A Bayesian Synthesis Approach to Data Fusion

Using Augmented Data-Dependent Priors

by

Katerina M. Marcoulides

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2017 by the
Graduate Supervisory Committee:

Kevin Grimm, Co-Chair
Roy Levy, Co-Chair
David Mackinnon
Hye Won Suk

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

The process of combining data is one in which information from disjoint datasets sharing at least a number of common variables is merged. This process is commonly referred to as data fusion, with the main objective of creating a new dataset permitting more flexible analyses than the separate analysis of each individual dataset. Many data fusion methods have been proposed in the literature, although most utilize the frequentist framework. This dissertation investigates a new approach called Bayesian Synthesis in which information obtained from one dataset acts as priors for the next analysis. This process continues sequentially until a single posterior distribution is created using all available data. These informative augmented data-dependent priors provide an extra source of information that may aid in the accuracy of estimation. To examine the performance of the proposed Bayesian Synthesis approach, first, results of simulated data with known population values under a variety of conditions were examined. Next, these results were compared to those from the traditional maximum likelihood approach to data fusion, as well as the data fusion approach analyzed via Bayes. The assessment of parameter recovery based on the proposed Bayesian Synthesis approach was evaluated using four criteria to reflect measures of raw bias, relative bias, accuracy, and efficiency. Subsequently, empirical analyses with real data were conducted. For this purpose, the fusion of real data from five longitudinal studies of mathematics ability varying in their assessment of ability and in the timing of measurement occasions was used. Results from the Bayesian Synthesis and data fusion approaches with combined data using Bayesian and maximum likelihood estimation methods were reported. The results illustrate that Bayesian Synthesis with data driven priors is a highly effective approach, provided that the sample sizes for the fused data are large enough to provide unbiased estimates.

Bayesian Synthesis provides another beneficial approach to data fusion that can effectively be used to enhance the validity of conclusions obtained from the merging of data from different studies.

DEDICATION

I am eternally grateful to my wonderful parents for their constant encouragement, support, and unconditional love. I am so lucky you have always made my education and happiness a priority. I appreciate all the sacrifices you made throughout your lives to make mine easier. You are both such incredible role models and without your guidance throughout my life I would not be where I am today and I never could have finished my dissertation without you.

I also want to thank my amazing fiancé, Joe Barbour, for his love, empathy, support, and encouragement, not only throughout this dissertation process, but every day. Without our countless study dates, afternoon coffees, and sushi nights in, I never could have finished this dissertation.

While he will never know how much he has helped me complete this dissertation, I want to acknowledge my dog, Sheftali, for all of his kisses and late night cuddles that helped me get through so many all-nighters.

Finally, I want to thank my grandfather for all of his encouragement and interest in my education. I am so appreciative of your pride in me, and how you are always excited to hear about my accomplishments throughout graduate school.

ACKNOWLEDGMENTS

I would like to express my great appreciation to my advisor Kevin Grimm and my co-chair Roy Levy for their countless hours of help and guidance, both with this dissertation as well as my academic career. Without Roy's classes in Bayesian methods I would not have had the idea for this dissertation, and without Kevin's class in item response theory and the opportunity to be his teaching assistant for longitudinal growth modeling I would not have the knowledge and the tools to complete this dissertation. I would also like to thank my committee members Dave MacKinnon and Hye Won Suk for their valuable advice and feedback during the various stages of preparing this dissertation. Finally, I want to express my appreciation to Keith Widaman for his help and guidance throughout my graduate education, both here at ASU and while at UC Davis, and to Emilio Ferrer who taught my first seminar on longitudinal growth modeling introducing me to the topic and for serving as a committee member for my graduate degree milestones at UC Davis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Statement of the Problem

Recent concerns that social and behavioral science studies may suffer from a lack of replicability have prompted researchers to modify their thinking and seek out new data analytic strategies to extract information from a body of related work. Extracting information from multiple sources and merging them together may provide different information than that provided by each separate source. In order to update beliefs in light of new evidence from the different sources requires a process of learning, or acquiring information, that is sequential (Jackman, 2009).

The process of sequential learning to arrive at conclusions that might be different to those from each separate source has a long history in the social and behavioral sciences. An early example is an analysis performed by Pearson (1904) in a study in which he tried to determine the relationship between mortality and inoculation with a vaccine for enteric (intestinal) fever. Using data gathered from five small sample studies, he averaged the calculated estimates of the correlation coefficient in each study and compared this value to typical correlation values computed for other vaccines. Pearson (1904) justified this action by indicating that this averaging was necessary because the samples from each study were "far too small to allow any definitive opinion being formed [and it was better] to group them" (pg. 1243) into a larger data series. Although it seemed natural to integrate data across the different studies, what was not exactly clear to him was "how much weight is to be attributed to the different results" (pg. 1243). Thus, even though social and behavioral scientists have long sought to find ways to combine information from data, a key concern in this process is what strategy is best.

Data repositories have grown at an explosive rate over the last few decades due to new information technologies, particularly those supporting World Wide Web applications. This growth has also led to an exponential increase in the amount of data collected on individuals, creating numerous opportunities for examining new theories and developing new methods of analysis. At the same time, the number of different sources over which this information is divided continues to grow, creating additional obstacles for effectively combining such data so that it can be explored. At its most basic level, the process of combining data can be thought of as one in which information from different datasets sharing at least some common variables is merged. The whole process for combining and analyzing such data from multiple sources is referred to henceforth as *data fusion* (Wilderjans, Bernal, Galindo-Villardón, & Ceulemans, 2015; Marcoulides & Grimm, 2016).

Although the literature contains numerous different terms that have been used to describe data fusion, including statistical matching, data matching, file concatenation, data integration, multi-source imputation and ascription, and data merging to name a few, the most commonly used term across the various disciplines has been *data fusion*. A unique characteristic of data fusion as an integrative data analysis method is the individual-level matching of different datasets. Rodgers (1984) indicated that although individual-level matching can be considered as somewhat related to the exact matching technique known as record linkage, the approaches are only equivalent in cases where the data contain identical individuals that have been matched (using for example an individual's name, social security number, etc.). However, having data from multiple sources on identical individuals is rarely encountered in research settings. This is because individual data are ideally obtained from random samples selected from a large population and the probability of the same individual appearing in the different samples is extremely small. In most cases, data from different individuals and sources is available, which can be combined and collectively analyzed to obtain the best estimate of the population effect. Therefore, the general objective of data fusion can be thought of as the creation of a new dataset that allows for even more flexible analyses than the separate analysis of individual datasets.

The general methodology of data fusion first appeared in the scientific literature in the 1960s as a mathematical model for data manipulation; although, its origins date back to military work in code breaking, information analysis, and cryptology during World War II (Bhattacharya & Saha, 2015). It became very popular in the 1970s in the defense industry, and in 1980 the U.S. Department of Defense established the Data Fusion Sub-Panel of the Joint Directors of Laboratories to unify its terminology and procedures (U.S. Department of Defense, 1991). In addition to military uses, current applications of data fusion cover a wide range of fields including bioinformatics, business, computer and information systems, data mining, law enforcement, medicine, and traffic control (Bhattacharya & Saha, 2015). Following the call made by Glass (2000) for the replacement of meta-analysis of aggregated data with an integration across studies of individual participant data, the field of psychology gained renewed interest in the topic of data fusion (Cooper & Patall, 2009).

It should be noted that various other names are also used in the psychology literature to refer to data fusion, such as data pooling, integrative data analysis, individual participant data analysis, mega-analysis, meta-data analysis, pooled meta-analysis, and raw meta-analysis (Bond, Wiitala, & Richard, 2003; Cooper & Patall, 2009; McArdle & Horn, 2002, 2005; Piccinin & Hofer, 2008). Although the above terms are generally used to refer to the method of data fusion, two different approaches are sometimes inferred; the *one-step* and the *two-step* approach (Cooper & Patall, 2009; Jones, Riley, Williamson, & Whitehead, 2009). With the two-step approach, data from each study are first analyzed and then parameter estimates from each study are synthesized using available methods for meta-analysis of aggregated data. The two-step approach is closely connected to classical meta-analysis for analyzing aggregated data. In the one-step approach, which is completely in line with the original ideology and definition of data fusion, the data from each included study are combined and analyzed simultaneously as a *single* dataset (Wald, 1999). Analyses using the one-step method of data fusion are not common in psychology (Cooper & Patall, 2009; Curran & Hussong, 2009; Marcoulides & Grimm, 2016; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009), even though it has been suggested that this specific method of data fusion can be very useful when studying psychological processes that are difficult to manipulate directly (e.g., changes in cognitive ability over time).

There are a number of important benefits to using the data fusion method. First, new data collection may be unnecessary. Several researchers are often working on similar topics so data that has already been collected from multiple sources can be combined. Additionally, data fusion does not solely rely on published studies; therefore the "file-drawer problem" (Rosenthal, 1979) of meta-analysis is less of a concern (Sharpe, 1997).

Another major benefit of data fusion is that combining data from multiple studies can create a more heterogeneous sample, which can increase the generalizability of results (Curran & Hussong, 2009). This limits threats to external validity because a number of different sub-populations are combined, increasing the likelihood that the results will generalize across different populations. Furthermore, combining data from multiple sources increases sample size so that more precise estimates of effects can be obtained and the increase in statistical power enables researchers to detect effects that are typically underpowered in single studies (Hedges & Pigott, 2001; Sansone, Morf, & Panter, 2008), such as mediating effects (Fritz & MacKinnon, 2007).

In spite of the many benefits of the data fusion approach for integrating data, there are some limitations. One challenge is the need to obtain raw data from the various sources. Researchers may not want to share their data for a number of reasons. This can limit how many studies are available to collectively analyze and can therefore limit the generalizability of the results. Additionally, because the actual raw data are being combined and analyzed, common measures of the constructs of interest are necessary. This can sometimes be problematic because many studies may use different measurement tools to measure the same construct. If different tests are used to use measure the same construct, then researchers must find a way to link the different tests, which can be difficult. There can also be potential problems if there are large amounts of missing data; although if the amount of missing data is small compared to the sample size, and if the data are missing completely at random or missing at random, then specific methods of estimation can be used that provide results similar to those obtained from an analysis of complete data (Moustaki & Knott, 2005). Consequently, determining appropriate methods to use in order to combine independent samples from sources that are to be analyzed simultaneously is an extremely important line of research.

The purpose of this dissertation is to examine the effectiveness of a newly proposed Bayesian approach to analyzing fused data and to determine whether it can provide accurate parameter estimates. To date most data fusion methods proposed in the literature are employed within the *frequentist* framework (e.g., Piccinin & Hofer, 2008). In contrast to the commonly used frequentist methods, this novel approach is based on a Bayesian framework and is henceforth called *Bayesian Synthesis*. In summary, this method employs several separate but well-defined steps performed within a traditional Bayesian framework. However, instead of combining datasets at once as is traditionally done (see Marcoulides & Grimm, 2016), information obtained from one dataset serves to provide prior information for the analysis of the next data set. This process continues sequentially until a single posterior distribution is created using all available data. It is hypothesized that the inclusion of informative data-dependent priors provides an extra source of information to estimate model parameters and that this additional information can effectively aid in the accuracy of the estimation and thus in the interpretation of results.

The remainder of the dissertation is organized in the following way. Chapter 2 provides an overview of the most widely used data fusion framework along with a description of popular frequentist data fusion methods as they are commonly applied to practical problems within the field of psychology. Where necessary, specific technical details of the approaches and algorithms are presented. Chapter 3 provides an overview of currently used Bayesian data fusion methods and presents the newly proposed

Bayesian Synthesis method. Chapter 4 describes the analyses and procedures that will be used to examine the new Bayesian Synthesis method; a simulation study to evaluate the proposed Bayesian Synthesis method and an empirical example to illustrate its use. Chapter 5 presents the results of the simulation study and the empirical example. Chapter 6 provides a discussion of the implications of the obtained results and provides suggestions for further research.

Chapter 2

OVERVIEW OF DATA FUSION

Data fusion encompasses many different things and covers a wide range of activities that makes it difficult to provide a precise definition (Wald, 1999). Adding to this difficulty is the fact that the field of data fusion was conceptualized and developed from a variety of viewpoints. For this reason, it is necessary to first describe an adaptation of one of the most widely used data fusion frameworks, the *Joint Directors of Laboratories (JDL) Data Fusion Framework* (U.S. Department of Defense, 1991). This framework is more general in structure than others offered in the literature and effectively highlights the main points and processes that are essential when conducting data fusion.

A Data Fusion Framework

The data fusion community decided in 1998 to adopt the following definition for data fusion (Wald, 1999):

> "….data fusion is a formal framework in which are expressed the means and tools for the alliance of data originating from different sources…." (pg.1191).

The above definition was a modification of the original definition provided by the *Joint Directors of Laboratories* and the U.S. Department of Defense (1991), which stated that:

> "…data fusion is a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from multiple sources…." (pg. 5).

Although a number of different frameworks to support the development of data fusion systems based on the above definitions have been proposed over the past few years

(Boström, Andler, Brohede, Johansson, Karlsson, van Laere, Niklasson, Nilsson, Persson, & Ziemke, 2007; Dubois, Liu, Ma, & Prade, 2016; National Research Council 1992), according to Esteban, Starr, Willetts, Hannah, and Bryanston-Cross (2004) the most widely used framework is the *JDL Data Fusion Framework* (U. S. Department of Defense, 1991). The *JDL Data Fusion Framework* was originally proposed to support the development of military applications, but it can also be generalized for a variety of other applications. It reflects a process to follow for a generic data fusion system, and is designed to establish a common language and model within which data fusion techniques can be implemented. There are multiple levels of processing that do not need to occur in any particular order and include:

Source pre-processing: This level creates preliminary information from the data that serves to interface it better with other levels of processing.

Object refinement: This level refines the identification of individual objects.

Situation refinement: Once individual objects are identified, their relationships to observed events need to be ascertained.

Process refinement: This level is not so much concerned with the data, but rather with how well the other levels have performed and whether they need to be improved.

Using the original *JDL Data Fusion Framework* from above, the following adapted framework is introduced for all data fusion applications within the social and behavioral sciences. As in the original *JDL Data Fusion Framework*, the adapted framework is also categorized into different hierarchical process levels and for additional clarity broken down into various sub-level processes. It is important to note that the described process levels are not meant to be handled in a strict order and can even be executed concurrently.

Level 1. Source pre-processing: Data from multiple sources are first selected and evaluated for possible data alliance. The characteristics of the population and samples included in the data are taken into account. The time metric is examined along with the specific details of all the variables and constructs measured.

Level 2. Object refinement: Using the information provided by Level 1, determine whether the data originating from the different sources are compatible. For example, determine whether data contain identical individuals or different individuals, or if they contain common variables. If so, any one or more of the following activities may be undertaken: record linkage, statistical matching, proximity estimation among variables and/or data units (e.g. nearest neighbor approaches), data imputation, or data linking.

Level 3. Situation refinement: Interpret the results from Level 2 in terms of the possible opportunities for data fusion operations and analyses. Evaluate the advantages and disadvantages of taking one course of action over another, focusing, for example, on aspects such as model selection, parameter estimation, and best-fit function criteria. Examine the obtained results and render a decision. If necessary, proceed to the next level.

Level 4. Process refinement: This is the refinement process, which basically loops around the other three levels to monitor and improve performance. It is only activated when additional sources or methods of information enhancement are available or needed to complete the necessary operations and analyses.

It is evident from the above descriptions of the various levels and sub-levels involved in the data fusion process that the bulk of the methodological difficulties are at Level 2 and Level 3, where one must determine if the data from different sources are compatible, and if so, how to best integrate these datasets for concurrent analysis.

For example, when studying within-person change across studies, the measurement of the rate of change must be equivalent across studies. This means that the *time-metric* used to track change and the *scaling* of the outcome must be equivalent. However, longitudinal studies often vary in the *number* and *timing* of assessments, which creates a potential confound when attempting to summarize research findings focused on within-person changes. The first part of this challenge involves the time metric and different studies, depending on the original goals of the study, may use different time metrics (e.g., age, measurement occasion, grade, time since the beginning of the study, time since puberty) to track change against. The second part of this challenge is the scaling of the outcome measure, which must also be equivalent across studies. Studies may use different scales (tests, surveys) to measure the same construct and this makes the results of longitudinal studies more difficult to synthesize because there is no good way to alter the rate of change to be scale-free – akin to using a standardized effect size in cross-sectional studies.

Approaches at Level 2: Data Combination

As a result of the many potential difficulties, a wide variety of methodological approaches have been applied over the past several decades across numerous disciplines (e.g., bioinformatics, business, computer and information systems, data mining, defense industry, engineering, law enforcement, medicine, and traffic control) in an attempt to tackle these challenges (Ahmed, Sutton, & Riley, 2012; Bhattacharya & Saha, 2015). The vast majority, and currently most popular, of these methodological approaches to handling potential difficulties at Level 2 can be categorized as using some form of missing data imputation, such as regression imputation, clustering, or nearest neighbor matching strategy based on selected similarity metrics (e.g., Euclidean distance, squared Euclidean distance, city block distance; D'Orazio, Di Zio, & Scanu, 2006), or latent variable modeling with full information maximum likelihood depending on the amount of overlapping information across studies. Much of the popularity of these methods comes from their widespread availability and ease of use in most commercially available computer software programs (e.g., SAS, SPSS, Stata). Some recent data fusion research with latent variable models has employed item response theory strategies, focusing on the overlap of information at the item level (e.g., see details below - Curran, Hussong, Cai, Huang, Chassin, Sher, & Zucker, 2008; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Marcoulides & Grimm, 2016). Many data fusion methods for combining datasets have been proposed in the literature and are further described below. Although most data fusion approaches in the literature are employed within the frequentist framework, data fusion approaches within the Bayesian framework are introduced in the next chapter.

The Nearest Neighbor Approach to Combining Datasets

One of the most commonly used approaches for object refinement in order to combine datasets for data fusion is based on the notion of a *nearest neighbor*. In simple terms, the choice of the record from one dataset fused to the other dataset is based on a distance measure metric, calculated on the basis of the common variables in both datasets. For example, consider a situation in which data from two different sources (A and B) are to be fused, with none of the same individuals appearing in both datasets. Certain variables (denoted X) are present in both datasets and are referred to as "common variables." Additional variables, denoted Y, are present in only the data from source A, and an additional set of variables, denoted Z, is only present in the data from source B. These variables are referred to as "unique variables" or "specific variables." The data fusion literature that relies on the nearest neighbor approach for combining datasets views the problem as one of creating a single combined dataset with observations on all three types of variables (X, Y, and Z), whereby the created dataset does not contain any missing data. The combining occurs by having the data from one source act as the recipient sample and data from the other source serve as the donor sample. This process is often referred to as the *marriage process* (Rassler, 2002). To minimize the amount of information that is discarded during the marriage process, the larger sample is generally selected as the recipient sample (Rassler, 2002). In situations where multiple donor samples are used to complete the recipient sample dataset, the marriage process has been termed polygamy (Rassler, 2002).

In order to determine the nearest neighbor, a similarity metric is computed based on a distance measure metric. A variety of distance measures can be used for this purpose. The most popular and simplest distance measure is the Euclidean distance (James, Witten, Hastie, & Tibshirani, 2013). The Euclidean distance between observations *i* and *j* and between observations *i* and *k* on variables X and Y is defined as

$$D(i,j) = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$$

and

$$D(i, k) = \sqrt{(X_i - X_k)^2 + (Y_i - Y_k)^2} .$$

Using these distance measures, observation i is declared to be more similar to observation j than to observation k if $D(i, j) < D(i, k)$, where $D(i, j)$ is the distance measure between individuals i and j and $D(i, k)$ is the distance measure between individuals i and k (in this case D is the Euclidian distance).

A simple illustration of the nearest neighbor approach is provided in Figure 1. In this example, observations from two datasets A and B are considered. In dataset A, information from individuals on the variables gender, education, and age were collected ("common variables"). In dataset B, information on the variables gender, education, and age were collected ("common variables") along with a mathematics achievement score ("unique variables"). Determined by their Euclidean distance measure (with the categorical variables recoded to numerical values), person 1 in dataset A and person 3 in dataset B are most similar and are considered nearest neighbors. The same is true for person 2 in dataset A and person 6 in dataset B. Consequently, if dataset A is the recipient and dataset B is the donor, the value of the mathematics achievement score for person 3 and person 6 is fused (or duplicated) to the data for person 1 and person 2 respectively.

Alternative distance measures that can also be used include the squared Euclidean distance and the City Block distance (sometimes referred to as the Manhattan distance), defined respectively as

$$D(i, j) = (X_i - X_j)^2 + (Y_i - Y_j)^2$$

and

$$D(i,j) = |X_i - X_j| + |Y_i - Y_j|.$$

A generalization of the Euclidean distance and its variants is called the Minkowski distance (Liu, 2007) and is defined as follows (with $q > 0$):

$$D(i,k) = \sqrt[q]{|X_i - X_k|^q + |Y_i - Y_k|^q}.$$

When $q = 2$, the Minkowski distance is equal to the Euclidean distance, and when $q = 1$, it is the same as the City Block distance metric. It is important to note that the above mentioned distance measures represent only a few of the many other distance measures that have been introduced in the data fusion and related data mining literature. For example, other distance measures include Chebyshev, cosine, Tanimoto, Hamming, and Mahalanobis (Kuhn & Johnson, 2013; McArdle & Ritschard, 2014; Resenda & Sousa, 2003). The literature on selecting one distance measure over another is inconclusive. Most researchers generally choose to use the more simple Euclidean distance measure as a starting and then compare obtained results to those from other distance measures (El-Sayed & Hamed, 2015).

Despite its simplicity and ease of use, research has shown that a very important aspect of the nearest neighbor approach is the availability of a sufficient number of common variables in both datasets to assist in the determination of the distance measure. If too few common variables are available then it might not be possible to compute a distance measure (Kuhn & Johnson, 2013). Additionally, different selections of variables can lead to a completely different fused dataset. For example, if only education and age were used in the above illustration to determine the nearest neighbor, the mathematics achievement scores from either person 3 or person 6 might be fused to the data for person 1 and person 2. In such a case one option for determining which score to use would be to be randomly select between the two possible values. It is important to select variables that are related to the unmeasured variables in order to assist in the creation of the missing value. Similarly, the choice of the distance measure that is calculated on the common variables in the selected datasets can influence results and be potentially problematic (D'Orazio, Di Zio, & Scanu, 2006). Consequently, despite the popularity of the nearest neighbor approach, its major disadvantage with using it for the purpose of data fusion is the precision with which the duplication of data from one dataset to the other is conducted. Kamakura and Wedel (1997) have indicated that because the choice of the type of distance measure and variables is clearly subjective, it can critically affect the overall quality of the obtained fused dataset. Another issue that can impact the effectiveness of using the nearest neighbor approach for data fusion is the scale of measurement of the variables included in the analysis. Data with variables that are on completely different scales will generate distances that are weighted towards variables that have the larger scales. Although some strategies have been recommended for standardizing the variables rather than using them in their original scale of measurement, these still suffer from the other above indicated limitations (Kuhn & Jonson, 2013).

An Item Response Theory Approach to Combining Datasets

In light of the limitations of the nearest neighbor approach and other missing data approaches to data combination, especially those associated with the scale of measurement of the variables, recent research has considered the application of psychometric models to combining datasets for data fusion. Specifically, some recent research has proposed the use of common-item linking methods available through item response theory (IRT) modeling (Curran et al., 2008; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Marcoulides & Grimm, 2016).

A commonly used item response model for dichotomous items is the one-parameter logistic model (1PL). The 1PL model can be written as

$$P(X_{in} = 1 | \theta_n, \beta_i) = \frac{e^{(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}$$

where $P(X_{in} = 1 | \theta_n, \beta_i)$ is the probability of getting item i correct given person n's ability level θ and item i's difficulty β . In the 1PL model, items only differ on their difficulty parameter, β_i (De Ayala, 2009). If multiple datasets have some items measuring the same construct that are common and some items that are unique, then these items can be subjected to the 1PL model. In the 1PL model, the common items have the same item parameters, which scale the latent variable θ, and the unique items are scaled appropriately. The latent variable, θ, can be estimated for each person and these scores are comparable even though all of the items were not the same across studies.

Using this common-item linking method allows multiple datasets to be combined or fused and subsequently analyzed simultaneously as a single dataset. This approach has been primarily advocated due to the accepted benefits of IRT modeling that the items used in an analysis do not need to come from the same scale or the same study (Curran et al., 2008). By employing IRT to estimate the item difficulties, comparable latent variable estimates can be created (Hambleton, Swaminathan, & Rogers, 1991). Additionally, by using the 1PL model, an individual's total score is a sufficient statistic for calculating their ability score (θ). Therefore, future studies do not need to necessarily provide individual item data. This makes it easier to incorporate additional datasets where only total scores but not individual item data are available.

Two main types of linking procedures are available in the IRT literature, those based on traditional practices and those based on augmented practices (e.g., Mislevy, 1988; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Sheehan, & Wingersky, 1993). Augmented linking practices were first introduced by Mislevy and his colleagues as a way to apply linking procedures when some necessary datasets cannot be obtained. Thus, a fundamental characteristic of augmented linking practices is that item characteristics (such as an item's difficulty) determined from available individual response data are augmented with item characteristics derived from an assumed theoretical model. In this manner, two different sources of item information are used: information from observed responses in the available data, and information generated from an assumed theoretical model. Mislevy, Sheehan, and Wingersky, (1993) have referred to the information generated from an assumed model as "collateral information", and emphasized that such information "is limited by the strength of its relationship to item operating characteristics" (pg. 56). To date, augmented linking practices have not been applied for purposes of data fusion, as the main goal of data fusion is to combine information from disjoint datasets sharing at least a number of common variables and not to artificially generate information on the basis of an assumed model. In contrast, because traditional linking practices rely exclusively on available data, they are very useful and more appropriately used for purposes of data fusion.

Several different data collection designs are often considered within traditional linking practices. One popular data collection design used is the common-item nonequivalent group design (Kolen, 2006; Kolen & Brennan, 2004). In the common-item nonequivalent group design two (or more) forms of a test or measurement tool that have a subset of items common to both versions are each administered to different groups (Kolen & Brennan, 2004). Using these common items, the different forms can then be linked and placed on the same scale. Another name given to this design is the nonequivalent group with anchor test (NEAT) design (von Davier, Holland, & Thayer, 2004). This name is used mainly when the set of common items make up an actual test, the so-called anchor test.

Two popular calibration procedures are also considered within the traditional linking practices, the concurrent and separate calibration approaches (Hanson & Beguin, 2002). Concurrent calibration generally involves the estimation of parameters using all the available data simultaneously to obtain a common scale. In contrast, the separate calibration procedure involves estimating item parameters separately and then using the linear relationship of the parameter estimates to transform one set of parameter estimates to the scale of the other form. Past research within the IRT literature has compared the two procedures and determined that both can be effectively used for item calibration (e.g., Petersen, Cook, & Stocking, 1983; Cook,-Eignor, & Wingersky, 1987; Kim & Cohen, 1998; Béguin, Hanson, & Glas, 2000; Béguin & Hanson, 2001; Hanson & Béguin, 2002; Kim & Kolen, 2006), however, there is still no conclusive evidence on which method to prefer (Lee & Ban, 2010). Additionally, no research has been conducted to date comparing the concurrent and separate calibration procedures in data fusion applications. Doing so would provide insight into the various issues and aspects of the linking process in data fusion applications and help researchers choose the most appropriate linking method.

There are some researchers that have criticized attempts to link items from different tests and studies. For example, Feuer, Holland, Green, Bertenthal, and Hemphill (1999) question if is even possible to link data obtained from different tests. Specifically, they questioned whether it is realistic to link state or commercial test data to data from the National Assessment of Educational Progress (NAEP) or the Third International Mathematics and Science Study (TIMMS). They believe that because the NAEP or TIMSS have low stakes for test takers, perhaps their scores will not accurately represent individuals. They believe that when stakes are high, respondents are usually more motivated and try harder. However, when items from different studies measuring the same construct and with similar stakes are to be linked (even if the items or the respondents are different), this criticism may not be valid. In situations where there may be no common items or too few common items to confidently link measures together, some researchers have suggested the use of a bridging study (Hussong, Curran, & Bauer, 2013). The main idea behind a bridging study is to conduct a new primary data collection for the specific purpose of linking together the different measures used in the original set of studies. Unfortunately, because this involves the recruitment of new participants (ideally from a similar population to that sampled in the contributing studies intended for data fusion) and administering items from all of the original studies to these new participants, conducting a bridging study is quite unrealistic in most situations.

Approaches at Level 3: Analysis

Once the datasets have been fused using one of the above data combination procedures, the next step in the data fusion process is to conduct the analysis of interest (e.g. regression, longitudinal growth modeling, etc.). As previously mentioned, in traditional frequentist data fusion methods the fused datasets are analyzed all at once as though they form a single dataset.

Concluding Remarks

As indicated at the beginning of this chapter, all the approaches described above can be classified as frequentist approaches. The next chapter describes alternative data fusion approaches that can be implemented within the Bayesian framework. The chapter also introduces a proposed new method that may be able to tackle some of the problems and limitations of the methods described above.

Chapter 3

THE BAYESIAN APPROACH TO COMBINING DATA FOR DATA FUSION

The Bayesian approach offers a clear alternative to the frequentist approach. It is distinguished by its use of probability distributions to describe uncertain quantities, which often leads to solutions to many difficult estimation problems (Bijak & Bryant, 2016; Hill, 1970). The application of the Bayesian approach to data analysis has been encouraged for a number of years by researchers across many different fields. Some examples include education (Muthén & Asparouhov, 2012; Kaplan & McCarty, 2013), engineering (Yee, Hoffman, Branch, Ungar, Malo, Ek, & Bourgouin, 2014), medicine (Smith, Speigelhalter, & Thomas, 1995; Bennett, Crowe, Price, Stamey, & Seaman, 2013), psychometrics (Rupp, Dey, & Zumbo, 2004), and statistics (Gelman, Carlin, Stern, & Rubin, 2004; Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2007). Additionally, theoretical and methodological evidence has emerged illustrating the many benefits of using a Bayesian approach (Bijak & Bryant, 2016; Hill, 1970). For example, Muthén and Asparouhov (2012) advocated the following four key points to motivate researchers to use the Bayesian approach: (i) more information about parameter estimates and model fit can be learned, particularly because specific distributional assumptions do not have to be met (e.g. does not assume a normal distribution), (ii) excellent small-sample performance, (iii) many analyses are not as computationally demanding, and (iv) a variety of different models can be analyzed.

The rationale behind the general Bayesian approach has its foundation in Bayes theorem, indicating that the notion of probability can be applied to the degree of belief or knowledge in a hypothesis (*H*) given observed data or evidence (*E*). The degree of belief in hypothesis *H* given the observed evidence *E* according to Bayes theorem is

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

where *P(H|E),* the probability of *H* given *E*, is called the posterior degree of belief in *H* (in other words, the updated belief in the hypothesis once the data or evidence is observed). *P*(*H*) represents the probability of a hypothesis (which is commonly called the prior degree of belief in *H*), *P*(*E*) represents the probability of the observed data or evidence, and *P(E|H)* is the probability of the observed data or evidence given a hypothesis. In terms of parameter estimation, once data are observed, prior information on the parameters is combined with information from the data (i.e., the likelihood). This provides a distribution of parameter information, but because this combination occurs after the data are observed, the distribution of possible parameter values is therefore considered the *posterior* parameter distribution (Fox, 2010). Thus, the posterior distribution specifies the probability that each parameter will equal a certain value or might lie within a range of values. A central issue then, of the Bayesian approach to parameter estimation, is estimating the posterior distribution using the prior degree of belief in *H*. It is important to note that the role played by the priors is a point of controversy and criticism in the literature because the use of different priors can result in different conclusions (Gelman et al., 2004; Jackman, 2009).

Noting in the above equation that *P*(*E*) does not depend on *H*, the equation can also be rewritten to indicate that the posterior distribution is simply proportional to a combination of the information contained in the data and the prior, such that

$$P(H \mid E) \propto P(E \mid H)\, P(H)$$

where $\propto$ is used to denote proportionality. Thus, the posterior distribution results from updating the prior from information contained in the data. If no dependable prior information is available, then non-informative or diffuse priors are used, which leads to posterior distributions that are determined only by the observed data. No dependable prior information can also include situations for which only partial or very little information about the parameters to be estimated is used (Levy & Choi, 2013).

In contrast, if specific and informed prior knowledge is available and used to determine the posterior distribution, even if it is subjective, then the priors are considered informative priors. For example, an informative prior can be constructed using information based on knowledge obtained from past studies (Levy & Choi, 2013). Consequently, the posterior distribution from one study can be potentially used as prior information for another study. Such an approach implies that the use of a prior from another study actually permits the inclusion of past research findings into the current analysis (Edwards et al., 1963; Jackman, 2009; Kass & Wasserman, 1996; Levy & Choi, 2013; Lindley, 2004; Tanner & Wong, 1987). Thus, the choice of the prior has the potential to greatly impact the resulting posterior distribution and determining what is the best prior has been an important topic of research (Gelman et al., 2004). One strategy involves the use of sensitivity analyses in which obtained solutions using different priors are compared (Hoeting, Madigan, Raftery, & Volinsky, 1999). Another strategy focuses on decreasing the dependence of the posterior distribution on the priors by increasing the sample size of the data (Jackman, 2009). As will be discussed later, it is this aspect of the Bayesian approach, along with the ability to include information from previous findings into the current analysis that will be implemented within a data fusion process proposed in this dissertation. Specifically, it is the key advantage of the Bayesian approach in terms of its ability to integrate information from multiple sources and to describe uncertainty coherently that will be exploited.

Accurately estimating the posterior distribution is not always possible because of various analytical and numerical complexities. One useful alternative to numerical integration and analytical approximation is called Markov Chain Monte Carlo (MCMC) estimation (Gelman & Rubin, 1992; Geman & Geman, 1984; Gilks et al, 1996; Levy, 2009; Rubin, 1987). The general idea behind MCMC is to sample from the posterior distribution and obtain sample estimates of the quantities of interest. MCMC estimation can therefore be considered a method of sampling, and relies on the Monte Carlo simulation principle that knowledge about anything can be obtained simply by sampling many times from the probability distribution of the parameter of interest (Jackman, 2009). In other words, if one wishes to learn about the posterior distribution, they should repeatedly sample from the posterior. However, as the distribution is not known, obtaining a sample without explicit knowledge of the distribution is challenging. In the MCMC approach, sampling from an unknown distributional form can be achieved by using a Markov Chain, which is basically a sequence of random numbers (Levy & Choi, 2013). The Markov Chain is expected, under general conditions, to converge to its stationary distribution, which will be equivalent to the posterior distribution of interest. In this manner, once the converged Markov chain is constructed, it is expected to represent the posterior distribution and each point in the chain will represent a sample of the posterior distribution of interest.

A number of different methods for creating a Markov chain have been proposed in the literature. These include the Metropolis algorithm, the Metropolis–Hastings algorithm, and the Gibbs sampler (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Metropolis & Ulam, 1949; Gilks, Richardson, & Spiegelhalter, 1996; Geman & Geman, 1984). Although often thought of as separate algorithms, according to Chib and Greenberg (1995) the distinction is rather artificial as each can be regarded as a special case of the Metropolis-Hastings algorithm with varying sets of conditional distributions. There are also a number of different ways that these algorithms can be combined to make them more efficient (e.g., in terms of computing time, storage space, or estimate variability – Tierney, 1994) and some popular and free software programs like BUGS (Bayesian Analysis Using Gibbs Sampling; Spiegelhalter, Thomas, Best, & Gilks, 1996) or R (R Development Core Team, 2010) package R2WinBugs (Sturtz, Ligges, & Gelman, 2005) are also available for use. Other software options for MCMC estimation include SAS PROC MCMC (SAS Institute Inc., 2014). Although there are many practical advantages to using MCMC algorithms in the Bayesian approach (e.g., it is not limited by the number and estimation of different parameters), of special importance under consideration for this dissertation is their ability to be extended to a variety of modeling frameworks, including data fusion (Muthén & Asparouhov, 2012).

The Bayesian approach has also received some attention in the general data fusion literature (e.g., Gilula, McCulloch, & Rossi, 2006; Rässler, 2002, 2003). Within this segment of the literature the application of the Bayesian approach has focused on the issue of data combination, and as discussed in Chapter 2, treats data combination as a missing data problem. In this context, the treatment of missing data first focuses on the underlying mechanism that generated the missing data and then evaluates them as being either ignorable or non-ignorable (Rubin, 1976; Rubin & Little, 2002). A missing data mechanism is considered ignorable when the process that generated the missing data does not affect any inferences about the variables. Two specific types of missing data mechanisms can be considered ignorable, those referred to as missing completely at random (MCAR) and those missing at random (MAR). For example, if age and education were observed on some sample of participants and there is missing data on education, and if the missing data on education is not related in any way to the observed values on both age and education, then the missing data would be considered missing completely at random. If the missing data on education is not related to the observed values of education, but might be related to age (e.g., if younger people do not report their level of education), the missing data would be considered missing at random. The use of auxiliary variables (i.e., variables that are secondary to the analyses, but may still be correlates of a missing variable) has also been recommended in the literature as a potential strategy to reduce or eliminate any possible bias when it is not evident whether the missing data mechanism is MCAR or MAR (Collins, Schafer, & Kam, 2001; Enders, 2010). However, since data fusion involves the matching of different datasets that generally do not share any common participants but have some variables in common, it is believed that the missing data can be reasonably considered to be MCAR or MAR (Rässler, 2002, 2003). Therefore, in data fusion the missing data issue becomes one of nonresponse, whereby any missing information is simply considered as MCAR or MAR because the missingness is brought about by the study design of the different participants in each sample (Gelman, King, and Liu, 1998). Thus, the missing data are mainly a consequence of unasked questions and the underlying mechanism that generated the missing data is ignorable (Rässler, 2002, 2003).

By viewing data fusion from the perspective of a missing data problem, it is assumed that a variety of modern data imputation techniques that take advantage of Bayesian modeling can be utilized (Rubin, 1987). To date, a number of different data imputation techniques based on Bayesian modeling have been proposed in the literature. Next, three popular approaches, multiple imputation, non-iterative Bayesian based imputation (NIBAS), and data augmentation, are summarized. Although these approaches can all be considered multiple imputation approaches because they all utilize the same three-step process (i.e., an imputation phase that creates multiple copies of the dataset, an analysis phase that analyzes the filled in datasets, and a pooling phase that combines everything into a single set of results), they differ in terms of the specific computational algorithm used during the imputation phase (Enders, 2010). Thus, although the approaches in general terms can be conceptualized as involving similar strategies for generating multiple copies of the data and then imputing (filling in) each copy with different estimates of the missing values, the approaches apply distinct algorithms. These different algorithms have been shown to perform well with all types of situations (e.g., categorical versus continuous data, cross-sectional versus longitudinal data, and normally distributed data), although no single approach works in every situation (Ender, 2010; Tanner & Wong, 1987).

Level 2: Data Combination in a Bayesian Framework

The practice of imputing missing data originated from the idea that every missing data point can be replaced by a best estimate of what the observed value would have been if it had not been missing. Thus, data imputation is an attractive strategy because it yields a complete data set for subsequent analysis (Enders, 2010). Initial implementation of this idea used single imputation, in which the missing data were replaced by a single value, often the estimated mean values from the available data or by using regression estimates. However, these replacement values were often found to be poor estimates of the unobserved values and produced biased parameter estimates (Rubin, 1987). In contrast, multiple imputation techniques are considered very effective and flexible tools for analyzing data with missing values. Unlike single imputations, where every missing data point is filled in with a single estimated plausible value, in multiple imputation each missing value is replaced with several plausible values. The general idea behind these multiple imputation techniques is that for each missing value in the fused dataset several values, according to some distributional assumptions concerning the missing data, can be imputed under an explicit Bayesian model. As indicated by Rubin (1987), it is this ability of the Bayesian framework to use prior information that makes it especially attractive for all kinds of statistical matching or data fusion tasks. For example, if an $n \times p$ dataset (with n = observations and p = variables) containing 20% missing values is encountered, the unknown missing data can be inferred under an explicit Bayesian model by computing the posterior distributions for the unobserved data given the observed data. However, in order to form the posterior distributions, simulated random samples of the dataset need to be generated. Each simulated random sample provides imputed values for every missing data point so that the entire $n \times p$ dataset no longer contains any missing values. This simulation process is repeated several times to produce the multiple imputations values, with each simulated dataset representing a possible realization of what the entire $n \times p$ dataset might have been like if there were no missing data. Accordingly, multiple imputation yields k different values (say 5 or more, although the choice of k is normally at the discretion of the researcher) for each missing data point based on the predictive distribution given the observed data. Using these complete data, statistical procedures of interest would be applied k times for the k datasets and parameter estimates would then be obtained by averaging over each of the files created with the imputed data. In cases that two datasets are to be fused, the approach would give k versions of the fused dataset. For example, consider the data fusion illustration present in Figure 2. If a dataset *A* containing *X* unique variables and *Z* common variables were to be fused with another dataset *B* containing *Y* unique and *Z* common variables, the multiple imputation technique would provide *k* different versions of the combined data where the X and the Y missing values are imputed to create a single dataset with no missing values.

The non-iterative Bayesian based imputation (NIBAS) approach is very similar to the multiple imputation approach. The difference is that random draws are performed on model generated parameter values instead of just taking the estimates from the data (Rässler, 2003). For example, the elements of each column in the common matrix *Z* in Figure 2 would function as predictors in different selected linear regression models and the values for the missing *X* in dataset *A* and for the missing *Y* in dataset *B* would be computed using randomly drawn regression coefficients provided by the differently generated models (with the resulting imputations influenced by the preceding choice of regression models; Rässler, 2003). Thus, the observed data posterior distribution is used to obtained values for the parameters and the posterior predictive distribution is used to fill in the missing data. Accordingly, it is the k different parameter values that are involved in the random draws. Using these data in the analyses and pooling phases, statistical procedures of interest would be applied for the k data files and parameter estimates would then be obtained by averaging over the files created with the imputed data as is done in the multiple imputation approach.

Another method based on MCMC estimation that is ideally suited for such missing data imputation activities is the data augmentation algorithm by Tanner and Wong (1987). This algorithm was originally developed to simulate the posterior distribution of any parameters of interest by means of Markov chains and is flexible enough for use in all kinds of estimation activities. The data augmentation algorithm specifically utilizes a variant of the Gibbs sampler used in MCMC estimation. The algorithm proceeds by first using the common variables to obtain estimates of the missing data in each of the datasets to be fused but then randomly supplements their value with a small deviation. This initial step is called the imputation step and utilizes the predictive distribution of the missing data. Next, using the values of the missing data in the data sets to be fused, new parameter estimates are obtained from the complete posterior distribution. It should be noted that although these sequential steps appear to be similar to those used in multiple imputation, the main difference is that the imputation and posterior steps are based on augmented data before generating a Markov chain that can be iterated until it converges to a stationary distribution for both the missing values and the parameters of interest. As with the previously described methods, once these data are generated, statistical procedures of interest can be applied to the data file and parameter estimates obtained.

Level 3: A Bayesian Approach to the Analysis of Fused Datasets

Although the Bayesian approach to data fusion has been predominantly advocated from the missing data perspective, where the intent is merely to impute missing data in order to create a single pooled dataset, the Bayesian approach has not been applied as a strategy to actually perform the combining or fusing of the datasets to then be analyzed. In other words, the Bayesian approach has not been used as a method to guide the very process of integrating and analyzing the data, it has only been used as a method to tackle characteristics related to the data itself. One way to view this distinction is to consider the multiple levels of processing within the *JDL Data Fusion Framework* (U.S. Department of Defense, 1991) introduced in the previous chapter. Currently available Bayesian approaches to data fusion can be thought of as having focused on the Object Refinement (Level 2) activity, whereas the proposed Bayesian approach will instead focus on the Situation Refinement (Level 3) activity. However, in order to effectively make use of such a Bayesian implementation and apply it to data fusion, a slightly different and novel perspective of viewing the issue of data fusion is also needed. The new perspective and methodology may be thought of as one in which several separate and well-defined steps are performed as part of the Bayesian implementation. Specifically, instead of simply combining datasets and analyzing them at once, information obtained from the analysis of one dataset acts as evidence for the next analysis. This process continues sequentially for all candidate datasets considered for data fusion until a final posterior distribution is obtained that incorporates information from each dataset. It is anticipated that such a data fusion approach conducted within the Bayesian framework will not only provide more accurate parameter estimates than approaches typically used for data fusion, but will also lead to more accurate interpretations of obtained results. This process is referred to as Bayesian Synthesis and is described next.

Bayesian Synthesis

In order to effectively apply a Bayesian approach to data analysis, the description of any model that is to be estimated must involve the specification of the prior distribution of the parameters. Two kinds of priors can be specified; subjective priors and objective priors (Kass & Wasserman, 1996). With a subjective prior, a researcher simply quantifies their personal degree of belief that is to be adjusted by the data. In contrast, an objective prior implies that the priors are specified according to some predetermined rule. For example, using Jeffrey priors, separation-strategy priors, distribution based priors (e.g., Inverse Wishart priors), single-unit priors (i.e., a prior that includes as much information as a single observation), or priors that maximize entropy (Jaynes, 1968; Jeffreys, 1961; Kass & Wasserman, 1995; Liu, Zhang, & Grimm, 2016). Additionally, depending on the amount of information provided about the properties of the parameters (e.g., a normal distribution with bounds $-\infty$, $+\infty$), the specified priors can also be classified as being uninformative, minimally informative, or informative (Levy & Choi, 2013). Minimally informative priors are meant to offer more information than uninformative priors, whereas informative priors offer extra information that may aid in parameter estimation.

The central role played by the priors was emphasized by Jeffreys (1961), who advocated that a prior should always be chosen "by convention," as a "standard of reference". However, specifying more subjective priors has been criticized in the literature because a researcher can pick and choose any prior that would allow them to bias at will the process of scientific inference (Efron, 1986). Objective priors on the other hand are viewed by many researchers as being independent of the person who performs the analysis, and for this reason are believed to be more advantageous than subjective priors. Although researchers may elect to use distinct objective priors with predetermined rules, given the data and the same assumptions regarding the model, different researchers ideally should arrive at the same conclusions (Efron, 1986). This is especially important in models with many parameters, as it skips the need to consider personal beliefs for every single parameter that is to be estimated (Bijak & Bryant, 2016).

Thus, because the choice of priors has the potential to greatly impact the resulting posterior distribution, determining what is the best prior to use is a very important topic of research (Gelman et al., 2004). Although a number of different strategies have been proposed in the literature (Hoeting, Madigan, Raftery, & Volinsky, 1999; Jackman, 2009; Kass & Wasserman, 1996), this dissertation focuses specifically on what is believed to be one of the main strengths of the Bayesian approach; namely, the ability to include information obtained from previous findings into the current analysis. Given that a major benefit of data fusion is that combining data from multiple sources increases sample size so that more precise estimates of overall effects can be determined, implementing such a strategy seems a natural extension to the data fusion approach.

The very notion of systematically using combined or pooled information can be traced back to the early work by Alan Turing in cryptology (Good, 1979). Although Turing is most famous for his code breaking work at Bletchley Park during World War II and his later contributions to computer science via his Turing machine, he also made important contributions to the field of probability theory (Good, 1979). While working on code breaking and cryptology, Turing became very interested in being able to estimate the probability of a hypothesis, but allowing for the prior probability to be updated when new information arrived piecemeal (Good, 1979). Turing thought of this approach as a sequential data analysis, using what he called weights of evidence, an approach that was in fact eventually applied to the actual code breaking of enemy messages generated during World War II.

Over the next several decades, both the ideas of data fusion and that of sequential data analysis became very popular in a variety of disciplines. Data fusion became popular in bioinformatics, business, computer and information systems, data mining, law enforcement, medicine, and traffic control (see detailed discussion provided in Chapter 1), whereas sequential data analysis received attention in fields like multi-sensor target location and tracking, and statistical and quality control (Bhattacharya & Saha, 2015; Ghosh, 1991). Sequential data analysis became mostly prevalent in situations where the selection of sample sizes could not be fixed in advanced and data needed to be analyzed as they were collected sequentially, rather than as a set. Examples of such situations include, target detection in multi-sensory radar systems, pattern recognition and machine learning, and fault detection in quality control processes (Fu, 1968; Ghosh, 1991; Lai, 1995). Edwards, Lindman, and Savage (1963) were possibly the most enthusiastic advocates of sequential data analysis when they declared that "…the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proved or disproved, or until the data collector runs out of time, money, or patience" (pg. 193).

Although both data fusion and sequential data analysis continue independently to be very popular approaches, to date they have not been combined into a single strategy that can be used to tackle data fusion situations in which multiple sources of data need to be sequentially combined in order to effectively utilize information from all datasets of interest. Although this activity is closely connected to the code breaking work of Turing, the only suggestion that can be found in the existing literature concerning the actual implementation of such a combining strategy can be found in the writings of Jackman (2009) who stated that, "Bayesian procedures are often equivalent to combining the information in one set of data with another set of data. In fact, if prior beliefs represent the result of a previous data analysis (or perhaps many previous data analyses), then Bayesian analysis is equivalent to pooling information" (pg. 19).

It is also important to note that in Bayesian theory the concept of including information obtained from data has existed in the literature for several decades and is referred to as using data-dependent priors (Darnieder, 2011). The general strategy behind this type of data-dependent priors is to use observed data to inform values for priors estimated initially through frequentist maximum likelihood methods (Casella, 2001). However, these data-dependent priors use the same data twice, once to obtain the initial maximum likelihood estimates and once to obtain the Bayesian estimates (Darnieder, 2011). In contrast, the approach proposed in this dissertation never reuses the same dataset. Instead, all specified priors correspond explicitly to data-dependent priors that are constructed through the use of different realized data than those used to obtain the parameter estimates of interest. In this manner, no data are ever reused in a Bayesian analysis as the data are being sequentially fused. Specifically, information obtained from the analysis of one dataset acts as priors or evidence for the next analysis. This process continues sequentially for all candidate datasets considered for data fusion until a final posterior distribution is obtained that incorporates information from each available dataset. It is anticipated that such a data fusion approach conducted within the Bayesian framework will not only provide more accurate parameter estimates than approaches typically used for data fusion, but will also lead to more accurate interpretations of obtained results.

It is important to note that this proposed Bayesian method does not strictly follow the exact definition of data fusion given previously. In this proposed Bayesian data fusion approach, although it is the individual datasets that are being analyzed, each analysis incorporates the data or the evidence from the previous studies. Therefore it still closely follows the general goals of data fusion, in that information from multiple datasets is being combined and analyzed. The main point is that the final posterior distribution will include information from all the datasets of interest. Because of this unique feature, the approach proposed in this dissertation is referred to as Bayesian Synthesis.

The proposed Bayesian Synthesis approach also bears some similarities to a technique referred to as *parallel analysis,* sometimes referred to as coordinated analysis with replication (Hofer & Piccinin, 2009; Piccinin & Hofer, 2008). The main purpose of a parallel analysis is the collaborative analysis of multiple independent datasets so that comparisons of results across studies can be made. To achieve this goal, raw data from multiple studies are analyzed individually using the same, or as similar as possible, analytic model and the results of these analyses are then synthesized or aggregated (often using traditional meta-analytic approaches). A *parallel analysis* is also sometimes categorized as a two-step approach that is closely connected to classical meta-analysis for analyzing aggregated data (Cooper & Patall, 2009; Jones, Riley, Williamson, & Whitehead, 2009). However, because there is no best way to summarize results obtained from a parallel analysis, simply aggregating research findings can be very problematic. For example, Marcoulides & Grimm (2016) recently examined the effectiveness of the parallel analysis approach to fitting different growth models and determined that the models did not always converge when individually fitted to separate datasets. In contrast, when the same datasets were combined into a single dataset and the growth models were fit to this fused dataset, the models were able to converge. Thus, providing additional evidence that data fusion methods can allow for the fitting of more complex models that might not otherwise have been possible to fit in a single dataset. It is believed that the proposed Bayesian Synthesis will also prove to be such an approach.

It was noted above that Bayes theorem can be thought of as a method for combining prior information with the currently available data in order to subsequently obtain parameter estimates. A simple way to express this idea is to restate Bayes theorem as

$$P(Unknowns \mid Data) \propto \frac{P(Data \mid Unknowns)P(Unknowns)}{P(Data)}$$

where the 'Unknowns' may be any parameter estimates (such as regression coefficients or some other quantity). As noted above, this equation can also be rewritten to indicate that the posterior distribution is simply proportional to a combination of the information contained in the data and the prior as

$$P(Unknowns \mid Data) \propto P(Data \mid Unknowns)P(Unknowns)$$

where *P(Unknowns)* represents the prior information about the unknown parameters, *P(Data| Unknowns)* represents information about the data given the unknown parameters, and *P(Unknowns| Data)* denotes the merging of the two sources of information into the posterior distribution for the unknown parameters.

Now if the combining of two datasets is considered, one can then view the prior information about the unknown parameters as equivalent to a data set that when merged with the current data would support the same kind of Bayesian inference to be made, which can be expressed as

$$P(Unknowns \mid Data_1, Data_2) \propto P(Data_2 \mid Unknowns) P(Unknowns \mid Data_1).$$

Subsequently, and as k additional datasets become available, these can be sequentially added to update the priors and the posterior distribution in the same manner. This can be expressed as

$$P(Unknowns \mid Data_1, Data_2, ...Data_{k+1}) \propto$$

$$P(Data_{k+1} \mid Unknowns) P(Unknowns \mid Data_1, Data_2, ...Data_{k+1}) P(Data_2 \mid Unknowns) P(Unknowns \mid Data_1).$$

It is this unique feature of combining information from multiple data sets and updating the priors in subsequent data analyses for the purpose of data fusion that will be the main focus of this dissertation. It is hypothesized that the sequential inclusion of informative data-dependent priors provides an extra source of information to estimate model parameters and that this additional information can effectively aid in the accuracy of parameters estimation and in the interpretation of results.

The next chapter discusses the various details of the analyses, designs, methods, and the data used to examine the effectiveness of the proposed Bayesian Synthesis approach. Although the benefits of using fused datasets have already been shown to be beneficial (e.g., Marcoulides & Grimm, 2016), what has not been determined or evaluated is whether a Bayesian approach based on a sequentially obtained final posterior distribution can be effectively applied as a new data fusion method.

Chapter 4

ANALYSES AND PROCEDURES

This chapter includes a description of the details of the proposed analyses for the empirical example and simulations to be conducted, both in terms of data generation and program implementation, the study design and computer programs used, and the indices to be used to examine parameter recovery.

Overview of Analyses

To examine the performance of the proposed Bayesian Synthesis approach described in the previous chapter, first Bayesian Synthesis results of simulated data with known population values under a variety of conditions will be examined. Next, these results will be compared to those from the traditional maximum likelihood approach to data fusion. In this approach, the individual data sets will be combined into one large data set and analyzed all at once. Additionally, to disentangle whether any differences in results are because of the Bayesian/frequentist difference, or the data fusion/synthesis difference, results from the data fusion approach analyzed via Bayes will also be compared. In this approach, the individual data sets will again be combined into one large data set and analyzed all at once as in the traditional maximum likelihood approach to data fusion, however, non-informative priors will be utilized for this fused analysis. Subsequently, empirical analyses with real data will be conducted. For this purpose, the fusion of real data from five longitudinal studies of mathematics ability varying in their assessment of ability and in the timing of measurement occasions will be used. Results from the Bayesian Synthesis and data fusion approaches with combined data using Bayesian and maximum likelihood estimation methods will be reported.

Monte Carlo Simulations

Paxton et al. (2001) presented different steps that should be followed by researchers when planning and performing Monte Carlo simulations. Based on their guidelines, the following key steps will be dealt with in this dissertation: (1) creating a valid model with specific conditions to fulfill the model, (2) choosing realistic population parameter values, (3) selecting a software package to conduct the simulations and the analyses, and (4) executing the simulations and summarizing results. Each of these steps is described next with the exception of step 4, which will be addressed in the next chapter. Although each step is presented separately, as Paxton et al. (2001) indicated, the steps are quite interconnected because decisions made in one step can influence another step. For example, selecting a certain statistical package for estimation can potentially limit one's ability to create a valid model (Paxton et al., 2001).

Creating a Valid Model with Specific Conditions to Fulfill the Model

The Monte Carlo simulations were modeled after the Marcoulides and Grimm (2016) study in which data from six longitudinal studies of mathematics ability were analyzed. The data in each of these six longitudinal studies varied in terms of sample size (e.g., n = 383 to n = 3,563), and in the timing and number of measurement occasions (e.g., observations taken at age 2 versus age 16, with anywhere from 3 to 10 measurement occasions). The six longitudinal studies were (1) The National Institute of Child Health and Human Development's (NICHD) Study of Early Child Care and Youth Development (SECCYD; NICHD Early Child Care Research Network, 2002), (2) National Center for Early Development and Learning's (NCEDL) Multi-State Pre-K Study (LoCasale-Crouch et al., 2007), (3) NCEDL's Study of State-Wide Early

Education Programs SWEEP; LoCasale-Crouch, et al., 2007), (4) Morrison's Longitudinal Study (MLS; Connor, Morrison, & Slominski, 2006), (5) Welfare, Children, Families: A Three City Study (WCF; Winston et al., 1999), and (6) the Panel Study of Income Dynamics Child Development Supplement (PSID; Hill, 1992) (for complete details of each study see Marcoulides & Grimm, 2016). The overall findings showed that as the children got older, their mathematics ability increased and the rate of change appeared to vary both within children over time and between children at any particular point in time.

Using the Marcoulides and Grimm (2016) study as a general guide of a valid model, the specific conditions selected to fulfill the various features of the model in order to generate the simulated data are described next. Before doing so, however, it is important to emphasize some fundamental assumptions that are needed in order to generate the synthetic data. Specifically, a fundamental assumption for generating the simulated data is that the observations in each synthetic dataset are sampled from the same population. Additionally, it is assumed that the time metric used to track change and the scaling of ability in each synthetic dataset are equivalent. Of course, in real data applications these assumptions may not be always fulfilled. For example, if studies use different instruments to measure the same construct, it is necessary to first scale the items to a common metric, potentially using an IRT approach (e.g., Curran et al., 2008, Marcoulides & Grimm, 2016; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). A prerequisite to implementing one of these IRT-based methods is that there are at least some common items across the scales used. Similarly, if it is unclear whether the data stem from a homogeneous population or from two or more subpopulations of individuals, then one of many different multilevel models can be used (e.g., Raykov et al., 2016; Rost & von Davier, 1993; Vermunt, 2003). These models can also effectively determine whether dissimilar groupings of item or ability parameters apply to different subpopulations.

Linear Growth Model. To generate the simulated data, a linear growth model will be used in which individuals are measured on an outcome y across multiple measurement occasions. Although the linear growth model may be considered to be of simple form, it has been widely used in developmental research due to its clear interpretation of model parameters (Liu et al., 2016). As will be described in more detail in the next section, the structure of the models to be used in this dissertation will exhibit positive growth over time with varying intercept and slope variances, and covariances that will include small, medium, and large values.

A linear growth model for the mathematics ability score can be written as

$$y_{\mathrm{tn}} = \eta_{0n} + \left(\frac{t - k_1}{k_2} \right) \cdot \eta_{1n} + e_{tn}$$

where y_{tn} is the outcome of interest at time t for individual n, η_{0n} is the latent variable intercept for individual n when $t = 0$, η_{1n} is the latent variable slope for individual n when $t = 0$, k_1 and k_2 are used to center the intercept and scale the slope, and e_{tn} is the residual at time t for individual n. The latent variables are assumed to follow a multivariate normal distribution with means, variances and covariances, $[\eta_{0n}, \eta_{1n}]' \sim N(\boldsymbol{\beta}, \boldsymbol{\Psi})$, and the residuals are assumed to follow a normal distribution with a mean of 0 and constant variance, $e_{tn} \sim N(0, \sigma_e^2)$. Figure 3 presents an example diagram of this model as a two-level model.

This linear growth model also follows the traditional convention of assuming there is a single residual variance (σ_e^2) over time and has covariances between measurement occasions that are zero. Such a constraint is quite common and reasonable whenever the same measuring instrument is used across measurement occasions and is expected to be consistent across time (McArdle & Grimm, 2010; Grimm & Widaman, 2010). Although the structure of residual variance is often seen as unimportant in growth modeling, Grimm and Widaman (2010) emphasized that bias in the latent variable variance-covariance matrix $\boldsymbol{\Psi}$ can sometimes arise from the specification of residuals. This is because there is an inverse relationship between the growth rate reliability and the magnitude of residual variance, thus the higher the growth rate reliability the smaller the residual variance (Willett, 1989). For this reason, and following the recommendations of Grimm and Widaman (2010), the population parameter values for the invariant residual variance used in the simulations in this dissertation were selected to ensure patterns of growth reliability that ranged between 0.8 and 1.0 (see also additional details provided in the section Choosing Realistic Population Values). Growth reliability estimates for each occasion of measurement can be computed using either $\boldsymbol{diag}\left[\frac{\boldsymbol{\Lambda\,\Psi\,\Lambda'}}{\boldsymbol{\Lambda\,\Psi\,\Lambda' + \Theta}}\right]$ as given by Grimm and Widaman (2010), or using Willett's (1989) equation $\boldsymbol{\rho_\theta} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \left[\frac{12\sigma_e^2}{t(t^2-1)}\right]}$, where $\boldsymbol{\rho_\theta}$ is the growth reliability, σ_θ^2 the slope variance, t the occasion of measurement, and σ_e^2 the residual variance. For example, using the growth model presented in Figure 3 above with $\boldsymbol{\Psi} = \begin{bmatrix} 0.60 & 0.50 \\ 0.50 & 0.50 \end{bmatrix}$ and a single residual variance σ_e^2 = 0.10 (implying a diagonal $\boldsymbol{\Theta}$ matrix), provides growth reliability estimates equal to 0.86, 0.95, 0.98, 0.99, 0.99, 0.99 respectively for each occasion of measurement. As indicated by Willett (1989), with sufficient measurement occasions, the influence of measurement error reduces to zero while growth reliability approaches 1.

Parameter estimation in a linear growth model involves the parameters $\boldsymbol{\beta}$, the variance-covariance matrix $\boldsymbol{\Psi}$, and the residual variance σ_e^2. Therefore, in order to apply a Bayesian analysis to the observed data that is to be sequentially fused, priors for these model parameters will be needed. For the purpose of this dissertation, non-informative priors for these parameters will be used for the initial analysis of the first data set. Based on results reported by Asparouhov and Muthén (2010), it is expected that the selection of these non-informative priors will not introduce bias in the computed parameter estimates, even in small sample size situations. Informative priors based on information obtained from the analysis of the previous dataset will be used for the parameters $\boldsymbol{\beta}$, the variance-covariance matrix $\boldsymbol{\Psi}$, and the residual variance σ_e^2 in the sequential analyses of the subsequent dataset. The final synthesized prior distribution will only be determined once all candidate data sets considered for data fusion have been used.

<u>Choosing Realistic Population Parameter Values</u>

The population parameter values for the intercept and slope means in $\boldsymbol{\beta}$, the slope and intercept variances and their covariance in the matrix $\boldsymbol{\Psi}$, and the common residual variance σ_e^2 were all selected to be similar to growth commonly seen in the developmental literature, where growth begins at early time points and continues linearly as time progresses. Using the Marcoulides and Grimm (2016) study as a general guide for selecting population parameter values, the following values were chosen. For all situations, 6 different datasets will be generated to implement the data fusion process, each with three different sample sizes considered, $N = 50$, $N = 250$, and $N = 1{,}000$

respectively. The sample sizes were selected to reflect small, medium, and large sample studies commonly encountered in the longitudinal literature (Harring, Weiss, & Hsu, 2012; McNeish, 2016; Paxton et al., 2001). Based on the empirical data analysis and with all estimates of ability scores specified to be normally distributed, the intercept and slope means will be fixed at $\boldsymbol{\beta_{Intercept}}$ = -2 and $\boldsymbol{\beta_{Slope}} = 0.4$ for all simulated conditions. Asparouhov and Muthén (2010) and McNeish (2016) indicated that in a linear growth model, the choice of prior distributions for the elements of the factor covariance matrix $\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}$ is vitally important compared to other parameters of the model. Accordingly, the magnitudes of the slope and intercept variances and the covariance in matrix $\boldsymbol{\Psi}$ will be varied and set at different levels to reflect small, medium, and large magnitudes, and the covariance will be set to reflect zero and small magnitudes. Thus the following three $\boldsymbol{\Psi}$ matrices will be used: $\boldsymbol{\Psi_1} = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$, $\boldsymbol{\Psi_2} = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$, and $\boldsymbol{\Psi_3} = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$. Finally, the common residual variance σ_e^2 will be fixed at 0.10 to reflect small amounts of residuals.

To begin the sequential process, initial non-informative priors will be used for the parameters in the first considered data set. For the intercept and slope means a Normal prior of the form N(mean, variance) will be used. Specifically, N(0, 10^{10}) will be used for the first data set, as this is also the default for a non-informative Normal prior in *Mplus* (Muthén & Muthén, 2012). For the parameters in the $\boldsymbol{\Psi}$ matrix, the Inverse Wishart prior of the form IW(S, *d*) will be used, where *d* is the pseudo-sample size and *S* is the scale matrix $\begin{bmatrix} d(\sigma^2_{Intercept}) & d(\sigma_{I\,S}) \\ d(\sigma_{I\,S}) & d(\sigma^2_{Slope}) \end{bmatrix}$, with the estimated intercept ($\sigma^2_{Intercept}$) and slope

(σ^2_{Slope}) variances and their covariance ($\sigma_{I\,S}$). The *Mplus* default non-informative prior, IW(0,-3), is utilized for the first data set in this dissertation (Muthén & Muthén, 2012). For the residual variance an Inverse Gamma of the form IG(α, β) will be used; $\alpha = \nu_0/2$ and $\beta = \nu_0\sigma_0^2/2$, where σ_0^2 can be interpreted as the best estimate of the variance and ν_0 can be interpreted as a pseudo-sample size. The default non-informative prior in *Mplus* that is utilized for the first data set is IG(-1, 0) (Muthén & Muthén, 2012).

After analyzing the first data set using the priors indicated above, posterior point summary estimates will be substituted into the respective priors for the next analysis, thus making the priors in the subsequent analyses informative priors. For the intercept and slope means in the second data set, the prior will be specified as $N(\beta_{\text{Intercept}\ dataset\_1}, \sigma^2_{\beta_{\text{Intercept}\ dataset\_1}})$, and in the third data set the prior will be specified as $N(\beta_{\text{Intercept}\ dataset\_2}, \sigma^2_{\beta_{\text{Intercept}\ dataset\_2}})$, thus specifying the prior for the next data set with the posterior point estimate of the mean and variance of the previous data set. This same prior specification will also be applied for the slope mean. For the parameters in the $\mathbf{\Psi}$ matrix in the second data set, the prior will be specified as $\text{IW}\left(\sigma^2_{\text{Intercept}\_dataset\_1} \cdot (\text{sample size}_{dataset\_1} - 3)\ , \text{sample size}_{dataset\_1} - 3\right)$, and in the third data set the prior will be specified as

$\text{IW}\left(\sigma^2_{\text{Intercept}\_dataset\_1} \cdot (\text{sample size}_{dataset\_1} + \text{sample size}_{dataset\_2} - 3)\ , \text{sample size}_{dataset\_1} + \text{sample size}_{dataset\_2} - 3\right)$,

thus increasing the pseudo-sample size of the current data set by the sample size of the previous data set. This same prior specification will also be applied for the slope variance and their covariance. For the residual variance in the second data set, the prior will be specified as IG $\left(\frac{samplesize_{dataset\_1}-2}{2},\frac{(samplesize_{dataset\_1}-2)\cdot\sigma^2_{e_{dataset\_1}}}{2}\right)$. And in the third data set the prior will be specified as

IG $\left(\frac{samplesize_{dataset\_1}+samplesize_{dataset\_2}-2}{2},\frac{(samplesize_{dataset\_1}+samplesize_{dataset\_2}-2)\cdot\sigma^2_{e_{dataset\_2}}}{2}\right)$, thus increasing the pseudo-sample size by the sample size of the previous data set, and using the posterior point estimate of the residual variance from the previous data set as the best estimate of the variance. This process of updating the prior for the next data set with the posterior point estimate from the previous dataset, and increasing the pseudo sample size of the next data set by the sample size of the previous data set continues for the fourth, fifth, and sixth datasets.

Additionally, the data sets will vary in number of measurement occasions, the spacing of those measurements, as well as the time at which the first measurement occurred. The decision to vary the timing of the first measurement occasion, the number of measurement occasions, and their spacing in each of the six datasets was again based on the data patterns observed in the Marcoulides and Grimm (2106) study. The number of measurement occasions in the six longitudinal datasets analyzed by Marcoulides and Grimm (2016) ranged from a low of 2 to a high of 10 occasions, with some measurements occurring at age 2 and others at age 16, for example. Furthermore, as in the Marcoulides and Grimm (2016) study, the ability scores for individuals on each occasion are on the theta scale and assumed to originate from fitting a one-parameter logistic model (1PL) to the item-level data.

Table 1 presents a list of the patterns of measurement occasions to be used in the simulated data. For example, dataset 1 will include 5 measurement occasions with the first starting at age 4.5, measured every two years, whereas dataset 3 will only have 2 measurement occasions starting at age 4, but measured twice per year. It is important to emphasize again that these data will be generated under the assumption that the observations in each synthetic dataset are sampled from the same population and that the time metric used to track change and the scaling of the ability in each dataset is equivalent.

Although sequential data analysis relies entirely on the updating of knowledge about an effect as new studies become available (e.g., updating ordered by publication date; Scheibehenne et al., 2016), it is also very possible that the actual order in which the datasets are used to compute priors may bias the proposed Bayesian Synthesis approach to data fusion. For this reason, the order in which the data are analyzed and the priors computed will also be varied in this dissertation. Specifically, two different data fusion sequences will be examined, one in which the datasets are analyzed and fused according to their rank order (sequentially from first to last; 1 to 6 – reflecting an ordering by publication date) and one in which the datasets are analyzed and fused in reverse rank order. However, according to the exchangeability assumption in Bayesian statistics, if the populations are the same, then the order in which the actual integration gets implemented in the data fusion process should not impact the final result (de Finetti, 1972, 1974). Where one starts should not in any way impact where one ends.

Parameter recovery criteria. The assessment of parameter recovery based on the proposed Bayesian Synthesis approach for all linear growth models will be evaluated using four criteria to reflect measures of raw bias, relative bias, accuracy, and efficiency (Bandalos & Gangé, 2012; Bandalos & Leite, 2013). These measures are some of the most common criteria used in the literature to quantify the accuracy of parameter recovery (Bandalos & Gangé, 2012; Bandalos & Leite, 2013).

Raw bias indicates the average deviation of an estimate from the true population parameter. Using the notation y to correspond to the population value of a parameter that is to be estimated, the raw bias ($B(\hat{y})$) for a parameter estimate $\hat{y}$ of the parameter y across replications is defined as follows,

$$B(\hat{y}) = \frac{\sum_{r=1}^{R}(\hat{y}_r - y)}{R}$$

where R corresponds to the total number of simulation replications conducted. The value of $B(\hat{y})$ indicates overestimation or underestimation when it is equal to a nonzero positive or negative value, respectively.

Using again the notation y to correspond to a parameter that is to be estimated in the model, the relative bias for a parameter estimate $\hat{y}$ for parameter y within a single replication ($RB(\hat{y})$) is defined as a percentage like scale as follows,

$$RB(\hat{y}) = \left(\frac{\hat{y} - y}{y}\right) \times 100$$

so that if the true parameter y is set at a value of 0.50 and the estimated parameter $\hat{y}$ is determined to be 0.60, then relative bias would be equal to 20, which can be simply interpreted as 20% positive bias. It is generally recognized in the literature that relative bias less than 5% is ignorable, between 5% and 10% is moderately biased, and values above 10% indicate substantial bias (Flora & Curran, 2004; Muthén & Muthén, 2002). An excellent feature of determining relative bias is that it can be used for comparisons of the magnitude of bias across different conditions, as all values are interpreted on the same percentage scale. Because multiple simulation replications will be conducted (in this case

250 replications will be analyzed for all linear growth models examined), average values of relative bias for each evaluated parameter are generally reported across all the replications. The average value for $RB(\hat{y})$ across all replications will be denoted as $\overline{RB(\hat{y})}$ and in the same manner, with values less than 5% reflecting ignorable bias, between 5% and 10% moderate bias, and values above 10% substantial bias (Flora & Curran, 2004; Muthén & Muthén, 2002).

The accuracy of parameter estimates can be defined as RMSE$(\hat{y})$, and is computed as follows,

$$RMSE\ (\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{y}_r - y)^2}{R-1}}$$

where R corresponds to the total number of simulation replications conducted, which in this dissertation will be set at 250. This accuracy criterion reflects the square root of the average deviation of the sample estimates from their population value squared, with lower values of accuracy corresponding to more precise estimates of the parameters or estimates of parameters that exhibit a smaller range of error (Bandalos & Gangé, 2012).

The efficiency of parameter estimates can be computed as

$$Efficiency(\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}\left(\hat{y}_r - \overline{\hat{y}}\right)^2}{R-1}}$$

where $\hat{y}$ is the parameter estimate, $\overline{\hat{y}}$ is the mean of the parameter estimates, and R is again the total number of simulation replications conducted. This statistic reflects variability around the sample mean. Values closer to zero correspond to more efficient estimates of the parameters. In other words, smaller values correspond to a smaller range of variability, or higher consistency of estimation.

Software

All simulated data will be generated in the publically available software program R (R Core Team, 2010) and all growth models will be fit using the commercially available software program *Mplus* (Muthén & Muthén, 2012). First, the R program code will specify the population parameter values. From this population, two hundred and fifty replications of six different datasets will be generated according to each of the specified design conditions and sample sizes previously described. Next the MplusAutomation package (Hallquist & Wiley, 2014) in R will be used to create *Mplus* input files that are then automatically executed in *Mplus*. Appendix B presents example code.

The analyses conducted in *Mplus* were specified to have no thinning, using the Gibbs (PX1) algorithm with a minimum of 50,000 iterations, using the Potential Scale Reduction (PSR) convergence criteria, a median summarized posterior, and running two chains (Muthén & Muthén, 2012). The selection of these program options was directed by past research findings using Bayesian estimation routines. For example, Link and Eaton (2012) determined that thinning is not necessary as long as the specified chains are allowed to run long enough, which in this case were set at two chains and a minimum of 50,000 iterations. Two chains were selected as past research by Asparouhov and Muthén (2010) has indicated that this number should be sufficient for most applications (which is also the reason the default value in the *Mplus* program is set at 2 chains). To improve estimation, each chain begins from a different start value using different seeds to make the random draws and then independent sequences of iterations occur at each chain (Muthén & Muthén, 2012). Another *Mplus* default for Bayesian estimation based on past research findings is the use of the median summarized posterior (although this can easily be changed to a mean summarized posterior value using the POINT=MEAN option in *Mplus*; Muthén & Muthén, 2012). Finally, convergence of the Gibbs (PX1) algorithm is assessed using the Potential Scale Reduction (PSR) convergence criterion. This convergence criterion examines the within chain and the between chain variability of the computed parameter estimates. As indicated by Gelman and Rubin (1994) and Gelman et al. (2004), convergence occurs when the within chain variance and the between chain variance are similar, with obtained PSR values approximately equal to 1, indicating convergence. Based on this research, the default PSR value used to assess Bayesian convergence in *Mplus* is set at 1.10 (although there is also an option available to change this option and make the criterion more or less stringent). For the analyses conducted in this dissertation, PSR values of 1.10 will be used to establish convergence.

Following the specifications of all the above described models and conditions, and once all the simulations are executed, the next activity will be to summarize the results and then evaluate whether the proposed Bayesian Synthesis approach based on a sequentially obtained final posterior distribution can be effectively applied as a new data fusion method. This will be the topic of the next chapter.

Real Data Analyses

To illustrate the use of the proposed Bayesian Synthesis approach in realistic settings, fused analyses with real data will also be evaluated. For this purpose, the fusion of real data from five longitudinal studies of mathematics ability will be performed. By conducting such an empirical analysis, it is hoped that the practical value of the proposed approach will be demonstrated.

The data for this empirical example come from five cohorts of the Head Start Family and Child Experiences Survey (FACES), which longitudinally study a number of outcomes over time. Each cohort consists of a nationally representative sample of 3 to 4 year old children enrolled in Head Start programs: (1) FACES 1997 included about 3,200 children who were enrolled in 40 Head Start programs, (2) FACES 2000 included about 2,800 children enrolled in 43 Head Start programs, (3) FACES 2003 included about 2,400 children who were enrolled 63 Head Start programs, (4) FACES 2006 included about 3,500 children who were enrolled in 60 Head Start programs, and (5) FACES 2009 included about 3,300 children who were enrolled in 60 Head Start programs. Data were collected in the fall and spring of the study's first year when all children attended Head Start; the spring of the study's second year when some children were in their second year of Head Start and some children were in kindergarten; and in the FACES 1997, there was a fourth assessment in the spring of the study's third year when some children were in kindergarten and some students were in first grade.

To examine mathematics ability, all individuals in these five studies were measured using the Applied Problems (AP) subtest of the Woodcock–Johnson Psycho-Educational Battery. The AP subtest measures early mathematics reasoning and problem-solving abilities, which requires children to analyze and solve mathematics problems while performing simple calculations. The FACES 1997 and 2000 cohorts used the Woodcock–Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock & Johnson, 1990) and the FACES 2003, 2006, and 2009 cohorts used the Woodcock–Johnson Psycho-Educational Battery–III (WJ-III; Woodcock, McGrew, & Mather, 2001). The AP test from the WJ-R contains 60 items, the AP test from the WJ-III contains 63 items, and the two versions have 39 items in common.

Before conducting any analyses, and to address the issue of different versions of the AP subtest, the same equating strategy applied in the Marcoulides and Grimm (2016) study was used. Marcoulides and Grimm (2016) fit a one-parameter logistic (1PL) item response model to the item-level data to estimate a latent variable score that was not dependent on the version of the subtest. By fitting a 1PL model to the item-level data from each of the AP subtest, Marcoulides and Grimm (2016) were able to analyze the data as though they formed a single test, where all items that were not administered because the items only appeared in one version of the subtest were considered missing. Based on this 1PL model specification, items that were common to both versions had a single set of item parameter estimates, which scaled the latent variable in the same metric regardless of the version of the Woodcock–Johnson used in a study. Using a translation table of the expected a posteriori estimates of latent ability, the AP scores from each study were placed on the same scale and were thus comparable across studies. It is these values that are used in the Bayesian Synthesis, the maximum likelihood approach to data fusion, as well as the Bayesian data fusion approach.

Chapter 5

RESULTS

This chapter presents the results of the simulations as well as the results from the real data example. The analyses of the simulated data are presented first and the analyses of the real data are presented second. For the simulated data, the results are organized by (i) magnitude of the population parameters of the slope and intercept variances and the covariance in the matrix $\boldsymbol{\Psi}$ (reflecting zero, small, medium, and large magnitudes), (ii) sample sizes considered (N = 50, N = 250, and N = 1000), and (iii) data fusion approach (Bayesian Synthesis, Bayesian data fusion, and maximum likelihood data fusion). For the real data, model parameter estimates based on Bayesian Synthesis and data fusion approaches with combined data using Bayesian and maximum likelihood estimation methods are also reported. Because with the real data the population parameters are unknown, no conclusions can be made about which approach gave more reliable estimates. The empirical analyses are only provided as a demonstration and an examination of the estimation effectiveness of the three different data fusion approaches. For the Bayesian Synthesis approach, results are also summarized from analyses in which the sequential ordering of the data integration process was varied. These analyses were conducted to examine the exchangeability assumption in Bayesian statistics and determine whether the order in which the data integration gets implemented in Bayesian Synthesis impacts the obtained final results.

Simulation Results

The simulation results for examining the performance of the proposed Bayesian Synthesis approach and the data fusion approaches with the combined data based on

Bayesian and maximum likelihood estimation approaches are presented in Tables 2 through 10. Parameter estimates in the specified linear growth model include the latent variable means $\boldsymbol{\beta}$, the variance-covariance matrix $\boldsymbol{\Psi}$, and the residual variance σ_e^2. Results (raw bias, average relative bias, accuracy, and efficiency) are presented for each estimated parameter based on each data fusion approach. As described in detail in the previous chapter, for the Bayesian Synthesis approach non-informative priors for these parameters were used for the analysis of the first data set and informative priors, based on information obtained from the results from the previous dataset, were used for the priors of the model parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, σ_e^2) in the analyses of the subsequent dataset.

The results for the first specified covariance matrix $\boldsymbol{\Psi_1} = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$ for a sample size $N = 50$, $N = 250$, and $N = 1{,}000$ are presented in Tables 2, 3, and 4. As can be seen by examining Table 2, the relative bias for the estimate of the variance of the intercept under the Bayesian Synthesis approach was above 10%, indicating substantial positive bias with respect to this parameter. In contrast, the relative bias for the intercept variance when using either Bayesian data fusion or the maximum likelihood data fusion approaches was less than 5% and were thus ignorable. The relative bias for the slope variance parameter estimate determined by the Bayesian Synthesis, Bayesian data fusion and maximum likelihood data fusion approaches were all less than 5% and thus ignorable for all three approaches. Note that the relative bias for estimates of the intercept-slope covariance cannot be computed, as the population value was zero. The magnitudes of the relative bias for the latent variable means $\boldsymbol{\beta}$, and residual variance σ_e^2, were found to be ignorable for all three approaches.

The observed substantial positive bias for the variance of the intercept using the Bayesian Synthesis approach suggests that sample size can play an important role in its ability to estimate parameters when compared to the other examined data fusion approaches. Despite this finding, the observed differences between the three data fusion approaches are to some degree expected as the criterion measures of the parameter estimates reported for the Bayesian Synthesis were based on a $N = 50$ sample size, whereas those for the Bayesian and maximum likelihood data fusion approaches were based on the combining of observations from six data sets, which results in a sample size of 300. However, by incorporating information from the other data sets in the Bayesian Synthesis approach, the pseudo-sample size is effectively increased by the respective sample size of each data set incorporated in the analysis. Therefore, the bias may be due to having an initial small sample of N=50 (McNeish, 2016). When the first data set is small (e.g. $N = 50$), the posterior point estimates that are produced can contain some bias. Then, by using the informative priors based on these potentially biased estimates, we maintain this bias across the additional samples. Further evidence of this inference can be seen by examining the magnitude of the measures of relative bias for the Bayesian Synthesis approach in Tables 3 and 4, which are based on results from $N = 250$ and $N = 1{,}000$ for the same first specified covariance matrix $\mathbf{\Psi_1}$. With larger sample sizes, the obtained relative bias criterion clearly shows that all three approaches provide estimates that are close to the true parameters. It is important to note that with $N = 250$ and $N = 1{,}000$, the Bayesian and maximum likelihood data fusion approaches are in fact based on results for $N = 1{,}500$ and $N = 6{,}000$ sample sizes, respectively. Nevertheless, even under such disparate sample size comparisons, the Bayesian Synthesis approach provides accurate parameter estimates when applied to larger sample sizes.

With respect to the other criterion measures, as can be seen by examining Tables 2, 3, and 4 the values of the raw bias (indicating the average deviation of an estimate from the population parameter - with nonzero positive or negative values indicating overestimation or underestimation, respectively), for the Bayesian synthesis, the Bayesian data fusion, and the maximum likelihood data fusion approach were all small and not substantially different from zero, indicating that the bias is ignorable. Similarly, examining the values for the accuracy and efficiency criteria for all parameter estimates provided in Tables 2, 3, and 4 it can be seen that these were also extremely small and correspond to precise estimates of the model parameters. Because these three criteria (raw bias, accuracy, and efficiency) indicate that the Bayesian Synthesis approach is performing as well as the Bayesian data fusion and the maximum likelihood data fusion approaches, then the substantial relative bias found for the intercept variance parameter using the Bayesian synthesis approach may not be as problematic as one might initially expect. Overall these findings illustrate that the three data fusion approaches provided relatively consistent parameter estimates under the various design conditions examined.

The results for the second specified covariance matrix $\mathbf{\Psi_2} = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ with sample sizes $N = 50$, $N = 250$, and $N = 1{,}000$ are presented in Tables 5, 6, and 7. As can be seen by examining Table 5, the relative bias for the estimate of the variance of the intercept under the Bayesian Synthesis approach was 5.434% - moderately biased. The relative bias for the intercept variance parameter estimate determined by both the Bayesian data fusion and the maximum likelihood data fusion approaches were less than

5% and thus ignorable. The relative bias for the slope variance parameter estimate determined by all three data fusion approaches was less than 5% and thus ignorable. In contrast, the relative bias of the Bayesian Synthesis approach estimate for the intercept-slope covariance was -44.904% indicating substantial negative bias, while at -9.728% it was moderately biased for the Bayesian data fusion approach. The relative bias of the maximum likelihood data fusion approach estimation of the same parameter was found to be only -2.712% and thus was considered ignorable. The magnitudes of the relative bias for the $\boldsymbol{\beta}$ parameters and σ_e^2 were again found to be ignorable for all three data fusion approaches.

The observed relative bias for the variance of the intercept and the intercept-slope covariance using the Bayesian Synthesis approach once again highlights the importance that sample size plays in the estimation of parameters when implementing data fusion strategies. Interestingly, for the intercept-slope covariance parameter estimate even the Bayesian data fusion approach (which was based on a combined sample size of $N = 300$) was also moderately negatively biased, whereas the maximum likelihood data fusion approach was only marginally biased (and considered ignorable). These results suggest that with small population intercept-slope covariance or correlation values (in this case r = 0.188), sample size can play a key role in impacting the accurate estimation of the parameter values. This finding is in line with past research that has determined that small sample sizes do not always permit a researcher to accurately estimate low valued coefficients until much larger sample sizes are used (Chin & Newsted, 1999; Muthén & Muthén, 2002). This observation is further supported when examining the results presented in Table 8, where estimating a population correlation value set at r = .50 the

Bayesian and maximum likelihood data fusion approaches do not exhibit any sizeable bias. Examining the measures of relative bias for the three data fusion approaches presented in Tables 6 and 7 (based on results from $N = 250$ and $N = 1000$) it can been seen that once larger sample sizes are used, all three approaches provide estimates that are close to the true parameters. As can also be seen by examining Tables 5, 6, and 7 the values of the raw bias and those for the accuracy and efficiency criterion for all parameters estimated with the Bayesian synthesis, the Bayesian data fusion, and the maximum likelihood data fusion approach are again all small and ignorable. Once again, while there was substantial bias indicated by the relative bias criterion, the other three criteria indicated that the Bayesian Synthesis approach was performing equally well as Bayesian data fusion and maximum likelihood approach in estimating the various parameters. This suggests that Bayesian Synthesis is in fact performing well and is another viable method for performing data fusion activities.

Finally, the results for the third specified covariance matrix $\boldsymbol{\Psi}_3 = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$ with sample sizes $N = 50$, $N = 250$, and $N = 1{,}000$ are presented in Tables 8, 9, and 10. As can be seen by examining Table 8, the relative bias for the estimate of the variance of the intercept under Bayesian Synthesis, Bayesian data fusion, and the maximum likelihood data fusion approach were less than 5% and are thus ignorable. Additionally, the relative bias for the slope variance parameter estimate determined by all three data fusion approaches were also less than 5% and thus ignorable. Only the relative bias for the intercept-slope covariance provided by the Bayesian Synthesis approach at a value of -9.208% was found to be moderately biased when using a sample size of $N = 50$. The relative bias of the Bayesian data fusion and the maximum likelihood data fusion approaches for the same parameters were found to be ignorable. The magnitudes of the relative bias for the parameters $\boldsymbol{\beta}$ and σ_e^2 were also found to be small and ignorable for all three data fusion approaches. Examining the measures of relative bias for the three data fusion approaches presented in Tables 9 and 10 (based on results obtained from $N = 250$ and $N = 1000$) it can been seen that once larger sample sizes are used, all three approaches provide estimates that are close to the true parameters. As can also be seen by examining Tables 8, 9, and 10, the values of the raw bias and those for the accuracy and efficiency criterion for all parameters estimated with the Bayesian synthesis, the Bayesian data fusion, and the maximum likelihood data fusion approach are again all small and ignorable. As with the first two $\boldsymbol{\Psi}$ matrix conditions, this one moderately biased parameter does not necessarily indicate a clear problem with the Bayesian Synthesis approach. In fact, because the other three criteria (raw bias, accuracy, and efficiency) indicate that Bayesian Synthesis is performing as well as the other two approaches, this can be seen as evidence in support of the Bayesian Synthesis approach as a new and viable method for data fusion activities.

Real Data Results

To illustrate the use of the proposed Bayesian Synthesis approach and compare its effectiveness to the other data fusion approaches based on Bayesian and maximum likelihood estimation methods, the different approaches were applied to the analysis of empirical data collected as part of five cohorts of the Head Start Family and Child Experiences Survey (FACES). Previous analyses of the data (Zill et al., 2003) and their trajectory plots (found in Appendix B) suggested that a linear growth model was plausible for the current data and, therefore, a linear growth model was fit to the data.

Model parameter estimates based on Bayesian Synthesis, Bayesian data fusion and the maximum likelihood data fusion approaches are presented in Table 11. Estimation in this linear growth model involves the parameters $\boldsymbol{\beta}_{intercept}$, $\boldsymbol{\beta}_{slope}$, the variance-covariance matrix $\boldsymbol{\Psi}$ (with elements $\sigma^2_{intercept}$, σ^2_{slope}, and σ_{is}) and the residual variance σ^2_e.

To get the Bayesian Synthesis estimates, results obtained from the analysis of one dataset acted as priors in the next data analysis. This process was applied sequentially for all candidate datasets until a final posterior distribution was produced that incorporated the information from each available dataset. This final posterior distribution was then used to determine the Bayesian Synthesis point summary estimates. The same sequential process was applied to the empirical data for a second time, however, using a varied ordering of the datasets. The obtained values using this process are also presented in Table 11 using the notation BS_R to denote the Bayesian Synthesis with a reversed order of data integration. As described in detail in the *Analyses* and *Procedures* chapter, initial non-informative priors will be used for the parameters in the first considered data set, and posterior point summary estimates will be substituted into the respective priors for the next analysis, thus making the priors in the subsequent analyses informative priors. In order to get the Bayesian data fusion and the maximum likelihood data fusion estimates, the five datasets were combined to form a single dataset and then analyzed. For the Bayesian data fusion approach, non-informative priors were selected for all estimated parameters ($\boldsymbol{\beta}$, σ^2_e, and $\boldsymbol{\Psi}$).

Based on the linear growth model fit to the data in the FACES studies using the Bayesian Synthesis approach, the Bayesian data fusion approach, and the maximum likelihood data fusion approach, the average mathematics ability score for a 4-year-old was -1.558, -1.563, and -1.56, and the mean annual rate of change was 0.478, 0.482, and 0.478 points per year respectively. These theta scores correspond to an average raw score of about 5 on the WJ-R and an average raw score of about a 6 on the WJ-III for a 4-year-old. Examining the intercept and slope variances, children significantly differed in their mathematics ability at age 4 and in their rate of growth over time using all three data fusion approaches.

As can be seen by examining Table 11, the values for the parameter estimates and corresponding standard deviations obtained when using the Bayesian Synthesis (with both sequential and varied ordering of the data sets), the Bayesian data fusion, and the maximum likelihood data fusion approach were all very similar. The results presented in Table 11 also provide support for the exchangeability assumption and demonstrate that the order in which data integration process in the Bayesian Synthesis approach does not impact the final results. These results illustrated that where one starts does not substantially impact where one ends. Based on these results it appears that the proposed Bayesian Synthesis approach does provide accurate parameter estimates that lead to similar interpretations of obtained results as the other examined data fusion methods.

Chapter 6

DISCUSSION

Data fusion methodology is an emerging field and determining the most appropriate integration method to use is critical. The decision concerning which method to use is even more difficult when studying processes longitudinally and when different measurement tools are used to assess the same construct. Although some recent research has demonstrated that data fusion methods can be notably beneficial for studying developmental processes that are difficult to model (e.g., Marcoulides & Grimm, 2016), so far these methods have not been systematically examined or commonly applied in the psychological sciences. The purpose of this dissertation was to examine the effectiveness of a newly proposed Bayesian Synthesis approach for analyzing fused longitudinal data and to determine whether it can provide accurate parameter estimates. This chapter provides a discussion of the implications of the obtained results, highlights some limitations, and offers suggestions for further research.

Overview and Implications of Results

To date most data fusion approaches suggested in the literature operate by combining data sets into a single unit before any analyses are conducted. In contrast to these approaches, the proposed Bayesian Synthesis method is based on a Bayesian framework in which information obtained from one dataset serves to provide prior information for the analysis of the next data set. This process continues sequentially until a single posterior distribution is created using all available data. Although some benefits of using fused datasets have been shown in the literature (e.g., Marcoulides & Grimm, 2016), what has not been determined or evaluated is whether estimates computed via a sequentially obtained final posterior distribution can be effectively applied as a data fusion method. It was hypothesized that the sequential inclusion of informative augmented data-dependent priors would provide an extra source of information to estimate model parameters and that this additional information could effectively aid in the accuracy of the estimation process and thus in the proper interpretation of results. The proposed Bayesian Synthesis approach can broadly be considered a combination of data fusion strategies with sequential data analysis methodology. To date such a combination has not been merged into a single analysis strategy that can be used to handle situations in which multiple sources of data are investigated. To examine the exchangeability assumption in Bayesian statistics and determine whether the order in which the data integration gets implemented in the proposed Bayesian Synthesis approach impacts the results, analyses in which the sequential ordering of the real data integration process was varied were also conducted.

This dissertation examined the proposed Bayesian Synthesis approach compared to other approaches using both real and simulated data. For the real data, five publicly available longitudinal studies of mathematics ability in children were compiled with the goal of modeling individual change over time. Although the studies varied slightly in their assessment of ability as well as in the timing and number of measurement occasions, these issues were dealt with by utilizing an item response model to estimate latent variable scores that were not dependent on the measurement instrument and by estimating a linear growth model in the multilevel modeling framework. The simulated data were based on longitudinal data used in a study by Marcoulides and Grimm (2016); synthetic data for six longitudinal studies of mathematics ability were generated and analyzed using a linear growth model. The general structure of these simulated data resemble those encountered in real longitudinal studies, but with known population parameter values.

Analyses of the real data sets using Bayesian Synthesis (with both sequential and non-sequential ordering of the data sets), and the Bayesian data fusion and the maximum likelihood data fusion approaches indicated that all three approaches provided very similar parameter estimates. The results also provided support for the exchangeability assumption in Bayesian statistics and demonstrated that the order in which the data integration process occurs in the Bayesian Synthesis approach does not impact the final results. These findings clearly illustrate that where one starts in the Bayesian Synthesis approach does not substantially impact where one ends. Based on these outcomes with the real data, it can be concluded that the proposed Bayesian Synthesis approach provides accurate parameter estimates and leads to similar interpretations of obtained results as the other two examined data fusion methods, both of which are based on analyses that occur after combining the data sets into a single data unit.

Analyses of the simulated data sets using Bayesian Synthesis (with both sequential and non-sequential ordering of the data sets), and the Bayesian data fusion and the maximum likelihood data fusion approaches provided two general patterns of results: with large sample sizes Bayesian Synthesis performed similarly to the other data fusion approaches, whereas with smaller samples there was some bias with certain model parameters. The similarity of parameter estimates determined by the Bayesian Synthesis and the other examined data fusion approaches when large samples were involved was also observed with the empirical datasets, given that each examined dataset included data from approximately 3,000 observations.

When the sample sizes of the simulated data were small (N = 50), the Bayesian Synthesis approach occasionally provided estimates with sizeable relative bias, particularly when estimating low valued growth parameters. As larger sample sizes were used, the accuracy of the obtained parameter estimates increased considerably. This finding was in line with past research that has determined that small sample sizes do not always permit a researcher to obtain accurate parameter estimates, especially when estimating low valued coefficients (Chin & Newsted, 1999; Muthén & Muthén, 2002).

It is commonly acknowledged in the literature that sample size plays an important role in just about every statistical technique applied in practice. There is also complete agreement among researchers that the larger the sample the more stable the parameter estimates. Muthén and Muthén (2002), for example, clearly demonstrated how parameter estimates greatly improve and their error rates diminish as sample sizes increase. The results obtained in this dissertation corroborate these commonly accepted views about sample size and confirm that when using Bayesian Synthesis the posterior point estimates of the parameters consistently converge to the true parameters of the model when applied to data sets that contain 250 or more observations. Researchers considering the use of Bayesian Synthesis in practical applications will need sufficiently large data sets in order to make sure that the parameter estimates are accurate. However, bias in certain parameters estimated via the Bayesian Synthesis approach was only found with one criterion, the relative bias. The other three criteria (raw bias, accuracy, and efficiency) used to evaluate the Bayesian Synthesis performance indicated that it was performing as well as the Bayesian data fusion and the maximum likelihood data fusion approaches. Therefore, while there is some evidence for bias in small sample sizes using the Bayesian

Synthesis approach, there is in fact more evidence that Bayesian Synthesis is providing accurate, efficient, and unbiased estimates even with small sample sizes.

Although the Bayesian Synthesis approach relies entirely on the updating of information as new studies become available, and so it is very possible that the actual order in which the data sets are used to compute priors may impact the proposed approach to data fusion, results from the analyses of the simulated data provided additional support for the exchangeability assumption in Bayesian statistics. Given that near identical results were also observed with the real data, the order in which the data integration process occurs in the Bayesian Synthesis approach does not appear to have a substantial impact the final results. Thus, where one starts in the Bayesian Synthesis approach does not substantially affect where one ends.

Based on the obtained results with both the real data and the simulated data it can be concluded that with large enough sample sizes the proposed Bayesian Synthesis approach does provide accurate parameter estimates. The inclusion of informative augmented data-dependent priors does in fact provide an extra source of information to estimate model parameters and this additional information does effectively aid in the accuracy of the estimation and thus in the interpretation of results. The newly proposed Bayesian Synthesis approach based on sequentially updating information can be effectively applied as a new data fusion method.

Limitations and Directions for Future Research

In spite of the many benefits of the proposed Bayesian Synthesis approach for integrating data, there are some limitations that must be kept in mind. The first general issue with data fusion methods is the need to obtain *raw* data from the multiple studies, which can be problematic if researchers are not willing to share their data. The real data used in this study were publicly available in a data repository. Hopefully, as more researchers become willing to share their data and the various granting agencies continue to require collected data to be deposited in a public repository, this issue will ultimately become less problematic. However, in light of this limitation, the Bayesian Synthesis approach can provide an alternative to obtaining raw data through the incorporation of prior information from published studies. The estimates from a published study can be converted into priors and utilized as prior information for the subsequent analysis of the next data set of interest. Using just the estimates and summary statistics from published studies eliminates the necessity for obtaining raw data in order to incorporate a study as prior information for the Bayesian Synthesis approach.

A fundamental assumption for generating the simulated data examined in this dissertation was that the observations in each dataset were sampled from the same population and that the scaling of ability in each synthetic dataset was equivalent. Of course, in real data applications these assumptions may not be always fulfilled. If it is unclear whether the data stem from a homogeneous population or from two or more subpopulations of individuals, then before applying Bayesian Synthesis one of many different multilevel models introduced in the literature may first have to be used (e.g., Raykov et al., 2016; Rost & von Davier, 1993; Vermunt, 2003). These models can also effectively determine whether dissimilar groupings of item or ability parameters apply to different subpopulations. Additionally, if studies use different instruments to measure the same construct, it will be necessary to first scale the items to a common metric –

potentially using an IRT approach (e.g., Curran et al., 2008, Marcoulides & Grimm, 2016; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009).

Another potential limitation is that the data fusion analyses in this dissertation only examined growth parameter estimates for a linear growth curve model. Although the linear growth curve model is widely used in developmental research due to its clear interpretation of model parameters, future research may also want to consider the effectiveness of the Bayesian Synthesis approach with nonlinear trajectories. Furthermore, the linear growth models utilized in this dissertation were based on the traditional convention of assuming that there is a single residual variance (σ_e^2) over each time and has covariances between measurement occasions that are zero. Although such a restriction is quite common and reasonable whenever the same measuring instrument is used across measurement occasions and is expected to be consistent across time (McArdle & Grimm, 2010; Grimm & Widaman, 2010), future research might also consider examining situations in which this specification of residuals is varied. Additionally, covariates were not included in the models in this dissertation. Future research should also be conducted to determine the best way to incorporate covariates in the Bayesian Synthesis approach.

This dissertation utilized point summary estimates of the posterior distributions instead of the actual full posterior distributions as required by a fully Bayesian execution of this Bayesian Synthesis approach. While this may be seen as a limitation of this approach, using point summary estimates of the posterior distributions can be seen as a benefit by increasing ease of execution. By utilizing point summary estimates of the prior distributions instead of the actual distributions, the Bayesian Synthesis approach can be executed using the commonly used software program *Mplus* (or the MplusAutomation package in R that calls *Mplus*; Hallquist & Wiley, 2014). This will allow more researchers to easily implement this Bayesian Synthesis approach in their research. The price of approximating the final posterior distribution by using point summary estimates outweighs the difficulties of incorporating the full distributions in the sequential analysis.

Finally, future research should also simulate these conditions using a variety of parameter values to determine whether the results generalize to other population models. Future research should also explore how small the sample size can be in the Bayesian Synthesis approach to ensure sufficiently stable parameter estimates. It was determined that with $N = 50$ the parameter estimates provided by Bayesian Synthesis can be biased, whereas $N = 250$ accurate parameter estimates were obtained. This future work could investigate the question: What is the smallest sample size that can be used in Bayesian Synthesis before the magnitude of bias is substantial? Given that a major benefit of the Bayesian Synthesis approach is that data from multiple sources can be analyzed to obtain estimates of overall effects, examining the conditions under which this approach does not operate well is a natural extension to this sequential data fusion approach.

Concluding Remarks

The process of sequential learning to arrive at conclusions that might not always be attainable from each separate source has a long history in the social and behavioral sciences. The newly proposed Bayesian Synthesis was shown to be a valuable data fusion approach that can be effectively used to help researchers address questions that may not always be achievable with a single study. An important challenge to data analysis for the foreseeable future will involve applying this new data fusion approach to diverse data and across different disciplines in a way that is accurate and accessible to a broad range of researchers. Although additional research needs to be done regarding when the Bayesian Synthesis approach is most useful and when it might prove to be problematic, the foundations for the continued evolution of this data fusion method have been formed.

References

Ahmed, I., Sutton, A.J., & Riley, R.D. (2012). Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ, 344,* 1-10.

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis of latent variable models using *Mplus*, Version 2. Unpublished manuscript. Retrieved from http://www.statmodel.com.

Bandalos, D.L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In. R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 92-110). New York, NY: Guildford Press.

Bandalos, D.L. & Leite, W.L. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd Ed.). (pp. 564-666). Greenwich, CT: Information Age Publishing.

Béguin, A.A., & Hanson, B.A. (2001, April). Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle.

Béguin, A.A., Hanson, B.A., & Glas, C. A.W. (2000, April). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

Bennett, M.M., Crowe, B.J., Price, K.L., Stamey, J.D., & Seaman Jr, J.W. (2013). Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data. *Journal of Biopharmaceutical Statistics*, *23*, 129-145.

Berkey, C. S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics*, 38, 953-961.

Bhattacharua, B., & Saha, B. (2015). Community model: A new data fusion filter paradigm. *American Journal of Advanced Computing, 2,* 25-31.

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies, 70,* 1-19.

Bond, C.F., Wiitala, W.L., & Richard, F.D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406-418.

Boström, H., Andler, S.F., Brohede, M., Johansson, R., Karlsson, A., van Laere, J., Niklasson, L., Nilsson, M., Persson, A., & Ziemke, T. (2007). *On the definition of information fusion as a field of research*. Technical Report, Skövde, Sweden: University of Skovde, School of Humanities and Informatics.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics, 2*, 485–500.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49,* 327-335.

Chin, W.W., and Newsted, P.R. (1999). Structural equation modeling analysis with small samples using partial least squares. In R.H. Hoyle (Ed.). *Statistical strategies for small sample research (pp. 307-341).* Thousand Oaks, CA: Sage Publications.

Collins, L.M., Schafer, J.L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Cook, L.L., Eignor, D.R., & Wingersky, M.S. (1987). Specifying the characteristics of linking items used for item response theory item calibration (Research Bulletin No. RB-87-24). Princeton, NJ: Educational Testing Service.

Cooper, H.M., & Patall, E.A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14,* 165-176.

Curran, P.J., & Hussong, A.M. (2009). Integrative data analysis: the simultaneous analysis of multiple datasets. *Psychological Methods, 14*, 81-100.

Curran, P.J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R.A. (2008). Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology, 44*, 365-380.

Darnieder, W.F. (2011). Bayesian methods for data-dependent priors (unpublished doctoral dissertation). Ohio State University, Ohio.

de Finetti, B. (1972). *Probability, induction, and statistics*. New York, NY: John Wiley & Sons.

de Finetti, B. (1974). *Theory of probability (Vol. I and Vol. II).* New York, NY: John Wiley & Sons.

De Ayala, R.J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

DeMars, C.E. (2012). A comparison of limited-information and full-information methods in Mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling, 19*, 610-632.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.

Dubois, D., Liu, W., Ma, J., & Prade, H. (2016). The basic principles of uncertain information fusion. An organised review of merging rules in different representation frameworks. *Information Fusion, 32,* 12-39.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Bulletin, 70,* 193-242.

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*, 1-5.

El-Sayed, M., & Hamed, K. (2015). Study of similarity measures with linear discriminant analysis for face recognition. *Journal of Software Engineering and Applications, 8,* 478-488.

Enders, C. K. (2005). *Applied missing data analysis*. New York, NY: Guilford Press.

Esteban, J., Starr, A., Willetts, R., Hannah, P., and Bryanston-Cross, P. (2004). A review of data fusion models and architectures: towards engineering guidelines. *Neural Computing and Applications, 14,* 273-281

Feuer, M.J., Holland, P.W., Green, B.F., Bertenthal, M.W., & Hemphill, F.C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.

Fox, J.P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fritz, M.S. and MacKinnon, D.P. (2007). Required sample size to detect the mediated effect. *Psychological science, 18,* 233-239.

Fu, K.S. (1968). *Sequential methods in pattern recognition and machine learning*. New York, NY: Academic Press.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). Bayesian data analysis. CRC. *Chapman and Hall, Boca Raton, FL*.

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys (with discussion). *Journal of the American Statistical Association, 93,* 846–874.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

Ghosh, B.K. (1991). A brief history of sequential analysis. In B.K. Ghosh and P.K. Sen (Eds.), *Handbook of sequential analysis* (1-19). New York, NY: Dekker.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996). Introducing Markov Chain Monte Carlo. *Markov chain Monte Carlo in practice*, *1*, 19.

Gilula, Z., McCulloch, R.E., & Rossi, P.E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, *43*, 73-83.

Glass, G.V. (2000, March). The future of meta-analysis. Paper presented at the University of California, Berkeley–Stanford University Colloquium on Meta-Analysis, Department of Psychology, University of California, Berkeley.

Good, I.J. (1979). Studies in the history of probability and statistics: XXXVII. Turing's statistical work in World War II. *Biometrika, 66*, 393–396.

Grimm, K.J. (2012). Intercept centering and time coding in latent difference score models. *Structural Equation Modeling, 19,* 137-151.

Grimm, K.J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches.* New York, NY: Guilford Press.

Grimm, K.J., & Widaman, K.F. (2010). Residual structures in latent growth curve modeling, *Structural Equation Modeling, 17*, 424-442.

Hallquist, M., & Wiley, J. (2014). MplusAutomation: Automating Mplus model estimation and interpretation. R package version 0.06-3. Retrieved fromhttp://CRAN.R-project.org/package=MplusAutomation.

Hambleton, R.K., Swaminathan, H., & Rogers, J.R. (1991). *Fundamentals of item response theory.* Newbury Park, California: Sage Publications, Inc.

Hanson, B.A., & Beguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.

Harring, J.R., Weiss, B.A., & Hsu, J-C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods, 17,* 193-214.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97-109.

Hedges, L.V., & Pigott, T.D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*, 203-217.

Hill, B.M. (1970). Some contrasts between Bayesian and classical influence in the analysis of variance and testing of models. In D.L. Meyer & R.O. Collier, Jr. (Eds.). *Bayesian Statistics* (pp.29-36). Itasca, IL: F.E. Peacock Publishers.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14,* 382-417.

Hofer, S.M. & Piccinin, A.M. (2009). Integrative data Analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14,* 150-164.

Hussong, A.M., Curran, P.J., & Bauer, D.J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9,* 61-89.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*, 227-241.

Jeffreys, H. (1961). Theory of probability (3rd Ed.). London: Oxford University Press.

Jones, A.P., Riley, R.D., Williamson, P.R., & Whitehead, A. (2009). Meta-analysis of individual patient data vs aggregate data from longitudinal clinical trials. *Clinical Trials, 6,* 16-27.

Kamakura, W.A., & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, *34*, 485-498.

Kaplan, D., & McCarty, A.T. (2013). Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys. *Large-scale Assessments in Education*, *1*, 1-26.

Kass, R.E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928- 934.

Kass, R.E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343-1370.

Kim, S. & Cohen, A.S. (1995). A comparison of Lord's Chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*, 291-312.

Kim, S. & Kolen, M.J. (2006). Robustness of format effects of IRT linking methods for mixed format tests. *Applied Measurement in Education, 19,* 357–381.

Kolen, M.J. (2006). Scaling and norming. *Educational Measurement*, *4*, 156-186.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (pp. 389-400). New York: Springer.

Lai, T.L. (1995). Sequential changepoint detection in quality control and dynamical systems (with discussion). *Journal of the Royal Statistical Society Series B,* **57**, 613-658.

Lee, W.C. & Ban, J.C. (2010). A Comparison of IRT Linking Procedures. *Applied Measurement in Education, 23,* 23-48.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, 1-18.

Levy, R., & Choi, J. (2013). Bayesian Structural Equation Modeling. In *Structural Equation Modeling: A Second Course* (2nd ed., pp. 563-623). Information Age Publishing.

Lindley, D.V. (2004). That wretched prior. *Significance, 1,* 85-87.

Link, W.A., & Eaton, M.J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3,* 112-115.

Liu, X. (1992) Entropy, Distance Measure and Similarity Measure of Fuzzy Sets and Their Relations. *Fuzzy Sets and Systems*, *52,* 305-318.

Marcoulides, K.M. & Grimm, K. J. (2016). Data integration approaches to longitudinal growth modeling. *Educational and Psychological Measurement.* Advance online publication. DOI: 10.1177/0013164416664117.

McArdle, J.J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29*, 403-435.

McArdle, J.J. & Epstein, D.B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110-133.

McArdle, J.J, & Grimm, K.J. (2010). Five steps in latent curve and latent change score modeling with longitudinal data. In K. Montfort, J. H.L. Oud, A. Satorra (Eds.), *Longitudinal research with latent variables* (245-273). New York, NY: Springer.

McArdle, J.J., & Nesselroade, J.R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.

McArdle, J.J., Grimm, K.J., Hamagami, F., Bowles, R.P., & Meredith, W. (2009). Modeling lifespan growth curves of cognition using longitudinal data with changing measures. *Psychological Methods, 14*, 126-149.

McArdle, J.J. & Horn, J.L. (2002). The benefits and limitations of mega-analysis with illustrations for the WAIS. Paper presented at the *International meeting of CODATA*. Montreal, Quebec, Canada.

McArdle, J.J., & Horn, J.L. (2005). *A mega analysis of the WAIS: Adult intelligence across the life-span.* Mahwah, NJ: Erlbaum (under contract).

McArdle, J.J., & Ritschard, G. (2014). *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*. New York, NY: Routledge.

McNeish, D. (2016). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling, 23,* 750-773.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107-122.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemistry and Physics, 21,* 1087-1091.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association, 44*, 335-341.

Mislevy, R. (1988). Exploiting auxiliary information about items in the estimation of Rasch item parameters. *Applied Psychological Measurement, 12,* 281-296.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-78.

Moustaki, I., & Knott, M. (2005). *Computational aspects of the EM and Bayesian estimation in latent variable models.* In: A. van der Ark, M. Croon, K. Sijtsma (Eds.) *New developments in categorical data analysis for the social and behavioral Sciences* (p. 103-124). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B.O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17,* 313-335.

Muthén, L.K., & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599-620.

Muthén, L.K., & Muthén, B.O. (2012). *Mplus User's Guide* (7th Edition). Los Angeles, CA: Muthén & Muthén.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, D.C.: National Academy Press.

O'Rourke, H.P. (2016). Time metric in latent difference score models (unpublished doctoral dissertation). Arizona State University, Arizona.

Paxton, P., Curran, P.J., Bollen, K.A., Kirby, J., Chen, F. (2001) Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*, 287-312.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3,* 1243-1246.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). Item response theory versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8, 137-156.

Piccinin, A.M. & Hofer, S.M. (2008). Integrative analysis of longitudinal studies on aging: Collaborative research networks, meta-analysis, and optimizing future studies. In S. M. Hofer and D. F. Alwin (Eds.), *Handbook on cognitive aging: Interdisciplinary perspectives* (446-476). Thousand Oaks, CA: Sage Publications.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York, NY: Springer.

Rässler, S. (2003). A Non-Iterative Bayesian Approach to Statistical Matching. *Statistica Neerlandica, 57,* 58-74.

Raykov, T., Marcoulides, G.A., & Chang, C. (2016) Examining population heterogeneity in finite mixture settings using latent variable modeling, *Structural Equation Modeling, 23*, 726-730.

Resenda, M. & Sousa, J.P. (2003). (Eds.), *Metaheuristics: Computer decision-making*. Boston, MA: Kluwer Academic Publishers.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics, 2,* 91-102.

Rost, J., & von Davier, M. (1993). Measuring different traits in different populations with the same items. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.). *Psychometric Methodology: Proceedings of the 7th European Meeting of the Psychometric Society* (p. 446-450). Stuttgart/New York: Gustav Fischer Verlag.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*, 543-546.

Rubin, D.B., & Little, R.J. (2002). *Statistical analysis with missing data*. Hoboken, NJ: J Wiley & Sons.

Rupp, A.A., Dey, D.K., & Zumbo, B.D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11,* 424-451.

Sansone, C., Morf, C.C., & Panter, A.T. (Eds.). (2003). *The SAGE handbook of methods in social psychology*. Thousand Oaks, CA: Sage Publications, Inc.

SAS Institute Inc. 2014. SAS/STAT® 13.2 User's Guide. Cary, NC: SAS Institute Inc.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*, 881-901.

Smith, T.C., Spiegelhalter, D.J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine, 14*, 2685-2699.

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (1996). *BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii).* Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software,12,* 1-16.

Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics, 22,* 1701-1728.

U. S. Department of Defense (1991). Data fusion lexicon. Data Fusion Subpanel of the Joint Directors of Laboratories. Technical Report for C3. San Diego, CA.

Vermunt, J. K. (2003) Multilevel latent class models. *Sociological Methodology, 33,* 213-239.

von Davier, M., Holland, P., & Thayer (2004). *The Kernel method of test equating*. Springer, New York.

Wald, L. (1999). Some terms of reference in data fusion. *IEEE Transactions on Geosciences and Remote Sensing, 37*, 1190-1193.

Wilderjans, T. F., Bernal, E. F., Galindo-Villardón, P., & Ceulemans, E. (2015, July). Data fusion of heterogeneous datasets. *Paper presented at the International Meeting of the Psychometric Society.* Beijing, China.

Willett, J.B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49,* 587-602.

Yee, E., Hoffman, I., Branch, R. P., Ungar, K., Malo, A., Ek, N., & Bourgouin, P. (2014). Bayesian inference for source term estimation: Application to the International Monitoring System radionuclide network.

Zhang, Z., Hamagami, F., Wang, L.L., Nesselroade, J.R., & Grimm, K.J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*, 374-383.

Zill, N., Resnick, G., Kim, K., O'Donnell, K., Sorongon, A., & McKey, R.H. (2003). *Head Start FACES 2000: A whole-child perspective on program performance, fourth progress report.* Child Care Bureau, Washington, DC.

Table 1

List of the patterns of measurement occasions to be used in the simulated data

Dataset	Number of Assessments	Years between Assessments	Starting Age
1	5	2	4.5
2	4	0.5	4
3	2	0.5	4
4	10	0.5	2.5
5	3	3	7
6	3	5	6

Table 2

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$ *and N=50*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	0.0234	0.0062	-0.0016	**11.7420**	3.1240	-0.834	0.0392	0.0237	0.0219	0.0313	0.0228	0.0218
σ^2_{slope}	0.0004	0.0004	-0.0001	3.7200	3.680	-1.200	0.0121	0.0122	0.0122	0.0015	0.0012	0.0012
σ_{is}	-0.0037	-0.0014	0.0004	--	--	--	0.0059	0.0042	0.0038	0.0046	0.0039	0.0037
$\beta_{intercept}$	-0.0125	0.0004	0.0009	0.6264	-0.0178	-0.0448	0.0341	0.031	0.0309	0.0317	0.0309	0.0309
β_{slope}	-0.0017	0.0006	0.0005	-0.4270	0.1650	0.1290	0.0077	0.0075	0.0075	0.0075	0.0075	0.0075
σ^2_e	0.0001	-0.0001	0.0001	0.0640	-0.1280	0.0640	0.0063	0.0047	0.0063	0.0063	0.0048	0.0047

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. The relative bias for estimates of the intercept-slope covariance cannot be computed, as the population value was zero. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 3

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$ *and N=250*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	BF	ML	BS	BF	ML	BS	BF	ML	BS	BF	ML
$\sigma^2_{intercept}$	-0.0035	0.0003	-0.0005	-1.7920	0.1660	-0.2680	0.0135	0.0107	0.0105	0.0131	0.0107	0.0105
σ^2_{slope}	0.0003	0.0000	0.0000	3.3600	-0.1200	-0.0400	0.0010	0.0006	0.0006	0.0009	0.0006	0.0006
σ_{is}	-0.0009	0.0000	0.0000	--	--	--	0.0028	0.0018	0.0018	0.0089	0.0018	0.0018
$\beta_{intercept}$	-0.002	0.0057	0.0007	0.1042	-0.2862	-0.0388	0.0138	0.0145	0.0134	0.0137	0.0134	0.0134
β_{slope}	0.0004	-0.0007	0.0008	0.0100	-0.1970	0.2200	0.0033	0.0032	0.0032	0.0033	0.0031	0.0032
σ^2_e	-0.0003	-0.0002	0.0000	-0.3880	-0.2520	0.0200	0.0026	0.0021	0.0021	0.0026	0.0021	0.0021

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. The relative bias for estimates of the intercept-slope covariance cannot be computed, as the population value was zero. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 4

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$ *and N=1000*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	0.0002	0.0008	0.0003	0.1360	0.4060	0.1320	0.0060	0.0050	0.0048	0.0059	0.0049	0.0049
σ^2_{slope}	0.0000	0.0001	0.0000	-0.6400	0.4800	0.200	0.0005	0.0002	0.0003	0.0005	0.0002	0.0003
σ_{is}	0.0000	-0.0002	0.0000	--	--	--	0.0013	0.0009	0.0009	0.0013	0.0009	0.0009
$\beta_{intercept}$	-0.0054	-0.0022	-0.0002	0.2722	0.1136	0.0142	0.0088	0.0071	0.0068	0.0070	0.0068	0.0068
β_{slope}	-0.0005	-0.0002	-0.0002	-0.1390	-0.068	-0.0390	0.0018	0.0016	0.0016	0.0017	0.0016	0.0016
σ^2_e	0.0001	0.0002	0.0000	0.2720	0.2440	0.1040	0.0015	0.0011	0.0011	0.0015	0.0011	0.0011

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. The relative bias for estimates of the intercept-slope covariance cannot be computed, as the population value was zero. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 5

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=50*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	0.0380	0.0125	-0.0034	***5.4340***	1.7834	-0.4908	0.0809	0.068	0.0654	0.0713	0.0668	0.0653
σ^2_{slope}	0.0003	0.0017	-0.0009	0.3200	1.7280	-0.9800	0.0093	0.0094	0.0089	0.0093	0.0093	0.0089
σ_{is}	-0.0224	-0.0048	-0.0013	**-44.904**	***-9.7280***	-2.7120	0.028	0.0180	0.0167	0.0167	0.0174	0.0167
$\beta_{intercept}$	-0.0213	0.0003	0.0084	1.0670	-0.0178	-0.4216	0.0554	0.0520	0.0513	0.0511	0.0506	0.0506
β_{slope}	-0.0075	0.0006	-0.0003	-1.8830	0.1650	-0.0720	0.0222	0.0209	0.0207	0.0209	0.0209	0.0207
σ^2_e	0.0005	-0.0001	0.00002	0.4920	-0.1280	0.1800	0.0063	0.0047	0.005	0.0063	0.005	0.005

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 6

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=250*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	-0.0125	-0.0004	-0.0022	-1.7840	-0.0691	-0.3211	0.0294	0.0259	0.0257	0.0266	0.0259	0.0257
σ^2_{slope}	0.0014	0.0004	0.0002	1.4040	0.4360	-0.1760	0.0045	0.0041	0.0041	0.0043	0.0041	0.0041
σ_{is}	0.0009	0.0009	0.0008	1.9360	1.9440	1.6960	0.0088	0.0080	0.0079	0.0088	0.0079	0.0079
$\beta_{intercept}$	-0.0067	0.0057	-0.0012	0.3346	-0.2862	0.0616	0.0244	0.00238	0.0236	0.0235	0.0236	0.0236
β_{slope}	0.0011	-0.0008	0.0005	0.2730	-0.1970	0.1270	0.0093	0.0092	0.0091	0.0093	0.0091	0.0091
σ^2_e	-0.0003	-0.0002	0.0000	-0.4920	-0.2320	0.0480	0.0029	0.0021	0.0022	0.0029	0.0023	0.0023

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 7

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=1000*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	0.0012	0.0006	0.0002	0.1720	0.8970	0.0302	0.0148	0.0146	0.0145	0.0148	0.0147	0.0145
σ^2_{slope}	-0.0018	0.0004	0.0000	-1.8320	0.3680	0.0480	0.0029	0.0022	0.0021	0.0022	0.0022	0.0021
σ_{is}	-0.0010	-0.0006	0.0000	-2.1040	1.3600	-0.1360	0.0043	0.0042	0.0041	0.0043	0.0042	0.0042
$\beta_{intercept}$	-0.0063	-0.0022	-0.0007	0.3134	0.1136	0.0388	0.0132	0.0072	0.0111	0.0116	0.0111	0.0129
β_{slope}	-0.0004	-0.0003	0.0000	-0.1040	-0.0680	-0.0020	0.0044	0.0045	0.0044	0.0044	0.0044	0.0044
σ^2_e	0.00002	0.0002	0.0000	0.1640	0.2440	-0.0040	0.0014	0.0011	0.0012	0.0014	0.0119	0.0012

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 8

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_s = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=50*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{Intercept}$	0.0172	0.0049	-0.0007	4.3130	1.2220	-0.1800	0.0470	0.0405	0.0396	0.0437	0.0403	0.0396
σ^2_{slope}	0.0006	0.0096	0.0006	0.1710	2.400	0.1490	0.0379	0.0394	0.0374	0.0379	0.0382	0.0374
σ_{is}	-0.0184	-0.0010	0.0011	***-9.2080***	-0.5320	0.5740	0.0384	0.0302	0.0300	0.0337	0.0303	0.0299
$\beta_{Intercept}$	-0.0288	-0.0006	-0.0022	1.4428	0.0332	0.1106	0.0516	0.0422	0.0421	0.0428	0.0422	0.0421
β_{slope}	-0.0091	-0.0006	0.0022	-2.2780	-0.1610	0.5610	0.0384	0.0360	0.0362	0.0374	0.0359	0.0361
σ^2_e	-0.0001	0.0000	-0.0000	-0.5600	0.0920	-0.3440	0.0067	0.0055	0.0055	0.0067	0.0055	0.0055

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{Intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 9

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_s = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=250*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**	**BS**	**BF**	**ML**
$\sigma^2_{intercept}$	-0.0058	0.0014	0.0009	-1.4540	0.3710	0.2440	0.0185	0.0165	0.0165	0.0175	0.0164	0.0165
σ^2_{slope}	0.0040	0.0005	-0.0009	1.0170	0.1420	-0.2450	0.0158	0.0149	0.0148	0.0153	0.0149	0.0148
σ_{is}	0.0053	0.0007	0.0002	2.6760	0.3540	0.1260	0.0141	0.0124	0.0122	0.0131	0.0124	0.0123
$\beta_{intercept}$	-0.0035	0.0036	-0.0001	0.1758	-0.1822	0.0052	0.0179	0.0174	0.0170	0.0176	0.0170	0.0170
β_{slope}	0.0017	-0.0022	0.0014	0.4360	-0.5620	0.3540	0.0169	0.0159	0.0157	0.0168	0.0158	0.0157
σ^2_e	-0.0002	-0.0005	-0.0000	-0.1920	-0.5480	-0.1640	0.0029	0.0024	0.0023	0.0029	0.0230	0.0023

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 10

Linear Growth Model Parameter Estimates for Variance-Covariance Matrix $\Psi_s = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$ *and N=1000*

Parameter	Raw Bias			Relative Bias			Accuracy			Efficiency		
	BS	BF	ML	BS	BF	ML	BS	BF	ML	BS	BF	ML
$\sigma^2_{intercept}$	-0.0024	0.0008	0.0008	0.5980	0.2210	0.2010	0.0090	0.0085	0.0084	0.0087	0.0085	0.0084
σ^2_{slope}	-0.0075	0.0014	0.0003	-1.8850	0.3690	0.0650	0.0105	0.0075	0.0074	0.0074	0.0074	0.0074
σ_{is}	-0.0034	-0.0005	-0.0000	-1.7400	-0.2800	-0.0240	0.0071	0.0061	0.006	0.0062	0.0061	0.0061
$\beta_{intercept}$	-0.0042	-0.0033	-0.0012	0.2118	0.1676	0.0612	0.0104	0.0097	0.0092	0.0095	0.0092	0.0092
β_{slope}	-0.0022	-0.0013	-0.0004	-0.5640	-0.3320	-0.1210	0.0098	0.0095	0.0094	0.0096	0.0094	0.0094
σ^2_e	0.0000	0.0002	0.0000	0.2280	0.2160	0.0280	0.0014	0.0012	0.0010	0.0014	0.0011	0.0011

Note: BS = Bayesian Synthesis, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion. For $\beta_{intercept}$, negative bias corresponds to overestimation and positive bias corresponds to underestimation. For all other parameters, negative bias corresponds to underestimation and positive bias corresponds to overestimation.

Table 11

Linear Growth Model Parameter Estimates for Real Data

Parameter	Estimate			
	BS	**BS$_R$**	**BF**	**ML**
$\sigma^2_{intercept}$	0.138	0.097	0.122	0.117
σ^2_{slope}	0.006	0.008	0.007	0.004
σ_{is}	-0.023	-0.009	-0.018	-0.014
$\beta_{intercept}$	-1.558	-1.557	-1.563	-1.560
β_{slope}	0.478	0.479	0.482	0.478
σ^2_e	0.086	0.076	0.078	0.081

Note: BS = Bayesian Synthesis, BS$_R$ = Bayesian Synthesis with reversed order of data integration, BF = Bayesian Fusion, ML = Maximum Likelihood Fusion.

Dataset A

Person	Gender	Education	Age	Achievement Score
1	M	High School	17	Score Not Collected
2	F	High School	17	Score Not Collected

Dataset B

Person	Gender	Education	Age	Achievement Score
3	M	High School	17	25
4	F	College	22	50
5	M	College	18	20
6	F	High School	17	20

Figure 1. A simple illustration of the nearest neighbor approach.

	Person	Common Variables Z	Unique Variables X	Unique Variables Y
A	1_A : : : n_A	Z_1 Z_2 .... Z_{P_A}	Missing Data	Y_1 Y_2 .... Y_{P_A}
B	1_B : : : n_B	Z_1 Z_2 ....Z_{P_B}	X_1 X_2 .... X_{P_B}	Missing Data

Figure 2. Data combination for data fusion of datasets A and B for *n* people on variables Z, X, and Y.

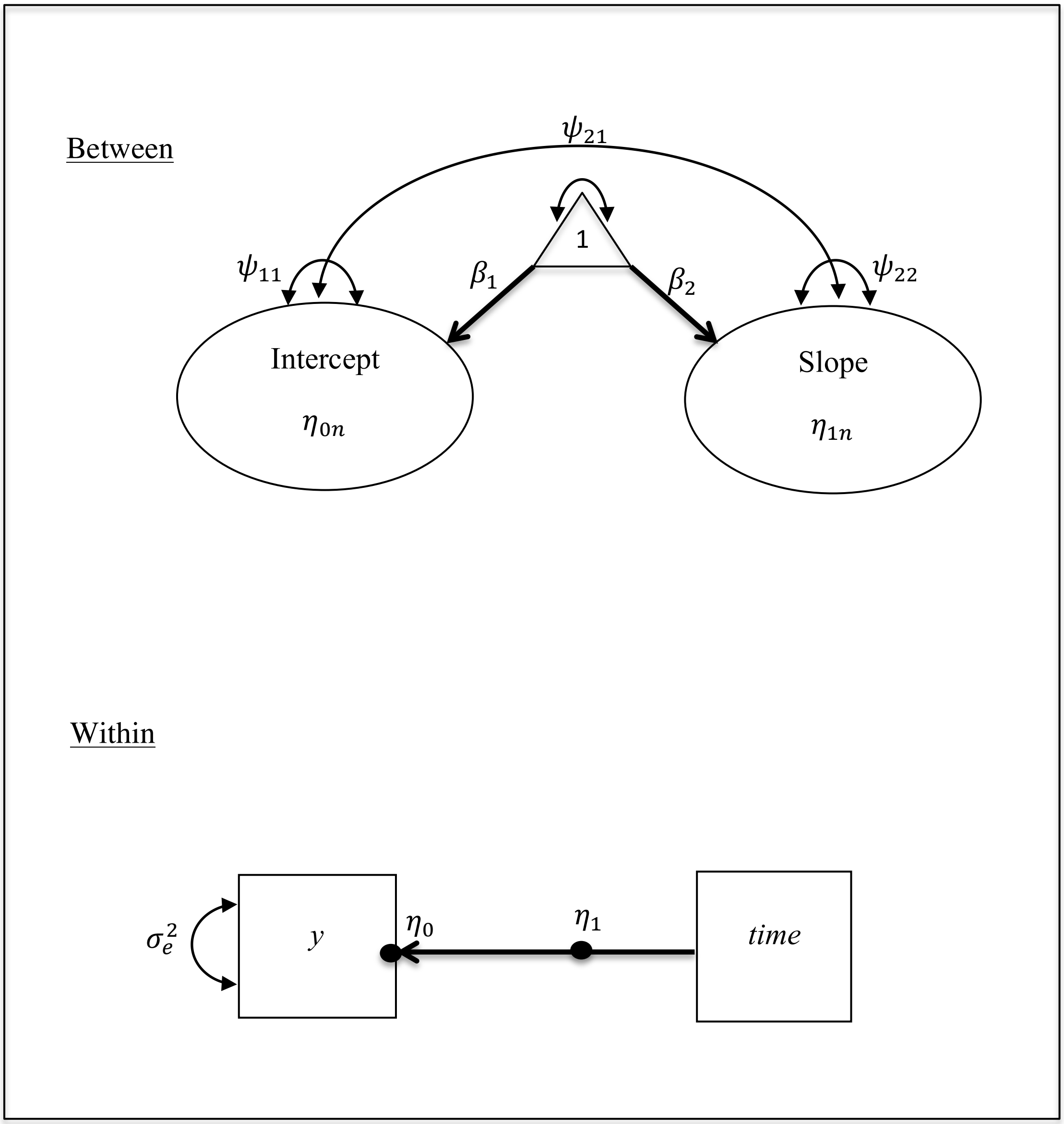
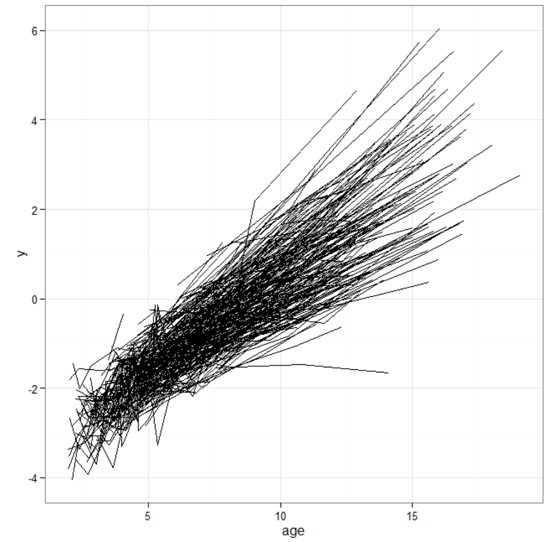


Figure 3. Example path diagram of a growth model specified in the multilevel modeling framework.

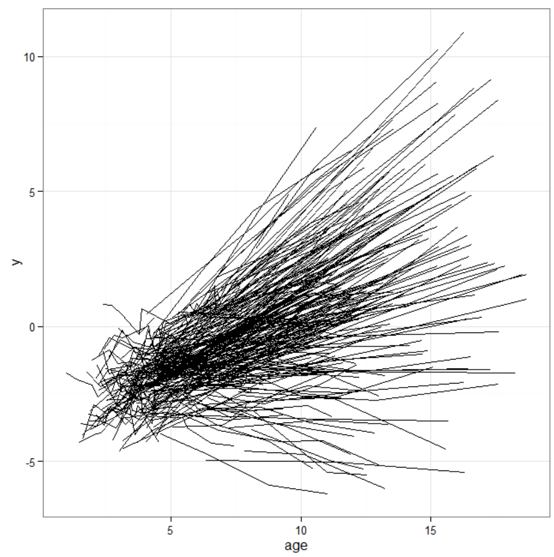
APPENDIX A

SIMULATED DATA THETA SCORE PLOTS FOR N=50

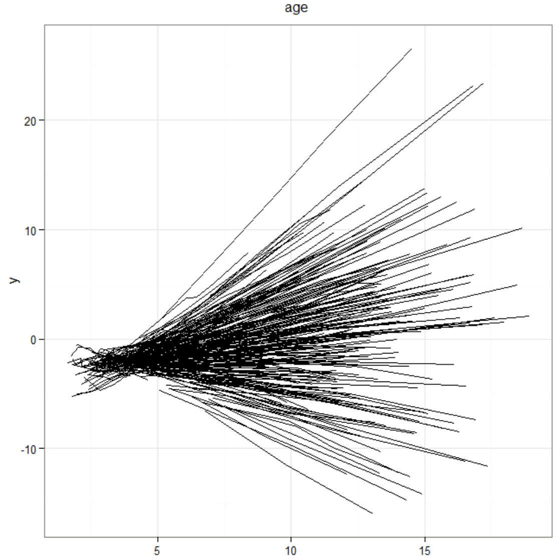
$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.20 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$$



$$\boldsymbol{\Psi}_2 = \begin{bmatrix} 0.70 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$$

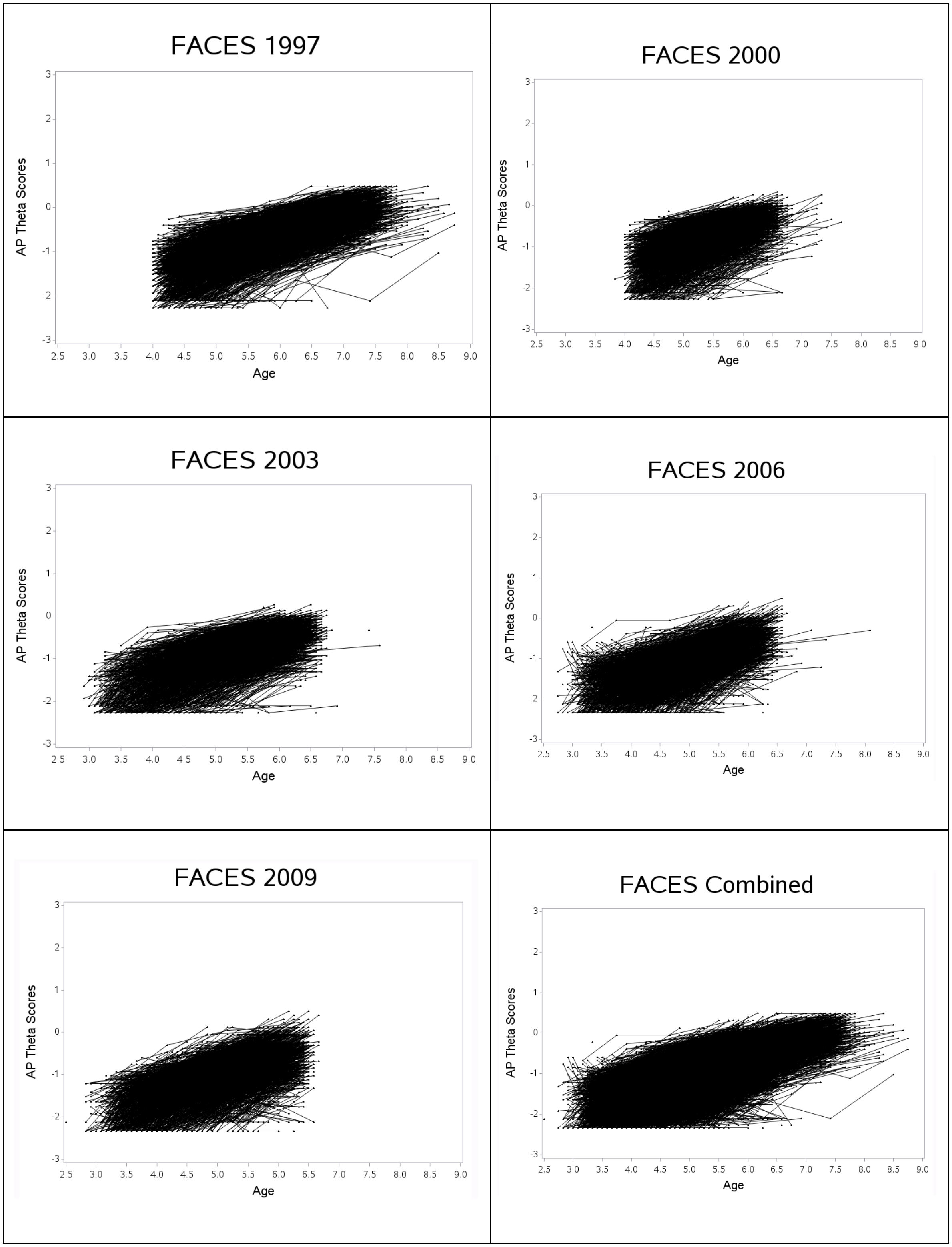


$$\boldsymbol{\Psi}_3 = \begin{bmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{bmatrix}$$



APPENDIX B

REAL DATA AP THETA SCORE PLOTS


FACES 1997
AP Theta Scores
Age
FACES 2000
AP Theta Scores
Age
FACES 2003
AP Theta Scores
Age
FACES 2006
AP Theta Scores
Age
FACES 2009
AP Theta Scores
Age
FACES Combined
AP Theta Scores
Age

APPENDIX C

EXAMPLE R CODE FOR Ψ_1 N=1000

```
#Code for Psi 1 N = 1000#
#Create empty matrices to be filled with the results#

bs_residual_est=matrix(NA,1,250)
bs_int_slp_cov_est=matrix(NA,1,250)
bs_intercept_mean_est=matrix(NA,1,250)
bs_slope_mean_est=matrix(NA,1,250)
bs_int_var_est=matrix(NA,1,250)
bs_slp_var_est=matrix(NA,1,250)

bs_residual_sd=matrix(NA,1,250)
bs_int_slp_cov_sd=matrix(NA,1,250)
bs_intercept_mean_sd=matrix(NA,1,250)
bs_slope_mean_sd=matrix(NA,1,250)
bs_int_var_sd=matrix(NA,1,250)
bs_slp_var_sd=matrix(NA,1,250)


bf_residual_est=matrix(NA,1,250)
bf_int_slp_cov_est=matrix(NA,1,250)
bf_intercept_mean_est=matrix(NA,1,250)
bf_slope_mean_est=matrix(NA,1,250)
bf_int_var_est=matrix(NA,1,250)
bf_slp_var_est=matrix(NA,1,250)

bf_residual_sd=matrix(NA,1,250)
bf_int_slp_cov_sd=matrix(NA,1,250)
bf_intercept_mean_sd=matrix(NA,1,250)
bf_slope_mean_sd=matrix(NA,1,250)
bf_int_var_sd=matrix(NA,1,250)
bf_slp_var_sd=matrix(NA,1,250)


ml_residual_est=matrix(NA,1,250)
ml_int_slp_cov_est=matrix(NA,1,250)
ml_intercept_mean_est=matrix(NA,1,250)
ml_slope_mean_est=matrix(NA,1,250)
ml_int_var_est=matrix(NA,1,250)
ml_slp_var_est=matrix(NA,1,250)

ml_residual_sd=matrix(NA,1,250)
ml_int_slp_cov_sd=matrix(NA,1,250)
ml_intercept_mean_sd=matrix(NA,1,250)
ml_slope_mean_sd=matrix(NA,1,250)
ml_int_var_sd=matrix(NA,1,250)
ml_slp_var_sd=matrix(NA,1,250)
```

```
for(h in 1:250){

## Simulate First Dataset ##
b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
N=1*N_1
T_1 = 5
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))


      for(i in 1:N_1){
        id[i]=i
        b_0i[i]=rnorm(1,0,1)
        b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
        b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
        b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

        age_1i =rnorm(1,4.5,sqrt(.2))

     for(t in 1:T_1){
      age_ti[i,t] = age_1i + (t-1) * rnorm(1,2,sqrt(.02))
      y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
     }
      }

data.1=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))
names(data.1)=c('id','age1','age2','age3','age4','age5',
              'y1','y2','y3','y4','y5',
              'b_0i','b_1i')

## Transform Data to Long ##
data.1.long = reshape(data.1, idvar='id',
     varying=c('age1','y1','age2','y2','age3','y3','age4','y4','age5','y5'),
     times=c(1,2,3,4,5),
     v.names=c('age','y'), direction='long')

data.1.long = data.1.long[order(data.1.long$id, data.1.long$age),]
```

```
## Plot ##
require(ggplot2)

plot_obs <- ggplot(data=data.1.long, aes(x=age, y=y, group=id)) +
              geom_line() +
               theme_bw()

print(plot_obs)

## Center Age ##
data.1.long$agec = data.1.long$age - 4

## Write Mplus Script ##

library(MplusAutomation)

###### Goals ######
# 1. Write Mplus Code for Data #1
# 2. Run model #1
# 3. Take parameter estimates from #2 and create syntax with
#    those estimates as prior values
# 4. Run model #3
# 5. Repeat with newly created dataset

# Replace function is needed to fill in parameters in Mplus script with specified values
loopReplace <- function(text, replacements) {
        for (v in names(replacements)){
          text <- gsub(sprintf("\\[\\[%s\\]\\]", v), replacements[[v]], text)
        }
        return(text)
}

# Step #1
data.1.script <- mplusObject(
TITLE = "Growth Model for Data #1;",
VARIABLE = "USEVARIABLES = y agec;
          CLUSTER=id;
          WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
      ESTIMATOR = BAYES;",
MODEL = "
%WITHIN%
    b1 | y ON agec;
    y (sigma2_u);
```

```
%BETWEEN%
   [y]  (beta_0);
   [b1] (beta_1);

    y   (sigma2_0);
    b1  (sigma2_1);
    y WITH b1 (sigma_10);
MODEL PRIORS:
sigma2_u ~ IG(-1, 0);
beta_0 ~ N(0, 10^10);
beta_1 ~ N(0, 10^10);
sigma2_0 ~ IW(0,-3);
sigma2_1 ~ IW(0,-3);
sigma_10 ~ IW(0,-3);
",
usevariables = c("y", "agec", "id"),
rdata = data.1.long)

data.1.result = mplusModeler(data.1.script, "data.1.long.dat", modelout = "data.1.inp",
run = 1L)

parms = data.1.result$results$parameters

parms

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est)
names(df_parms.1) <- c("param","est")
df_parms2.1 <- t(df_parms.1)
df_parms2.1 <- as.data.frame(df_parms2.1)

df2.1 <- df_parms2.1[2,]
df2.1

df_parms.2 <- data.frame(parms[[1]]$param, parms[[1]]$posterior_sd)

names(df_parms.2) <- c("param","sd")

df_parms.2[7,2] = (N - 2)/2  #PriorIG_1 sample size/2
df_parms.2[7,2]

df_parms.2[8,2] = ((N - 2)*(df_parms.1[1,2]))/2 #PriorIG_2 sample size*variance/2
df_parms.2[8,2]

df_parms.2[9,2] = (df_parms.2[3,2])^2
df_parms.2[9,2]
```

```
df_parms.2[10,2] = (df_parms.2[4,2])^2
df_parms.2[10,2]

df_parms.2[11,2] = df_parms.1[5,2]*(N - 3)
df_parms.2[11,2]

df_parms.2[12,2] = df_parms.1[6,2]*(N - 3)
df_parms.2[12,2]

df_parms.2[13,2] = df_parms.1[2,2]*(N - 3)
df_parms.2[13,2]

df_parms.2[14,2] = (N - 3)
df_parms.2[14,2]

df_parms2.2 <- t(df_parms.2)
df_parms2.2 <- as.data.frame(df_parms2.2)

df2.2 <- df_parms2.2[2,]

df.f = cbind(df2.1,df2.2)
df.f

names(df.f) <-
c("SIGMA2_U","SIGMA_10","BETA_0","BETA_1","SIGMA2_0","SIGMA2_1",
"SIGMA2_U_SD","SIGMA_10_SD","BETA_0_SD","BETA_1_SD","SIGMA2_0_SD",
                "SIGMA2_1_SD","PriorIG_1","PriorIG_2","PriorN_2_1","PriorN_2_2",
                "Prior_IW_1","Prior_IW_2","Prior_IW_3", "Prior_IW_N")


####################################
#
#    Generate Second Data Set
#
####################################
b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
N=2*N_1
T_1 = 4
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
```

```
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))

        for(i in 1:N_1){
          id[i]=i + 1000
          b_0i[i]=rnorm(1,0,1)
          b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
          b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
          b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

          age_1i =rnorm(1,4.5,sqrt(.2))

       for(t in 1:T_1){
        age_ti[i,t] = age_1i + (t-1) * rnorm(1,2,sqrt(.02))
          y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
       }
        }

data.2=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))
names(data.2)=c('id','age1','age2','age3','age4',
               'y1','y2','y3','y4',
               'b_0i','b_1i')

## Transform Data to Long ##

data.2.long = reshape(data.2, idvar='id',
       varying=c('age1','y1','age2','y2','age3','y3','age4','y4'),
       times=c(1,2,3,4),
       v.names=c('age','y'), direction='long')

data.2.long = data.2.long[order(data.2.long$id, data.2.long$age),]

## Plot ##
require(ggplot2)

plot_obs <- ggplot(data=data.2.long, aes(x=age, y=y, group=id)) +
             geom_line() +
             theme_bw()

print(plot_obs)

## Center Age ##

data.2.long$agec = data.2.long$age - 4

## Write Mplus Script ##
```

```
library(MplusAutomation)

# Replace function is needed to fill in parameters in Mplus script with specified values

loopReplace <- function(text, replacements) {
        for (v in names(replacements)){
           text <- gsub(sprintf("\\[\\[%s\\]\\]", v), replacements[[v]], text)
        }
        return(text)
}

####### Create Syntax for Second Data Set #######

data.2.script <- mplusObject(
TITLE = "Growth Model for Data #2;",
VARIABLE = "USEVARIABLES = y agec;
          CLUSTER=id;
          WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
      ESTIMATOR = BAYES;",
MODEL = loopReplace("
%WITHIN%
    b1 | y ON agec;
    y (sigma2_u);

%BETWEEN%
   [y]  (beta_0);
   [b1] (beta_1);

    y   (sigma2_0);
    b1  (sigma2_1);
    y WITH b1 (sigma_10);

MODEL PRIORS:
sigma2_u ~ IG([[PriorIG_1]], [[PriorIG_2]]);
beta_0 ~ N([[BETA_0]], [[PriorN_2_1]]);
beta_1 ~ N([[BETA_1]], [[PriorN_2_2]]);
sigma2_0 ~ IW([[Prior_IW_1]], [[Prior_IW_N]]); !mean_estimate*(sample_size - p -1),
sample_size)
sigma2_1 ~ IW([[Prior_IW_2]], [[Prior_IW_N]]);
sigma_10 ~ IW([[Prior_IW_3]], [[Prior_IW_N]]);", df.f)
,
usevariables = c("y", "agec", "id"),
rdata = data.2.long)
```

```
data.2.result = mplusModeler(data.2.script, "data.2.long.dat", modelout = "data.2.inp",
run = 1L)

parms = data.2.result$results$parameters

parms

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est)
names(df_parms.1) <- c("param","est")
df_parms2.1 <- t(df_parms.1)
df_parms2.1 <- as.data.frame(df_parms2.1)

df2.1 <- df_parms2.1[2,]
df2.1

df_parms.2 <- data.frame(parms[[1]]$param, parms[[1]]$posterior_sd)
names(df_parms.2) <- c("param","sd")

df_parms.2[7,2] = (N - 2)/2  #PriorIG_1 sample size/2
df_parms.2[7,2]

df_parms.2[8,2] = ((N - 2)*(df_parms.1[1,2]))/2 #PriorIG_2 sample size*variance/2
df_parms.2[8,2]

df_parms.2[9,2] = (df_parms.2[3,2])^2
df_parms.2[9,2]

df_parms.2[10,2] = (df_parms.2[4,2])^2
df_parms.2[10,2]

df_parms.2[11,2] = df_parms.1[5,2]*(N - 3)
df_parms.2[11,2]

df_parms.2[12,2] = df_parms.1[6,2]*(N - 3)
df_parms.2[12,2]

df_parms.2[13,2] = df_parms.1[2,2]*(N - 3)
df_parms.2[13,2]

df_parms.2[14,2] = (N - 3)
df_parms.2[14,2]

df_parms2.2 <- t(df_parms.2)
df_parms2.2 <- as.data.frame(df_parms2.2)

df2.2 <- df_parms2.2[2,]
```

```
df.f = cbind(df2.1,df2.2)
df.f

names(df.f) <-
c("SIGMA2_U","SIGMA_10","BETA_0","BETA_1","SIGMA2_0","SIGMA2_1",
"SIGMA2_U_SD","SIGMA_10_SD","BETA_0_SD","BETA_1_SD","SIGMA2_0_SD",
                "SIGMA2_1_SD","PriorIG_1","PriorIG_2","PriorN_2_1","PriorN_2_2",
                "Prior_IW_1","Prior_IW_2","Prior_IW_3", "Prior_IW_N")

####################################
#
#   Generate Third Data Set
#
####################################

b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
N=3*N_1
T_1 = 2
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))

     for(i in 1:N_1){
       id[i]=i + 2000
       b_0i[i]=rnorm(1,0,1)
       b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
       b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
       b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

       age_1i =rnorm(1,4,sqrt(.2))

     for(t in 1:T_1){
      age_ti[i,t] = age_1i + (t-1) * rnorm(1,.5,sqrt(.02))
       y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
     }
     }

data.3=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))
```

```
names(data.3)=c('id','age1','age2','y1','y2',
                'b_0i','b_1i')

## Transform Data to Long ##
data.3.long = reshape(data.3, idvar='id',
      varying=c('age1','y1','age2','y2'),
      times=c(1,2),
      v.names=c('age','y'), direction='long')

data.3.long = data.3.long[order(data.3.long$id, data.3.long$age),]


## Plot ##
plot_obs <- ggplot(data=data.3.long, aes(x=age, y=y, group=id)) +
            geom_line() +
            theme_bw()

print(plot_obs)

# Center Age ##

data.3.long$agec = data.3.long$age - 4

####### Create Syntax for Third Data Set #######
data.3.script <- mplusObject(
TITLE = "Growth Model for Data #3;",
VARIABLE = "USEVARIABLES = y agec;
        CLUSTER=id;
        WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
     ESTIMATOR = BAYES;",
MODEL = loopReplace("
%WITHIN%
    b1 | y ON agec;
    y (sigma2_u);

%BETWEEN%
   [y]  (beta_0);
   [b1] (beta_1);

    y   (sigma2_0);
    b1  (sigma2_1);
    y WITH b1 (sigma_10);
MODEL PRIORS:
sigma2_u ~ IG([[PriorIG_1]], [[PriorIG_2]]);
```

```
beta_0 ~ N([[BETA_0]], [[PriorN_2_1]]);
beta_1 ~ N([[BETA_1]], [[PriorN_2_2]]);
sigma2_0 ~ IW([[Prior_IW_1]], [[Prior_IW_N]]); !mean_estimate*(sample_size - p -1),
sample_size)
sigma2_1 ~ IW([[Prior_IW_2]], [[Prior_IW_N]]);
sigma_10 ~ IW([[Prior_IW_3]], [[Prior_IW_N]]);", df.f)
,
usevariables = c("y", "agec", "id"),
rdata = data.3.long)

data.3.result = mplusModeler(data.3.script, "data.3.long.dat", modelout = "data.3.inp",
run = 1L)

parms = data.3.result$results$parameters

parms

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est)
names(df_parms.1) <- c("param","est")
df_parms2.1 <- t(df_parms.1)
df_parms2.1 <- as.data.frame(df_parms2.1)

df2.1 <- df_parms2.1[2,]
df2.1

df_parms.2 <- data.frame(parms[[1]]$param, parms[[1]]$posterior_sd)

names(df_parms.2) <- c("param","sd")

df_parms.2[7,2] = (N - 2)/2  #PriorIG_1 sample size/2
df_parms.2[7,2]

df_parms.2[8,2] = ((N - 2)*(df_parms.1[1,2]))/2 #PriorIG_2 sample size*variance/2
df_parms.2[8,2]

df_parms.2[9,2] = (df_parms.2[3,2])^2
df_parms.2[9,2]

df_parms.2[10,2] = (df_parms.2[4,2])^2
df_parms.2[10,2]

df_parms.2[11,2] = df_parms.1[5,2]*(N - 3)
df_parms.2[11,2]

df_parms.2[12,2] = df_parms.1[6,2]*(N - 3)
df_parms.2[12,2]
```

```
df_parms.2[13,2] = df_parms.1[2,2]*(N - 3)
df_parms.2[13,2]

df_parms.2[14,2] = (N - 3)
df_parms.2[14,2]

df_parms2.2 <- t(df_parms.2)
df_parms2.2 <- as.data.frame(df_parms2.2)

df2.2 <- df_parms2.2[2,]

df.f = cbind(df2.1,df2.2)
df.f

names(df.f) <-
c("SIGMA2_U","SIGMA_10","BETA_0","BETA_1","SIGMA2_0","SIGMA2_1",
"SIGMA2_U_SD","SIGMA_10_SD","BETA_0_SD","BETA_1_SD","SIGMA2_0_SD",
                "SIGMA2_1_SD","PriorIG_1","PriorIG_2","PriorN_2_1","PriorN_2_2",
                "Prior_IW_1","Prior_IW_2","Prior_IW_3", "Prior_IW_N")


#####################################
#
#   Generate Fourth Data Set
#
#####################################

b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
N=4*N_1
T_1 = 10
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))

      for(i in 1:N_1){
        id[i]=i + 3000
        b_0i[i]=rnorm(1,0,1)
        b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
```

```
        b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
        b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

        age_1i =rnorm(1,2.5,sqrt(.2))

      for(t in 1:T_1){
       age_ti[i,t] = age_1i + (t-1) * rnorm(1,.5,sqrt(.02))
         y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
      }
       }

data.4=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))

names(data.4)=c('id','age1','age2','age3','age4','age5','age6','age7','age8','age9','age10',
                'y1','y2','y3','y4','y5','y6','y7','y8','y9','y10',
                'b_0i','b_1i')

## Transform Data to Long ##
data.4.long = reshape(data.4, idvar='id',
      varying=c('age1','y1','age2','y2','age3','y3','age4','y4','age5','y5',
                'age6','y6','age7','y7','age8','y8','age9','y9','age10','y10'),
      times=c(1,2,3,4,5,6,7,8,9,10),
      v.names=c('age','y'), direction='long')

data.4.long = data.4.long[order(data.4.long$id, data.4.long$age),]

## Plot ##
plot_obs <- ggplot(data=data.4.long, aes(x=age, y=y, group=id)) +
             geom_line() +
             theme_bw()

print(plot_obs)

## Center Age ##

data.4.long$agec = data.4.long$age - 4

####### Create Syntax for Fourth Data Set #######
data.4.script <- mplusObject(
TITLE = "Growth Model for Data #4;",
VARIABLE = "USEVARIABLES = y agec;
        CLUSTER=id;
        WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
     ESTIMATOR = BAYES;",
MODEL = loopReplace("
```

```
%WITHIN%
    b1 | y ON agec;
    y (sigma2_u);

%BETWEEN%
   [y]  (beta_0);
   [b1] (beta_1);

    y   (sigma2_0);
    b1  (sigma2_1);
    y WITH b1 (sigma_10);

MODEL PRIORS:
sigma2_u ~ IG([[PriorIG_1]], [[PriorIG_2]]);
beta_0 ~ N([[BETA_0]], [[PriorN_2_1]]);
beta_1 ~ N([[BETA_1]], [[PriorN_2_2]]);
sigma2_0 ~ IW([[Prior_IW_1]],[[Prior_IW_N]]); !mean_estimate*(sample_size - p -1),
sample_size)
sigma2_1 ~ IW([[Prior_IW_2]], [[Prior_IW_N]]);
sigma_10 ~ IW([[Prior_IW_3]], [[Prior_IW_N]]);", df.f)
,
usevariables = c("y", "agec", "id"),
rdata = data.4.long)

data.4.result = mplusModeler(data.4.script, "data.4.long.dat", modelout = "data.4.inp",
run = 1L)

parms = data.4.result$results$parameters

parms

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est)
names(df_parms.1) <- c("param","est")
df_parms2.1 <- t(df_parms.1)
df_parms2.1 <- as.data.frame(df_parms2.1)

df2.1 <- df_parms2.1[2,]
df2.1

df_parms.2 <- data.frame(parms[[1]]$param, parms[[1]]$posterior_sd)
names(df_parms.2) <- c("param","sd")

df_parms.2[7,2] = (N - 2)/2  #PriorIG_1 sample size/2
df_parms.2[7,2]

df_parms.2[8,2] = ((N - 2)*(df_parms.1[1,2]))/2 #PriorIG_2 sample size*variance/2
```

```
df_parms.2[8,2]

df_parms.2[9,2] = (df_parms.2[3,2])^2
df_parms.2[9,2]

df_parms.2[10,2] = (df_parms.2[4,2])^2
df_parms.2[10,2]

df_parms.2[11,2] = df_parms.1[5,2]*(N - 3)
df_parms.2[11,2]

df_parms.2[12,2] = df_parms.1[6,2]*(N - 3)
df_parms.2[12,2]

df_parms.2[13,2] = df_parms.1[2,2]*(N - 3)
df_parms.2[13,2]

df_parms.2[14,2] = (N - 3)
df_parms.2[14,2]

df_parms2.2 <- t(df_parms.2)
df_parms2.2 <- as.data.frame(df_parms2.2)

df2.2 <- df_parms2.2[2,]

df.f = cbind(df2.1,df2.2)
df.f

names(df.f) <-
c("SIGMA2_U","SIGMA_10","BETA_0","BETA_1","SIGMA2_0","SIGMA2_1",
"SIGMA2_U_SD","SIGMA_10_SD","BETA_0_SD","BETA_1_SD","SIGMA2_0_SD",
                 "SIGMA2_1_SD","PriorIG_1","PriorIG_2","PriorN_2_1","PriorN_2_2",
                 "Prior_IW_1","Prior_IW_2","Prior_IW_3", "Prior_IW_N")


######################################
#
#    Generate Fifth Data Set
#
######################################

b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
```

```
N=5*N_1
T_1 = 3
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))


   for(i in 1:N_1){
     id[i]=i + 4000
     b_0i[i]=rnorm(1,0,1)
     b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
     b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
     b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

     age_1i =rnorm(1,7,sqrt(.5))

  for(t in 1:T_1){
   age_ti[i,t] = age_1i + (t-1) * rnorm(1,3,sqrt(.2))
     y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
  }
   }

data.5=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))

names(data.5)=c('id','age1','age2','age3',
             'y1','y2','y3',
             'b_0i','b_1i')

## Transform Data to Long ##
data.5.long = reshape(data.5, idvar='id',
     varying=c('age1','y1','age2','y2','age3','y3'),
     times=c(1,2,3),
     v.names=c('age','y'), direction='long')

data.5.long = data.5.long[order(data.5.long$id, data.5.long$age),]

## Plot ##

plot_obs <- ggplot(data=data.5.long, aes(x=age, y=y, group=id)) +
          geom_line() +
          theme_bw()

print(plot_obs)
```

```
## Center Age ##
data.5.long$agec = data.5.long$age - 4

####### Create Syntax for Fifth Data Set #######
data.5.script <- mplusObject(
TITLE = "Growth Model for Data #5;",
VARIABLE = "USEVARIABLES = y agec;
        CLUSTER=id;
        WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
     ESTIMATOR = BAYES;",
MODEL = loopReplace("
%WITHIN%
   b1 | y ON agec;
   y (sigma2_u);

%BETWEEN%
  [y]  (beta_0);
  [b1] (beta_1);

   y   (sigma2_0);
   b1  (sigma2_1);
   y WITH b1 (sigma_10);

MODEL PRIORS:
sigma2_u ~ IG([[PriorIG_1]], [[PriorIG_2]]);
beta_0 ~ N([[BETA_0]], [[PriorN_2_1]]);
beta_1 ~ N([[BETA_1]], [[PriorN_2_2]]);
sigma2_0 ~ IW([[Prior_IW_1]], [[Prior_IW_N]]); !mean_estimate*(sample_size - p -1),
sample_size)
sigma2_1 ~ IW([[Prior_IW_2]], [[Prior_IW_N]]);
sigma_10 ~ IW([[Prior_IW_3]], [[Prior_IW_N]]);", df.f)
,
usevariables = c("y", "agec", "id"),
rdata = data.5.long)

data.5.result = mplusModeler(data.5.script, "data.5.long.dat", modelout = "data.5.inp",
run = 1L)

parms = data.5.result$results$parameters
parms

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est)
names(df_parms.1) <- c("param","est")
df_parms2.1 <- t(df_parms.1)
df_parms2.1 <- as.data.frame(df_parms2.1)
```

```
df2.1 <- df_parms2.1[2,]
df2.1

df_parms.2 <- data.frame(parms[[1]]$param, parms[[1]]$posterior_sd)
names(df_parms.2) <- c("param","sd")

df_parms.2[7,2] = (N - 2)/2  #PriorIG_1 sample size/2
df_parms.2[7,2]

df_parms.2[8,2] = ((N - 2)*(df_parms.1[1,2]))/2 #PriorIG_2 sample size*variance/2
df_parms.2[8,2]

df_parms.2[9,2] = (df_parms.2[3,2])^2
df_parms.2[9,2]

df_parms.2[10,2] = (df_parms.2[4,2])^2
df_parms.2[10,2]

df_parms.2[11,2] = df_parms.1[5,2]*(N - 3)
df_parms.2[11,2]

df_parms.2[12,2] = df_parms.1[6,2]*(N - 3)
df_parms.2[12,2]

df_parms.2[13,2] = df_parms.1[2,2]*(N - 3)
df_parms.2[13,2]

df_parms.2[14,2] = (N - 3)
df_parms.2[14,2]

df_parms2.2 <- t(df_parms.2)
df_parms2.2 <- as.data.frame(df_parms2.2)

df2.2 <- df_parms2.2[2,]

df.f = cbind(df2.1,df2.2)
df.f

names(df.f) <-
c("SIGMA2_U","SIGMA_10","BETA_0","BETA_1","SIGMA2_0","SIGMA2_1",
"SIGMA2_U_SD","SIGMA_10_SD","BETA_0_SD","BETA_1_SD","SIGMA2_0_SD",
              "SIGMA2_1_SD","PriorIG_1","PriorIG_2","PriorN_2_1","PriorN_2_2",
              "Prior_IW_1","Prior_IW_2","Prior_IW_3", "Prior_IW_N")
```

```
####################################
#
#    Generate Sixth Data Set
#
####################################

b_0i=NULL
b_1i=NULL
id=NULL
age_1i=NULL
N_1 = 1000
T_1 = 3
age_ti=matrix(NA,N_1,T_1)
y_ti=matrix(NA,N_1,T_1)
Psi_11 = .2
Psi_22 = .01
Psi_21 = 0
R = Psi_21/(sqrt(Psi_11*Psi_22))


     for(i in 1:N_1){
       id[i]=i + 5000
       b_0i[i]=rnorm(1,0,1)
       b_1i[i]=R * b_0i[i] + rnorm(1,0,sqrt(1-R^2))
       b_0i[i]=(b_0i[i] * sqrt(Psi_11))- 2
       b_1i[i]=(b_1i[i] * sqrt(Psi_22))+ .4

       age_1i =rnorm(1,6,sqrt(.8))

    for(t in 1:T_1){
     age_ti[i,t] = age_1i + (t-1) * rnorm(1,5,sqrt(.2))
       y_ti[i,t]   = b_0i[i]+b_1i[i]*((age_ti[i,t]-4))+rnorm(1,0,sqrt(.1))
    }
     }

data.6=as.data.frame(cbind(id,age_ti,y_ti,b_0i,b_1i))

names(data.6)=c('id','age1','age2','age3',
              'y1','y2','y3',
              'b_0i','b_1i')

## Transform Data to Long ##
data.6.long = reshape(data.6, idvar='id',
    varying=c('age1','y1','age2','y2','age3','y3'),
    times=c(1,2,3),
    v.names=c('age','y'), direction='long')
```

```
data.6.long = data.6.long[order(data.6.long$id, data.6.long$age),]

## Plot ##
plot_obs <- ggplot(data=data.6.long, aes(x=age, y=y, group=id)) +
            geom_line() +
            theme_bw()

print(plot_obs)

## Center Age
data.6.long$agec = data.6.long$age - 4

####### Create Syntax for Sixth Data Set #######
data.6.script <- mplusObject(
TITLE = "Growth Model for Data #6;",
VARIABLE = "USEVARIABLES = y agec;
        CLUSTER=id;
        WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
     ESTIMATOR = BAYES;",
MODEL = loopReplace("
%WITHIN%
   b1 | y ON agec;
   y (sigma2_u);

%BETWEEN%
  [y]  (beta_0);
  [b1] (beta_1);

   y   (sigma2_0);
   b1  (sigma2_1);
   y WITH b1 (sigma_10);

MODEL PRIORS:
sigma2_u ~ IG([[PriorIG_1]], [[PriorIG_2]]);
beta_0 ~ N([[BETA_0]], [[PriorN_2_1]]);
beta_1 ~ N([[BETA_1]], [[PriorN_2_2]]);
sigma2_0 ~ IW([[Prior_IW_1]], [[Prior_IW_N]]); !mean_estimate*(sample_size - p -1),
sample_size)
sigma2_1 ~ IW([[Prior_IW_2]], [[Prior_IW_N]]);
sigma_10 ~ IW([[Prior_IW_3]], [[Prior_IW_N]]);", df.f)
,
usevariables = c("y", "agec", "id"),
rdata = data.6.long)
```

```
data.6.result = mplusModeler(data.6.script, "data.6.long.dat", modelout = "data.6.inp",
run = 1L)

parms = data.6.result$results$parameters

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est, parms[[1]]$posterior_sd)

bs_residual_est[1,h] = df_parms.1[1,2]
bs_int_slp_cov_est[1,h] = df_parms.1[2,2]
bs_intercept_mean_est[1,h] = df_parms.1[3,2]
bs_slope_mean_est[1,h] = df_parms.1[4,2]
bs_int_var_est[1,h] = df_parms.1[5,2]
bs_slp_var_est[1,h] = df_parms.1[6,2]

bs_residual_sd[1,h] = df_parms.1[1,3]
bs_int_slp_cov_sd[1,h] = df_parms.1[2,3]
bs_intercept_mean_sd[1,h] = df_parms.1[3,3]
bs_slope_mean_sd[1,h] = df_parms.1[4,3]
bs_int_var_sd[1,h] = df_parms.1[5,3]
bs_slp_var_sd[1,h] = df_parms.1[6,3]

###################### Data Fusion ########################

data.fused = rbind(data.1.long,data.2.long,data.3.long,data.4.long,data.5.long,data.6.long)

### BAYES ###
data.fused.script.bayes <- mplusObject(
TITLE = "Growth Model for Combined Data Bayes;",
VARIABLE = "USEVARIABLES = y agec;
          CLUSTER=id;
          WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
      ESTIMATOR = BAYES;",
MODEL = "
%WITHIN%
    b1 | y ON agec;
    y (sigma2_u);

%BETWEEN%
   [y]  (beta_0);
   [b1] (beta_1);

    y   (sigma2_0);
    b1  (sigma2_1);
    y WITH b1 (sigma_10);
MODEL PRIORS:
```

```
sigma2_u ~ IG(-1, 0);
beta_0 ~ N(0, 10^10);
beta_1 ~ N(0, 10^10);
sigma2_0 ~ IW(0,-3);
sigma2_1 ~ IW(0,-3);
sigma_10 ~ IW(0,-3);
",
usevariables = c("id","b_0i","b_1i","time","age","y","agec"),
rdata = data.fused)

data.fused.result.bayes = mplusModeler(data.fused.script.bayes, "data.fused.dat",
modelout = "data.fused.bayes.inp", run = 1L)

parms = data.fused.result.bayes$results$parameters

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est, parms[[1]]$posterior_sd)

bf_residual_est[1,h] = df_parms.1[1,2]
bf_int_slp_cov_est[1,h] = df_parms.1[2,2]
bf_intercept_mean_est[1,h] = df_parms.1[3,2]
bf_slope_mean_est[1,h] = df_parms.1[4,2]
bf_int_var_est[1,h] = df_parms.1[5,2]
bf_slp_var_est[1,h] = df_parms.1[6,2]

bf_residual_sd[1,h] = df_parms.1[1,3]
bf_int_slp_cov_sd[1,h] = df_parms.1[2,3]
bf_intercept_mean_sd[1,h] = df_parms.1[3,3]
bf_slope_mean_sd[1,h] = df_parms.1[4,3]
bf_int_var_sd[1,h] = df_parms.1[5,3]
bf_slp_var_sd[1,h] = df_parms.1[6,3]

#### ML ####

data.fused.script.ml <- mplusObject(
TITLE = "Growth Model for Combined Data ML;",
VARIABLE = "USEVARIABLES = y agec;
        CLUSTER=id;
        WITHIN=agec;",
ANALYSIS = "TYPE=TWOLEVEL RANDOM;
     ESTIMATOR = ML;",
MODEL = "
%WITHIN%
   b1 | y ON agec;
   y (sigma2_u);

%BETWEEN%
```

```
    [y]  (beta_0);
    [b1] (beta_1);

     y   (sigma2_0);
     b1  (sigma2_1);
     y WITH b1 (sigma_10);
",
usevariables = c("id","b_0i","b_1i","time","age","y","agec"),
rdata = data.fused)

data.fused.result.ml = mplusModeler(data.fused.script.ml, "data.fused.dat", modelout =
"data.fused.ml.inp", run = 1L)

parms = data.fused.result.ml$results$parameters

df_parms.1 <- data.frame(parms[[1]]$param, parms[[1]]$est, parms[[1]]$se)

ml_residual_est[1,h] = df_parms.1[1,2]
ml_int_slp_cov_est[1,h] = df_parms.1[2,2]
ml_intercept_mean_est[1,h] = df_parms.1[3,2]
ml_slope_mean_est[1,h] = df_parms.1[4,2]
ml_int_var_est[1,h] = df_parms.1[5,2]
ml_slp_var_est[1,h] = df_parms.1[6,2]

ml_residual_sd[1,h] = df_parms.1[1,3]
ml_int_slp_cov_sd[1,h] = df_parms.1[2,3]
ml_intercept_mean_sd[1,h] = df_parms.1[3,3]
ml_slope_mean_sd[1,h] = df_parms.1[4,3]
ml_int_var_sd[1,h] = df_parms.1[5,3]
ml_slp_var_sd[1,h] = df_parms.1[6,3]

}

###################~~RESULTS~~#########################

###Bayesian Synthesis Results###
bs_residual_est
bs_int_slp_cov_est
bs_intercept_mean_est
bs_slope_mean_est
bs_int_var_est
bs_slp_var_est

bs_residual_sd
bs_int_slp_cov_sd
bs_intercept_mean_sd
```

```
bs_slope_mean_sd
bs_int_var_sd
bs_slp_var_sd

###Bayesian analysis of Fused Data###
bf_residual_est
bf_int_slp_cov_est
bf_intercept_mean_est
bf_slope_mean_est
bf_int_var_est
bf_slp_var_est

bf_residual_sd
bf_int_slp_cov_sd
bf_intercept_mean_sd
bf_slope_mean_sd
bf_int_var_sd
bf_slp_var_sd

###ML Data Fusion###
ml_residual_est
ml_int_slp_cov_est
ml_intercept_mean_est
ml_slope_mean_est
ml_int_var_est
ml_slp_var_est

ml_residual_sd
ml_int_slp_cov_sd
ml_intercept_mean_sd
ml_slope_mean_sd
ml_int_var_sd
ml_slp_var_sd

results_output_all_250 = rbind(bs_residual_est,bs_int_slp_cov_est,
bs_intercept_mean_est,bs_slope_mean_est,bs_int_var_est,bs_slp_var_est,
bs_residual_sd,bs_int_slp_cov_sd,bs_intercept_mean_sd,bs_slope_mean_sd,bs_int_var_s
d,bs_slp_var_sd, bf_residual_est,bf_int_slp_cov_est,bf_intercept_mean_est,
bf_slope_mean_est,bf_int_var_est,bf_slp_var_est,bf_residual_sd,bf_int_slp_cov_sd,
bf_intercept_mean_sd,bf_slope_mean_sd,bf_int_var_sd,bf_slp_var_sd,ml_residual_est,
ml_int_slp_cov_est,ml_intercept_mean_est,ml_slope_mean_est,ml_int_var_est,
ml_slp_var_est,ml_residual_sd,ml_int_slp_cov_sd,ml_intercept_mean_sd,
ml_slope_mean_sd,ml_int_var_sd,ml_slp_var_sd)
```