w_{it}x'_{it}V^{-1}x_{it}\right]^{-1} \left(\frac{n}{n-1}\left(\sum_i \left(\theta_i - \bar{\theta}\right)\left(\theta_i - \bar{\theta}\right)'\right)\right) \left[\sum_{i\in s} w_{it}x'_{it}V^{-1}x_{it}\right]^{-1} \tag{4.18}$$

where i is the cluster (individual in a growth model), and n is the number of clusters in the sample, θ_i is the $\hat{\beta}_{wi}$ for observation i and $\bar{\theta}$ is the average of the θ_i over the sample.

4.3 Weighting for Selection and Nonresponse in Longitudinal Panel Surveys

Sample designs that include unequal selection of individuals into the sample are common for panel surveys. This often results from the need to oversample particular subpopulations of interest to obtain adequate sample size or due to a multiple stage selection designs. Individuals may enter the study at differential rates also due to sampling

frame deficiencies and nonparticipation of the respondent. Unequal selection of individuals that are a result of the sample design itself provide known probabilities of inclusion for a given respondent. Inclusion probabilities that result from sample frame problems and nonresponse are not known with certainty and must be modeled. The unequal selection of individuals may be MNAR or nonignorable for many models and outcomes, therefore, sampling statisticians provide probability (sampling) weights, which are applied in estimation to counteract bias (Cochran,1977; Kish, 1965; Lohr, 1999). Weights due to the sampling design are "base weights" and the modeled weights for nonparticipation are generally applied as adjustments to the base weights (Biemer & Christ, 2007; Biemer & Christ, 2008). The adjusted weight for the probability of inclusion of observation i into the study will be denoted w_i.

Once an individual enters a panel study, there is also the potential that the individual is not observed in one or more subsequent waves of data collect. This is generally due to nonresponse for which the probabilities of inclusion are not known with certainty and must be modeled. Sampling statisticians generally create a new weight for each wave to account for changes in the sample due to wave-specific nonresponse. Weights that allow for proper inference to the cohort population at a given wave are created such that the initial weight used for inclusion into the sample at the study initiation, w_i is adjusted for the wave-specific nonresponse. The nonresponse adjustment to the weight for entry into the study at time t is

$$\lambda_{it} = \frac{1}{\pi_{it}|\pi_i} \tag{4.19}$$

where $\pi_{it}|\pi_i$ is the conditional probability that observation i responds at time point t given their selection into the sample. These adjustments are multiplied by the weights for inclusion into the study to obtain the time-specific, cross-sectional weight for observation i at time point t:

$$w_{it} = w_i\lambda_{it} = \frac{1}{\pi_i} \cdot \frac{1}{\pi_{it}|\pi_i} \tag{4.20}$$

where w_i is the probability of inclusion into the study due to selection, frame deficiencies,

and nonresponse and where $\lambda_{it} = 1$ when $t = 1$.

Weights for longitudinal analyses are also typically provided with panel survey data. These "panel weights" are used with the sample of individuals who have complete data for all data collection waves included in a particular analysis. The weight is therefore the inverse of the probability that the individual was observed at every time point included in the longitudinal (repeated measures) analysis. Assuming that nonresponse at time point t is independent of nonresponse at time point $t+1$ conditioned on the model generating the nonresponse probabilities, the panel weight for observation i is

$$w_i^p = \frac{1}{\pi_i} \cdot \frac{1}{\pi_{i1}|\pi_i} \cdot \frac{1}{\pi_{i2}|\pi_i} ... \frac{1}{\pi_{it}|\pi_i} \tag{4.21}$$

Any combination of waves of data may be included in the panel weight and need not be consecutive. For example, if one were analyzing observations from waves 1, 3, and 5, the w_i would be multiplied by λ_{i1}, λ_{i3}, and λ_{i5}.

The weights outlined in (4.20) and (4.21) may be used in growth curve models using the estimator in (4.17) several ways: 1) The analyst may apply w_i to individual i at each and every observed time point for individual i used in the model. In this application, the weight is time invariant, 2) The w_{it} may be applied to individual i at time t for every observed time point for individual i used in the model. In this application, the weight is time varying, and 3) The panel weight, w_i^p, may be applied to individual i at every time point used in the model where only individuals observed at every time point are included in the model. These three approaches to weighting: time invariant weighting, time varying weighting, and panel weighting, may be used in a mixed or marginal modeling framework. A fourth approach that may be used in the mixed modeling framework is multilevel weighting (Pfeffermann, et al., 1998; Skinner & Holmes, 2003). In this approach the λ_{it} are applied to sums or integrals over time points and the w_i are applied to sums or integrals over the individuals in the algorithm used to estimate the parameters in (4.17). Whereas for weighting approaches 1) - 3), the weights would be applied to summing over individuals. Additionally, the λ_{it} are scaled to sum to the actual within

person sample size in this method. Table 4.1 summarizes the weighting / modeling approaches that will be evaluated in this chapter.

It is unknown which of the proposed weighting methods would provide the optimal estimates in terms of consistency and efficiency given that the weights are correct. Panel weighting is probably the most common approach to weighting in longitudinal analysis because most publicly available longitudinal survey data that includes sampling weights also include and recommend the use of panel weights in longitudinal analysis. However, this is not a very efficient use of data since it uses fewer individual-by-time observations. The time invariant weighting approach may be naively applied in many situations. This weighting should correct for bias due to unequal inclusion of individuals into the study, but would not necessarily correct for bias due to time-specific, intermittent nonresponse. The time varying weighting approach has not been explicitly addressed in the literature, but should be a method that corrects for bias due to time-specific unequal inclusion. Additionally, the performance of probability weighting methods likely differ depending on whether precision weighting is used with the GLS estimator with weight matrix V or whether the traditional marginal modeling approach of no precision weighting using an independence assumption is used. These three weighting methods in combination with the two estimation approaches can also be compared to the multilevel modeling case, which is specifically designed for probability weighting in the case where GLS with weight matrix V is used.

The weights described here are specific to cohort panel data where inference is to the cohort who entered the study at initiation and does not include inference to a dynamic population that is changing over time with the exception of changes due to deaths. Other longitudinal data intended to observe a dynamic population, such as the Current Population Survey (CPS), which includes resampling or the Panel Study of Income Dynamics (PSID), which includes sample refreshing are not directly evaluated in this chapter. With cohort panel studies, unequal selection of observations may occur at the first time point

due to sample design, non-participation, and other sampling deficiencies while unequal inclusion of observations at follow up waves is solely due to intermittent nonresponse or attrition. Therefore, the weights described in this chapter would not apply to a sample that is refreshed with selection of additional observations at a later wave of data collection. Examples of survey data with cross-sectional and panel weights like those outlined in this section include the National Longitudinal study of Youth (NLSY79 and NLSY97), ADD Health, and the Early Childhood Longitudinal Study (ECLS).

It is also important to note that probability weights provided to analysts using panel survey data are designed to correct for unit nonresponse and do not take into consideration additional missing data due to nonresponse to particular items on the survey. Therefore, individuals who are observed at a given time point, but are not observed on every analysis variable at that time point would not be accommodated by the probability weights if that individual-by-time observation is not included in the analysis. This is not to say that weights for this type of situation may not be developed, but that the sampling weights made available for an entire data set do not correct for item missing. Other methods are available for dealing with item missingness including multiple imputation (Rubin, 1987; Schafer, 1997) and direct maximum likelihood (Arbuckle, 1996; Schafer & Graham, 2002).

4.4 Simulation

A simulation analysis was undertaken in order to evaluate the various weighting methods outlined in Table 4.1. A finite population with 1 million level two observations and 10 level one observations (time points) per level two observation was generated using random effects generating models. Model values were chosen to be somewhat comparable to the model results from one of the empirical examples from the NSCAW application undertaken in the next section. Two outcomes were generated. One outcome is a function of an unconditional model with fixed and random intercept and slope following equation (4.4). The other outcome is a function of a conditional model with fixed and random

intercepts, fixed and random slopes, fixed level one and level two covariates, and fixed interactions of the covariates with time following equation (4.10). The unconditional generating models is

$$y_{it} = \mu_\alpha + \mu_\beta T + \delta_{\alpha i} + \delta_{\beta i} T + \varepsilon_{it} \tag{4.22}$$

Where T represents time and $t = 0, 1, 2, ..., 10$, $\mu_\alpha = 20$, $\mu_\beta = 2$, $var(\delta_{\alpha i}) = 70$, $var(\delta_{\beta i}) = 1$, $var(\varepsilon_{it}) = 12.96$, and $cov(\delta_{\alpha i}, \delta_{\beta i}) = -5$. The conditional generating model is

$$y_{it} = \mu_\alpha + \gamma_{\alpha 1} X_{it} + \gamma_{\alpha 2} W_i + \mu_\beta T + \gamma_{\beta 1} X_{it} T + \gamma_{\beta 2} W_i T + \delta_{\alpha i} + \delta_{\beta i} T + \varepsilon_{it} \tag{4.23}$$

Where T represents time and $t = 0, 1, 2, ..., 10$, X_{it} is a time varying covariate $\sim N(0, 1)$, W_i is a time invariant covariate $\sim N(0, 1)$, $\mu_\alpha = 5$, $\mu_\beta = 4$, $\gamma_{\alpha 1} = 2$, $\gamma_{\alpha 2} = 1$, $\gamma_{\beta 1} = -0.3$, $\gamma_{\beta 2} = 0.5$, $var(\delta_{\alpha i}) = 16$, $var(\delta_{\beta i}) = 2$, $var(\varepsilon_{it}) = 12.25$, and $cov(\delta_{\alpha i}, \delta_{\beta i}) = 1$. Random effects are orthogonal to the fixed effects in both models. The fixed and random effects parameter values used in equations (4.22 and (4.23) mimic the values from the Kaufman Brief Intelligence Test (K-BIT) Matrices scale example used later in this chapter (Section 12). The K-BIT Matrices scale measures the ability to perceive relationships and complete analogies. The parameter values used for the generated variables, X and W, mimic the values for child age and other race, respectively. The y_{ij} in these models are quasi-continuous scale score assessments that are sums of multiple items. The K-BIT scale is normally distributed in this example and ranges from 0 to 50 with a mean and median of 26.

4.4.1 Simulated Sample Selection and Simulation Weights:

Samples were selected from the generated finite population with unequal probability. First, level two observations (individuals) were selected as a function of the random slopes where selection is proportional to $\exp(\delta_{\beta i})$. This selection mechanism results in overselection of observations with larger slope values. It also affects the intercept values since these are correlated with the slopes. Five hundred replicates with three different

level two sample sizes were selected. The level two sample sizes are 50, 200, and 500. Next, from each of the replicate level two samples, level one observations (time points) were selected as a function of the random error terms. Only the first 5 time points are included in the level one selection of samples to be more comparable with many panel studies and the empirical example provided in this paper. Errors at each time point are stratified into two strata where for stratum one $\varepsilon_{it} > 0$ and for stratum two $\varepsilon_{it} \leqq 0$. Level one observations (time points) are selected using stratified sampling in the proportions presented in Table 4.2.

This selection mechanism results in overselection of level one observations that have values greater than the average value as given by the population intercept and slope. The selection results in nonignorable (MNAR) missing values at level one since observations are dependent on the missing values of y_{it}. The majority (two-thirds) of cases have at least four time points and around one-third have three or fewer time points. The particular selection design used in this simulation is designed to affect the average intercept and slope values, $E(\alpha_i)$ and $E(\beta_i)$ where the average slope value should be inflated for both models and the average intercept value should be deflated in the unconditional model and inflated in the conditional model due to the negative and positive covariances between the random effects. Random intercepts and slopes may also be affected by such a design, but are not necessarily affected. The random variation around the newly biased fixed estimates may remain the same as the random variation in the population for the true parameters. This is indeed the case for most of this simulation data as we shall observe in the results. The covariate effects should not be affected by the selection design because these effects are homogenous across the different levels of the slope and level one error values.

Unequal selection of observations was done in two stages in the SAS procedure SURVEYSELECT. Weights for the individuals selected into the samples and the conditional weights for the repeated measures selected into the samples are calculated automatically

by the SURVEYSELECT procedure as the inverse of the probability of selection of each observation. Time invariant weights are equal to the inverse of the probability of selection for level two observations, $w_i = \frac{1}{\pi_i}$ where π_i is proportional to $\exp(\delta_{\beta i})$. These weights correct for unequal inclusion of individuals into the sample. The conditional weights for the selection of repeated measures is the inverse of the conditional probability of selection of level one observations given selection of the level two observation, $\lambda_{it} = \frac{1}{\pi_{it}|\pi_i}$ where $\pi_{it}|\pi_i$ is the conditional probability of level one observation at time t given level two observation i is selected into the sample and is proportional to values presented in Table 4.2.

Time varying weights were calculated as the inverse of the product of the probability of selection of level two observations and the probability of selection of level one observations. Specifically, the probability of inclusion into the sample is multiplied by the conditional probability of inclusion for each time point separately. Time varying weights are therefore:

$$\omega_{it} = \frac{1}{\pi_i} \cdot \frac{1}{\pi_{it}|\pi_i}$$

where w_{it} is the inverse of the unconditional probability of selection of level one observation at time t. The w_{it} are the traditional single level weights for cross-sectional analysis at each time point separately. Multilevel weights are equal to w_i for level two and $w_t = \frac{1}{\pi_{it}|\pi_i}$ for level one. Panel weights, w_i^p, are derived as the inverse of the product of the level two probability of inclusion and the conditional probability of inclusion for every level one time point because selection probabilities are independent at each time point, i.e.,

$$w_i^p = \frac{1}{\pi_i} \cdot \frac{1}{\pi_{i0}|\pi_i} \cdot \frac{1}{\pi_{i1}|\pi_i} \cdot \frac{1}{\pi_{i2}|\pi_i} \cdot \frac{1}{\pi_{i3}|\pi_i} \cdot \frac{1}{\pi_{i4}|\pi_i}$$

The degree of unequal selection may be evaluated using the unequal weighting effect (UWE), which measures the amount of noise added to estimates. The UWE for a mean estimate is equal to $1 + cv_w^2 = 1 + \frac{var(w)}{(\bar{w})^2}$ where the standard error of the mean is increased

by a factor of $\sqrt{\text{UWE}}$. Table 4.3 presents the $\sqrt{\text{UWE}}$ for each of the weights used by generating model and level two sample size. From the table it can be seen that the panel weights have the most variation relative to the other weights. The time varying weights have the second greatest variability as these combine unequal selection from both levels one and two. Finally, the selection of level two observations as indicated by $\sqrt{\text{UWE}}$ for the time invariant weights is more variable than the selection of level one observations conditional on the level two selection.

4.4.2 Evaluation:

First, the mixed and marginal models with both independent and uniform correlation weight matrix are estimated for the data with unequal selection at level two (individuals), but with complete and balanced data at level one (time) with all 10 time points generated by the population models. In this situation, only weighting of level two observations using the time invariant weight, w_i, is necessary for correcting estimates. This was done to show the equivalence of the estimation methods for this situation as well as to determine the degree of bias due to nonignorable unequal selection of the individuals.

Following this, the models are estimated for each final replicate samples that suffer from intermittent nonresponse using the eleven methods outlined in Table 4.1. Within the eleven methods there are two model types, 1) mixed model and 2) marginal model. There are two estimation types within the marginal model using 1) an independent correlation weight matrix or 2) an uniform correlation weight matrix. Finally, there are five weighting approaches using 1) the time invariant weight, 2) the panel weight, 3) the time varying weight, 4) the multilevel weights, and 5) using no weight. Note that the specified correlation weight matrix does not affect results for the panel weight analysis since observations that are missing a time point are excluded from the analysis and the data is therefore balanced (i.e., complete case analysis). The multilevel weights only apply to the mixed model. Finally, the results for the unweighted marginal models with uniform correlation matrix match very closely with the other unweighted models,

especially the results for the unweighted mixed model, therefore those results are not presented here. The multilev and surveyglim procedures in LISREL v.8.8 software are used to estimate the mixed and marginal models, respectively. SUDAAN v.9.0 is used to estimate the marginal models with uniform correlation matrix.

Quality of the weighting methods was evaluated according to the bias, efficiency, and coverage of fixed and random effects parameters. Degree of bias is estimated by the difference in the finite population parameter values given for equations (4.22) and (4.23) the expected value of the parameter estimates:

$$E\left(\hat{\theta}_r\right) = \frac{\sum_{r=1}^{500} \hat{\theta}_r}{500}$$

where $\hat{\theta}_r$ is the parameter estimate for replicate r. Efficiency is judged by the standard deviation of the $\hat{\theta}_r$ across the 500 replicates. The expected values of the standard error estimates, $E\left(se_r\hat{\theta}_r\right)$, were also calculated where $\left(se_r\hat{\theta}_r\right)$ is the standard error for estimate $\hat{\theta}_r$ in replicate r. These were compared to the standard deviation of the $\hat{\theta}_r$. Finally, coverage rates were calculated as the proportion of finite population parameters that fall within the 95% confidence region for each sample. The 95% confidence region for each replicate was calculated using

$$\hat{\theta}_r \pm 1.96\left(se_r\hat{\theta}_r\right). \tag{4.24}$$

4.5 Results

4.5.1 Balanced Data with 10 Time Points

Tables 4.4 - 4.6 presents the weighted and unweighted results for the unconditional model. Each of the three models give identical results for both the weighted and unweighted fixed effects estimates, however the standard error estimates differ marginally when level two sample sizes are 50 (Table 4.4). The standard error estimates in the weighted case are generally underestimating variability as measured by the $std_r\left(\hat{\theta}_r\right)$, but improve with increases in level two sample size. The opposite is true for the unweighted

standard error estimates where the larger sample sizes result in underestimation of the standard deviation of the estimates.

The unweighted and weighted results reveal that the intercept and slope parameters are markedly biased due to the unequal selection of level two observations. The unweighted intercept is estimated around 15 compared to the parameter value of 20 and the unweighted slope is estimated around 3 compared to the parameter value of 2. The use of time invariant weights in this case results in essentially unbiased estimates of the population parameter values for the fixed intercept and slope. The weighted and unweighted random effects estimates presented in Table 4.5 show that the selection mechanism used here does not affect the random effects parameter estimates because the unweighted estimates are unbiased. The weighted estimates for the larger sample sizes are also converging on the population parameter values. Comparison of the standard deviations of the weighted random effects estimates to the unweighted random effects estimates reveals that the weighted estimates are much more volatile due to the weights and hence require the larger sample size for asymptotic properties to be fulfilled. Finally, coverage rates given in Table 4.6 reveal that there is zero coverage for the unweighted fixed effects estimates However, coverage rates for the generally unbiased random effects estimates are much superior in the unweighted analysis. Coverage rates for the weighted random intercepts and random slopes for a sample size of 500 approaches the coverage rates for the comparable unweighted estimates with a sample size of 50 indicating that the UWE for random effects parameters is very large and far greater than the UWE for a mean as presented in Table 4.3 for the time invariant weight.

Results for the conditional model with complete and balanced level one data with 10 data points per level two observation using the time invariant weights are presented in Tables 4.7 through 4.9. For the larger conditional growth model, larger sample sizes are required for convergence of the weighted fixed intercept and slope estimates on the parameter values (Table 4.7). The unweighted estimates reveal that the selection of level

two observations results in intercept estimates close to 6 compared to the population parameter value of 5. As well, the expected value of the unweighted fixed slope estimates are nearly two points higher than the parameter value of 4. Estimates for the effects of covariates (Table 4.7) are good for both the unweighted and weighted analysis since the selection design did not affect these relationships. Again, the mixed model and marginal models with independent and uniform correlation matrix produce the same results within sample size and weighting method indicating that unequal selection of level two observations only may be addressed equally in the various estimation types. Standard error estimates are again lower than the standard deviations for the fixed effects estimates (Table 4.8. - covariate effects exhibit the same pattern but are not shown in this table). This is true for all of the fixed effects in the weighted analysis, but the estimates improve with level two sample size. The standard error estimates for the unweighted fixed effects are unbiased except for the unweighted slope estimates, which are themselves biased. Random effects results for the conditional model are similar to those of the unconditional model (Table 4.9), though the unweighted analysis reveals that there may be a slight underestimation of the random slopes indicating that unequal selection may have decreased the variance of the slopes. In general, the weighted estimates require much larger sample sizes for convergence on the random effects population parameters. Coverage rates (Table 4.10) for the weighted estimates in the conditional model are poorer for biased parameters such as the fixed slope than the coverage rates for similar parameters in the unconditional model. For example, the fixed slope estimate coverage reaches only 0.78 with a level two sample of size of 500. As expected, the coverage rates for the biased parameters, the fixed intercept and slope, are very poor for the unweighted estimates. However, the coverage for unbiased fixed and random effects is better in the unweighted analysis than the weighted analysis. The coverage rate for the slightly biased random slope of the unweighted analysis gets worse with increased sample size probably due to smaller standard error estimates.

Results for the mixed and marginal models for the case where there is unequal selection of level two observations but no unequal selection of time points, reveal that using the time invariant weight corrects for biases due to the selection design. This is true for both the mixed and marginal models with either correlation matrix, all of which give very similar results within weighting approach and level two sample size. The weighted analysis requires fairly large samples sizes to obtain decent coverage rates, particularly for random effects estimates measured in a mixed model. Next, we turn to analysis of data with both unequal inclusion of level two and level one observations. In this situation, there are many more weighting options and the type of model and estimation techniques have more variable results.

4.5.2 Unbalanced Data with 5 Time Points

The eleven weighting and modeling combinations outlined in Section 1.2 are applied to the data with unequal selection of level two and level one observation per the selection mechanisms outlined above. Only five time points are included in this evaluation. Consider first the unconditional model. Unequal selection of level two observations resulted in an expected value of 15 and 3 for the unweighted estimates of the fixed intercept and slope (Table 4.4 above). Looking at the unweighted estimates in Table 4.11, the unweighted intercept estimates are about the same because very little unequal selection occurred at the first time point; however, the slopes are inflated even further by the overselection of positive level one error terms. The unweighted fixed slope estimates are now approximately 3.2 versus the population parameter of 2. For unweighted method 10 and 11, the mixed and marginal models produce identical results. The method of applying the time invariant weight also produces the same results across the three modeling and estimation methods (method 1, 2, and 3). These results reveal that the bias due to level two unequal inclusion is corrected, however the 0.2 increase in the slope due to unequal inclusion of time points remains. The traditional panel weighting methods (4 and 5) perform the same across the model types because data are balanced and weights

are applied at level two. This method performs well with larger complete data sample sizes. For example, after listwise deletion of level two observations with missing time points, a sample size of approximately 137 (averaged across the replicates) give decent results using the panel weights. Even with sample sizes as small as approximately 55, the panel weighting method performs better than all but one other method. Using the time varying weight in combination with an independence correlation weight matrix in a marginal model seems to perform the best of all the methods for estimation of fixed effects. Even with level two sample sizes as small as 50, the expected value of estimates are fairly unbiased. Performance is great with level two sample sizes of 200 and 500. The benefit of the marginal method is that it does not require large level one sample sizes like the multilevel modeling method does with level one weighting. Larger level one sample sizes are often not available for trajectory type analysis using panel data. Using the time varying weight in the mixed model performs about as well as using the time invariant weight in the marginal models because it weights at level two only. Finally, there remains some bias in the marginal models with uniform correlation matrix and time varying weights. This may be due to the correlation matrix weighting counteracting the sampling weights. Standard error estimates are underestimated for the sample size with 50 level two observations.

Table 4.12 gives the results for random effects estimates for the unconditional model. Once again, the unweighted models give the best estimates of the random effects because these were not affected by the sample design in the unconditional model. Panel weights perform poorly since level two sample sizes are much smaller after listwise deletion, however, with approximately 137 level two observations the random effects estimates are converging on the population parameter values. Coverage rates (Table 4.13) corroborate the other tables and show that the marginal model method with time varying weights and an independent correlation matrix give the best coverage for the fixed intercept and slope estimates. Panel weighting comes in second when listwise sample sizes are larger,

e.g., 137. However, coverage rates for the random effects are worst for the panel weighting due to the smaller level two sample size. Other weighting methods in the mixed model do not differ much in terms of coverage performance. Coverage for the level one random error is worse than the other random effects estimates.

Similar results are revealed for the conditional model (Tables 4.14 - 4.17) in terms of estimation of the fixed intercept and slope. However, larger sample sizes are required for convergence on the population parameter values. For example, the marginal method with time varying weights and an independence correlation matrix is still the best performing method relative to the others, yet it converges more slowly on the population values in the conditional model relative to the unconditional model. It is unclear why method 9 does not perform better given previous work on this method. It may require larger level one sample sizes as has been indicated in Pfeffermann, et al. (1998). Using the time invariant weight is again not a good method since the bias due to time-specific missing values is uncorrected. The panel weighting method performs alright as well, but with the conditional model approximate sizes of 137 are not large enough. The marginal model with uniform correlation matrix, method 8, underestimates the intercept and again does not perform as well as the marginal model with assumed independence between time points. In the nonignorable missing data case, using the correlation information does not improve estimates. Covariate effects are essentially unbiased for all methods since they are not affected by the selection mechanism. Random effects estimates (Table 4.16) are not estimated well for any method in the smaller sample sizes.

4.6 NSCAW Empirical Example

4.6.1 Methods

The various weighting methods were evaluated in an empirical example using data from the National Survey of Child and Adolescent Well-Being (NSCAW). The NSCAW is a panel survey of children in the child welfare system in the United States. The target

population of the NSCAW Child Protection Services (CPS) sample is "all children in the U.S. who are subjects of child abuse or neglect investigations (or assessments) conducted by CPS and who live in states not requiring agency first contact." (Dowd, et al., 2006). The NSCAW sample design is a complex design that includes stratification, clustering, and unequal selection probabilities. The NSCAW Child Protection Services (CPS) cohort includes 5,501 children, ages birth to 14 (at the time of sampling), who had contact with the child welfare system within a fifteen-month period which began in October, 1999. Face-to-face interviews were administered at three points in time: Wave 1, 18 months post-Wave 1, and 36 months post-Wave 1.

Linear growth models were fit to three NSCAW measures observed for three waves of data collection that occurred at 18 month intervals. The first measure is a depression scale, the Children's Depression Inventory (CDI), which measures depression by asking various questions of children about their engagement in certain activities or their experience of certain feelings (Kovacs, 1992). The second and third measures are the Math and Verbal scores from the Kaufman Brief Intelligence Test (K-BIT), which is a brief, individually administered measure of verbal and nonverbal intelligence for children, adolescents, and adults (Kaufman & Kaufman, 1990). An unconditional and a conditional linear growth model with three time points was analyzed for continuous outcomes. In the conditional growth model, intercepts and slopes are conditioned upon child age, child sex, and child race/ethnicity. There are no item missing for these independent variables.

The NSCAW wave one weight was used for time invariant weighting, this weight is the inverse of the probability of selection for an individual child at wave 1 and represents the probability of entering the study. The cross-sectional weights for waves one, two, and three were used for time varying weighting, these weights are the inverse of the probability of inclusion at waves 1, 2, and 3 separately where the weight at waves 2 and 3 are equal to the wave 1 weight with additional time specific, nonresponse adjustments. NSCAW panel weights, which were developed for use with children observed at all three

waves were used for panel weighting. Multilevel weights were derived using the wave one weight at level two and the wave two and wave three weights divided by the wave one weight at level one.

The NSCAW data was selected with a complex sample design that included clusters that are comparable to counties and stratification where strata are larger geographic regions. For the purposes of this study, the clustering at the PSU (county) level and the stratification are ignored. The data are treated as if the only nesting is at the level of the individual where time is nested within individual. This was done to avoid confounding of the way that the different methods may treat additional clustering and stratification. If county level clustering were accounted for, the standard error estimates would likely be larger because the between county variance would be used rather than the between child variance. The regional strata would also affect the standard error estimates because variances would be calculated within stratum and subsequently aggregated across strata.

Table 4.18 presents the missing data patterns for the outcomes. For each measure, around 55% have complete data across all waves. The NSCAW sampling weights only correct for that portion of the missing data that is a result of unit missing due to non-participation at the particular wave of data collection. Therefore, it is important to look at why the data are missing. All missing data from wave one are a result of item missing because all children were observed on some measures at that wave. About 25% of the sample is missing for each of the scales at wave one. At waves two and three, some of the missing is due to nonparticipation of the child at that wave and another portion of the missing is due to children who were not observed for the specific scale. At wave two, around 10% of the sample are missing units and around 15-16% are missing items. At wave three, most (around 11% of the sample) are unit missing and approximately 4% are item missing. The majority of item missing could be considered MCAR because children were not observed on the scale due to their age. For all three scales, very young children were not assessed. Therefore, as the children aged between waves, they were

later assessed on the scales. For each of the three scales, the number of respondents who were observed at least once on the measure is large $\geqq 2914$. The unequal weighting effect for the estimate of a mean is 2.9 for the wave 1 weight, 3.0 for the time varying weight, and 3.2 for the panel weight. This indicates that the standard error estimate for a mean would increase by a factor of approximately 1.7 - 1.8 due to the variance in the probability weights. The weights are not that informative for these measures where the correlation between the outcomes and the various types of weights ranges from an absolute value of 0.01 to 0.03. This low degree of informativeness means that the unequal inclusion is nearly MCAR for these outcomes.

The multilev and surveyglim procedures in LISREL v.8.8 software are used to estimate the mixed and marginal models, respectively. SUDAAN v.9.0 is used to estimate the marginal models with uniform correlation matrix. LISREL multilev results are comparable with results from Mplus v5.0. For the marginal models with independent correlation matrix, the SUDAAN and LISREL surveyglim results are identical.

4.6.2 Results

Tables 4.19 through 4.33 present the results for the growth models for the three NSCAW outcomes. Results for the eleven methods listed in Table 4.1 are evaluated, however, the results are grouped by model type. Tables 4.19 through 4.21 present the fixed effects estimates from the unconditional models. There are little differences in estimates for the CDI across models and weighting methods. Most differences appear with the KBIT measures. Weighting procedures used with the mixed model and all available observations produce similar results with the exception of some differences for the unweighted results. Only the results for panel weighting, which utilize a different sample, differs markedly from the other results. The fact that not weighting at all provides similar, though more efficient, estimates indicates that the level of bias in fixed effect estimates due to unequal selection is rather low for these outcomes.

Marginal model methods that use an independent correlation weight matrix have

similar results for the time invariant and time varying weighting methods. There are more differences in results between weighted analyses and unweighted analyses for the marginal models. The panel weighting method produces quite different results as compared to the other weighting methods for the marginal models where the panel weighting results are sometimes closer in size to the results from mixed models, for example, the KBIT slopes. Marginal models with uniform correlation weight matrix give results that are more similar to the mixed model results for the KBIT measures. The fact that the sampling weights are only accounting for that part of the missing due to child nonparticipation at a given wave and not very informative, as well as the fact that there are so few time points likely results in the greater impact of the correlation weight matrix versus the sampling weights. It is likely that the use of the correlation matrix weighting is increasing the efficiency of these estimates by utilizing the association between repeated measures over time. The fixed effects estimates and their standard errors for the mixed and marginal models are the same when using the sample observed at all waves with panel weights. This is a situation where the data are perfectly balanced with three observations per person and no missing data and was corroborated in the simulation analysis.

Tables 4.22 through 4.24 present random effects results for the various mixed models. Random effects estimates for the mixed models differ mostly for the unweighted and panel weighted methods. It is difficult to determine the reason for these differences. The panel weighting method is certainly less efficient due to the smaller sample sizes, but also may results in biased random effects for some of the outcomes.

Fixed intercepts and slopes for the conditional growth models seem to be more affected by weighting methods (Tables 4.25 through 4.27). For example, ignoring the weights results in much different estimates as compared to methods utilizing weights. And, applying time varying weights produces some differences for the mixed models. For example, the CDI fixed slope and the KBIT Verbal fixed intercept. Panel weighting also stands out as having very different estimates compared to the other methods.

The marginal models with independent correlation weight matrix have estimates that are sometimes very different from the mixed models and the marginal models with uniform correlation weight matrix again tend to be closer to the mixed models.

Covariate effects on intercepts and slopes reveal some similar patterns (Tables 4.28 through 4.30 and Tables 4.31 through 4.33). With the exception again of panel weighting and not weighting, the other weighting methods within the mixed modeling framework produce similar results. On occasion, the time varying weights have differences. For example, some of the race effects on the intercepts differ for the CDI outcome. Marginal models with independent correlation weight matrix have covariate results that differ from the mixed model for many of the effects. Again, as expected the marginal models with uniform correlation weight matrix produce results that are often more in line with the mixed models than the comparable marginal model with independent weight matrix. Time varying weighting and using time invariant weights also show different results within the marginal models. Although not presented here, the results for panel weighting with the marginal model and uniform correlation weight matrix gives identical estimates and standard errors to the other models that use panel weighting. Random effects estimates for the conditional models are presented in Tables 4.34 through 4.36. With the exception of panel weighting and not weighting the other methods produce comparable random effects estimates.

It is difficult to discern which estimation methods are best for this empirical example. The best guess is based on what is known about the missing data. In this example, it is likely that the methods that utilize the correlation of the observations over time in the estimation process (the mixed model and the marginal model with uniform correlation matrix) performs better since much of the wave-specific missing data can be considered MCAR and it is plausible that the rest is MAR conditioned on observed outcomes for individuals at other waves. Efficiency of estimates should be improved using this correlation especially given the small number of time points. Bias ought to also be decreased

since information from the other waves is utilized. The differences for some of the slopes in the two estimation methods is quite marked indicating the potential benefits to using correlation information in the estimation process.

4.7 Conclusions

Probability weighting is a method that may be used to correct for bias due to non-ignorable missing data in growth models. Weighting is used with survey data to account for unequal selection of observations into the data, but it is a method that may also be used to address missing data issues in general. Weighting methods require larger sample sizes to meet asymptotic properties because the weights add variation to parameter estimates. However, with large sample sizes and survey software that properly handles the weights, this approach is easily applied. In longitudinal analyses, weights may be used to address both unequal selection probabilities and intermittent dropout over time.

Weights have traditionally been applied in a marginal modeling framework, but may also be used in a mixed modeling framework. In both the mixed and marginal modeling frameworks with continuous or quasi-continuous outcomes, a generalized least squares estimator (GLS) may be used. This GLS estimator weights the fixed effects parameters using the covariance matrix of within cluster correlations. In the mixed model this matrix is estimated using the random effects estimates. In the marginal modeling framework, GLS estimation utilizing the covariance matrix is termed generalized estimation equations (GEE) (Liang & Zeger, 1986). The GLS estimator will differ in the mixed and marginal modeling frameworks only when the specified covariance weight matrix or its estimation differ (Zeger, Liang, Alberts, 1988; Skinner & Vieira, 2007). This type of estimation will render missing repeated measures MAR when the missing data are related to other observed repeated measures and will improve efficiency when repeated measures are correlated. In this case, weights that correct for time specific nonresponse are unneccessary.

In this chapter several weighting approaches are applied to linear growth models

estimated in the mixed and marginal modeling frameworks. It is important to consider alternative weighting methods for growth models since there are several options, some of which have not been considered in the literature, e.g., treating weights as time varying. Also, the various weighting methods have not been compared and it is unknown which combination of weighting and model or estimation methods result in the best estimates. It has been traditional to apply panel weights and casewise delete any observations that do not have complete data across time points. This method seems to perform well for nonignorable unequal inclusion of observations, yet it is far less efficient due to the decreases in sample size. Unless the casewise deleted sample is large, it is better to use all available person-by-time observations along with a time varying weight in the marginal model or the multilevel weights in a mixed model. Though the latter requires large level one sample sizes. Also, for the mixed model generally and especially with weighted analysis, random effects estimates require much larger sample sizes as compared to the fixed effects. The time varying weights and the multilevel weights importantly include probability information for both missing level two (individuals) and level one (time) observations.

It is always important to understand the unequal selection and nonresponse well before choosing a method for dealing with it. Sampling weights available for an entire data set are not designed to handle the item missing problem. In most instances, survey data will suffer from both unit selection and nonresponse as well as missing items. Sampling weights are used when missingness is MNAR or nonignorable. Other methods, such as utilizing the correlation weight matrix can improve estimates since in most situations it is probably fair to assume that outcomes are related over time.

Method Label	Model framework	correlation matrix (V_0)*	weighting method	weight used
1. mixed time invariant weight	mixed	uniform / exchangeable	time-invariant	w_i
2. marginal time invariant weight	marginal	independent	time-invariant	w_i
3. marginal time invariant weight (exch.)	marginal	uniform / exchangeable	time-invariant	w_i
4. mixed panel weight	mixed	uniform / exchangeable	panel	w_i^p
5. marginal panel weight	marginal	independent	panel	w_i^p
6. mixed time varying weight	mixed	uniform / exchangeable	time varying	w_{it}
7. marginal time varying weight	marginal	independent	time varying	w_{it}
8. marginal time varying weight (exch.)	marginal	uniform / exchangeable	time varying	w_{it}
9. mixed multilevel weights	mixed	uniform / exchangeable	multilevel	w_i and λ_{it}
10. mixed no weight	mixed	uniform / exchangeable	no weighting	n/a
11. marginal no weight	marginal	independent	no weighting	n/a

* this matrix may be estimated differently for the mixed and marginal models

Table 4.1: Estimation Methods Evaluated

unconditional outcome	50 clusters	200 clusters	500 clusters
5 observations	24.1	27.6	27.3
4 observations	41.8	42.0	42.0
3 observations	26.1	23.8	24.0
2 observations	7.2	6.1	6.1
1 observation	0.8	0.6	0.6
conditional outcome	50 clusters	200 clusters	500 clusters
5 observations	24.1	27.9	27.4
4 observations	41.8	41.7	41.9
3 observations	26.0	23.7	24.0
2 observations	7.3	6.1	6.1
1 observation	0.7	0.6	0.6

Table 4.2: Percent of Level Two Observations with Level One Sample Size Averaged Across Sample Replicates

	√UWE 50 clusters	√UWE 200 clusters	√UWE 500 clusters
unconditional model			
time invariant weight	1.66	1.64	1.65
panel weight	1.91	1.87	1.88
time varying weight	1.76	1.70	1.71
conditional level 1 weights	1.06	1.04	1.04
conditional model			
time invariant weight	2.68	2.68	2.90
panel weight	4.11	2.67	3.14
time varying weight	2.93	2.78	2.99
conditional level 1 weights	1.06	1.04	1.04

Table 4.3: Unequal Weighting Effects Averaged Across Sample Replicates

	Level Two n	Level One n	intercept	slope	intercept s.e.	slope s.e.	intercept std.	slope std.
WEIGHTED								
population values			**20**	**2**				
mixed model	50	10	19.72	2.04	1.71	0.23	2.09	0.31
both marginal models			19.72	2.04	1.73	0.24	2.09	0.31
mixed & both marginal models	200	10	19.91	2.02	1.02	0.14	1.15	0.17
mixed & both marginal models	500	10	19.93	2.01	0.68	0.10	0.71	0.11
UNWEIGHTED								
population values			**20**	**2**				
mixed model	50	10	15.01	3.00	1.19	0.15	1.19	0.16
both marginal models			15.01	3.00	1.20	0.15	1.19	0.16
mixed & both marginal models	200	10	14.98	3.00	0.61	0.08	0.75	0.12
mixed & both marginal models	500	10	14.96	3.01	0.38	0.05	0.71	0.13

Table 4.4: Expected Values of Fixed Effects Estimates for the Unconditional Trajectory Model with No Missing Time Points

	Level Two n	Level One n	random intercept	random slope	cov. random	error	random intercept std.	random slope std.	cov. random std.	error std.
WEIGHTED										
population values			**70**	**1**	**-5**	**12.96**				
mixed model	50	10	62.21	0.88	-4.24	12.99	23.46	0.43	2.70	1.39
mixed model	200	10	68.16	0.96	-4.80	12.95	14.01	0.29	1.75	0.73
mixed model	500	10	68.71	0.99	-4.88	12.97	9.14	0.19	1.14	0.45
UNWEIGHTED										
population values			**70**	**1**	**-5**	**12.96**				
mixed model	50	10	66.91	0.99	-4.83	13.00	14.59	0.28	1.71	0.94
mixed model	200	10	69.42	0.99	-4.94	12.97	8.16	0.18	1.01	0.45
mixed model	500	10	69.41	0.99	-4.96	12.99	5.54	0.14	0.75	0.29

Table 4.5: Expected Values of Random Effects Estimates for the Unconditional Trajectory Model with No Missing Time Points

	Level Two n	Level One n	intercept	slope	random intercept	random slope	cov. random	error
WEIGHTED								
mixed model	50	10	0.87	0.82	0.74	0.67	0.68	0.84
both marginal models			0.88	0.82				
mixed model	200	10	0.89	0.88	0.85	0.81	0.82	0.81
both marginal models			0.90	0.89				
mixed & both marginal models	500	10	0.93	0.93	0.90	0.89	0.88	0.82
UNWEIGHTED								
mixed & both marginal models	50	10	0.01	0.00	0.90	0.89	0.92	0.94
mixed & both marginal models	200	10	0.00	0.00	0.92	0.91	0.92	0.97
mixed & both marginal models	500	10	0.00	0.00	0.93	0.93	0.93	0.95

Table 4.6: 95% Coverage Rates for Unconditional Trajectory Model Estimates with No Missing Time Points

	Level Two n	Level One n	intercept	slope	intercept on x	intercept on w	slope on x	slope on w
WEIGHTED								
population values			**5**	**4**	**2**	**1**	**-0.3**	**0.5**
mixed model	50	10	5.17	4.27	2.00	0.95	-0.31	0.47
marginal model			5.16	4.29	2.00	0.95	-0.31	0.47
marginal exchangeable model			5.09	4.29	2.03	0.97	-0.31	0.47
mixed model	200	10	5.02	4.10	2.00	1.00	-0.30	0.50
marginal model			5.01	4.11	1.96	1.00	-0.31	0.50
marginal exchangeable model			4.97	4.11	2.02	1.00	-0.31	0.50
mixed model	500	10	4.99	4.04	2.00	0.97	-0.30	0.51
marginal model			4.98	4.05	1.99	0.97	-0.29	0.51
marginal exchangeable model			4.96	4.05	2.02	0.97	-0.30	0.51
UNWEIGHTED								
population values			**5**	**4**	**2**	**1**	**-0.3**	**0.5**
mixed model	50	10	5.95	5.99	1.99	0.97	-0.30	0.49
marginal model			5.95	5.99	1.99	0.97	-0.31	0.50
marginal exchangeable model			5.95	5.99	1.99	0.97	-0.31	0.49
mixed model	200	10	5.97	6.00	2.00	0.98	-0.30	0.50
marginal model			5.97	6.00	1.97	0.98	-0.30	0.50
marginal exchangeable model			5.97	6.00	2.03	0.98	-0.31	0.50
mixed model	500	10	5.98	6.01	1.99	0.99	-0.30	0.49
marginal model			5.98	6.01	1.97	0.99	-0.30	0.49
marginal exchangeable model			5.98	6.01	2.00	0.99	-0.30	0.49

Table 4.7: Expected Values of the Fixed Effects Estimates for the Conditional Trajectory Model with No Missing Time Points

	Level Two n	Level One n	mean intercept s.e.	mean slope s.e.	intercept std.	slope std.
WEIGHTED						
mixed model	50	10	0.95	0.31	1.23	0.53
marginal model			0.95	0.30	1.23	0.52
marginal exchangeable model			0.97	0.30	1.24	0.51
mixed model	200	10	0.65	0.25	0.74	0.33
marginal model			0.64	0.24	0.73	0.33
marginal exchangeable model			0.66	0.24	0.75	0.33
mixed model	500	10	0.46	0.20	0.51	0.29
marginal model			0.46	0.19	0.51	0.28
marginal exchangeable model			0.47	0.19	0.52	0.29
UNWEIGHTED						
mixed & both marginal models	50	10	0.62	0.20	0.63	0.31
mixed & both marginal models	200	10	0.32	0.10	0.31	0.21
mixed & both marginal models	500	10	0.20	0.06	0.23	0.24

Table 4.8: Standard Deviations and Expected Values of Standard Errors for Fixed Effects Estimates for the Conditional Trajectory Model with No Missing Time Points

	Level Two n	Level One n	random intercept	random slope	cov. random	error	random intercept std.	random slope std.	cov. random std.	error std.
WEIGHTED										
population values			**16**	**2**	**1**	**12.25**				
mixed model	50	10	12.61	1.30	0.64	11.87	6.56	0.56	1.47	1.68
mixed model	200	10	14.85	1.71	0.89	12.23	4.16	0.51	1.35	1.09
mixed model	500	10	15.45	1.88	0.96	12.19	3.17	0.58	0.96	0.68
UNWEIGHTED										
population values			**16**	**2**	**1**	**12.25**				
mixed model	50	10	15.03	1.90	0.94	12.26	4.01	0.44	0.87	0.83
mixed model	200	10	15.79	1.94	0.97	12.28	1.99	0.31	0.45	0.43
mixed model	500	10	15.94	1.96	0.95	12.28	1.32	0.32	0.31	0.26

Table 4.9: Expected Values of Random Effects Estimates for the Conditional Trajectory Model with No Missing Time Points

	Level Two n	Level One n	intercept	slope	intercept on x	intercept on w	slope on x	slope on w	random intercept	random slope	cov. random	error
WEIGHTED												
mixed	50	10	0.86	0.63	0.85	0.83	0.85	0.74	0.69	0.41	0.82	0.60
marginal			0.85	0.61	0.81	0.83	0.77	0.76				
marginal exchangeable			0.86	0.63	0.83	0.83	0.81	0.73				
mixed	200	10	0.92	0.78	0.93	0.88	0.91	0.84	0.85	0.68	0.89	0.49
marginal			0.92	0.77	0.87	0.89	0.86	0.84				
marginal exchangeable			0.93	0.77	0.86	0.88	0.84	0.83				
mixed	500	10	0.91	0.79	0.93	0.92	0.94	0.92	0.88	0.74	0.90	0.46
marginal			0.91	0.78	0.90	0.92	0.90	0.92				
marginal exchangeable			0.91	0.78	0.88	0.92	0.89	0.92				
UNWEIGHTED												
mixed & both marginal models	50	10	0.65	0.00	0.94	0.92	0.94	0.93	0.90	0.88	0.96	0.95
mixed & both marginal models	200	10	0.12	0.00	0.95	0.93	0.95	0.93	0.94	0.89	0.95	0.96
mixed & both marginal models	500	10	0.00	0.00	0.94	0.93	0.95	0.95	0.94	0.77	0.93	0.95

Table 4.10: 95% Confidence Rates for Estimates for the Conditional Trajectory Model with No Missing Time Points

	Level Two n	L. One mean n	intercept	slope	intercept s.e.	slope s.e.	intercept std.	slope std.
population values			**20**	**2**				
1. mixed time invariant weight	50	3.8	19.88	2.32	1.76	0.33	2.12	0.41
2. marginal time invariant weight	50	3.8	19.86	2.31	1.79	0.38	2.14	0.45
3. marginal time invariant weight (exch.)	50	3.8	19.87	2.32	1.77	0.33	2.13	0.41
4. mixed panel weight	12*	5	18.73	2.32	2.71	0.45	3.86	0.71
5. marginal panel weight	12*	5	18.73	2.32	2.83	0.47	3.86	0.71
6. mixed time varying weight	50	3.8	19.83	2.33	1.76	0.33	2.11	0.41
7. marginal time varying weight	50	3.8	19.70	2.05	1.82	0.42	2.14	0.49
8. marginal time varying weight (exch.)	50	3.8	17.47	2.18	1.83	0.34	2.52	0.43
9. mixed multilevel weights	50	3.8	19.62	2.23	1.76	0.34	2.11	0.42
10. mixed no weight	50	3.8	15.18	3.27	1.23	0.23	1.21	0.22
11. marginal no weight	50	3.8	15.18	3.27	1.26	0.27	1.20	0.26
1. mixed time invariant weight	200	3.9	20.05	2.23	1.05	0.19	1.17	0.22
2. marginal time invariant weight	200	3.9	20.06	2.22	1.07	0.23	1.20	0.25
3. marginal time invariant weight (exch.)	200	3.9	20.05	2.23	1.05	0.20	1.17	0.22
4. mixed panel weight	55*	5	19.60	2.08	1.75	0.31	2.20	0.41
5. marginal panel weight	55*	5	19.60	2.08	1.77	0.31	2.20	0.41
6. mixed time varying weight	200	3.9	20.01	2.24	1.05	0.19	1.16	0.22
7. marginal time varying weight	200	3.9	19.92	2.01	1.08	0.24	1.21	0.26
8. marginal time varying weight (exch.)	200	3.9	17.81	2.12	1.06	0.19	1.37	0.23
9. mixed multilevel weights	200	3.9	19.85	2.15	1.05	0.20	1.17	0.22
10. mixed no weight	200	3.9	15.13	3.21	0.62	0.12	0.76	0.15
11. marginal no weight	200	3.9	15.12	3.21	0.64	0.14	0.78	0.16
1. mixed time invariant weight	500	3.9	20.06	2.22	0.70	0.13	0.73	0.14
2. marginal time invariant weight	500	3.9	20.07	2.22	0.71	0.15	0.75	0.16
3. marginal time invariant weight (exch.)	500	3.9	20.06	2.22	0.70	0.13	0.73	0.14
4. mixed panel weight	137*	5	19.77	2.02	1.29	0.24	1.51	0.29
5. marginal panel weight	137*	5	19.77	2.02	1.30	0.24	1.51	0.29
6. mixed time varying weight	500	3.9	20.02	2.23	0.70	0.13	0.72	0.14
7. marginal time varying weight	500	3.9	19.93	2.00	0.71	0.16	0.75	0.17
8. marginal time varying weight (exch.)	500	3.9	17.83	2.11	0.71	0.14	0.87	0.15
9. mixed multilevel weights	500	3.9	19.87	2.14	0.70	0.14	0.73	0.14
10. mixed no weight	500	3.9	15.09	3.22	0.40	0.07	0.71	0.14
11. marginal no weight	500	3.9	15.09	3.22	0.40	0.09	0.72	0.14

* mean n averages over 500 replicates

Table 4.11: Expected Values of Weighted Fixed Effects Estimates for the Unconditional Trajectory Model

	Level Two n	L. One mean n	random intercept	random slope	cov. random	error	random intercept std.	random slope std.	cov. random std.	error std.
population values			**70**	**1**	**-5**	**12.96**				
1. mixed time invariant weight	50	3.8	62.42	0.78	-4.19	12.39	25.29	0.93	3.93	2.75
4. mixed panel weight	12*	5	50.60	0.32	-2.66	12.91	34.10	1.05	4.68	4.51
6. mixed time varying weight	50	3.8	62.44	0.78	-4.19	12.39	25.12	0.93	3.93	2.74
9. mixed multilevel weights	50	3.8	61.51	0.83	-3.86	12.23	25.01	0.96	3.93	2.81
10. mixed no weight	50	3.8	67.28	0.97	-4.91	12.40	15.68	0.60	2.61	1.67
1. mixed time invariant weight	200	3.9	67.78	0.89	-4.65	12.68	14.84	0.53	2.41	1.50
4. mixed panel weight	55*	5	61.75	0.69	-3.64	13.06	24.85	0.76	3.70	2.46
6. mixed time varying weight	200	3.9	67.78	0.89	-4.64	12.69	14.87	0.53	2.41	1.50
9. mixed multilevel weights	200	3.9	66.77	0.90	-4.31	12.61	14.78	0.53	2.39	1.54
10. mixed no weight	200	3.9	69.24	0.97	-4.89	12.60	8.49	0.30	1.36	0.91
1. mixed time invariant weight	500	3.9	68.79	0.98	-4.89	12.63	9.71	0.36	1.52	0.96
4. mixed panel weight	137*	5	65.41	0.92	-4.46	12.97	15.57	0.59	2.55	1.61
6. mixed time varying weight	500	3.9	68.77	0.98	-4.88	12.63	9.71	0.36	1.52	0.96
9. mixed multilevel weights	500	3.9	67.89	1.00	-4.59	12.55	9.75	0.38	1.56	0.95
10. mixed no weight	500	3.9	69.57	1.00	-5.00	12.63	5.80	0.21	0.95	0.59

* mean n averages over 500 replicates

Table 4.12: Expected Values of Weighted Random Effects Estimates for the Unconditional Trajectory Model

	Level Two n	L. One mean n	intercept	slope	random intercept	random slope	cov. random	error
1. mixed time invariant weight	50	3.8	0.89	0.69	0.76	0.79	0.77	0.62
2. marginal time invariant weight	50	3.8	0.90	0.74				
3. marginal time invariant weight (exch.)	50	3.8	0.89	0.68				
4. mixed panel weight	12*	5	0.76	0.67	0.59	0.55	0.62	0.66
5. marginal panel weight	12*	5	0.78	0.68				
6. mixed time varying weight	50	3.8	0.88	0.69	0.76	0.78	0.77	0.61
7. marginal time varying weight	50	3.8	0.89	0.88				
8. marginal time varying weight (exch.)	50	3.8	0.57	0.78				
9. mixed multilevel weights	50	3.8	0.87	0.76	0.76	0.83	0.76	0.59
10. mixed no weight	50	3.8	0.01	0.00	0.91	0.94	0.92	0.92
11. marginal no weight	50	3.8	0.02	0.01				
1. mixed time invariant weight	200	3.9	0.92	0.68	0.85	0.85	0.84	0.63
2. marginal time invariant weight	200	3.9	0.92	0.73				
3. marginal time invariant weight (exch.)	200	3.9	0.92	0.68				
4. mixed panel weight	55*	5	0.84	0.81	0.74	0.71	0.68	0.61
5. marginal panel weight	55*	5	0.84	0.82				
6. mixed time varying weight	200	3.9	0.92	0.65	0.84	0.85	0.85	0.63
7. marginal time varying weight	200	3.9	0.91	0.92				
8. marginal time varying weight (exch.)	200	3.9	0.42	0.81				
9. mixed multilevel weights	200	3.9	0.90	0.77	0.84	0.89	0.82	0.62
10. mixed no weight	200	3.9	0.00	0.00	0.92	0.95	0.93	0.92
11. marginal no weight	200	3.9	0.00	0.00				
1. mixed time invariant weight	500	3.9	0.95	0.54	0.91	0.91	0.91	0.60
2. marginal time invariant weight	500	3.9	0.95	0.64				
3. marginal time invariant weight (exch.)	500	3.9	0.95	0.55				
4. mixed panel weight	137*	5	0.88	0.86	0.84	0.82	0.81	0.55
5. marginal panel weight	137*	5	0.88	0.86				
6. mixed time varying weight	500	3.9	0.95	0.51	0.91	0.91	0.91	0.60
7. marginal time varying weight	500	3.9	0.94	0.93				
8. marginal time varying weight (exch.)	500	3.9	0.21	0.79				
9. mixed multilevel weights	500	3.9	0.92	0.75	0.91	0.95	0.88	0.59
10. mixed no weight	500	3.9	0.00	0.00	0.93	0.92	0.92	0.91
11. marginal no weight	500	3.9	0.00	0.00				

* mean n averages over 500 replicates

Table 4.13: 95% Coverage Rate for Weighted Estimates for the Unconditional Trajectory Model

	Level Two n	L. One mean n	intercept	slope	intercept on x	intercept on w	slope on x	slope std.
population values			**5**	**4**	**2**	**1**	**-0.3**	**0.5**
1. mixed time invariant weight	50	3.8	5.23	4.61	1.99	0.92	-0.31	0.51
2. marginal time invariant weight	50	3.8	5.24	4.64	1.94	0.91	-0.29	0.50
3. marginal time invariant weight (exch.)	50	3.8	5.23	4.69	1.93	0.95	-0.28	0.51
4. mixed panel weight	12*	5	5.22	4.98	2.02	0.86	-0.31	0.47
5. marginal panel weight	12*	5	5.21	5.03	2.02	0.86	-0.30	0.47
6. mixed time varying weight	50	3.8	5.20	4.62	1.99	0.92	-0.31	0.51
7. marginal time varying weight	50	3.8	5.06	4.45	1.94	0.91	-0.29	0.50
8. marginal time varying weight (exch.)	50	3.8	4.85	4.53	1.91	0.97	-0.29	0.51
9. mixed multilevel weights	50	3.8	5.02	4.54	1.98	0.92	-0.31	0.50
10. mixed no weight	50	3.8	6.10	6.25	1.98	0.98	-0.30	0.49
11. marginal no weight	50	3.8	6.11	6.24	1.96	0.97	-0.29	0.49
1. mixed time invariant weight	200	3.9	5.13	4.33	1.97	1.00	-0.29	0.50
2. marginal time invariant weight	200	3.9	5.14	4.35	1.97	0.99	-0.29	0.51
3. marginal time invariant weight (exch.)	200	3.9	5.12	4.35	1.95	0.98	-0.29	0.51
4. mixed panel weight	55*	5	4.99	4.39	1.97	0.96	-0.31	0.48
5. marginal panel weight	55*	5	5.03	4.41	1.96	0.95	-0.30	0.48
6. mixed time varying weight	200	3.9	5.09	4.34	1.97	1.00	-0.29	0.50
7. marginal time varying weight	200	3.9	5.00	4.17	1.97	0.98	-0.29	0.51
8. marginal time varying weight (exch.)	200	3.9	4.54	4.15	1.92	1.05	-0.30	0.52
9. mixed multilevel weights	200	3.9	4.94	4.27	1.97	1.00	-0.29	0.50
10. mixed no weight	200	3.9	6.11	6.21	2.01	0.97	-0.31	0.51
11. marginal no weight	200	3.9	6.11	6.20	2.00	0.96	-0.31	0.51
1. mixed time invariant weight	500	3.9	5.10	4.26	2.00	0.96	-0.31	0.50
2. marginal time invariant weight	500	3.9	5.12	4.26	1.99	0.95	-0.30	0.51
3. marginal time invariant weight (exch.)	500	3.9	5.11	4.27	2.01	0.96	-0.31	0.51
4. mixed panel weight	137*	5	4.99	4.21	1.96	0.97	-0.29	0.51
5. marginal panel weight	137*	5	5.00	4.23	1.99	0.96	-0.30	0.51
6. mixed time varying weight	500	3.9	5.07	4.27	2.00	0.96	-0.31	0.50
7. marginal time varying weight	500	3.9	4.98	4.07	1.99	0.95	-0.30	0.51
8. marginal time varying weight (exch.)	500	3.9	4.52	4.09	2.00	0.95	-0.31	0.50
9. mixed multilevel weights	500	3.9	4.92	4.20	2.00	0.96	-0.30	0.50
10. mixed no weight	500	3.9	6.12	6.21	1.99	0.98	-0.30	0.50
11. marginal no weight	500	3.9	6.12	6.21	1.99	0.98	-0.31	0.50

* mean n averages over 500 replicates

Table 4.14: Expected Values of Weighted Fixed Effects Estimates for the Conditional Trajectory Model

	Level Two n	L. One mean n	mean intercept s.e.	mean slope s.e.	mean intercept on x s.e.	mean intercept on w s.e.	mean slope on x s.e.	mean slope on w s.e.	intercept std.	slope std.	intercept on x std.	intercept on w std.	slope on x std.	slope on w std.
1. mixed time invariant weight	50	3.8	1.03	0.39	0.67	0.93	0.30	0.32	1.31	0.63	0.98	1.39	0.43	0.51
2. marginal time invariant weight	50	3.8	1.04	0.40	0.87	0.96	0.41	0.35	1.32	0.63	1.32	1.41	0.64	0.55
3. marginal time invariant weight (exch.)	50	3.8	1.08	0.39	0.83	0.99	0.39	0.34	1.31	0.62	1.48	1.38	0.64	0.52
4. mixed panel weight	12*	5	1.41	0.45	0.98	1.43	0.43	0.46	2.50	0.96	1.67	2.68	0.72	0.86
5. marginal panel weight	12*	5	1.46	0.48	1.24	1.49	0.55	0.49	2.50	0.93	2.15	2.73	0.94	0.89
6. mixed time varying weight	50	3.8	1.03	0.39	0.67	0.93	0.3	0.32	1.31	0.63	0.98	1.39	0.43	0.51
7. marginal time varying weight	50	3.8	1.07	0.41	0.91	0.98	0.43	0.36	1.38	0.67	1.38	1.45	0.67	0.58
8. marginal time varying weight (exch.)	50	3.8	1.13	0.42	0.87	1.02	0.41	0.37	1.6	0.73	1.51	1.45	0.67	0.65
9. mixed multilevel weights	50	3.8	1.04	0.39	0.69	0.94	0.31	0.33	1.33	0.63	0.99	1.4	0.44	0.52
10. mixed no weight	50	3.8	0.69	0.27	0.48	0.70	0.22	0.27	0.71	0.34	0.54	0.74	0.25	0.28
11. marginal no weight	50	3.8	0.72	0.30	0.63	0.72	0.33	0.29	0.73	0.36	0.69	0.76	0.35	0.32
1. mixed time invariant weight	200	3.9	0.71	0.29	0.46	0.63	0.21	0.25	0.81	0.38	0.53	0.82	0.26	0.36
2. marginal time invariant weight	200	3.9	0.71	0.30	0.60	0.64	0.30	0.26	0.81	0.40	0.76	0.83	0.40	0.37
3. marginal time invariant weight (exch.)	200	3.9	0.73	0.30	0.49	0.64	0.24	0.25	0.84	0.38	0.68	0.89	0.33	0.43
4. mixed panel weight	55*	5	1.02	0.36	0.66	0.94	0.28	0.31	1.36	0.59	0.87	1.29	0.38	0.50
5. marginal panel weight	55*	5	1.01	0.36	0.85	0.94	0.37	0.31	1.34	0.60	1.15	1.30	0.55	0.49
6. mixed time varying weight	200	3.9	0.71	0.29	0.46	0.63	0.21	0.25	0.81	0.38	0.53	0.82	0.26	0.36
7. marginal time varying weight	200	3.9	0.74	0.32	0.64	0.66	0.32	0.27	0.84	0.45	0.83	0.86	0.44	0.4
8. marginal time varying weight (exch.)	200	3.9	0.73	0.3	1.22	2.21	0.32	0.31	0.99	0.59	4.03	2.86	0.52	0.63
9. mixed multilevel weights	200	3.9	0.71	0.3	0.47	0.63	0.21	0.25	0.82	0.39	0.54	0.82	0.27	0.36
10. mixed no weight	200	3.9	0.35	0.13	0.24	0.35	0.11	0.13	0.34	0.23	0.24	0.35	0.11	0.14
11. marginal no weight	200	3.9	0.36	0.15	0.33	0.36	0.17	0.15	0.35	0.24	0.32	0.37	0.16	0.14
1. mixed time invariant weight	500	3.9	0.50	0.23	0.33	0.46	0.15	0.20	0.58	0.32	0.37	0.55	0.17	0.25
2. marginal time invariant weight	500	3.9	0.51	0.24	0.44	0.46	0.24	0.21	0.60	0.34	0.52	0.56	0.30	0.27
3. marginal time invariant weight (exch.)	500	3.9	0.51	0.23	0.36	0.46	0.19	0.20	0.59	0.32	0.45	0.56	0.23	0.26
4. mixed panel weight	137*	5	0.80	0.31	0.50	0.71	0.22	0.26	1.02	0.49	0.61	0.88	0.27	0.41
5. marginal panel weight	137*	5	0.79	0.30	0.66	0.71	0.31	0.26	1.02	0.47	0.85	0.88	0.44	0.40
6. mixed time varying weight	500	3.9	0.5	0.23	0.33	0.46	0.15	0.2	0.58	0.32	0.37	0.55	0.17	0.25
7. marginal time varying weight	500	3.9	0.54	0.26	0.47	0.48	0.25	0.22	0.64	0.36	0.56	0.59	0.32	0.29
8. marginal time varying weight (exch.)	500	3.9	0.52	0.24	0.39	0.47	0.21	0.2	0.74	0.39	0.47	0.57	0.26	0.28
9. mixed multilevel weights	500	3.9	0.5	0.23	0.34	0.46	0.16	0.2	0.58	0.32	0.38	0.55	0.17	0.26
10. mixed no weight	500	3.9	0.22	0.09	0.15	0.22	0.07	0.09	0.25	0.25	0.16	0.23	0.07	0.09
11. marginal no weight	500	3.9	0.23	0.09	0.21	0.23	0.11	0.09	0.25	0.25	0.21	0.23	0.11	0.10

Table 4.15: Standard Deviations and Expected Values of Standard Errors for Weighted Fixed Effects Estimates for the Conditional Trajectory Model with No Level One Weighting

	Level Two n	L. One mean n	random intercept	random slope	cov. random	error	random intercept std.	random slope std.	cov. random std.	error std.
population values			**16**	**2**	**1**	**12.25**				
1. mixed time invariant weight	50	3.8	12.14	0.98	0.96	11.07	8.10	1.10	2.25	3.94
4. mixed panel weight	12*	5	6.37	0.24	1.17	10.98	10.22	1.09	2.36	4.42
6. mixed time varying weight	50	3.8	12.10	0.98	0.96	11.08	8.10	1.10	2.24	3.94
9. mixed multilevel weights	50	3.8	11.25	0.97	1.26	10.90	8.13	1.10	2.32	3.74
10. mixed no weight	50	3.8	14.97	1.74	1.07	11.67	4.83	0.74	1.39	1.70
1. mixed time invariant weight	200	3.9	14.92	1.62	0.96	11.72	5.39	0.85	1.80	2.01
4. mixed panel weight	55*	5	12.33	0.98	1.10	11.94	8.24	0.94	2.09	2.74
6. mixed time varying weight	200	3.9	14.88	1.62	0.96	11.72	5.39	0.85	1.80	2.01
9. mixed multilevel weights	200	3.9	14.10	1.63	1.23	11.65	5.45	0.90	1.83	1.98
10. mixed no weight	200	3.9	15.82	1.87	1.02	11.93	2.46	0.41	0.73	0.87
1. mixed time invariant weight	500	3.9	15.58	1.81	0.98	11.73	3.63	0.82	1.28	1.37
4. mixed panel weight	137*	5	14.06	1.39	0.95	11.80	5.47	0.91	1.79	1.97
6. mixed time varying weight	500	3.9	15.54	1.81	0.99	11.74	3.63	0.83	1.28	1.38
9. mixed multilevel weights	500	3.9	14.74	1.80	1.28	11.67	3.68	0.83	1.33	1.39
10. mixed no weight	500	3.9	16.00	1.93	0.94	11.92	1.65	0.36	0.48	0.55

* mean n averages over 500 replicates

Table 4.16: Expected Values of Weighted Random Effects Estimates for the Conditional Trajectory Model

Weighting Method	Level Two n	L. One mean n	intercept	slope	intercept on x	intercept on w	slope on x	slope on w	random intercept	random slope	cov. random	error
1. mixed time invariant weight	50	3.8	0.85	0.52	0.83	0.81	0.83	0.79	0.74	0.49	0.85	0.42
2. marginal time invariant weight	50	3.8	0.83	0.54	0.80	0.80	0.82	0.78				
3. marginal time invariant weight (exch.)	50	3.8	0.87	0.48	0.80	0.85	0.83	0.79				
4. mixed panel weight	12	5	0.68	0.38	0.73	0.65	0.73	0.71	0.41	0.22	0.82	0.55
5. marginal panel weight	12	5	0.71	0.41	0.75	0.67	0.76	0.71				
6. mixed time varying weight	50	3.8	0.84	0.52	0.83	0.81	0.83	0.79	0.73	0.50	0.85	0.42
7. marginal time varying weight	50	3.8	0.83	0.61	0.81	0.79	0.79	0.78				
8. marginal time varying weight (exch.)	50	3.8	0.86	0.57	0.82	0.82	0.82	0.80				
9. mixed multilevel weights	50	3.8	0.83	0.55	0.83	0.81	0.83	0.80	0.73	0.52	0.87	0.43
10. mixed no weight	50	3.8	0.63	0.00	0.92	0.93	0.93	0.94	0.90	0.87	0.95	0.91
11. marginal no weight	50	3.8	0.65	0.00	0.93	0.92	0.92	0.92				
1. mixed time invariant weight	200	3.9	0.91	0.64	0.89	0.88	0.91	0.86	0.86	0.75	0.91	0.46
2. marginal time invariant weight	200	3.9	0.91	0.62	0.90	0.87	0.88	0.85				
3. marginal time invariant weight (exch.)	200	3.9	0.91	0.61	0.91	0.89	0.90	0.87				
4. mixed panel weight	55	5	0.86	0.61	0.86	0.83	0.85	0.80	0.66	0.41	0.84	0.45
5. marginal panel weight	55	5	0.85	0.60	0.82	0.83	0.82	0.81				
6. mixed time varying weight	200	3.9	0.92	0.63	0.89	0.88	0.91	0.86	0.86	0.75	0.91	0.46
7. marginal time varying weight	200	3.9	0.92	0.77	0.90	0.87	0.85	0.84				
8. marginal time varying weight (exch.)	200	3.9	0.87	0.72	0.90	0.87	0.86	0.85				
9. mixed multilevel weights	200	3.9	0.92	0.69	0.90	0.88	0.91	0.87	0.84	0.77	0.91	0.49
10. mixed no weight	200	3.9	0.12	0.00	0.94	0.94	0.94	0.94	0.94	0.90	0.93	0.91
11. marginal no weight	200	3.9	0.12	0.00	0.95	0.93	0.96	0.96				
1. mixed time invariant weight	500	3.9	0.90	0.60	0.93	0.94	0.91	0.89	0.91	0.79	0.94	0.43
2. marginal time invariant weight	500	3.9	0.89	0.63	0.92	0.93	0.90	0.88				
3. marginal time invariant weight (exch.)	500	3.9	0.90	0.61	0.91	0.94	0.88	0.88				
4. mixed panel weight	137	5	0.89	0.67	0.91	0.87	0.89	0.84	0.84	0.58	0.90	0.40
5. marginal panel weight	137	5	0.90	0.66	0.88	0.87	0.86	0.85				
6. mixed time varying weight	500	3.9	0.91	0.58	0.93	0.94	0.91	0.89	0.92	0.79	0.93	0.43
7. marginal time varying weight	500	3.9	0.92	0.80	0.92	0.93	0.91	0.89				
8. marginal time varying weight (exch.)	500	3.9	0.81	0.77	0.90	0.93	0.88	0.88				
9. mixed multilevel weights	500	3.9	0.92	0.68	0.93	0.95	0.92	0.89	0.90	0.81	0.92	0.46
10. mixed no weight	500	3.9	0.19	0.00	0.93	0.95	0.94	0.95	0.95	0.86	0.94	0.92
11. marginal no weight	500	3.9	0.19	0.00	0.93	0.94	0.94	0.94				

* mean n averages over 500 replicates

Table 4.17: 95% Coverage Rates for Weighted Estimates for the Conditional Trajectory Model

wave 1	wave 2	wave 3	CDI (n=2914)	KBIT Verbal	KBIT Math (n=3854)
X	X	X	55.1	55.4	56.2
X	X		6.5	6.4	6.4
X			7.1	8.0	7.9
X		X	6.4	7.5	7.6
	X	X	11.5	10.5	10.1
	X		1.7	0.9	0.8
		X	11.8	11.3	11.0

Table 4.18: Percent of Sample Observations with Missing Pattern by Outcome

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	2914	2.3	9.83	0.33	-1.08	0.19
mixed panel weight	1605	3	9.68	0.39	-1.22	0.20
mixed time varying weight	2914	2.3	9.77	0.33	-1.04	0.20
mixed multilevel weights	2914	2.3	9.83	0.33	-1.09	0.19
mixed no weight	2914	2.3	10.01	0.16	-1.03	0.09
marginal time invariant weight	2914	2.3	9.74	0.34	-1.03	0.20
marginal panel weight	1605	3	9.68	0.39	-1.22	0.20
marginal time varying weight	2914	2.3	9.67	0.34	-1.07	0.21
marginal no weight	2914	2.3	9.78	0.17	-0.90	0.09
marginal invariant weight exchangeable	2914	2.3	9.87	0.33	-1.11	0.19
marginal time varying weight exchangeable	2914	2.3	9.76	0.33	-1.12	0.19

Table 4.19: CDI Unconditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	3854	2.4	21.32	0.30	2.22	0.10
mixed panel weight	2167	3	23.49	0.33	2.49	0.11
mixed time varying weight	3854	2.4	21.14	0.30	2.21	0.10
mixed multilevel weights	3854	2.4	21.36	0.30	2.20	0.10
mixed no weight	3854	2.4	21.12	0.16	2.20	0.06
marginal time invariant weight	3854	2.4	23.55	0.29	0.92	0.15
marginal panel weight	2167	3	23.49	0.33	2.49	0.11
marginal time varying weight	3854	2.4	23.61	0.29	0.99	0.16
marginal no weight	3854	2.4	23.70	0.15	0.80	0.08
marginal invariant weight exchangeable	3854	2.4	21.74	0.29	1.98	0.11
marginal time varying weight exchangeable	3854	2.4	21.81	0.29	2.08	0.11

Table 4.20: KBIT Matrices Unconditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	3831	2.4	32.18	0.55	4.69	0.12
mixed panel weight	2124	3	37.96	0.61	4.42	0.14
mixed time varying weight	3831	2.4	31.74	0.56	4.72	0.12
mixed multilevel weights	3831	2.4	32.21	0.55	4.68	0.12
mixed no weight	3831	2.4	31.72	0.30	4.62	0.07
marginal time invariant weight	3831	2.4	37.55	0.50	1.64	0.24
marginal panel weight	2124	3	37.96	0.61	4.42	0.14
marginal time varying weight	3831	2.4	37.65	0.50	1.74	0.24
marginal no weight	3831	2.4	37.96	0.27	1.23	0.14
marginal invariant weight exchangeable	3831	2.4	33.45	0.51	3.97	0.13
marginal time varying weight exchangeable	3831	2.4	33.71	0.52	4.15	0.12

Table 4.21: KBIT Verbal Unconditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		random slope variance		covariance random effects		error variance	
mixed time invariant weight	2914	2.3	41.57	5.77	5.68	2.47	-10.71	2.96	24.47	1.43
mixed panel weight	1605	3	43.05	6.59	5.15	1.98	-10.77	2.79	22.74	1.07
mixed time varying weight	2914	2.3	39.54	5.75	5.43	2.73	-9.79	3.03	24.95	1.62
mixed multilevel weights	2914	2.3	40.24	6.07	5.02	2.51	-9.73	3.08	24.84	1.50
mixed no weight	2914	2.3	39.88	1.99	4.45	0.71	-8.13	0.97	25.56	0.83

Table 4.22: CDI Unconditional Trajectory Models Random Effects with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		random slope variance		covariance random effects		error variance	
mixed time invariant weight	3854	2.4	74.54	4.29	1.32	0.93	-5.12	1.44	13.14	0.66
mixed panel weight	2167	3	51.79	4.00	0.78	0.95	-5.25	1.43	13.47	0.74
mixed time varying weight	3854	2.4	75.20	4.36	1.24	0.93	-4.86	1.43	13.13	0.67
mixed multilevel weights	3854	2.4	73.31	4.38	1.10	0.97	-4.49	1.50	13.18	0.70
mixed no weight	3854	2.4	79.28	2.23	2.21	0.32	-5.29	0.62	13.12	0.37

Table 4.23: KBIT Matrices Unconditional Trajectory Models Random Effects with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		random slope variance		covariance random effects		error variance	
mixed time invariant weight	3831	2.4	306.68	13.44	5.80	1.25	-30.77	2.70	11.77	0.55
mixed panel weight	2124	3	188.22	12.42	5.05	1.36	-22.49	2.87	12.09	0.60
mixed time varying weight	3831	2.4	312.98	13.87	5.70	1.24	-30.87	2.70	11.72	0.55
mixed multilevel weights	3831	2.4	306.23	13.65	5.70	1.30	-30.50	2.78	11.83	0.57
mixed no weight	3831	2.4	332.53	8.04	7.73	0.41	-34.00	1.47	11.24	0.33

Table 4.24: KBIT Verbal Unconditional Trajectory Models Random Effects with Standard Errors

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	2905	2.3	9.00	1.58	-0.23	0.96
mixed panel weight	1601	3	8.64	1.91	-1.43	0.98
mixed time varying weight	2905	2.3	8.45	1.66	0.05	1.06
mixed multilevel weights	2905	2.3	9.00	1.60	-0.25	0.97
mixed no weight	2905	2.3	11.19	0.71	-0.80	0.39
marginal time invariant weight	2905	2.3	8.57	1.58	0.01	0.95
marginal panel weight	1601	3	8.64	1.91	-1.43	0.98
marginal time varying weight	2905	2.3	8.33	1.61	0.13	0.99
marginal no weight	2905	2.3	10.00	0.76	-0.15	0.43
marginal invariant weight exchangeable	2905	2.3	9.34	1.54	-0.44	0.92
marginal time varying weight exchangeable	2905	2.3	8.97	1.60	-0.27	0.98

Table 4.25: CDI Conditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	3843	2.4	5.13	0.51	4.22	0.28
mixed panel weight	2162	3	6.67	0.63	5.52	0.35
mixed time varying weight	3843	2.4	5.10	0.51	4.18	0.27
mixed multilevel weights	3843	2.4	5.24	0.52	4.17	0.28
mixed no weight	3843	2.4	4.68	0.30	4.16	0.16
marginal time invariant weight	3843	2.4	7.21	0.56	3.06	0.31
marginal panel weight	2162	3	6.67	0.63	5.52	0.35
marginal time varying weight	3843	2.4	7.30	0.56	3.11	0.32
marginal no weight	3843	2.4	6.82	0.31	3.04	0.17
marginal invariant weight exchangeable	3843	2.4	5.09	0.51	4.24	0.28
marginal time varying weight exchangeable	3843	2.4	5.14	0.52	4.33	0.28

Table 4.26: KBIT Matrices Conditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n	intercept		slope	
mixed time invariant weight	3820	2.4	1.15	0.63	8.61	0.25
mixed panel weight	2119	3	5.53	0.98	9.67	0.32
mixed time varying weight	3820	2.4	0.97	0.63	8.57	0.25
mixed multilevel weights	3820	2.4	1.14	0.63	8.63	0.25
mixed no weight	3820	2.4	0.06	0.40	8.81	0.18
marginal time invariant weight	3820	2.4	5.56	0.76	6.23	0.36
marginal panel weight	2119	3	5.53	0.98	9.67	0.32
marginal time varying weight	3820	2.4	5.65	0.77	6.34	0.36
marginal no weight	3820	2.4	4.96	0.45	6.21	0.23
marginal invariant weight exchangeable	3820	2.4	1.18	0.63	8.59	0.25
marginal time varying weight exchangeable	3820	2.4	1.17	0.64	8.80	0.26

Table 4.27: KBIT Verbal Conditional Trajectory Models Fixed Effects with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	2905	2.3	0.14	0.14	-1.17	0.64	-0.91	0.72	0.18	0.98	0.75	1.05
mixed panel weight	1601	3	0.22	0.17	-1.55	0.78	-1.67	0.85	-0.42	1.02	-0.19	1.13
mixed time varying weight	2905	2.3	0.18	0.15	-1.13	0.64	-0.73	0.73	0.27	0.98	0.89	1.06
mixed multilevel weights	2905	2.3	0.14	0.14	-1.17	0.64	-0.90	0.72	0.19	0.99	0.72	1.06
mixed no weight	2905	2.3	-0.11	0.06	-0.39	0.32	-0.44	0.38	0.84	0.47	0.53	0.61
marginal time invariant weight	2905	2.3	0.20	0.14	-1.44	0.66	-1.31	0.73	0.21	1.02	0.70	1.07
marginal panel weight	1601	3	0.22	0.17	-1.55	0.78	-1.67	0.85	-0.42	1.02	-0.19	1.13
marginal time varying weight	2905	2.3	0.22	0.15	-1.50	0.66	-1.16	0.74	0.34	1.03	0.73	1.06
marginal no weight	2905	2.3	0.00	0.07	-0.56	0.33	-0.46	0.39	0.67	0.48	0.55	0.58
marginal invariant weight exchangeable	2905	2.3	0.12	0.14	-1.17	0.64	-1.04	0.72	0.11	0.98	0.67	1.05
marginal time varying weight exchangeable	2905	2.3	0.14	0.14	-1.21	0.64	-0.82	0.72	0.28	1.00	0.81	1.05

Table 4.28: CDI Conditional Trajectory Models Fixed Effects on the Intercept with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	3843	2.4	2.06	0.05	0.00	0.36	-1.19	0.42	-0.73	0.52	-0.17	0.68
mixed panel weight	2162	3	1.88	0.06	0.24	0.42	-0.84	0.49	-0.54	0.64	0.11	0.91
mixed time varying weight	3843	2.4	2.07	0.05	-0.04	0.36	-1.21	0.41	-0.74	0.53	-0.22	0.67
mixed multilevel weights	3843	2.4	2.05	0.05	0.02	0.36	-1.21	0.41	-0.78	0.52	-0.16	0.68
mixed no weight	3843	2.4	2.07	0.03	0.15	0.19	-1.49	0.22	-0.27	0.27	-0.39	0.37
marginal time invariant weight	3843	2.4	1.85	0.05	0.14	0.38	-1.18	0.44	-0.57	0.53	-0.04	0.75
marginal panel weight	2162	3	1.88	0.06	0.24	0.42	-0.84	0.49	-0.54	0.64	0.11	0.91
marginal time varying weight	3843	2.4	1.84	0.05	0.16	0.38	-1.21	0.44	-0.65	0.53	-0.08	0.75
marginal no weight	3843	2.4	1.86	0.03	0.22	0.20	-1.51	0.24	-0.15	0.27	-0.25	0.40
marginal invariant weight exchangeable	3843	2.4	2.06	0.05	-0.01	0.36	-1.20	0.42	-0.74	0.52	-0.17	0.68
marginal time varying weight exchangeable	3843	2.4	2.08	0.05	0.00	0.36	-1.33	0.41	-0.89	0.52	-0.32	0.68

Table 4.29: KBIT Matrices Conditional Trajectory Models Fixed Effects on the Intercept with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	3820	2.4	4.15	0.07	0.41	0.47	-4.13	0.50	-4.89	0.80	-1.47	1.04
mixed panel weight	2119	3	3.75	0.09	-0.17	0.59	-3.74	0.66	-3.64	0.97	-0.99	1.09
mixed time varying weight	3820	2.4	4.17	0.07	0.39	0.47	-4.17	0.50	-4.88	0.80	-1.60	1.04
mixed multilevel weights	3820	2.4	4.15	0.07	0.42	0.47	-4.12	0.50	-4.92	0.80	-1.48	1.04
mixed no weight	3820	2.4	4.20	0.04	0.88	0.27	-4.19	0.31	-5.14	0.39	-1.36	0.52
marginal time invariant weight	3820	2.4	3.71	0.07	0.46	0.50	-3.86	0.54	-4.62	0.82	-1.21	1.14
marginal panel weight	2119	3	3.75	0.09	-0.17	0.59	-3.74	0.66	-3.64	0.97	-0.99	1.09
marginal time varying weight	3820	2.4	3.71	0.08	0.48	0.50	-3.86	0.54	-4.68	0.83	-1.27	1.15
marginal no weight	3820	2.4	3.73	0.04	0.79	0.29	-4.02	0.32	-4.74	0.47	-1.19	0.52
marginal invariant weight exchangeable	3820	2.4	4.14	0.07	0.43	0.47	-4.13	0.50	-4.87	0.79	-1.47	1.04
marginal time varying weight exchangeable	3820	2.4	4.17	0.07	0.38	0.48	-4.24	0.51	-5.01	0.81	-1.66	1.05

Table 4.30: KBIT Verbal Conditional Trajectory Models Fixed Effects on the Intercept with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	2905	2.3	-0.13	0.08	0.50	0.37	0.62	0.42	0.48	0.55	-0.90	0.58
mixed panel weight	1601	3	-0.02	0.09	0.35	0.39	0.76	0.44	0.67	0.61	-0.44	0.60
mixed time varying weight	2905	2.3	-0.15	0.09	0.59	0.38	0.57	0.44	0.40	0.57	-0.92	0.59
mixed multilevel weights	2905	2.3	-0.12	0.08	0.48	0.36	0.62	0.42	0.48	0.56	-0.85	0.58
mixed no weight	2905	2.3	-0.02	0.03	-0.18	0.18	0.16	0.21	-0.26	0.26	-0.39	0.35
marginal time invariant weight	2905	2.3	-0.15	0.08	0.64	0.39	0.73	0.44	0.34	0.61	-1.12	0.62
marginal panel weight	1601	3	-0.02	0.09	0.35	0.39	0.76	0.44	0.67	0.61	-0.44	0.61
marginal time varying weight	2905	2.3	-0.17	0.09	0.64	0.39	0.75	0.44	0.29	0.61	-1.01	0.62
marginal no weight	2905	2.3	-0.07	0.04	-0.12	0.19	0.12	0.22	-0.24	0.28	-0.49	0.34
marginal invariant weight exchangeable	2905	2.3	-0.11	0.08	0.49	0.36	0.70	0.42	0.54	0.55	-0.83	0.56
marginal time varying weight exchangeable	2905	2.3	-0.12	0.08	0.47	0.36	0.66	0.42	0.46	0.56	-0.79	0.57

Table 4.31: CDI Conditional Trajectory Models Fixed Effects on the Slope with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	3843	2.4	-0.24	0.03	0.25	0.18	0.04	0.21	0.08	0.27	0.50	0.31
mixed panel weight	2162	3	-0.35	0.03	0.32	0.20	-0.29	0.22	-0.05	0.28	0.46	0.34
mixed time varying weight	3843	2.4	-0.24	0.03	0.25	0.18	0.04	0.21	0.08	0.27	0.49	0.30
mixed multilevel weights	3843	2.4	-0.24	0.03	0.23	0.18	0.04	0.21	0.11	0.27	0.50	0.31
mixed no weight	3843	2.4	-0.21	0.01	0.09	0.10	0.01	0.12	0.05	0.15	0.19	0.21
marginal time invariant weight	3843	2.4	-0.13	0.03	0.23	0.20	0.10	0.23	-0.02	0.30	0.34	0.37
marginal panel weight	2162	3	-0.35	0.03	0.32	0.20	-0.29	0.22	-0.05	0.28	0.46	0.34
marginal time varying weight	3843	2.4	-0.12	0.03	0.20	0.20	0.00	0.23	-0.02	0.30	0.27	0.40
marginal no weight	3843	2.4	-0.11	0.02	0.12	0.11	0.07	0.13	0.00	0.16	0.07	0.21
marginal invariant weight exchangeable	3843	2.4	-0.24	0.03	0.26	0.18	0.05	0.21	0.08	0.27	0.50	0.31
marginal time varying weight exchangeable	3843	2.4	-0.25	0.03	0.24	0.18	-0.01	0.21	0.08	0.27	0.46	0.34

Table 4.32: KBIT Matrices Conditional Trajectory Models Fixed Effects on the Slope with Standard Errors

Method	Level Two n	Level One mean n		age		male		black		hispanic		other
mixed time invariant weight	3820	2.4	-0.50	0.03	0.30	0.18	0.21	0.20	0.06	0.30	-0.34	0.38
mixed panel weight	2119	3	-0.59	0.03	0.28	0.21	0.14	0.23	-0.16	0.35	-0.37	0.42
mixed time varying weight	3820	2.4	-0.50	0.03	0.30	0.18	0.20	0.20	0.07	0.30	-0.36	0.38
mixed multilevel weights	3820	2.4	-0.51	0.03	0.28	0.18	0.20	0.21	0.08	0.30	-0.32	0.37
mixed no weight	3820	2.4	-0.51	0.02	0.07	0.11	-0.22	0.13	0.22	0.17	0.14	0.23
marginal time invariant weight	3820	2.4	-0.26	0.03	0.17	0.24	0.11	0.25	0.03	0.41	-0.52	0.56
marginal panel weight	2119	3	-0.59	0.03	0.28	0.21	0.14	0.23	-0.16	0.35	-0.37	0.42
marginal time varying weight	3820	2.4	-0.26	0.04	0.09	0.24	-0.04	0.25	0.06	0.42	-0.60	0.55
marginal no weight	3820	2.4	-0.26	0.02	0.17	0.14	-0.23	0.16	0.09	0.22	0.08	0.26
marginal invariant weight exchangeable	3820	2.4	-0.50	0.03	0.28	0.18	0.21	0.21	0.07	0.30	-0.35	0.38
marginal time varying weight exchangeable	3820	2.4	-0.51	0.03	0.20	0.18	0.13	0.21	0.08	0.30	-0.35	0.37

Table 4.33: KBIT Verbal Conditional Trajectory Models Fixed Effects on the Slope with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		covariance random effects		random slope variance		error variance	
mixed time invariant weight	2905	2.3	40.57	5.65	-10.46	2.93	5.64	2.43	24.36	1.41
mixed panel weight	1601	3	41.50	6.11	-10.28	2.68	4.92	1.94	22.70	1.07
mixed time varying weight	2905	2.3	38.69	5.64	-9.62	2.99	5.47	2.68	24.74	1.58
mixed multilevel weights	2905	2.3	39.20	5.95	-9.45	3.06	4.96	2.47	24.76	1.48
mixed no weight	2905	2.3	39.43	1.98	-7.98	0.96	4.34	0.71	25.57	0.83

Table 4.34: CDI Conditional Trajectory Models Random Effects with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		covariance random effects		random slope variance		error variance	
mixed time invariant weight	3843	2.4	17.32	1.95	0.84	0.89	-0.04	0.87	13.20	0.76
mixed panel weight	2162	3	15.71	2.04	1.46	0.85	-0.57	0.80	13.50	0.69
mixed time varying weight	3843	2.4	17.05	1.94	0.97	0.88	-0.11	0.86	13.22	0.76
mixed multilevel weights	3843	2.4	16.41	1.99	1.41	0.92	-0.26	0.88	13.24	0.75
mixed no weight	3843	2.4	18.66	0.83	-0.41	0.37	1.07	0.29	13.10	0.37

Table 4.35: KBIT Matrices Conditional Trajectory Models Random Effects with Standard Errors

Method	Level Two n	Level One mean n	random intercept variance		covariance random effects		random slope variance		error variance	
mixed time invariant weight	3820	2.4	44.85	3.15	0.36	1.19	1.51	1.07	11.83	0.70
mixed panel weight	2119	3	41.37	3.68	0.36	1.31	1.39	1.17	12.11	0.76
mixed time varying weight	3820	2.4	44.56	3.14	0.55	1.18	1.47	1.06	11.79	0.71
mixed multilevel weights	3820	2.4	44.40	3.27	0.71	1.27	1.41	1.13	11.87	0.74
mixed no weight	3820	2.4	50.77	1.53	-1.95	0.50	3.36	0.31	11.28	0.33

Table 4.36: KBIT Verbal Conditional Trajectory Models Random Effects with Standard Errors

CHAPTER 5

Conclusions

Sampling and nonsampling error can have dramatic effects on estimates and are common problems for analysis of survey data. Random measurement error will have deleterious effects on estimates in many circumstances. Random measurement error decreases the efficiency of estimators and may result in biased estimates of associations. Longitudinal structural equation models may be used to assess the degree of random measurement error by providing estimates of reliable and unreliable variance components. Chapter 2 of this dissertation proposes models for measuring reliability of scale scores using panel survey data. Estimators may also be biased and inconsistent as a result of unequal probabilities of selection. Unequal probabilities of selection are common to survey data collected using complex sampling designs. The potential for bias was demonstrated in two separate studies in this dissertation presented in Chapters 3 and 4. Probability weighting is a method used to correct the biases due to unequal selection. Several weighting methods were assessed in Chapters 3 and 4 in the context of multilevel or mixed modeling as to their performance in producing consistent and efficient estimates under informative selection. The research in this dissertation adds to our understanding of evaluating and reducing the effects of survey errors that arise with analysis of complex samples.

Reliability estimates for a number of scale scores are obtained using longitudinal, latent variable models in Chapter 2. The models are very general in that they do not require model assumptions that are necessary for assessing reliability in single-item panel

measures, namely, assumptions of constant variance over time. In addition, the proposed models allow for the partitioning of item specific error variances as well as variance due to additional factors that are not attributable to the primary trait. One limitation of the item-level factor models proposed is that for scale scores with many items, the estimation of all three variance components: trait variance, specific error variance, and additional factor variance, can be unwieldy and lead to empirical underidentification. However, one of the benefits of the proposed models is that they may be reduced by placing further assumptions on the models without assuming constant variance over time. For example, the additional factor variance may include all possible additional item covariance or just a component of that covariance. The models are also useful in that the standard errors for the reliability estimates are provided using standard SEM software.

In the evaluation of reliability for the scales in the NSCAW, several of the scale scores contain substantial random measurement error. For example, many estimates from the item level models are below 0.7 and as low as 0.45. The analysis also revealed that the assumptions of constant error variance over time had little effect on reliability estimates as compared to the assumption of no specific error variance or no additional factor variance. Therefore, the ability to partition these additional variance components proved critical to accurate reliability estimates. However, while the models with estimates of additional variance are desirable in their generality, the number of estimated parameters can becomes unwieldy and empirical underidentification can occur. Further research should focus on the specific empirical identification problems with these models and whether they would be common in practice. As well, simpler models may produce the same reliability estimates in some conditions. For example, with two indicator factor models the additional factor variance will go into the residual shocks and therefore be included in stable variance without having to specify this variance separately with an additional factor. This model is more stable and provides the same reliability estimates, but specific model parameters are biased and the model fit is poorer. Whether the more

parsimonious models are equivalent to the more general models when there are more than two items per factor is unknown and the efficiency of the reliability estimates should also be compared across these approaches.

Probability weighting in multilevel or mixed models is a relatively new methodology that combines the traditional design-based weighting approach to dealing with unequal inclusion of observations with the traditional model-based approach to incorporating clustering. In Chapter 3 of this dissertation, several alternative weighting methods for analyzing a two-level model using data selected with unequal probabilities are evaluated. The main point of this study is to compare and contrast the current gold standard weighting approach to the alternative method of incorporating sample design variables directly in the model. The hypothesis is that the latter approach will result in estimates with lower total loss as measured by mean squared error and coverage. The primary limitation of the multilevel weighting approach is that the appropriate weights are often more volatile resulting in weighted estimators that are less efficient. While the upside of including sample design variables is that estimators are more efficient. This latter approach is also somewhat more feasible when weighting clusters because the number of sample design variables used in unequal selection are usually reasonable for selection of clusters in the common multistage sampling designs. In addition to the contrast of using multilevel weights versus incorporating design variables, is the assessment of using single-level weighting instead of the proper multilevel weights. This method is evaluated primary to assess the deficiencies of an approach that is probably employed by many analysts who do not have access to the multilevel weights.

Results of the simulation analysis indicate that RMSE and coverage rates can be improved by including design variables directly in the model compared to using multilevel weighting. Limitations of this method are that the analyst needs access to the sample design variables used in selecting clusters. In order to offset biases due to selection, specification of the model at the cluster level must also be correct in terms of the effects of

the design variables. Even if specification is not exact, it may be better to use the variable approach rather than the weighting approach when sample sizes are small. Alternatively, with large sample sizes at both levels, the multilevel modeling method is a good method that is easy to apply when the proper weights are available. Even so, with sample sizes as large as 150 level two units and 75 level one units per level two unit, the variable method can greatly improve RMSE and coverage as demonstrated in Chapter 3.

Chapter 4 also includes an evaluation of weighting methods for a mixed model; however, the focus is on longitudinal data and the linear growth curve model. Also, the methods compared in this chapter include weighting in the marginal model framework in addition to weighting in the mixed model framework. In the longitudinal situation, unequal inclusion of observations at the level of time is different from unequal inclusion due to the sample design. Intermittent dropout over time is common for panel studies and there is often information about individuals from time points where they are observed that can be used to specify a model that is robust to time specific dropout. Also, the weights provided to analysts for time specific nonresponse differ. For example, panel weights have traditionally been used to deal with intermittent nonresponse and there is no comparable weight for standard (not longitudinal) multilevel data. In the Chapter 4 analysis, different weighting approaches in the context of the mixed and marginal models are compared. And, it was shown that the GLS/ML estimators for fixed effects in the two types of models are equivalent when the same variance weight matrix is used. It was found that the best weighting approach depends on whether the variance weighting is used or not.

Results from the simulation analysis show that panel weighting performs well with adequate sample sizes and results are equivalent under the mixed and marginal models. However, the best performance resulted from time-varying weighting using all available person-by-time observations in the marginal model with no covariance weighting, although this method does not provide random effects estimates. Multilevel weights in the

mixed model do provide random effects estimates and perform well with large level one and level two sample sizes. Common results across Chapters 3 and 4 are that weighting in the multilevel or mixed model framework seems to require much larger samples for obtaining good random effects estimates, particularly when the model includes more than just a random intercept term. The empirical example results from Chapter 4 seem to confirm the simulation analysis results from Chapter 3. That is, if the model is specified such that the unequal inclusion is MAR, probability weighting is not necessary. In the longitudinal case, this assumption seems more plausible since the observed measures from the same individual at other time should provide more information for missing time points than observed measures from individuals within a cluster would provide for other individuals in that cluster. In both chapters, sampling weights prove necessary and useful when missingness is MNAR or nonignorable.

Future research should include evaluations of the intersection of design based and model based estimation approaches. Some research has begun which uses Bayesian methods for modeling both the sampling design and the analytic model (Pfeffermann, Da Silva Moura, & Silva, 2006; Little, 2004). The Bayesian model methods improve estimator efficiency without overburdening the theoretical model with sample design variables. But, these methods have not been used much in practice. Another modeling approach to informative selection that could be investigated and compared to weighting is the use of multiple imputation (Schafer, 1997; Little & Rubin, 1987). In some circumstances, for example drop out in longitudinal data analysis, it may be desirable to impute missing data using sample design variables in the imputation model rather than weight in the analytic model. Multiple imputation may prove more efficient under certain circumstances compared to probability weighting and would not overburden the analytical (or theoretical model). Additionally, weighting could be considered as an alternative to model based methods for dealing with missing data. Most missing data methods in the model based literature do not work for data that are MNAR. However, some modeling methods have

been proposed such as pattern mixture modeling (Hedeker & Gibbons, 1997) and these could be compared to probability weighting.

This dissertation research contributes to our ability to assess and correct for survey errors. Chapter 2 contributes to an historic and vast literature about reliability. In this chapter, a general, longitudinal, latent variable model useful for estimating scale score reliability is outlined. The model is used to test assumptions about constant error variance over time common to single-indicator panel models and to estimate reliability in a less restrictive model. This research contributes to the literature by specifying a new general model for reliability estimation as well as results from an additional testing situation for the assumptions of single-indicator reliability models. Chapters 3 and 4 contribute to a new area of literature that deals with probability weighting for unequal selection and nonresponse in models of multilevel data. Optimal weighting and estimation combinations are offered for two level models of data selected with PPS sampling and growth curve models with intermittent nonresponse. This research contributes new information about how estimation using multilevel weighting compares to unweighting estimation using modeling of the sample design for clusters. It also provides a comparison of alternative weighting approaches, some which have not been considered before, specific to growth curve modeling. These methods have not heretofore been evaluated relative to one another.

CHAPTER 6

Appendix

Child Behavior Checklist (CBCL), Youth Self-Report (YSR), and Teacher Report Form (TRF) Caregivers, youth, and teachers reported on children's competencies and problem behavior in the Child Behavior Checklist, Youth Self-Report and the Teacher Report Form (Achenbach 1991a, 1991b, 1991c, and 1991d). The problem scale is composed of eight syndromes (Withdrawn, Somatic Complaints, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquent Behavior, and Aggressive Behavior) and an Other Problems category. Behaviors are also categorized as Externalizing (containing the Delinquent and Aggressive Behavior syndromes) or Internalizing (containing the Withdrawn, Somatic Complaints, and Anxious/Depressed syndromes). A Total Problems score is derived from the total of the syndromes and Other Problems items (Achenbach, 1991a). Items for the CBCL, YSR, and TRF are on a 3-point Likert-type scale. Items are summed to produce the scales for internalizing, externalizing, and total problem behavior. Items for the CBCL were different for the 2 - 3-year-old age group and the 4 - 18-year-old age group. Scales were combined for the split-half models.

Social Skills Rating System (SSRS) and Vineland Adaptive Behavior Scale (VABS) Screener - Daily Living In the NSCAW, the Social Skills Rating System measures parent (SSRS) and teacher (SSRST) perception of the child's social skills in four domains: cooperation, assertion, responsibility, and self-control (Gresham and Elliott,

1990). Items are three-point ordinal scales, which are summed to create the SSRS scores. The Vineland Adaptive Behavior Scale (VABS) is was used to measure daily living skills among children aged 0 to 10 years as assessed by caregivers (Sparrow, et al., 1984). A 15-item Daily Living Skills domain was used and is part of the 261-item Vineland Adaptive Behavior Scale. This domain measures personal skills (e.g., how the child eats, dresses, and performs personal hygiene), domestic skills (household tasks the child performs), and community skills (how the child spends his or her time, and telephone skills). Separate scores were computed for the three age groups, 0-1, 2-5 and 6-10 as the sum of the 15 items in the domain. Items are measured on a 3 point ordinal scales.

Children's Depression Inventory (CDI) The Children's Depression Inventory (CDI) measures depression by asking various questions of children about their engagement in certain activities or their experience of certain feelings (Kovacs, 1992). CDI contains 27 items, each with a 3-point Likert-type scale (0 = absence of symptom, 1 = mild symptom, 2 = definite symptom) that addresses a range of depressive symptoms as indicated by five factors: Negative Mood, Interpersonal Problems, Ineffectiveness, Anhedonia, and Negative Self-Esteem. Scores are the sum of the items responses for each factor. We estimated the reliability of the total CDI scale, which includes all five depression factors.

Peer Loneliness and Social Dissatisfaction Questionnaire for Young Children The Peer Loneliness and Social Dissatisfaction questionnaire (Asher and Wheeler, 1985) is designed to measure peer relationships, including social rejection with questions about success in making and keeping friendships and school adjustment. The questionnaire is administered only to children in school. All items are measured on a three-point ordinal scale and are summed to produce a total score.

School Engagement School engagement scale is comprised of the sum of four-point ordinal scales measuring student's subjective achievement and their disposition toward

learning and school. Items were derived from the Drug Free Schools (DFSCA) Outcome Study Questionnaire (US Department of Education: Office of the Under Secretary). Engagement is measured for children 6 years and older who are in school.

The Home Observation for Measurement of the Environment Short Form (HOME-SF) HOME-SF measures the quality and quantity of stimulation and support in the home environment of children from birth to 10 years (Caldwell and Bradley, 1984; Bradley, 1994). Items address the mother's behaviors toward the child and various aspects of the physical environment (e.g., safe play environment, size of living space). During the caregiver interview, the interviewer indicates whether these conditions exist, do not exist, or were not observed. HOME-SF is a short form version of the HOME scale. Items making up the HOME-SF have varying ordinal scales and are subsequently dichotomized prior to summing into cognitive stimulation, emotional support, and total scales.

Research Assessment Package for Schools - Self-Report Instrument for Middle School Students (RAPS-SM) "A shorter version of the Relatedness scale from RAPS-S (IRRE, 1998) was used to measure children's feelings about their relationship with their primary and secondary caregivers. There were two sets of questions, one for each caregiver. Four subscales were used for NSCAW: Parental Emotional Security, Involvement, Autonomy Support, and Structure. Children answered how true each statement was (1 = not at all true, 2 = not very true, 3 = sort of true, and 4 = very true). Parental Emotional Security asked how true it was that the child felt good, mad, or happy with his or her caregiver. Involvement asked questions about the caregiver's interest in, time spent with, and things done to help the child. Autonomy Support inquired about the caregiver's trust of the child and whether the child was allowed to make his or her own decisions. Structure asked about the caregiver's fair treatment of the child, the caregiver's belief in the child's abilities, and the child's understanding of what the caregiver wants" (Dowd, et al., 2004: Appendix III, DFUM.) A mean rather than a summed relatedness score was created to account for the fact that not all children answered the

same number of questions.

Short-Form Health Survey (SF-12) SF-12, a shorter version of the SF-36 (12 versus 36 items), measures mental and physical health of the caregiver (Ware, et al., 1996, 1998). The same twelve items make up the physical and mental health raw scale scores. Items are dichotomous and ordinal in original form. Scales are created from the weighted sum of the items after they are dichotomized. The scoring steps were performed as described by the publisher. See the DFUM Appendix III and Ware, et al. (1998).

Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA) MBA is a brief, wide-range test of basic skills and knowledge, including tests of reading, mathematics, writing, and factual knowledge (science, social studies, and humanities). MBA may be used with children and adults aged 4 to over 90 years (Woodcock, McGrew, Werder, 1994). NSCAW utilized MBA with children aged 6 and older and administered only the Reading and Math tests. A file of raw scores was generated for all children who completed the section. Administered assessments included Reading letter-word identification (A), Vocabulary (B), and Comprehension (C), Mathematics (D) and General Knowledge (E). MBA reading is comprised of assessments A, B, and C and MBA math is comprised of assessments D and E. Raw scores for each section were computed as the sum of the correct items in the subtest plus a base score. For items in the analysis models, the basescore is distributed across all items equally. In the NSCAW sample, some items in the MBA reading and math scale did not discriminate between cases since very few or no children gave correct answers. Items with variances less than 0.005 were removed from the scale.

Preschool Language Scale-3 (PLS-3) PLS-3 measures language development of children from birth to 5 years (Zimmerman, Steiner, and Pond, 1992). The Auditory Comprehension subscale measures precursors of receptive communication skills with tasks focusing on attention abilities. The Expressive Communication subscale measures pre-

cursors of expressive communication skills with tasks that focus on social communication and vocal development. Items in the subscales are dichotomous (1=correct, 0=incorrect).

References

Achenbach, T. M. (1991a). *Manual for the child behavior checklist 2 - 3 and 1991 profile*. Burlington, Department of Psychiatry, University of Vermont.

Achenbach, T. M. (1991b). *Manual for the child behavior checklist 4 - 18 and 1991 profile*. Burlington, Department of Psychiatry, University of Vermont.

Achenbach, T. (1991c). *Manual for the youth self-report and 1991 profile*. Burlington, Department of Psychiatry, University of Vermont.

Achenbach, T. M. (1991d). *Manual for the teacher's report form and 1991 profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Agresti, A. (2002). *Categorical Data Analysis*, Second Edition. New York: Wiley.

Alanen, E., Leskinen, E., and Kuusinen (1998). Testing Equality of Reliability and Stability with Simple Linear Constraints in Multi-Wave, Multi-Variable Models. *British Journal of Mathematical and Statistical Psychology, 51*, 327-341.

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Anastasi, A. (1988). *Psychological Testing.* New York: MacMillan.

Anderson, T. W. (1959). Some Stochastic Process Models for Intelligence Test Scores. In K.J. Arrow, S. Karlin and P. Suppes, (Eds.) *Mathematical Methods in the Social Sciences.* Stanford: Stanford University Press.

Arbuckle, J. L. (1996). Full Information Estimation in the Presence of Incomplete Data. In G. A. Marcoulides and R. E. Schumacker (Eds.). *Advanced Structural Equation Modeling: Issues and Techniques.* Hillsdale, NJ: Erlbaum, 243-277.

Asher, S., and Wheeler, V. (1985). Children's loneliness: a comparison of rejected and neglected peer status. *Journal of Consulting and Clinical Psychology, 53*, 500-505.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods, 35*, 439-460.

Asparouhov, T. and Muthén, B. (2006a). *Multilevel modeling ofcomplex survey data*. Presented at the Joint Statistical Meeting in Seattle, August 2006.

Asparouhov, T. and Muthén, B. (2006b). Comparison of Estimation Methods for Complex Survey Data Analysis. online paper
http://www.statmodel.com/download/SurveyComp21.pdf

Bentler, P. M. (2005). Covariance Structure Models for Maximal Reliability of Unit-weighted Composites. In Lee S. (Ed.) *Handbook of Structural Equation Models.* Amsterdam: Elsevier.

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606

Biemer, P. P., Christ, S. L. (2008). Constructing the Survey Weights. in Levy, P. S. and Lemeshow, S. (eds) *Sampling of Populations: Methods and Applications.* 4th Ed. New York: Wiley.

Biemer, P. P., Christ, S. L. (2007). Weighting Survey Data, In Hox, J., de Leeuw, E. and Dillman, D.A. (eds) *The International Handbook of Survey Methodology.* New York: Lawrence Erlbaum Associates.

Biemer, P. P., Christ, S. L., and Wiesen, C. (under review). *Psychological Methods.* A General Approach for Estimating Scale Score Reliability for Longitudinal Data.

Binder, D. (1983). On the Variance of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review, 51*, 279-292.

Binder, D. A. and Roberts, G. A. (2003). Design-based and Model-based Methods for Estimating Model Parameters. In Chambers, R. L. and Skinner, C. J., Eds. *Analysis of Survey Data.* New York: Wiley.

Blalock, H. M. (1970a). A Causal Approach to Nonrandom Measurement Errors. *The American Political Science Review, 64*, 1099-1111.

Blalock, H. M. (1970b). Estimating Measurement Error using Multiple Indicators and Several Points in Time. *American Sociological Review, 35*, 101-111.

Bollen, K. A. (1980). Issues in the Comparative Measurement of Political Democracy. *American Sociological Review, 45*, 379-390.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: Wiley.

Bollen, K. A. and Curran, P. J. (2006). *Latent Curve Models : A Structural Equation Perspective.* New York: Wiley.

Bradley, R. (1994). The HOME Inventory: Review and reflections. In H. Reese (Ed.), *Advances in child development and behavior.* (pp. 241-288). San Diego, CA; Academic Press.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Browne, M. W. and Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In K. Bollen and J.S. Long (Eds.), *Testing Structural Equation Models*. (pp.136-359). New York: Plenum Press.

Caldwell, B., and Bradley, R. (1984). *Home Observation for Measurement of the Environment*. Little Rock, AR: University of Arkansas at Little Rock.

Campbell, D. T. and Cook, T. D. (1979). Quasi-Experimentation, Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

Chromy, J. R. (1978). Sequential Sample Selection Methods. Proceedings of the Survey Research Methods Section, 401-406.

Cochran, W. G. (1977). *Sampling Techniques*, Third Edition. New York: Wiley.

Coenders, Germà, Saris, W. E., Batista-Foguet, J. M., and Andreenkova, A. (1999). Stability of Three-Wave Simplex Estimates of Reliability. *Structural Equation Modeling, 6*, 135-157.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cudeck, R., Du Toit, S., and Sörbom, D. (Eds). (2001). Structural Equation Modeling: Present and Future – A Festschrift in honor of Karl G. Jöreskog. Lincolnwood: Scientific Software International.

Diggle P. J., Heagerty P. J., Liang K-Y, Zeger S. L. (2002). *Analysis of Longitudinal Data*, Second Edition, Oxford University Press.

Diggle, P. J., and Kenward, M. G. (1994). Informative Drop-out in Longitudinal Data Analysis. *Applied Statistics, 43*, 49-93.

Dowd, K., S. Kinsey, S. Wheeless, S. Suresh and the NSCAW Research Group (2004). National Survey of Child and Adolescent Well-being: Combined Waves 1-4 Data File User's Manual. Research Triangle Park, NC: RTI International.

Duncan, S. C., and Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research, 29*, 313-338.

du Toit, S. (2006). Notes on the Implementation of Weights in Level-2 Models. unpublished document.

du Toit, S. (2006). The Analysis of Multilevel Models with Continuous Outcomes in the Case of Data with Weight Variables. LISREL Technical Document. http://www.ssicentral.com/lisrel/techdocs/mlevelw.pdf

duToit, S. (1979). *The Analysis of Growth Curves. Doctoral Dissertation*, University of South Africa. Obtained from the author.

Ferrando, P. J. (2003). Analyzing Retest Increases in Reliability: A Covariance Structure Modeling Approach. *Structural Equation Modeling, 10*, 222-237.

Folsom, R. E., Potter, F. J., and Williams, S. R. (1987). Notes on a Composite Size Measure for Self-weighting Samples in Multiple Domains. American Statistical Association Meeting, Section of Survey Research Methods.

Gerbing, D. W., and Anderson, J. C. (1984). On the Meaning of within-Factor Correlated Measurement Errors. *The Journal of Consumer Research, 11*, 572-580.

Goldstein, H. (1986). Multilevel Linear Mixed Model Analysis Using Iterative Generalised Least Squares. *Biometrika, 73*, 43-56.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd ed. London: Arnold.

Goren, P. (2005). Party Identification and Core Political Values. *American Journal of Political Science, 49*, 881-896.

Graubard, B. I. and Korn, E. L. (1996). Modelling the Sampling Design in the Analysis of Health Surveys. *Statistical Methods in Medical Research, 5*, 263-281.

Gresham, F., and Elliott, S. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.

Grilli, L. and Pratesi, M. (2004). Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs. *Survey Methodology, 30*, 93-103.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association, 78*, 776-793.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sampling Survey Methods and Theory*. New York: Wiley.

Hedeker, D., and Gibbons, R. D. (1997). Application of Random-effects Pattern-mixture Models for Missing Data in Longitudinal Studies. *Psychological Methods, 2*, 64-78.

Heise, D. R. (1969). Separating Reliability and Stability in Test-Retest Correlation. *American Sociological Review, 34*, 93-101.

Heise, D. R. (1970). Comment on "The Estimation of Random Measurement Error in Panel Data." *American Sociological Review, 35*, 117.

Heise, D. R and Bohrnstedt. (1970). Validity, Invalidity, and Reliability. *Sociological Methodology, 2*, 104-129.

Hogan, T. P., Benjamin, A., and Brezinski, K. L. (2000). Reliability Methods: A Note on the Frequency of Use of Various Types. *Educational and Psychological Measurement, 60*, 523-531.

Howard, K. I. (1964). Differentiation of Individuals as a Function of Repeated Testing. *Education and Psychological Measurement, 24*, 875-894.

Howard, K. I. and Diesenhaus, H. (1965). Item Response Patterns as a Function of Repeated Testing. *Education and Psychological Measurement, 25*, 365-379.

Hoyt. C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

Jackson, R. W. B. (1939). Reliability of Mental Tests. *British Journal of Psychology, 29*, 267-287.

Jagodzinski, W. and Kühnel, S. M. (1987). Estimation of Reliability and Stability in Single-Indicator Multiple-Wave Models. *Sociological Methods and Research, 15*, 219-258.

Jagodzinski, W., Kühnel, S. M., and Schmidt, P. (1987). Is There a "Socratic Effect" in Nonexperimental Panel Studies? *Sociological Methods and Research, 15*, 259-302.

Jöreskog, K. G. (1970). Estimation and Testing of Simplex Models. *The British Journal of Mathematical and Statistical Psychology, 23*, 121-145.

Jöreskog, K. G. (1971). Statistical Analysis of Sets of CongenericTests. *Psychometrika, 36*, 109-133.

Jöreskog, K. G. (1979). Statistical Models and Methods for Analysis of Longitudinal Data. In Jöreskog and Sörbom (Eds.) *Advances in Factor Analysis and Structural Equation Models*. Cambridge: Abt Books.

Kish, L. (1965). *Survey Sampling*. Wiley: New York.

Korn, E. L. and Graubard, B. I. (2003). Estimating Variance Components by Using Survey Data. *Journal of the Royal Statistical Society, B, 65*, 175-190.

Kovačević, M. S. and Rai, S. N. (2003). A Pseudo Maximum Likelihood Approach to Multilevel Modelling of Survey Data. *Communications in Statistics, 32*, 103-121.

Kovačević, M. S., Rong, H., and You, Y. (2006). Bootstrapping for Variance Estimation in Multi-Level Models Fitted to Survey Data. Proceedings of the Survey Research Methods Section, ASA.

Kovacs, M. (1992). *Children's depression inventory*. North Tonawanda, NY, Multi-Health Systems, Inc.

Levy, P. S., and Lemeshow, S. (2008). *Sampling of Populations: Methods and Applications*. 4th ed. New York: Wiley.

Liang, K-Y. and Zeger, S. L., (1986). Longitudinal Analysis Using Generalized Linear Models. *Biometrika, 73*, 13-22.

Liang, K.-Y.,and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13-22.

LISREL (2008). Chapter 4: Multilevel Models.
http://www.ssicentral.com/lisrel/techdocs/mlevelw.pdf

Little, R. J. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association, 99*, 546-556.

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112-1121.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Long, S. J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Longford, N. T.(1995). Model-based Methods for Analysis of Data from 1990 NAEP Trial State Assessment. *Research and Development Report NCES 95-696*. Washington D.C. National Center for Education Statistics.

Longford, N. T.(1996). Model-based Variance Estimation in Surveys with Stratified Clustered Designs. *Australian Journal of Statistics, 38*, 333-352.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McArdle, J. J. (1986). Latent Variable Growth Within Behavior Genetic Models. *Behavior Genetics, 16*, 163-200.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum ASsociates, Inc.

McGuire, W. J. (1960). Congnitive Consistency and Attitude Change. *Journal of Abnormal and Social Psychology, 3*, 345-353.

Mehta, P. D., and West, S. G. (2000). Putting the Individual Back into Individual Growth Curves. *Psychological Methods, 5*, 23-43.

Meredith, W., and Tisak, J. (1990). Latent Curve Analysis. *Psychometrika, 55*, 107-122.

Muthén B. O. and Muthén, L. K. (1998-2006). *Mplus User's Guide*. Fourth Edition. Los Angeles: Muthén and Muthén.

Nathan, G. and Holt, D. The Effect of Data from Complex Surveys. *Journal of the Royal Statistical Society, Series A, 143*, 474-487.

Nathan, G. and Smith, T. M. F. (1989). The Effect of Selection on Regression Analysis. In Skinner, C. J., Holt, D., and Smith, T. M. F., Eds. *Analysis of Complex Surveys*. New York: Wiley.

Palmquist, B. and Green, D. P. (1992). Estimation of Models with Correlated Measurement Errors from Panel Data. *Sociological Methodology, 22*, 119-146.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review, 61*, 317-337.

Pfeffermann, Da Silva Moura, and Silva (2006). Multilevel Modelling under Informative Sampling. *Biometrika, 93*, 943-959.

Pfeffermann, D. and LaVange, L. M. (1989). Regression Models for Stratified Multi-Stage Cluster Samples. In Skinner, C.J., Holt, D., and Smith, T. M. F. (Eds.) *Analysis of Complex Surveys*. pp.237-260. New York: Wiley.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of Royal Statistical Society. Series B, Statistical Methodology, 60*, 23-40.

Prentice, R. L. (1988). Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics, 44*, 1033-1048.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel Modelling of Complex Survey Data. *Journal of Royal Statistical Society. Series A, Statistics in Society, 169*, 805-827.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69*, 167-190.

Raffalovich, L. E., and Bohrnstedt, G. W. (1987). Common, Specific, and Error Variance Components of Factor Models: Estimation with Longitudinal Data. *Sociological Methods and Research, 15*, 385-405.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods*. 2nd Ed. Thousand Oaks: Sage.

Raykov, T. (2004). Estimation of Maximal Reliability: A Note on a Covariance Structure Modelling Approach. *British Journal of Mathematical Statistical Psychology, 57*, 21-27.

Raykov, T. (2001). Studying Change in Scale Reliability for Repeated Multiple Measurements via Covariance Structure Modeling. In Cudeck, R., Du Toit, S., and Sörbom, D. (Eds). *Structural Equation Modeling: Present and Future – A Festschrift in honor of Karl G. Jöreskog*. Lincolnwood: Scientific Software International.

Raykov, T. (2001). Bias of Coefficient Alpha for Fixed Congeneric Measures with Correlated Errors. *Applied Psychological Measurement, 25*, 69-76.

Raykov, T. and Shrout, P. E. (2002). Reliability of Scales with General Structure: Point and Interval Estimation using a Structural Equation Modeling Approach. *Structural Equation Modeling, 9*, 195-212.

Raykov, T. and Tisak, J. (2004). Examining Time-invariance in Reliability in Mulit-wave, Multi-indicator Models: A Covariance Structure Analysis Approach Accounting for Indicator Specificity. *British Journal of Mathematical and Statistical Psychology, 57*, 253-263.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Saris, W. E., and Aalberts, C. (2003). Different Explanations for Correlated Distrubance Terms in MTMM Studies. *Structural Equation Modeling, 10*, 193-213.

Saris, W. and Andrews, F. M. (1991). Evaluation of Measurement Instruments using a Structural Modeling Approach. In Biemer, Groves, Lyberg, Mathiowetz, and Sudman (Eds.). *Measurement Errors in Surveys*. New York: Wiley.

Satorra, A. and Bentler, P. M. (1988). Scaling Corrections for chi-square Statistics in Covariance Structure Analysis. 1988 Proceedings of the Business and Economic Statistics Section of the ASA, 308-313.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Schafer, J. L., and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 147-177.

Scott, A. J. and Wild, C. J. (1989). Selection Based on the Response Variable in Logistic Regression. In Skinner, C. J., Holt, D., and Smith, T. M. F., Eds. *Analysis of Complex Surveys*. New York: Wiley.

Skinner, C. J. (1989). Domain Means, Regression and Multivariate Analysis. In C. J. Skinner and D. Holt and T. M. F. Smith (eds.), *Analysis of Complex Surveys*. pp. 59-87. New York: Wiley.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

Skinner, C. J., and Holmes, D. J. (2003). Random Effects Models for Longitudinal Survey Data. In Chambers, R. L. and Skinner, C. J., Eds. *Analysis of Survey Data*. New York: Wiley.

Skinner, C. J. and Vieira, M. D. T. (2007). Variance Estimation in the Analysis of Clustered Longitudinal Data. *Survey Methodology, 33*, 3-12.

Smith, T. M. F. and Holmes, D. J. (1989). Multivariate Analysis. In Skinner, C. J., Holt, D., and Smith, T. M. F., Eds. *Analysis of Complex Surveys*. New York: Wiley.

Sparrow, S. S., Balla, D. A., and Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales: Interview Edition, Survey Form Manual*. Circle Pines, MN: American Guidance Service.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Steiger, J. H., and Lind, J. C. (1980). Statistically-based tests for the number of common factors. Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City. May 30, 1980.

Ware, J., Kosinski, M., and Keller, S. (1998). *SF-12: How to score the SF-12 physical and mental health summary scales* (3 ed.). Lincoln, RI: Quality Metric Incorporated.

Ware JE, Kosinski M, and Keller SD. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220-233.

Werts, C. E., Jöreskog, K. G., and Linn, R. L. (1971). Comment on "the Estimation of Measurement Error in Panel Data." *American Sociological Review, 36*, 110-113.

Werts, C. E., Linn, R. L. (1970). Cautions in Applying Various Procedures for Determining the Reliability and Validity of Multiple-Item Scales. *American Sociological Review, 35*, 757-759.

Willett, J. B., and Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Psychological Bulletin, 116*, 363-381.

Wiley, D. E. and Wiley, J. A. (1970). The Estimation of Measurement Error in Panel Data. *American Sociological Review, 35*, 112-117.

Wiley, J. A. and Wiley, M. G. (1974). A Note on Correlated Errors in Repeated Measurements. *Sociological Methods and Research, 3*, 172-188.

Woodcock, R., McGrew, K., and Werder, J. (1994). *Woodcock-McGrew-Werder Mini-Battery of Achievement*. Itasca, IL: Riverside Publishing.

Zeger, S. L., and Liang K-Y., (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics, 42*, 121-130.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics, 44*, 1049-1060.

Zimmerman, I., Steiner, V., and Pond, R. (1992). *Preschool language scale-3: examiner's manual*. San Antonio, TX: The Psychological Corporation.