Title
Analysis Strategies for Planned Missing Data in Health Sciences and Education Research

Author
Harrell, Lauren Allison

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Analysis Strategies for Planned Missing Data in Health Sciences and Education Research

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

by

Lauren Allison Harrell

2015

Abstract of the Dissertation

Analysis Strategies for Planned Missing Data in Health Sciences and Education Research

by

Lauren Allison Harrell

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2015

Professor Thomas R. Belin, Chair

In health and educational research, planned-missing-data designs have been used to reduce the number of variables collected on participants, thus reducing respondent burden and the number of resources necessary for study. The purpose of this dissertation research is to develop and improve analysis strategies for planned-missing-data designs, with specific applications to partial mouth recording protocols in oral health studies and balanced incomplete block designs in large-scale educational survey assessments. For the oral-health examination, multidimensional item response theory models (MIRT) are investigated in addition to multiple imputation strategies from hierarchical normal models to recover information on periodontal disease status when data are collected on only half of the mouth. Using data from the National Assessment of Educational Progress (NAEP), complex MIRT models are investigated to improve the estimation of population ability characteristics as well as to explore the potential for other components of academic to be measured from the same data.

The dissertation of Lauren Allison Harrell is approved.

Catherine Crespi-Chun

Christina M. Kitchen

Vivek Shetty

Li Cai

Thomas R. Belin, Committee Chair

University of California, Los Angeles

2015

To Biscuit...

...I know.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My experience at UCLA was incredible thanks to my team of mentors. First I would like to thank my advisor, Dr. Tom Belin, who has been helping me pursue my goals since before my first day at UCLA when I had asked his help in finding a GSR. Dr. Belin has been outstanding as a mentor, storyteller, and advocate, providing guidance on everything ranging from my research, successful collaboration and communication, my career, and recognizing when "it's a trap." Through my training grant in Advanced Quantitative Methods in Education Research, I was fortunate to have a second advisor, Dr. Li Cai. Dr. Cai has been instrumental in my career trajectory and my research, helping me navigate the field of educational measurement and to pursue creative methodological solutions to my varied research interests. Dr. Vivek Shetty, in addition to supplying collaborative research experiences and incorporating my planned-missing-data designs into the data collection, has imparted invaluable wisdom. I would also like to thank my committee members Dr. Christina Ramirez and Dr. Kate Crespi for the helpful discussions and feedback. It has been an absolute pleasure working with you all.

To Mike, I cannot thank you enough (but I'll try anyways) for taking care of me, encouraging my ideas, keeping me happy and thinking positively, and helping in every way possible throughout my graduate career (and taking me to Disneyland the day after my defense). You are wonderful, and I look forward to our next adventures when I'm no longer a student. In addition, I would not have been successful without the love and encouragement from my family, Lizz, Wells, Mom, Dad, and Marty. I owe a debt of gratitude to many special people in my life, including (but not limited to) Kestrel, Jenga, Ritika, Christie, Sean, John, Bob and Jeanne Seibert, Heather and Hawk Arps, and the rest of my friends, for their support and faith in me. I would also like to thank my communities (dissertation writers group + Bonnie and Katie, PHSA, UCLA Biostatistics, ZSC, ERB, & PP) for being a part of my life. Abbey, thanks for the warmth and companionship.

I would not have made it through either the MS or PhD program without the staff in the Biostatistics and FSPH Student Affairs offices. Particularly, Roxana Naranjo has gone above and beyond in helping me navigate the degree process and financial aid. I also want to acknowledge Jason Clague for helping me with the dental health data cleaning and analysis.

Vita

2006–2007	Tutor/Grading Teaching Assistant, Claremont McKenna College
2007–2009	Research Assistant, College of Veterinary Medicine, Western University
2009	B.S. (Biology) with Sequence in Scientific Modeling, Claremont McKenna College
2009–present	Graduate Student Researcher, Department of Biostatistics/School of Dentistry, UCLA.
2010–2014	Special Reader, Department of Biostatistics, UCLA.
2011	M.S. (Biostatistics), UCLA.
2014–2014	Graduate Student Researcher, National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Publications and Presentations

Harrell, L, Belin, T, & Shetty, V. (2014) Developing a Propensity-Score Framework for Evaluating the Oral Health Consequences of Methamphetamine Use. (Submitted: Under Review).

Dye, B., Harrell, L., Murphy, D., Castro, T., Belin, T.R., & Shetty, V. (2014) Performance of a Quality Assurance Program for Assessing Dental Health in Metham-

phetamine Users. *BMC: Oral Health* (Submitted: Under Review).

Murphy, D, Harrell, L, Fintzy, R, Vitero, S., Gutierrez, A. & Shetty, V. (2014) Soda Consumption Among Methamphetamine Users Being Seen for a Dental Exam. *Oral Health and Preventative Dentistry*, (In Press).

Murphy, D.A., Harrell, L., Fintzy, R., Belin, T. R., Gutierrez, A., Vitero, S. J., & Shetty, V. (2014) A Comparison of Methamphetamine Users to a Matched NHANES Cohort: Propensity Score Analyses for Oral Health Care and Dental Service Need.*Journal of Behavioral Health Services & Research*, (In Press).

Robles, T., Sharma, R., Harrell, L,. Elashoff, D.A., Yamaguchi, M., & Shetty, V. (2013) Saliva Sampling Method affects Performance of a Salivary α-Amylase Biosensor. *American Journal of Human Biology.* 25.6: 719-724.

Harrell, L., and V. Shetty. (2012) Extended Antibiotic Therapy may Reduce Risk of Infection Following Orthognathic Surgery. *Journal of Evidence Based Dental Practice* 12.3: 144-145.

Robles, T. F., Sharma, R., Park, K. S., Harrell, L., Yamaguchi, M., & Shetty, V. (2012) Utility of a Salivary Biosensor for Objective Assessment of Surgery-Related Stress. *Journal of Oral and Maxillofacial Surgery.*

Young, S. D., Harrell, L., Jaganath, D., Cohen, A. C., & Shoptaw, S. (2012). Feasibility of recruiting peer educators for an online social networking-based health intervention. *Health Education Journal.*

Harrell, L., and L. Cai. (2015) "Multidimensional Item Calibration and Plausible

Value Imputation in Large-Scale Educational Assessments using the Metropolis-Hastings Robbins-Monro Algorithm." *Modern Modeling Methods (M3) Conference*, Storrs, CT.

Harrell, L., and L. Cai. (2015) "Plausible Value Imputation From Multidimensional IRT Models Using Stochastic Approximation Methods." *National Council on Education Measurement*, Chicago, IL.

Harrell, L. (2014) "The Use of Item Response Theory Models to Evaluate Oral Health Consequences of Methamphetamine Use." Presented at the *Southern California Chapter of the American Statistical Association*, Duarte, CA.

Harrell, L., T. Belin, and L. Cai. (2014) "Multiple Imputation and Multidimensionality in Large-Scale Educational Assessments." *Joint Statistical Meetings*, Boston, MA.

Harrell, L. (2014) "Item Response Theory Models for Periodontal Examination Data with Planned Missingness." Presented at *Modern Modeling Methods (M3) Conference*, Storrs, CT.

Harrell, L., and L. Cai. (2014) "Multidimensional Item Response Theory for Cognitive Items in Large-Scale Assessments: an Application of the Metropolis-Hastings Robins-Monro Algorithm." *The CATS Conference*, Redondo Beach, CA.

Harrell, L., and L. Cai. (2014) "Multidimensional Item Response Theory in Large-Scale Assessments." *National Council on Education Measurement*, Philadelphia, PA.

Harrell, L. (2013) "Analysis Strategies for Planned Missing Data in an Oral Health Study." *Joint Statistical Meetings*, Montreal, QC.

Harrell, L. (2013) "An Item Response Theory Model for Periodontal Examination Data with Missingness by Design." *Western North American Region of the International Biometrics Society*, Los Angeles, CA.

CHAPTER 1

Problem Statement

The purpose of the dissertation research presented here is to combine recent analysis methodologies in novel ways in order to combat the challenges posed by planned-missing-data designs in health and education research. The goal is to improve the methodologies such that planned-missing-data designs can be more readily applied without a critical loss of information for statistical inference.

Response burden is a problem confronting researchers interested in the complex relationships affecting outcomes in human subjects research, particularly when addressing vulnerable or low-income subjects. Response burden can induce failure to complete questionnaires and potential loss of research participants to follow-up in longitudinal studies. When data are incomplete, failure to account for relationships between variables can result in bias and loss of precision. While many statistical methods have been developed in response to missing data, there are strong scientific reasons to take steps in the study design to circumvent the reduction in participants. Edwards et al. [ERC02] showed that with mailed health surveys, shorter questionnaires roughly doubled the probability of response.

One tool for alleviating subject burden in educational statistics is the idea planned missing data, where an individual's test, survey, or examination includes only a sample of items or variables of interest. The National Assessment of Educational Progress (NAEP) has been administering examinations with incomplete block or matrix sampling design strategies since 1969 to appraise the knowledge of students in the United States in specific subject areas including mathematics,

writing, reading, geography, history, and science. The NAEP analysis model, in use since 1984, has undergone limited updates.

Planned-missing-data designs have become increasingly prevalent in recent years, particularly in longitudinal and psychological settings [PR10, GTC01]. There has been some exploration of the concept in health-related research, but there is considerable potential for expanding the scope of this idea by integrating it with modern statistical analysis techniques [RG95]. As will be discussed in further detail in Chapter 4, until 2009, the National Health and Nutrition Examination Study (NHANES) used partial-mouth examinations instead of full-mouth sampling in the periodontal examination due to patient time constraints, pain associated with exam probes, and cost. Full-mouth examinations were then implemented due to the underestimation of periodontal disease prevalence by partial-mouth recording protocols [EDW12]. For any study utilizing periodontal examinations, the time and cost, as well as pain to the study participants, has not gone away. It is hypothesized that analysis strategies can be developed to improve estimates of periodontal disease prevalence using half-mouth data.

The methodology for estimating population quantities from large-scale educational assessments was developed nearly 25 years ago using the available software [Mis91, MJM92b]. As many advancements have been made in computing and educational measurement, the existing framework for large-scale assessments, which was designed during a time when data storage was at a premium and processor speeds were low, should be re-evaluated in order to estimate a greater range of possible cognitive abilities.

CHAPTER 2

Background/Literature review

The background and context for the research are split into five major sections. Section 2.1 introduces the primary two datasets utilized in this research. In Section 2.2, a brief overview of item response theory models is presented, which will be applicable to both datasets. The large-scale educational assessment framework, with the generation and use of plausible values, is summarized in Section 2.3. The use of propensity scores to select a representative control sample of non-methamphetamine users is detailed in Section 2.4. Finally, the overall structure of the dissertation is described in Section 2.5.

2.1 Description of the datasets

The research presented in this dissertation is applied to two different datasets that each have planned-missing-data designs. Section 2.1, which describes these two data sources, is broken into two components. First, some preliminary background on the National Assessment of Educational Progress is discussed in Section 2.1.1. Then, the study of the oral health consequences of methamphetamine use is described in Section 2.1.2.

2.1.1 National Assessment of Educational Progress

In educational assessment, interest focuses on the ability of the student or respondent in a given subject area. Ability, however, can be conceptualized as an

individual's capacity to respond to any of an infinite number of potential challenges, and as such, cannot be measured directly. Therefore it makes sense to consider ability to be be considered a latent, or unobserved, variable. Mislevy, Johnson and Muraki (1992) detailed the analysis methods used in NAEP since 1984 which implement the balanced-incomplete-block-design approach to assess proficiency. Until 1984, the matrix design strategy involved booklets with alternating taped instructions to specify the subset of questions being administered in different testing sessions [NAE11]. The booklet for an individual session was the same, with different taped instructions for test-takers at a classroom level, but booklets between sessions shared no common response. In 1984, the design strategy switched to a balanced-incomplete-block design, in which dissimilar blocks of items appear in different booklets. Each booklet also contains a set of common items, and blocks are balanced such that the length of the exams is reasonable.

For the purposes of this research, data from 2011 NAEP Science Assessment will be considered. The NAEP assessment data are collected on a complex sample of 4th, 8th, and 12th graders, sampled within districts and schools. The Science assessment framework has three primary content domains, namely Life Sciences, Earth Sciences, and Physical Sciences, in addition to scientific practice and cognitive demands.

2.1.2 Oral Consequences of Methamphetamine Use

In a study of 574 methamphetamine users from the Los Angeles metropolitan area, data were collected on the lifetime history of methamphetamine use as well as on oral health outcomes from both self-reported and examination measures. The two oral-health examinations given were the DMFS, which indicates each tooth surface as decayed, missing, filled, or sound (i.e. intact, reflecting no apparent problems), and the periodontal examination, which measures gum disease through probing four sites per tooth, checking for bleeding, and quantifying re-

cession, pocket depth, and attachment loss. While bleeding is a dichotomous outcome, pocket depth is an ordinal variable that can take on integer values between 0 and 12 mm, and attachment loss may range between -4 and 15 mm, with negative values referring to a scenario when gum tissue appears above a target level associated with ideal attachment.

Subjects in this study were randomized to receive either a full-mouth periodontal exam, where all the teeth are probed, or one of two half-mouth exams where either the upper-right and lower-left quadrants or the upper-left and lower-right quadrants are observed. Diagrams of the quadrants of the mouth sampled are depicted in Figure 2.1. The four probe sites, mesiobuccal (M), buccal (B), distal-facial (D), and distal-lingual (DL), are marked in blue.

2.2 Item Response Theory Models

Item response theory models describe the relationship of categorical test items with a continuous latent trait, such as ability. Let θ be the parameter (or a vector of parameters) that summarizes a test-taker's ability in a given subject area(s). The values of θ are unobserved directly and are therefore considered latent variables. These ability parameters can be estimated through the answers a test-taker provides to each of the test items (and perhaps additional background characteristics). In item response theory, the probability of a given response to an item is modeled as a function of the ability parameter(s) θ.

2.2.1 Two Parameter Logistic Model

The two parameter logistic model (2PL), one of the more commonly used item response models, relates the latent ability to dichotomous items. The probability of a correct response is modeled as a function of two parameters, the slope and

location:

$$P(y_j = 1|\theta, a_j, b_j) = 1/(1 + exp[a_j(\theta - b_j)])$$
$$\equiv P_j(\theta).$$

The location parameter, b_j, is also known as the difficulty parameter and can be interpreted as the value of θ at which the probability of a correct response is 0.5. The slope parameter, a_j, relates the strength of the relationship between the item and the ability. The slope can also be interpreted as the value of the slope at b_j.

2.2.2 Three Parameter Logistic Model

The three parameter logistic model (3PL) is used in instances where there is a binary correct versus incorrect answer. Usually used for individual items on multiple choice exams, this model predicts the probability of a correct response on a question based on θ, the student's proficiency in the given subject area, and three other item-specific parameters quantifying sensitivity to proficiency, difficulty, and probability of random correct response. Let j be the index of a given question. The 3PL used in NAEP is characterized as follows:

$$P(x_j = 1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j)/(1 + exp[-1.7a_j(\theta - b_j)])$$
$$\equiv P_j(\theta), \tag{2.1}$$

where

- x_j indicates a correct response (1 if correct, 0 if incorrect)
- a_j is the slope for the item j, which describes the relationship between proficiency and probability of correct response for the given question j ($a_j >$

0)

- b_j characterizes the general difficulty of the question
- c_j is the probability of a correct response from students of low proficiency $(0 \leq c_j < 1)$

In multiple choice questions on NAEP, c_j was originally estimated by using the reciprocal of the number of possible answers [MJM92a]. For example, if a student is presented with five possible answers, c_j wwould be estimated by $\frac{1}{5}$, which is the probability of a correct response given a completely random guess. Unit scale can be chosen arbitrarily to ensure linearity for Equation (2.1). Now, c_j can be estimated by specifying a prior distribution on c_j, such as Beta(1,4) for a five possible answers.

2.2.3 Graded Response Model

Introduced by Samejima [Sam69], the graded response model is used to estimate the probability of response for items with i ordered categories. In cognitive assessments, the graded response model is often used for test items in which partial credit may be given. Let item j have K graded categories. The cumulative probabilities of response are

$$\begin{aligned} P(x_j \geq 0|\theta) &= 1.0 \\ P(x_j \geq 1|\theta) &= \frac{1}{1 + exp[-c_{j,1} + a_j\theta]} \\ &... \\ P(x_j \geq K-1|\theta) &= \frac{1}{1 + exp[-c_{j,K-1} + a_j\theta]} \\ P(x_j \geq K|\theta) &= 0, \end{aligned}$$

where a_j is the slope parameter for item j and $c_{j,k}$ are the item intercepts for the $k = 0, ..., K-1$ levels.

2.3 Large Scale Assessment Conceptual Framework

Plausible value methodology has long been the gold standard in large-scale educational assessments. Plausible values were introduced as a device to better approximate the population distribution of abilities, treating the values of the student ability parameters as missing data. Rather than a single point estimate of student-level proficiency, however, multiple estimates are imputed from a posterior distribution that includes the student's item responses and background characteristics [Mis91, MJM92b].

Let $X = (x_1, , x_N)$ be the matrix of demographic covariates, where x_i is a vector of demographic characteristics for individual i. Let $y_i = (y_{i1}, , y_{iD})$ be the individual outcomes on the D proficiency domains, comprised of item responses y_{id} with typical element y_{id}^l denoting item l within domain d. Of interest is estimating the unobserved proficiency in D domains, and the vector of these latent variables for student i is represented by θ_i=$(\theta_{i1}, , \theta_{iD})$. In this framework, students are considered independent of one another, and responses to different questions by an individual student are also assumed independent. This framework does not allow or consider conditional dependence between items across different domains, and item responses are theorized to only depend on the proficiency domain of which it was designed to measure [TG97].

Appealing to the idea of using conditional independence assumptions to represent salient features of measured outcomes in line with de Finetti's Theorem, it does simplify the construction of the probability distribution for student i's responses y_i, conditional on his or her proficiency vector θ_i to $\prod_{d=1}^{D}[f_d(y_{id}|\theta_{id})]$. The probability model of an observed response pattern for content domain d for stu-

dent i depends on the scoring or response format of the question. Dichotomously scored items are often represented using the classical two parameter logistic model or three parameter logistic model. For multiple-categorical items, more complex IRT models such as the graded response model [Sam69] or the generalized partial credit model [Mur92] may be used.

Let β_d represent the vector of all the item parameters that relate to the proficiency domain d such that $\beta = (\beta_1, , \beta_D)$. The latent proficiency vector θ_i, is assumed to be conditionally normally distributed with mean vector Γx_i and covariance matrix Σ, where Γ represents a matrix of unknown regression parameters [TG97]. Let $\phi(\theta_i; \Gamma x_i, \Sigma)$ represent the conditional normal density function of the latent proficiency vector. The likelihood function of the parameters β, , and Σ is constructed as

$$L(\beta, \Gamma, \Sigma | \theta, X, Y) \propto \prod_{i=1}^{N} \phi(\theta_i; \Gamma x_i, \Sigma) \prod_{d=1}^{D} f_d(y_{id} \theta_{id}; \beta_d),$$

where $Y = (y_1, , y_N)$ and $\theta = (\theta_1, , \theta_N)$, which is proportional to the posterior distribution of the latent proficiency vector $f(\theta | X, Y, \beta, \Gamma, \Sigma)$.

The posterior distribution is used to draw the plausible values in four major steps. First, the item parameters (β_d) are estimated separately for each proficiency domain. Note that this does not involve simultaneous estimation of the item parameters for the multivariate proficiency vector θ_i. Second, treating the estimated item parameters as if they are known, Γ and Σ are estimated using the observed item response data in an Expectation-Maximization (EM) algorithm [DLR77, BA81] by fixing the item parameters to their estimates from step one. In step three, the posterior distribution of β, Γ, and Σ is then approximated from these regression parameter and error covariance matrix estimates based on a large-sample normal approximation centered on the posterior mode emerging from the EM algorithm. And finally, for each of the generated regression parameter sets and covariance matrices, a value of θ_i for each student is drawn from a

normal distribution with mean Γx_i and covariance matrix Σ. This final step of generating imputations is generally repeated 5 times to produce 5 plausible values per student [TG97].

The existing plausible values framework produces data sets that are relatively user-friendly for secondary analysis, but generation of the plausible values is tedious and computationally intensive. The current framework defines the proficiency domains a priori and fails to account for any conditional dependence between items on different proficiency exams. The current formulation does not allow for investigators to explore alternative parameterizations of the latent proficiency space. Consider as a concrete example the case of NAEP science and mathematics assessments. There are a number of items related to data analysis and statistics on both the science and mathematics sections in the 8th and 12th grade NAEP exams. The existing plausible values assume that these questions are independent, conditional on only the examinees proficiency on the exam subject in which the question is presented, even though there may be residual dependence between the items, especially since the items on both sections relate to a meaningful domain that may be labeled data analysis and statistics. If an investigator would like to measure proficiency in data analysis and statistics using items from both science and mathematics assessments, the items would have to be recalibrated and then the plausible values imputation would have to be redone.

The plausible values estimates in this framework are also potentially biased from not accounting for local dependence and other confounding effects [Yen84]. For instance, there may be several questions (forming a testlet) following the same reading passage, and a student's responses to those questions can be expected to be more related to one another than to other items due to the effect of the common stimulus above and beyond what can be attributed to reading proficiency [Cai10, CSH11]. However the current models do not control for testlet-level dependence. The current operational framework simply combines correlated items

into a single graded item. The multilevel nature of the data, such as the variability among students within a school or among schools within a district, is another source of potential bias in the imputation. Thomas [Tho00] highlighted the sensitivity of the imputation model to assumptions such as homogeneous variance across subpopulations and noted that computational intensity was a limiting factor in expanding the existing models. The proposed model could resolve the potential biases existing in the current plausible value imputation methods by accounting for the multilevel data structure, a multivariate domain framework, sub-population level variability, as well as nuisance local dependence.

Recent research has been conducted indicating the feasibility of certain multidimensional models in large scale assessments, such as the bifactor model, where there is one general domain and several subdomains [RJD14a], or a simple structure where all the subdomains are estimated simultaneously and allowed to covary [SD05]. A stochastic approximation to the latent regression simultaneous with item parameter estimation, similar to what is being proposed here, was investigated by von Davier and Sinharay [DS10], with the results compared to the existing software for estimation used in NAEP. The resulting software, SGROUP, improved the estimates of the posterior standard deviations. While this approach was applied using simple multidimensional IRT models, such as simultaneous estimation of mathematics subscales, it was not applied to more complex, structured MIRT models. Nor has the impact on plausible value imputation been explored.

2.3.1 Summary of Plausible Values Imputation

When interested in population characteristics, consistent estimates can be obtained by maximizing the likelihood only using population parameters and the data. A statistic $t(\theta, Y)$ can be computed to estimate a population characteristic of interest, T, where Y are the responses of all sampled students to the background questions. The variance of $t(\theta, Y)$ around T can be estimated using a

jackknife or bootstrap estimate $U(\theta, Y)$. As θ is a latent, unobserved, variable, it can be treated as missing, and can be approximated given observed responses to questions X and background variables Y [MJM92a]:

$$\begin{aligned} t*(X,Y) &= E[t(\theta,Y)|X,Y] \\ &= \int t(\theta,Y)p(\theta|X,Y)d\theta \end{aligned} \tag{2.2}$$

When closed-form solutions cannot be obtained, Monte Carlo integration can be used to approximate by drawing randomly from $p(\theta|x_i, y_i)$. Thus the value of θ can be considered an imputation for a given subject i by randomly selecting from the conditional distribution of θ given the examinee's responses. If multiple random draws are performed for each student, estimates of uncertainty due to the missingness of θ can also be obtained [MJM92a, LR02].
To compute the conditional distribution of plausible values for θ, conditional independence on the background characteristics is assumed such that

$$p(\theta|x_i, y_i) \propto P(x_i|\theta)p(\theta|y_i) \tag{2.3}$$

The algorithms for estimating the posterior distribution of θ given the data in NAEP have been evolving since 1985, and the literature includes many discussions after Mislevy, Johnson, and Muraki's paper in 1992 on the appropriate model for plausible values imputation. The imputation methods in Mislevy, Johnson, and Muraki focus only on drawing values of the parameter from the posterior distribution of the observed data. Thus there is no imputation considered for the missing data due to questions not presented to each subject. Thomas and Gan [TG97] provided improvements on the multiple imputation methods such that multiple imputation is extended to the item level to generate complete imputed datasets.

2.3.2 Analysis based on Multiple Imputation

Given a number of imputations, M, for the individual θ from the conditional distribution $p(\theta_i|x_i, y_i)$, inferences on the population quantity of interest T can be based on the scalar statistic $t(\theta, Y)$ using multiple imputation analysis procedures. Let $m = 1, .., M$ denote the imputation number. From each set of $\hat{\theta}_m$, one can evaluate t yielding $\hat{t}_m$. The sampling variance U_m can be computed using a multiple weight jackknife approach. The estimate of t is given in Equation (2.4) [MJM92a]. The procedure for making inferences on T from t is consistent with the procedure specified by Little and Rubin (2002), which accounts for variability within and between imputations.

$$t^* = \sum_{m=1}^{M} \hat{t}_m / M \tag{2.4}$$

The average sampling variance is given in Equation (2.5), which is an estimate of the uncertainty about t due to sampling.

$$U^* = \sum_{m=1}^{M} U_m / M \tag{2.5}$$

The variance due to uncertainty from the unobserved θ is given in Equation (2.6).

$$B_M = \sum_{m=1}^{M} (\hat{t}_m - t^*)^2 / (M - 1) \tag{2.6}$$

The estimate of the total variance of t^* is specified in Equation (2.7).

$$V = U^* + \left(1 + \frac{1}{M}\right) B_M \tag{2.7}$$

At the time Mislevy, Johnson, and Muraki published their report on NAEP scaling procedures, $U*$ was approximated by the jackknife variance of only the first

set U_1 due to the computational intensity. However, now the publicly available AM Beta software can handle the full set of 62 replicate weights.

The statistic $(t^* - T)/V^{1/2}$ is approximately t distributed with v degrees of freedom, where v is given in Equation (2.8), where $f_M = (1 + M^{-1})B_m/V$ is the proportion of the total variance due to the missingness of θ.

$$v = \frac{1}{\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d}} \tag{2.8}$$

2.4 Using propensity scores to select a demographically representative control group

In the study of oral consequences of methamphetamine use, recruitment efforts focused on obtaining a local sample of methamphetamine users, so that a local control group was not measured. It would be impossible, as well as unethical, to randomize participants to use methamphetamine or not in order to estimate the effect of methamphetamine on oral health outcomes. The study was designed, however, to take many of the same measurements as the National Health and Nutrition Examination Study (NHANES) such that data from the general United States population could be utilized as a control sample. The study participants from the full NHANES database are demographically different than the group of Los Angeles methamphetamine users. If one were to do a comparison of outcomes between the full samples from both studies, the "treatment effect" of methamphetamine could not be isolated due to the presence of many potentially confounding variables. Thus propensity score matching is used to select a sample of individuals from the NHANES subject pool who are demographically similar to the MA users.

Table 2.1 displays the proportion of each sample of eligible subjects observed to have various levels of several demographic characteristics. The MA users are

more likely to be male, born in the United States, African-American, single, and to smoke cigarettes than the subjects sampled from the general population. For the purposes of comparison on oral health outcomes, the eligible pool of NHANES was narrowed by requiring that subjects have both the full caries examination as well as a completed periodontal examination. Subjects fail to complete the periodontal examination for many reasons, but the most common reasons are lack of any teeth and severe pain.

Propensity scores are defined as an estimate of the probability of unit being assigned to a treatment given a set of covariates [RR83]. In practice, propensity scores can be estimated through the use of logistic regression, and a unit i's estimated propensity score is simply the resulting $\hat{p}_i$ estimated from the fitted model. In order to estimate the average treatment effect without bias from confouding covariates, subjects from the MA and NHANES can be matched on their values of propensity scores [RR85], or the assessment can build on subclassification into groups based on the ranking of propensity scores [RR84].

For this analysis, the logistic regression model predicting the probability of being from the MA sample included the following variables: age, gender, ethnicity (white, African-American, non-white Hispanic, or other race), education (No high school diploma, high school graduate or GED, some college or associates degree, or bachelors degree or higher), marital status (married/living as married versus never married, divorced, separated), and cigarette smoking status (non-smoker, former smoker, or current smoker). The model also included all two way interactions for which sufficient groupings existed in the MA sample. While smoking status could be viewed as a post-treatment concomitant variable (we have no way of knowing whether cigarette smoking started before of after first MA use), it is here used as a proxy for any other potentially confounding demographic characteristics not measured by the study.

Once propensity scores were estimated using the logistic regression model, the

an appropriate number of subgroups and number of NHANES subjects to include was investigated. The goal is to classify study participants into subgroups in which the covariates are balanced based on the values of the propensity scores. For increased power, we would like to maximize the number of subjects included from the NHANES study. First, observations which were lower than the two minimum propensity scores from the MA sample and higher than the two largest propensity scores from the MA sample were discarded. This was done to ensure that all selected observations from the NHANES study fall within the reasonable range of propensity scores of the MA sample. Next, matching procedures were conducted using the Matchit package in R, pairing m NHANES subjects to each subject from the MA study (without replacement) for $m = 1, ...5$. Figure 2.2 displays the distribution of propensity scores used for each matching ratio. For each matched set and the set of all eligible NHANES subjects, all observations were ranked by propensity score and assigned to groups based on the quantile of the propensity score distributions for each of 4, 5 or 6 groups. Finally, balance between the covariates was checked.

For the purposes of this analysis, we defined balance using a regression model (logistic or linear) predicting the demographic characteristic by the categorized propensity score group and an indicator of being from the MA study. If the indicator for MA was a significant predictor of the characteristic after adjusting for the group, the covariate was determined to be out of balance. Table 2.2 displays the number of covariates out of balance for each matching ratio and the number of groups.

Because the goal is to maximize the number of NHANES subjects included in the analysis while still maintaining balance of covariates within each propensity score grouping, the optimal matching ratio appears to be five NHANES subjects to each MA subject. Five quantiles of propensity scores appeared to provide sufficient number of groups under this matching ratio. Table 2.3 contains the number

of subjects from each of the NHANES and MA studies within each propensity score group.

The procedure described here was also applied to select a cohort from the NHANES 2011-2012 sample with distributions of background characteristics similar to the MA users. During that study year, the protocol for the periodontal examination consisted of full-mouth examinations with 6 sites per tooth, allowing for comparable definitions for periodontal disease between the two samples. The propensity-score-matching procedure yielded a sample of 1090 NHANES 2011-2012 subjects and four propensity-score subgroups.

2.5 Overview of the Dissertation

The rest of the dissertation is organized as follows. An improved method for analyzing the complex large-scale educational assessment data is presented in Chapter 3. Characterization of the underestimation of periodontal disease when using partial mouth data is presented in Chapter 4. In addition, Chapter 4 introduces multiple imputation from hierarchical normal models as a potential solution for underestimation when data are missing by design. Chapters 5 and 6 introduce item response theory models for oral health data, modeling periodontal disease and caries disease respectively. Ideas for the future extensions of the research are listed in Chapter 7.

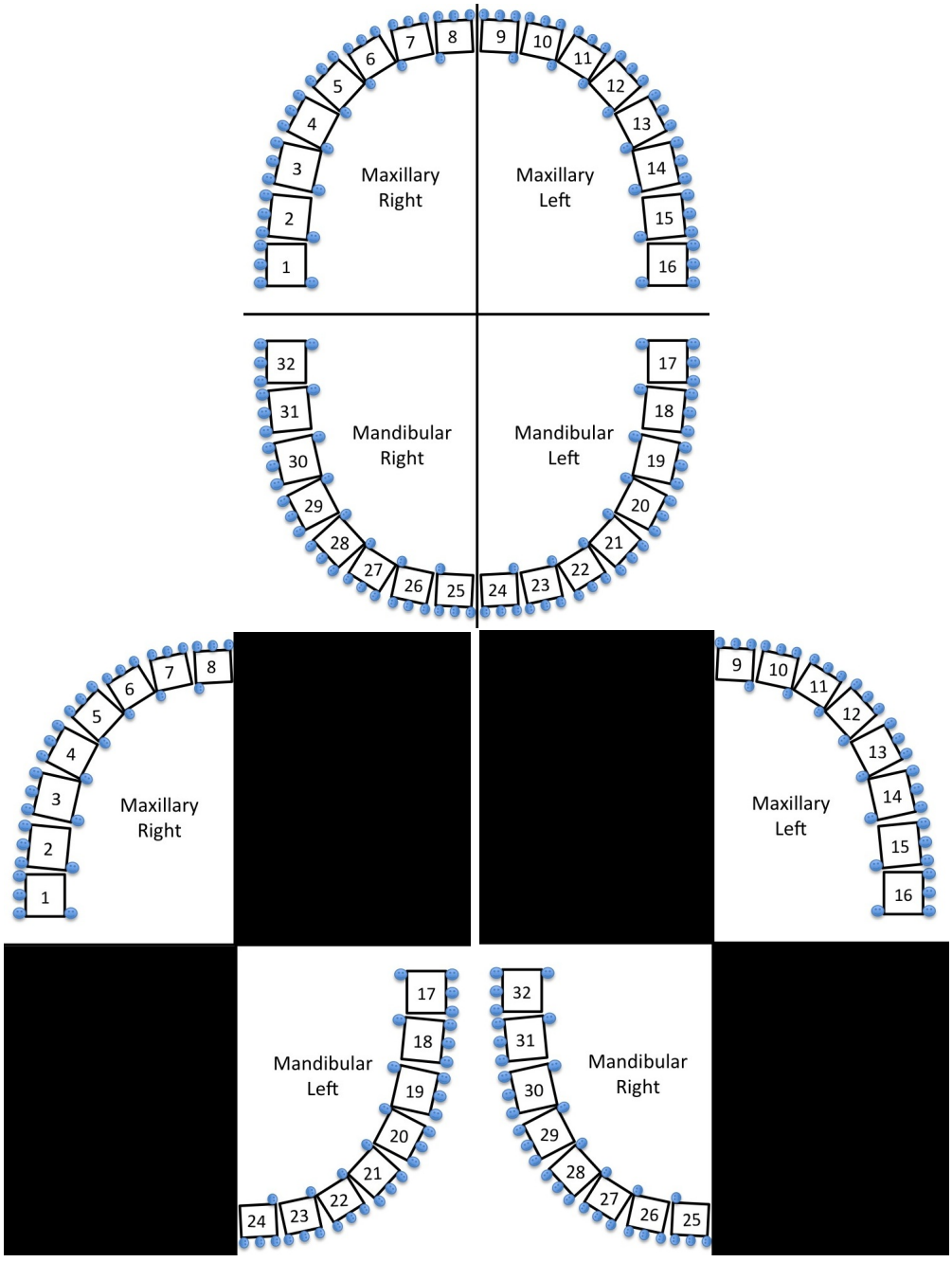

Figure 2.1: The three sampling mechanisms for the periodontal examination

	Meth Users n=551	NHANES n=9327
	Yes	Yes
Male	**446 (80.9%)**	**4472 (47.9%)**
Born in the US	**464 (84.2%)**	**6975 (74.8%)**
Born in Mexico	47 (8.5%)	1414 (15.2%)
Born outside of US or Mexico	40 (7.3%)	936 (10.0%)
White	**103 (18.7%)**	**4200 (45.0%)**
Black/African American	**234 (42.5%)**	**1914 (20.5%)**
Non-white Hispanic	176 (31.9%)	2848 (30.5%)
Other Race	38 (6.9%)	365 (3.91%)
Graduated High School or GED	391 (71.0%)	8559 (65.1%)
Some college/associates degree	155 (28.1%)	2604 (27.9%)
Bachelors degree or higher	**39 (7.1%)**	**1852 (19.9%)**
Married or living as married	**39 (7.1%)**	**7650 (60.4%)**
Former smoker	**53 (9.6%)**	**2824 (23.0%)**
Current smoker	**377 (68.4%)**	**3067 (25.0%)**

Table 2.1: Demographic Characteristics of MA and NHANES samples with both the caries and periodontal exams completed (Boldface font reflects $p < 0.05$ on χ^2 test of independence)

	N Groups = 4	N Groups = 5	N Groups = 6
1:1 Matching	0	0	0
2:1 Matching	1	1	0
3:1 Matching	2	1	1
4:1 Matching	0	0	0
5:1 Matching	3	0	0
All NHANES within P-score range	5	6	5

Table 2.2: Number of covariates out of balance after adjusting for propensity score subgroup by matching ration and number of subgroups

Propensity Score Range	Group 1 0.001-0.027	Group 2 0.027-0.054	Group 3 0.054-0.114	Group 4 0.114-0.268	Group 5 0.268-0.971	Total
NHANES	630	632	628	532	333	2755
Meth	31	29	34	129	328	551
Total	661	661	662	661	661	3306

Table 2.3: Resulting number of subjects in each subgroup

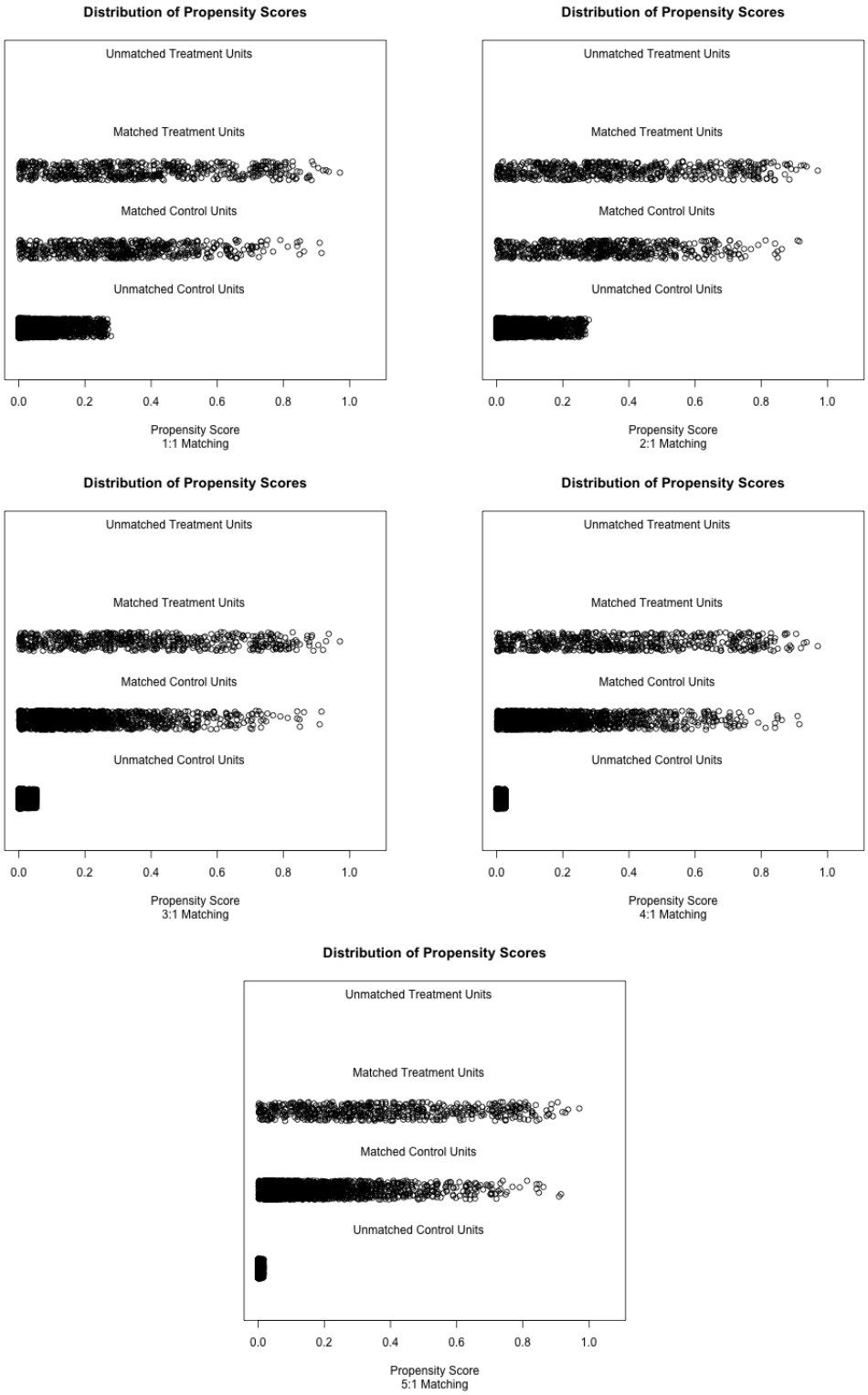

Figure 2.2: Distributions of propensity scores used for each matching ratio

CHAPTER 3

Multidimensional Plausible Value Imputation via the Metropolis-Hastings Robbins-Monro Algorithm

The National Assessment of Educational Progress (NAEP) and similar large-scale educational assessments currently use a multistage estimation process which calibrates test items separate from the latent regression on to background questionnaire responses. In this paper, we present an application of the Metropolis-Hastings Robbins-Monro algorithm to not only extend the item calibration to complex, multidimensional models but also allow for the conditioning regression of background characteristics to be included in the item calibration process. The resulting MCMC chain and estimated posterior distributions can be used to draw plausible values for the latent traits. A simulation study is conducted to demonstrate the method when data are generated under a two-tier model for the item response data with latent characteristics. The data generating model is compared to alternatives, including the simple three primary domain model calibrated in practice. The method is applied to the 2011 NAEP Science assessment through calibrating the data generating model assumed by the framework.

3.1 Background

The purpose of this research is to introduce a method to calibrate item parameters for complex multidimensional models simultaneously with performing a latent re-

gression of the ability parameters onto the matrix of background characteristics, and the extensions of this method to large-scale educational assessments such as the National Assessment of Educational Progress (NAEP). The goal is to demonstrate the use of these models for flexible recalibration of NAEP items as well as evaluating the applicability of multidimensional IRT models to large-scale educational assessments with planned-missing-data designs for the cognitive items.

3.1.1 Motivating Example: 2011 NAEP Science Assessment

We use the 2011 National Assessment of Educational Progress (NAEP) framework as a motivating example for our research. Three primary content domains are assessed (Life Sciences, Earth and Space Sciences, and Physical Sciences), and it is on these three areas that subscale scores are produced. In the 2011 Science Assessment, items are also calibrated onto a general science domain (separately from the calibration onto each content domain) [NAE12]. In the development of this assessment, guiding framework for test items specified that test item involve the use of one of the four following scientific practices:

1. Identifying science principles,
2. Using science principles,
3. Using scientific inquiry, and
4. Using technological design.

In addition, each item is designed to measure one of four cognitive demands:

1. Declarative knowledge ("Knowing that")
2. Schematic knowledge ("Knowing why")

3. Procedural knowledge ("Knowing How")

4. Strategic knowledge ("Knowing when and where to apply knowledge") .

Thus, if we follow the measurement model that is assumed by the 2011 NAEP Science Framework, the test items are assumed to correspond to one content domain, one scientific practice domain, and one cognitive demand. The potential data-generating model that can approximate this student response process is actually fairly complex. In operational practice, items are calibrated onto only one content domain in essentially a unidimensional IRT model, as well as all together in a unidimensional science scale, and plausible values are generated for only these content domains and a general science domain. The framework implies a considerably more complex data generating model than the IRT models calibrated in practice.

3.1.2 Plausible Value Methodology

Plausible value methodology has long been the gold standard in large-scale educational assessments. Plausible values were introduced as a device to better approximate the population distribution of abilities, treating the values of the student ability parameters as missing data. Rather than a single point estimate of student-level proficiency, however, multiple estimates are "imputed" from a posterior distribution that includes the student's item responses and background characteristics [Mis91, MJM92b].

In traditional educational assessment, interest focuses on the ability of the student or respondent in a given subject area. Ability, however, cannot be measured directly, and therefore can be considered a latent, or unobserved, variable. Mislevy, Johnson and Muraki [MJM92b] described in detail the analysis methods used in NAEP since 1984, which implement the balanced incomplete block design approach to assess proficiency. By treating the student scores as missing data,

population quantities and variances are adjusted to reflect the limited amount of information on the individual students.

3.1.3 Characterizing the NAEP framework

To characterize the NAEP framework, we start with some notation. Let $X = (x_1, ..., x_N)$ be the matrix of demographic covariates, where x_i is a vector of demographic characteristics for individual i. Let $y_i = (y_{i1}, ..., y_{iD})$ be the individual outcomes on the D proficiency domains, comprised of item responses y_{id} with typical element y_{id}^l denoting item l within domain d. Of interest is estimating the unobserved proficiency in D domains, and the vector of these latent variables for student i is represented by θ_i=$(\theta_{i1}, ..., \theta_{iD})$. In this framework, students are considered independent of one another, and responses to different questions by an individual student are also assumed independent. It should be noted that the assumption of independence between students is not made in producing the plausible values from the operational framework. In fact, the students are sampled within schools, and schools are sampled from primary sampling units (be it states, regions, or districts). The analysis using plausible values incorporates jack-knife variance estimation to estimate the variability at the primary cluster level, but the item parameter estimation of the cognitive test items do not fully adjust the complex sampling design. As will be discussed in further detail, this framework does not allow or consider conditional dependence between items across different domains, and item responses are theorized to only depend on the proficiency domain of which it was designed to measure [TG97].

While the underlying conditional independence assumptions may be untenable, it does simplify the construction of the probability distribution for student i's responses y_i, conditional on his or her proficiency vector θ_i to $\prod_{d=1}^{D}[f_d(y_{id}|\theta_{id})]$. The probability model of an observed response pattern for content domain d for stu-

dent i depends on the scoring or response format of the question. Dichotomously scored items are often represented using the classical two parameter logistic model or three parameter logistic model. For multiple-categorical items, more complex IRT models such as the graded response model [Sam69] or the generalized partial credit model [Mur92] may be used.

3.1.4 Likelihood functions

Let β_d represent the vector of all the item parameters that relate to the proficiency domain d such that $\beta = (\beta_1, \ldots, \beta_d, \ldots, \beta_D)$. The latent proficiency vector θ_i, is assumed to be conditionally normally distributed with mean vector Γx_i and covariance matrix Σ, where Γ represents a matrix of unknown regression parameters [TG97]. Let $\phi(\theta_i; \Gamma x_i, \Sigma)$ represent the conditional normal density function of the latent proficiency vector. The (complete data) likelihood function of the parameters β, Γ, and Σ is constructed as

$$L(\beta, \Gamma, \Sigma | \theta, Y, X) \propto \prod_{i=1}^{N} \phi(\theta_i; \Gamma x_i, \Sigma) \prod_{d=1}^{D} f_d(y_{id} | \theta_{id}; \beta_d), \tag{3.1}$$

where $Y = (y_1, ..., y_N)$ and $\theta = (\theta_1, ..., \theta_N)$. The complete data likelihood is proportional to the posterior distribution of the latent proficiency vector $f(\theta | Y, X, \beta, \Gamma, \Sigma)$. By contrast, the observed data likelihood function can be represented by

$$L(\beta, \Gamma, \Sigma | Y, X) = \prod_{i=1}^{N} \left[\int \prod_{j=1}^{n} f(y_{ij} | \theta, \beta) \Phi(d\theta | \Gamma x_i, \Sigma) \right]. \tag{3.2}$$

The observed data likelihood (Equation 3.2) requires the integration of the product of the item response function and prior across items over the latent distribution of θ. On the other hand, the complete data likelihood function (Equation 3.1 affords considerable simplifications. It can be separated into the product of two products:

one product of all item response functions for subject i and one product of all of the priors for each subject. The numerical integrations required can make the observed data likelihood burdensome to compute, let alone optimize, under more complex multidimensional models.

3.1.5 Approximating the posterior distribution and drawing plausible values

The posterior distribution is used to draw the plausible values in four major steps. First, the item parameters (β_d) are estimated separately for each proficiency domain. Note that this does not involve simultaneous estimation of the item parameters for the multivariate proficiency vector θ_i. Second, treating the estimated item parameters as if they are known, Γ and Σ are estimated using the observed item response data in an Expectation-Maximization (EM) algorithm [DLR77, BA81] by fixing the item parameters to their estimates from step one. In step three, the posterior distribution of β, Γ, and Σ is then approximated from these regression parameter and error covariance matrix estimates. And finally, for each of the generated regression parameter sets and covariance matrices, a value of θ_i for each student is drawn from a normal distribution with mean Γx_i and covariance matrix Σ. This final step of generating imputations is generally repeated 5 times to produce 5 plausible values per student [TG97].

3.1.6 Potential drawbacks of existing methodology

The existing plausible values frameworks produce data sets that are relatively user friendly for secondary analysis, but generation of the plausible values is tedious and computationally intensive. The current framework defines the proficiency domains a priori, and fails to account for any conditional dependence between

items on different proficiency exams. The current formulation does not allow for investigators to explore alternative parameterizations of the latent proficiency space, such as data generating model of the NAEP Science Framework, for example. The plausible values estimates in this framework are potentially biased from not accounting for local dependence and other confounding effects [Yen84]. For instance, there may be several questions (forming a testlet) following the same reading passage, and a student's responses to those questions are routinely more related due to the effect of the common stimulus above and beyond what can be attributed to reading proficiency [Cai10, CSH11]. However the current models do not control for testlet-level dependence. The multilevel nature of the data, such as the variability within a school within a district, is another source of potential bias in the imputation.

Additional bias is introduced when the plausible values are used as predictors in a model. A recent article showed that inference in which latent domains are independent predictors can be biased if the set of background covariates does not contain the outcome of interest [SJT14]. Thomas [Tho00] highlighted the sensitivity of the imputation model to assumptions such as homogeneous variance across subpopulations and noted that computational intensity was a limiting factor in expanding the existing models. The proposed model could resolve the potential biases existing in the current plausible value imputation methods by accounting for the multilevel data structure, a multivariate domain framework, sub-population level variability, as well as nuisance local dependence.

Recent research has been conducted indicating the feasibility of certain multidimensional models in large scale assessments, such as the bifactor model, where there is one general domain and several subdomains [RJD14a], or a simple structure where all the subdomains are estimated simultaneously and allowed to covary [SD05]. Trifactor models [BHB13, RJD14b] have also been proposed for account-

ing for residual correlations within subdomains, particularly for large-scale assessments. It should be noted that in Rijmen, et al., the method applied quadrature-based estimation. A stochastic approximation to the latent regression parameters was investigated by von Davier et al.[DS10], the results compared were to the existing software for estimation used in NAEP. The resulting software, SGROUP, solved some of the problems with the posterior standard deviations. While this approach was applied using simple multidimensional IRT models, such as simultaneous estimation of mathematics subscales, it was not applied to more complex, structured MIRT models. Nor has the impact on plausible value imputation been explored.

There have been other methods proposed for analyzing data from the NAEP framework that also include simultaneous calibration and regression. Scott and Ip [SI02] introduced an MCMC procedure to estimate both the regression and IRT parameters simultaneously while introducing random effects for item clusters, which improved estimates of the standard errors for subgroup means. Cohen and Jiang [CJ99] observed that the assumptions about the population distribution in the calibration of the measurement model are different than the assumptions in the estimation of population of characteristics. Their method interprets the observed categorical item responses as categorizations of some continuous, unobserved normal variable and uses a Monte Carlo EM variant for estimation. Similar to other past work, neither of these methods address the issue of multidimensionality within the cognitive assessment.

3.2 Methods

3.2.1 Metropolis-Hastings Robbins-Monro Algorithm

The Metropolis-Hastings Robbins-Monro Algorithm was introduced by [Cai08] to combat the issue of dimensionality that has made multidimensional IRT unfeasible

for realistically complex assessment situations. The MH-RM is a data augmented Robbins-Monro type stochastic approximation algorithm driven by random imputations produced by a Metropolis-Hastings sampler. It can be seen as an extension of the Stochastic Approximation EM algorithm. The guiding insight is that the practice of maximum marginal likelihood estimation in latent variable modeling is similar to the engineering application of the Robbins-Monro method for the identification and control of a dynamical system with observational noise. Finding the MLE amounts to finding the root of the likelihood equations, but because of missing data (latent variables), the marginal log-likelihood itself is difficult to evaluate directly. In contrast, the complete data log-likelihood takes a much simpler form. Therefore, we use Monte Carlo methods to impute just enough missing data so that the complete data log-likelihood can be optimized easily. Due to the purposefully injected Monte Carlo error, the ascent directions will be noisy. This is where the Robbins-Monro method plays an important role. It filters out the noise so that item parameter estimates converge with probability 1 to the MLE.

Cycle $j+1$ of the MH-RM algorithm for multidimensional IRT with covariates consists of the following three steps:

1. **Imputation.** The complete data are formed by imputing values for $\theta^{(t+1)}$. Using the latest iteration of estimates of item parameters, $\beta^{(t)}$, and latent regression parameters, $\Gamma^{(t)}$, from the previous cycle t, random samples of the individual latent traits $\theta^{(t+1)}$ are imputed using the Metropolis-Hastings sampler from a Markov chain having the posterior of the individual latent traits $\pi(\theta|Y, X, \beta^{(t)}, \Gamma^{(t)}, \Sigma^{(t)})$ as the unique invariant distribution. We then have the complete data formed as $(\theta^{(t+1)}, Y, X)$.

2. **Approximation.** Based on the imputed data, the complete data log-likelihood and its derivatives are evaluated so that the ascent directions for the item and latent density parameters can be determined later. For instance, the complete data score function for the item parameters is approxi-

mated as $s_{t+1} = dL(\beta^{(t)}|\theta^{(t+1)}, Y, X)/d\beta$, and the complete data information matrix for the item parameters is $H_{t+1} = -(d^2L(\beta^{(t)}|\theta^{(t+1)}, Y, X))/(d\beta d\beta')$.

3. **Robbins-Monro Update.** Robbins-Monro stochastic approximation filters are applied when updating the estimates of item and latent density parameters. Let ϵ_t be a vector of non-negative gain constants such that $\epsilon_t \in (0, 1]$ and $\sum_{t=0}^{\infty} \epsilon_t = \infty$, $\sum_{t=0}^{\infty} \epsilon_t^2 < \infty$. The Robbins-Monro filter will be applied to obtain a recursive stochastic approximation of the conditional expectation of the complete data information matrix: $I_{t+1} = I_t + \epsilon_t(H_{t+1} - I_t)$. The Robbins-Monro filter is applied again when updating the new parameter estimates: $\beta^{(t+1)} = \beta^{(t)} + \epsilon_t(I_{t+1})^{-1}s_{t+1}$. The iterations are started from some initial values $\beta^{(0)}$ and terminated when the estimates stabilize. Cai [Cai08] showed that the sequence of parameters converges with probability 1 to a local maximum of the observed data likelihood $L(\beta, \Gamma, \Sigma|Y, X)$.

3.2.2 Two Implementation Notes

In practical terms, implementation of the MH-RM algorithm for the model described here requires two sets of analytical results. First, the derivatives of the complete data likelihood with respect to the item parameters and the latent regression parameters must be derived. Fortunately, these are standard statistical results. For the item parameters, Cai [Cai08] provided analytical derivatives for major item response models useful for large-scale assessments. For the regression part, upon "observing" the imputed θ scores, the latent regression model becomes a multivariate linear regression, whose derivatives are straightforward results in standard multivariate analysis texts, e.g., Mardia et al.'s classic book [MKB79].

Second, one must be able to impute the θ values in order to run the MH-RM iterations. To do this, a random walk Metropolis algorithm is used. For each student, let θ_i^c be the current value of θ. Let $\theta_i^p = \theta_i^c + e$ be a proposal value,

where e is an independent draw from a multivariate normal with zero means and covariance matrix equal to a scalar multiple of the identity matrix. Acceptance or rejection of the proposal is determined by evaluating and comparing the complete data likelihood function in Equation (3.1) at the proposal and current values. The proposal is accepted if the likelihood ratio of proposal vs. current values exceeds 1, or if accepted by rejection sampling with the said likelihood ratio as the acceptance ratio. The Metropolis sampler should be tuned on a case-by-case basis to achieve optimal acceptance rates for high quality sampling of the posterior.

3.3 Simulation Study

The simulation study was designed to generate science assessment data under complex model assumptions. The structure applied most closely resembles a two-tier model [Cai10], in which items correspond to one of three general domains (content), which are correlated, and one subdomain (science practice). Within a content domain, the model resembles a bifactor model, where the subdomains are independent of one another after adjusting for the general domain. The structure for the generated data can be seen in Table 3.1. In total, parameters were generated for 150 possible dichotomous items following a 2PL model for consistency of estimation. Location parameters were generated from a standard N(0,1) distribution, while slope parameters for the primary content domains were drawn from a truncated normal(1.5,0.3) with a minimum slope of 0.4. The slope parameters for the secondary practice domains were drawn from a truncated N(0.8,0.2) with a minimum slope of 0.4. Thirty regression parameters for each of the primary domains were generated from N(0,2) with a correlation between domains of 0.7. All latent domains had the same variance, 1, but only the primary domains were allowed to correlate at 0.8, reflecting the correlations between subdomains found in practice.

Parameter (prm) files were generated in R and read into flexMIRT 2.0 [HC13]. From there, a dataset of 9000 was generated using these parameters. This dataset was read back into R to impose a balanced incomplete block design in which individual responses to all but two blocks of 15 questions were set to missing. This procedure produces data that is structurally similar to a NAEP Science or Mathematics Assessment. Finally, item calibrations were conducted on the resulting data under the following different models:

1. the data generating model, with three content domains (correlated), three practice domains (uncorrelated), and covariate on the full set of data with no missing responses as a reference
2. the data generating model, with three content domains (correlated), three practice domains (uncorrelated), and covariates
3. the data generating model, with three content domains (correlated), three practice domains (uncorrelated), and covariates with beta priors on all items
4. three content domains (correlated), three practice domains (uncorrelated), and no covariates
5. three content domains (correlated) with covariates, and
6. three content domains (correlated) with no covariates.

3.3.1 Evaluation of Models

Most of the proposed models can be shown to be nested within each other. For these models, we can use a likelihood ratio test to evaluate improved fit when freeing additional parameters. Additionally, we can compare the *expected a posteriori, EAP,* scores between the data generating model and any possible alternative models to gain a general sense of how scores may be biased under each model specification.

3.3.2 Results

3.3.2.1 IRT Parameter Recovery

The overall mean relative bias (bias/true value) of the content (primary) slope and location parameters as well as the average estimate of the coefficient for variation (CV) for each fitted model are presented in Table 3.4. We use the estimate of the coefficient of variation (RMSE/$|\lambda|$) for each item to standardize the estimate of the error for averages across all items to be compared. The distributions of bias for the slope and content parameters can be viewed in Figures 3.3 and 3.3 respectively.

Not surprisingly, the model which produced the least amount of bias in the content slope parameters was the two-tier data-generating model with covariates as calibrated on the full set of data rather than on the balanced-incomplete-block data (Model 1). However, of all the models calibrated on the BIB data, the model in which only the three content domains were estimated with covariates produced the least amount of bias in the slope parameters (Model 5). The average discrimination parameter bias for this model was slightly lower than the two-tier (data-generating) model with covariates. Due to the sparseness of the data (and the small sample size relative to what would be observed in practice), there is likely not enough information to compute the two-tier model with lower bias, and the lack of information is likely inflating the discrimination parameters.

The model with the greatest amount of bias in the slope parameters, on average, is the two-tier model without covariates incorporated (Model 4). The second poorest-performing model appears to be the model that is closest to the one estimated in practice: three content domains with no covariates incorporated (Model 6). The discrimination parameters in this model have the largest amount of bias relative as compared with any of the other models. When the covariates are not

included in the model, the additional information that the covariates provide is no longer incorporated. It should be noted that the inflation of the slope parameters for all models would likely be reduced with a greater number of observations. In this simulation, there were only 9000 observations per simulation. In practice, the test items are calibrated on over 100,000 study participants, increasing the amount of information about the response patterns.

The slope parameters, in general, have very little bias regardless of the model used. Model 1, the two-tier model with covariates on the full set of data, has very little bias with the exception of one outlier. The average bias of the location parameters is positive for the two-tier models on the BIB data (Models 2, 3, & 4), but the average bias is slightly negative for the location parameters in the one-tier model without covariates (Model 6). The one-tier model with covariates produces zero bias, on average.

3.3.2.2 Regression and Correlation Parameter Recovery

The average relative bias of the regression parameters is close to 50% for all models considered, including the two-tier model calibrated on the full set of data (see Table 3.4). Regression parameter estimates were not computed for models 4 and 6, because covariates were not included in the item calibration. Underestimation in the regression parameters is not unexpected at this stage. For calibration using MH-RM, only single imputation of θ is used at each iteration. However, any analysis of the data will be combining the information across multiple imputations from the posterior distribution.

The correlation between each of the content domains was 0.7 under the gen-

erating model. All 6 of the models considered allowed for correlation between the content domains during calibration, and thus the correlations estimated can be checked for bias (Table 3.5). All of the models produced little bias. Surprisingly, the highest degree bias of the correlations was found in the full data two-tier model (Model 1). The correlations between the first and second domains were slightly underestimated in all of the models on the BIB data (Models 2-6). The bias of the estimates of correlation between domains 1 and 3 were very small, ranging from 0.001 to 0.063 in the BIB data models, with the lowest biases being found in the models with covariates (2, 3, and 4), and a similar pattern was seen with the correlation estimates between dimensions 2 and 3.

3.3.2.3 Model Fit

All of the models are essentially various constrained versions of the two-tier model with covariates. We can compare the fit of Models 4, 5, and 6 to Model 2 through the likelihood ratio test. Table 3.3 displays the average -2 log likelihood, AIC, and BIC for each of these models across each simulation. On average, the two-tier data-generating model with covariates (Model 2) had the lowest values of each fit criteria. When examining these values by criteria, there are several instances in which the one-tier model with covariates (Model 5) produces lower values of -2log likelihood than Model 2. Given the sparse nature of the data and the limited information about each pattern, it is possible for the model with fewer dimensions to fit the data better. The poorest fitting model is still the one-tier model with no covariates, but the one-tier model with covariates has a much closer deviance to the two-tier model with covariates than any other model.

3.3.2.4 Comparing *Expected A Posteriori* Scores Between Models

In large-scale assessments with balanced-incomplete-block or other sparse test designs, single point estimates are not used for individual students. However, we can examine the *expected a posteriori* (or EAP) scores produced under each model to examine the consistency under each model. The EAP scores for each of the three primary content domains estimated from the first simulated dataset under Models 2, 4, 5, and 6 are plotted Figures 3.5-3.7. It can be seen that the EAP scores between Model 2 (the two-tier model with covariates) and Model 4 (the one-tier model with covariates) are the most similar ($\rho = 0.99$, $R^2 = 0.97$). The relationships between all other models follow a similar pattern: as *expected a posteriori* estimates move away from the mean, the relationship between the estimates from different models becomes weaker. There is greater dispersion towards the tail ends of the distribution of θ.

3.3.2.5 Number of imputations under complex models

The simulated datasets and estimated model parameters can be used examine the number of imputations that may be necessary to estimate the relationship between a content subscale and one of the background characteristics under complex models. While the operational framework for NAEP has switched from using 5 plausible value imputations to 20 for the most recent examinations, it is prudent to examine if the number of imputations should be increased under more complex latent structures.

Using the first simulated BIB dataset and the parameter estimates generated under the two-tier model with covariates (Model 2) as an example, the following number of plausible values were generated for the data: 2, 3, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, and 250. For each imputation,

a regression was performed of each of the three content subscales onto the first covariate, x1. The results were combined and analyzed across imputations using proc mianalyze in SAS. The estimates of the fraction of missing information for each number of imputations are plotted in Figure 3.8. The estimate of the fraction of missing information can be used as a monitoring tool to determine convergence based on the number of imputations; the number of imputations necessary can be determined when the estimate of the fraction of missing information stabilizes. For each of the content domains, the fraction of missing information estimates start to stabilize after just 5 imputations, but continues to fluctuate upwards until 40 imputations. Not surprisingly, the relative efficiency of the imputations approaches 0.99 right after 5 imputations.

3.4 Two-tier Calibration of the 2011 NAEP Science Assessment

We used the item and background questionnaire responses from the 2011 NAEP Science Assessment to demonstrate the application of the methodology in practice. The data was comprised of individual item responses from 124170 eighth-grade students. The 2011 NAEP Science Assessment contained 10 booklets (eleven including one bilingual booklet), of which each student received two for a total of 38 combinations of booklets. The number of items in each booklet ranged from 14 to 18 items.

3.4.1 Calibration of Models

The primary goal of applying the MH-RM estimation procedure to the 2011 NAEP Science assessment is to demonstrate how the method can be used in practice for both simple and complex IRT models. Beyond that, we can test if two-tier models

(as seen in Figure 3.11) assumed by the data-generating framework fit the data better than the operational model without covariates and only content domains. The item response functions chosen for the items in this calibration are the same as specified in the NAEP documentation. The majority of the items are modeled as 3PL, with several 2PL and graded items. Some of the items scored as graded with more than two categories are actually summed combinations of individual test items found to have high residual dependence, which is how the current operational NAEP handles items grouped in testlets with residual dependence. We assumed a Beta(1,4) prior distribution on the guessing parameter for the 3PL models, as most multiple-choice questions had five possible answers.

The number of items corresponding to each domain varied. Forty-four of the 144 questions pertained to Physical Sciences, while 59 and 41 pertained to Earth and Space Sciences and Life Sciences respectively. A plurality of items (59) asked students to identify scientific principles, and 52 questions involved using scientific principles. The use of scientific inquiry was measured in 23 items, and only 5 items pertained to the use of technological design.

For the purposes of demonstration, we only selected a handful of background characteristics to include in the regression model. If this method would be applied in an operational setting, we would want to include some set of principal components or other independent combinations of all necessary background characteristics. In this example, we selected 11 dichotomous variables to serve as the background characteristics in the model. The background variables are all binary indicators of the following traits: female, English Language Learner, classification of having a disability, eligible for free lunch, eligible for reduced-cost lunch, public school (versus private), and self-identified ethnicity (Black/African American, Asian American, Hispanic or Latino, American Indian, or Other).

3.4.2 Results

3.4.2.1 Model Fit

We calibrated six models to the data, which are shown graphically in Figures 3.2 and 3.11:

- Three primary content domains, uncorrelated, no covariates,
- Three primary content domains, uncorrelated, with 11 covariates,
- Three primary content domains, correlated, with 11 covariates,
- Two-Tier with 3 content and 4 practice domains, uncorrelated, with no covariates,
- Two-Tier with 3 content and 4 practice domains, uncorrelated, with 11 covariates on the 3 content domains, and
- Two-Tier with 3 content and 4 practice domains, correlated, with 11 covariates on the 3 content domains.

Some models failed to pass the second-order test (evaluating whether the solution is a maximum), specifically the two-tier model in which the primary domains are correlated as depicted in Figure 3.1. In addition, a one-tier model with three correlated content domains also failed to pass the second-order test. The fit indexes for each of these models is printed in Table 3.6. We can compare the models without the correlated content domains using the likelihood ratio test (we do not compare the models that did not converge to a maximum likelihood solution). The two-tier model with covariates fit the data significantly better (p¡0.00001 for each likelihood ratio test) than each of the competing nested models: the two-tier without covariates, the one-tier without covariates, and the one-tier with covariates. The one-tier model without covariates or correlation between the content

domains is essentially the model calibrated in practice, and the results show that even the same model with covariates provides a better fit than the operational model.

3.4.2.2 Item Parameters

In the simulation study, slope parameter estimates were typically much larger than the generating parameters. However, when complex models are calibrated from the 2011 NAEP Science Assessment, the item parameters do not deviate much from the item parameters calibrated in operation. In Figure 3.12, 3PL item parameters are compared between the two-tier model with and without correlated content domains when covariates are included in the calibration. Assumptions about the orthogonality of the content domains does not appear to impact the location parameters; the estimates of the location parameters are nearly identical between the two models. The estimates of the guessing parameters vary slightly. The slope parameters for the primary contents are consistently higher for the two-tier model in which the content domains are correlated, while the practice domain slopes are consistently higher in the model without correlated primary domains. When the primary content domains are constrained to be independent, it is likely that some additional information that would be captured by the correlations is incorporated into the secondary practice domains.

In Figures 3.13 and 3.14, the 3PL item parameters each of these two-tier models is compared to the operational item parameters given in the NAEP technical documentation. The guessing parameters appear to be consistent regardless of the model used. For both two-tier models, the location parameters estimated are slightly larger than the location parameters calibrated in practice. The differences in these parameters could reflect differences in the latent ability distribution used to calibrate the items. The content slope parameters in the two-tier model with correlated content domains are generally higher than the values calibrated in op-

eration, but this trend does not exist when the content domains are constrained to be orthogonal.

The content slope parameters calibrated from both two-tier models with covariates do not diverge much from the parameters calibrated in the operational setting with no covariates, and all estimated slope parameters values are less than 3. This result is distinct from the results of the simulation study where slope parameters were consistently estimated at values above 3 or 4 for any given model.

3.4.2.3 Evaluating the number of imputations

For each content domain under the two-tier (uncorrelated) model with covariates, 2, 3, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 150, and 200 plausible values were generated. The regression of each content domain onto the dichotomous variable indicating whether a student is female was conducted for each imputation, and the results were combined using PROC MIANALYZE in SAS 9.4.

The relative efficiency of the estimates for each number of imputations is presented in Figure 3.16. For each of the content domains, the relative efficiency (to an infinite number of imputations) reaches close to 99% by 30 imputations. The estimates of the fraction of missing information for the regression slope for female (i.e. the difference between males and females) by the number of imputations are plotted in Figure 3.15. For Physical Sciences as well as Life Sciences, the estimates appear to converge by 100 imputations. However, for Earth and Space Sciences, the estimate of the fraction of missing information decreases until 100 imputations and increases until 200 imputations. More imputations than 200 may be necessary for the convergence for this domain.

3.5 Concluding remarks and extensions

We have demonstrated here a novel method for simultaneously estimating the latent regression parameters while calibrating the cognitive assessment items. This procedure allows for more information to be incorporated from the set of background characteristics while calibrating the items, which may improve the estimates of the item parameters, especially when the amount of cognitive information on individual test-takers is limited.

The results of the simulation study showed that with a limited sample size and balanced-incomplete-block designs, the two-tier model with covariates does not necessarily fit better nor produce less bias than the one-tier model with covariates which only loads items onto the primary cognitive domains. However, in practice with a much larger sample size (as seen in the actual data analysis of the 2011 NAEP Science assessment), the more complex model assumed by the data-generating framework does improve the overall model fit. Based on the simulation study, even with the limited amount of information from a small sample size and BIB design, the models with covariates improve the estimation of the parameters as well as providing a better relative fit than calibrating the IRT models without covariates.

In practice with actual large-scale assessment data, some models were found to have difficulty converging to a maximum-likelihood solution. While some two-tier models were able to be estimated, models correlated domains with covariates failed to pass the second-order test. It is possible that changing the starting values may help, or the issue could be found with the way the complete data information matrix is estimated. Future research is necessary to improve the procedure for these models.

Optimizing the latent regression step is a popular area of research. While we present simple linear models in this paper, this method can be generalized to incorporate other models for the background characteristics. The method as described here assumes that the matrix of background characteristics has no missing values, which would be consistent with using principle components or factor scores based on the background characteristics as the independent variables in the regression. However, incorporating more complex latent regression models or background characteristics with missing observations should be a subject of continuing research.

The potential applications of this method extend beyond generation of plausible values. While we regressed only the content domains onto the background characteristics in this paper, we could also draw plausible values and inference for the scientific practice domains using this methodology. This method provides a means of estimating item parameters with the model containing covariates, and thus can be used to generate the plausible values through traditional methods involving a second latent conditioning step if deemed necessary. This procedure enables the secondary analyst to examine alternative scale formations from large-scale assessments and use different combinations of background characteristics to produce plausible values.

3.6 Figures

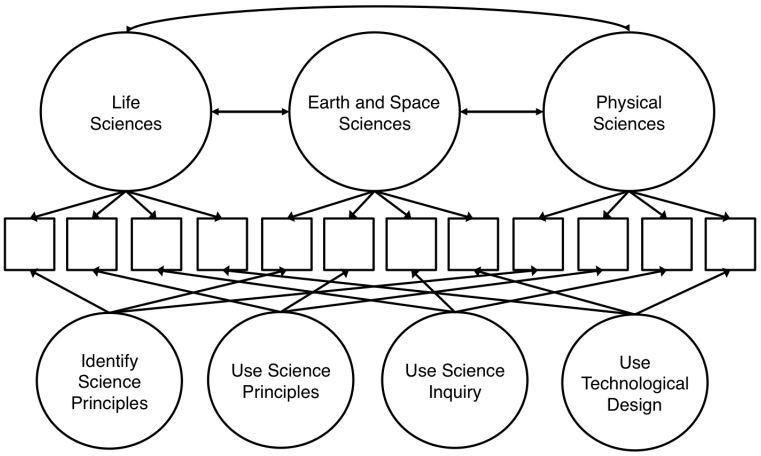

Figure 3.1: Two-Tier Model with Correlated Primary Content Domains

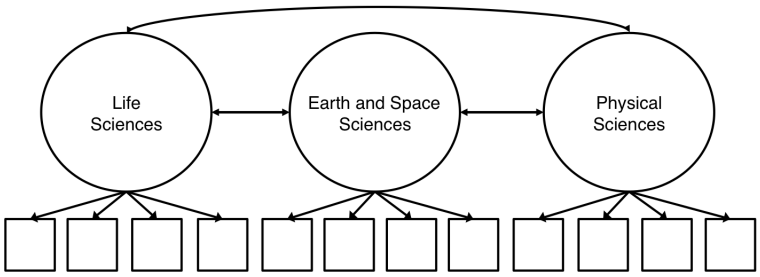

Figure 3.2: Model with correlated content domains (One Tier)

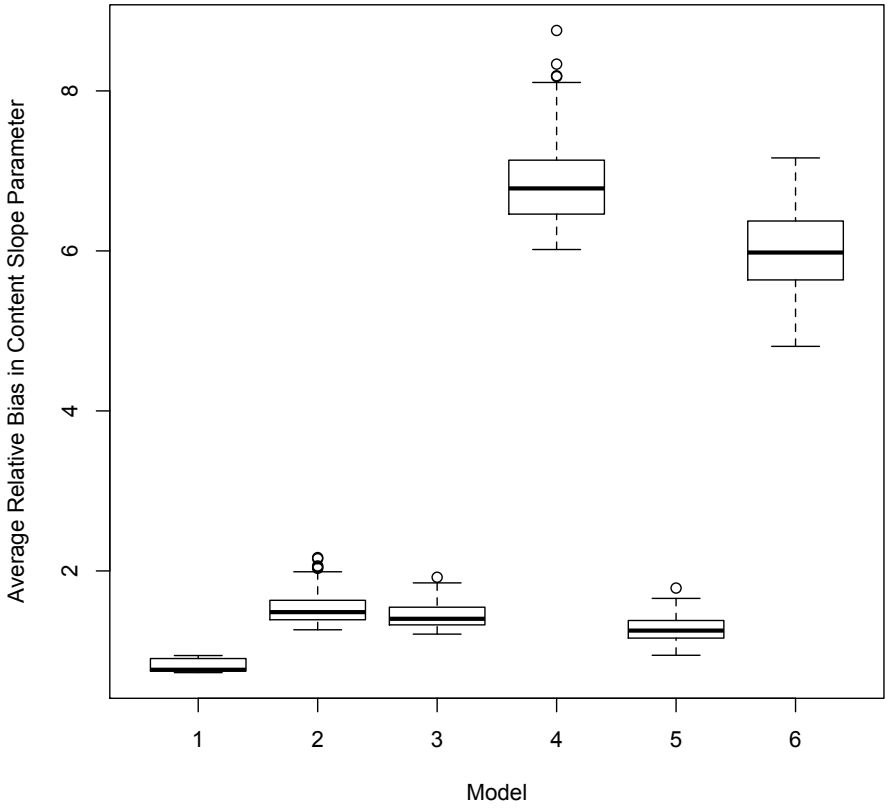

Figure 3.3: Distribution of Average Relative Bias of Content Slopes for 150 Items between Models

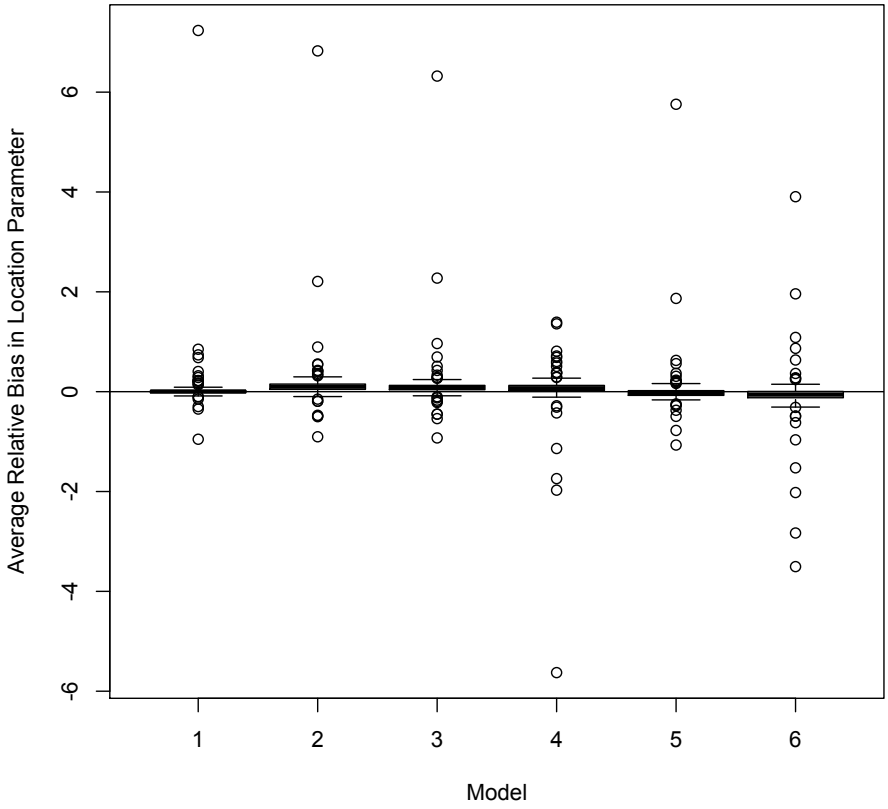

Figure 3.4: Distribution of Average Relative Bias for Location Parameters for 150 Items between Models

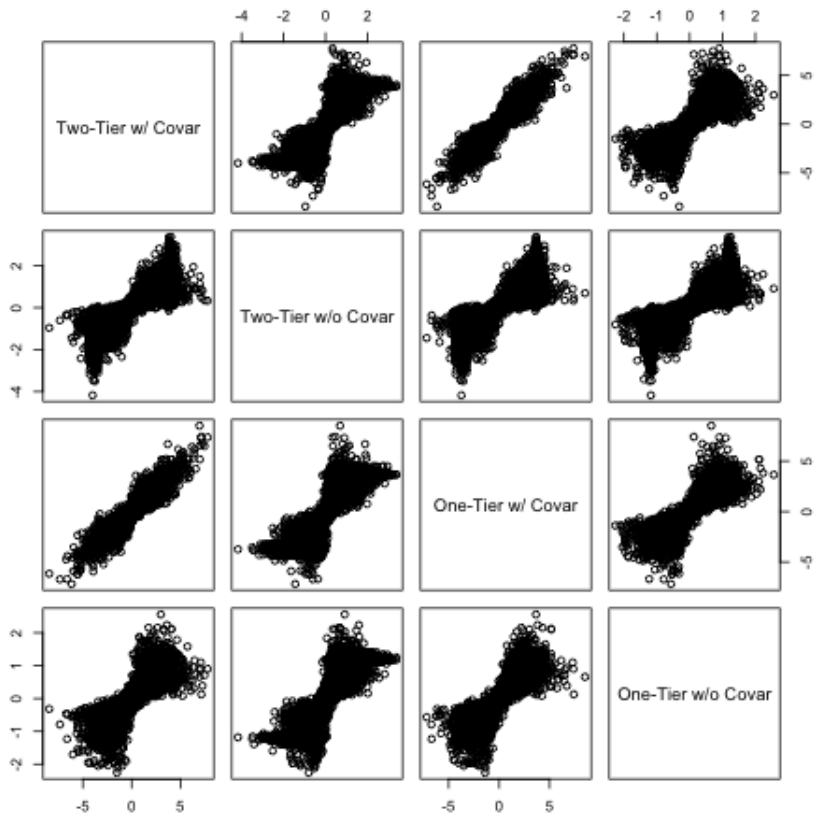

Figure 3.5: Distribution of EAP scores for content domain 1

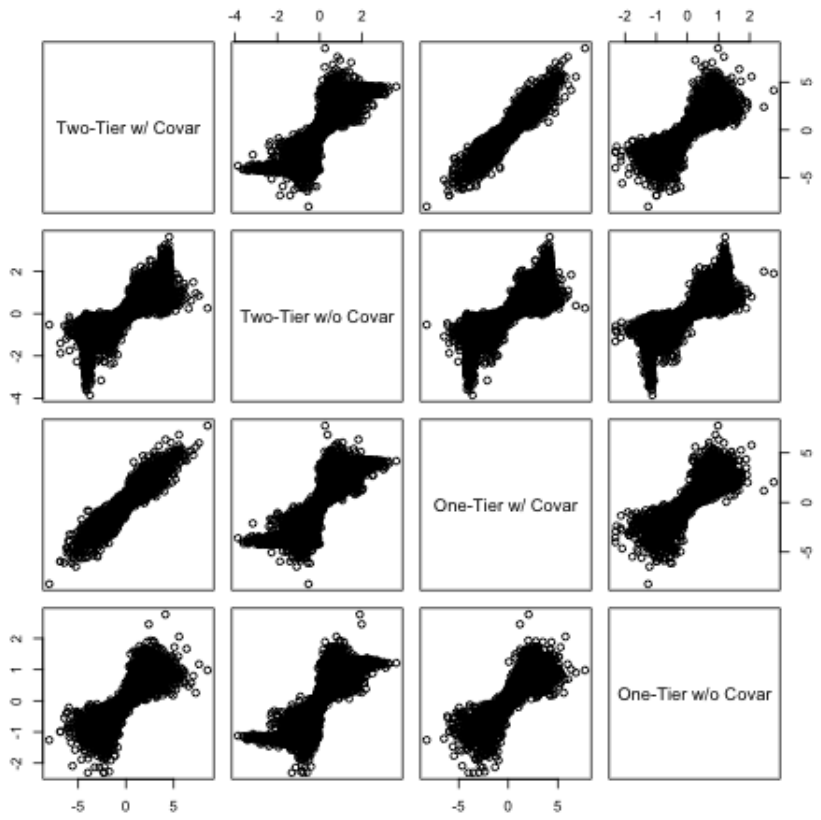

Figure 3.6: Distribution of EAP scores for content domain 2

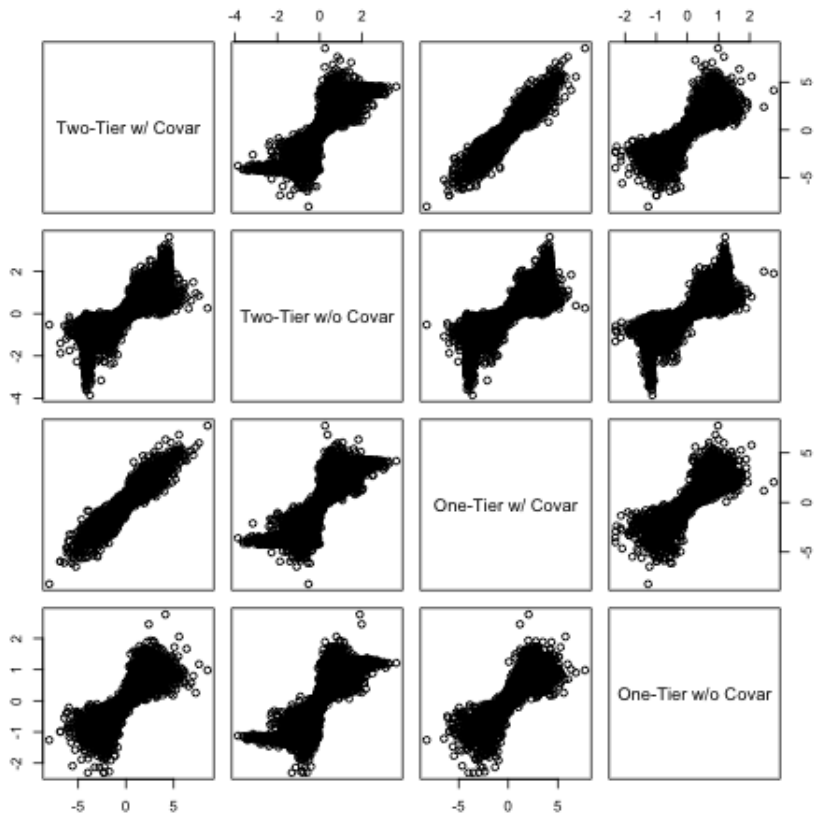

Figure 3.7: Distribution of EAP scores for content domain 3

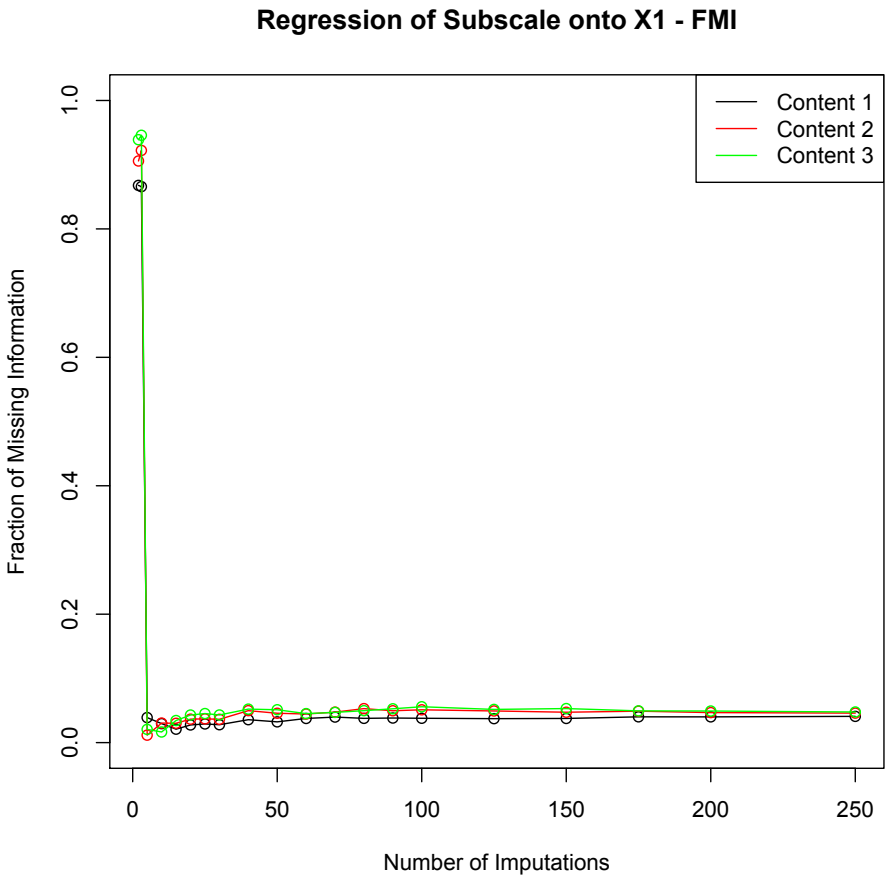

Figure 3.8: Fraction of missing information of regression of content subscale onto X1 from simulated data by the number of imputations

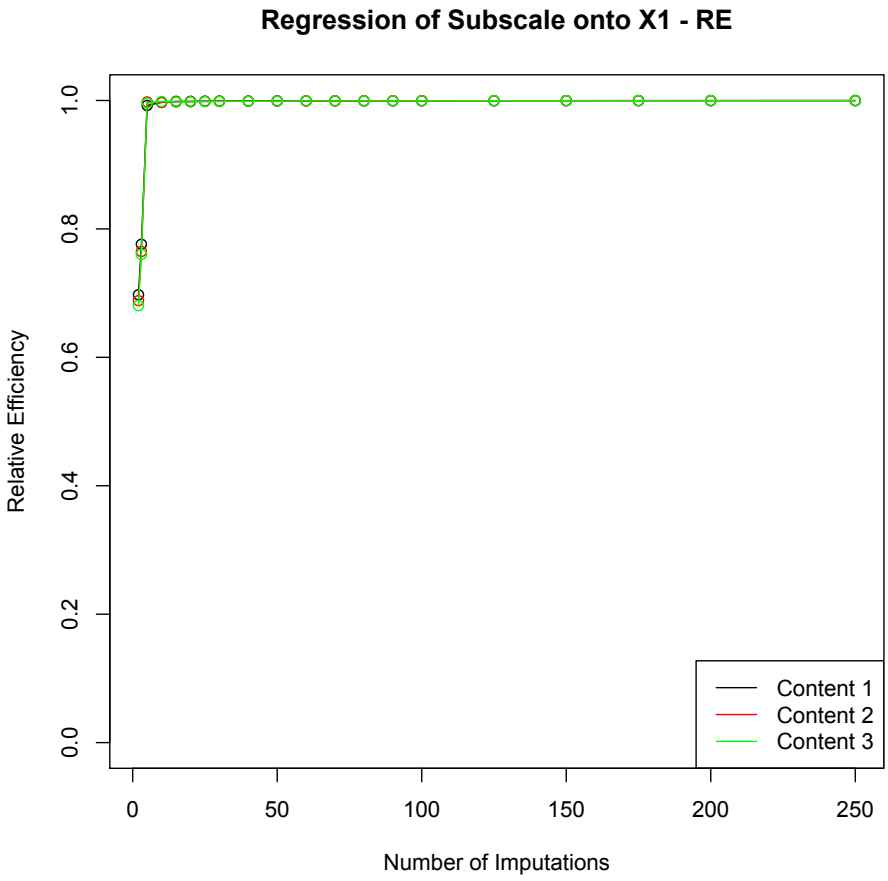

Figure 3.9: Relative efficiency (to theoretical infinite imputations) of regression of content subscale onto X1 from simulated data by the number of imputations

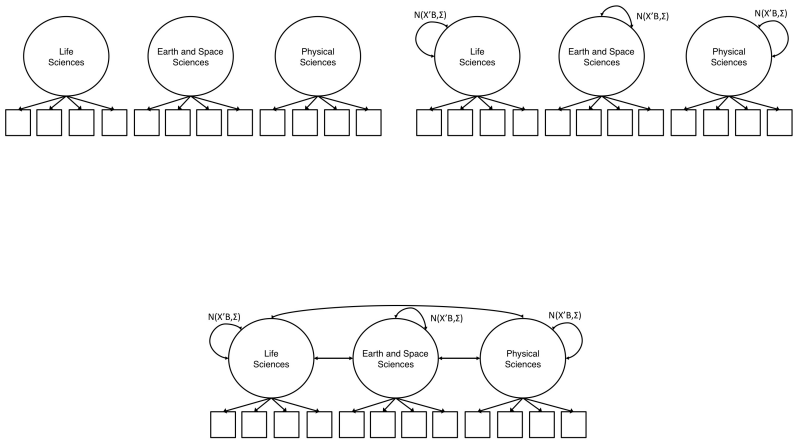

Figure 3.10: One-tier models calibrated on the 2011 NAEP Science Assessment Data

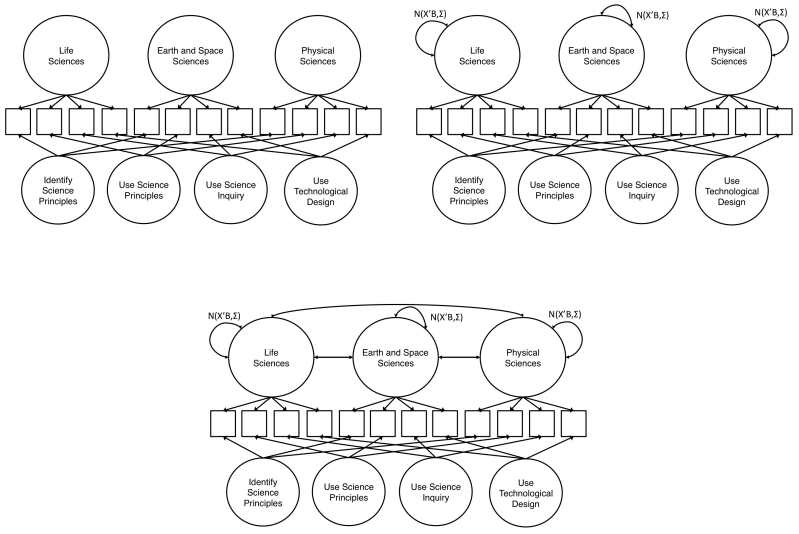

Figure 3.11: Two-tier models calibrated on the 2011 NAEP Science Assessment Data

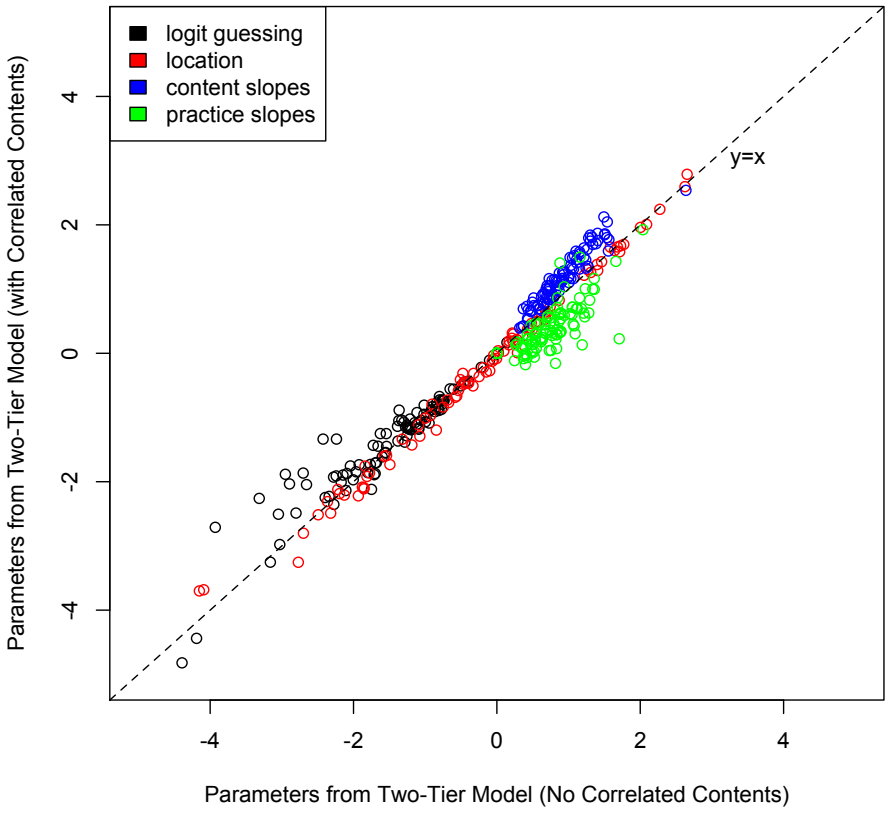

Figure 3.12: Item parameters for 3PL Items - two-tier covariate model with correlated content domains versus two-tier covariate model without correlated content domains

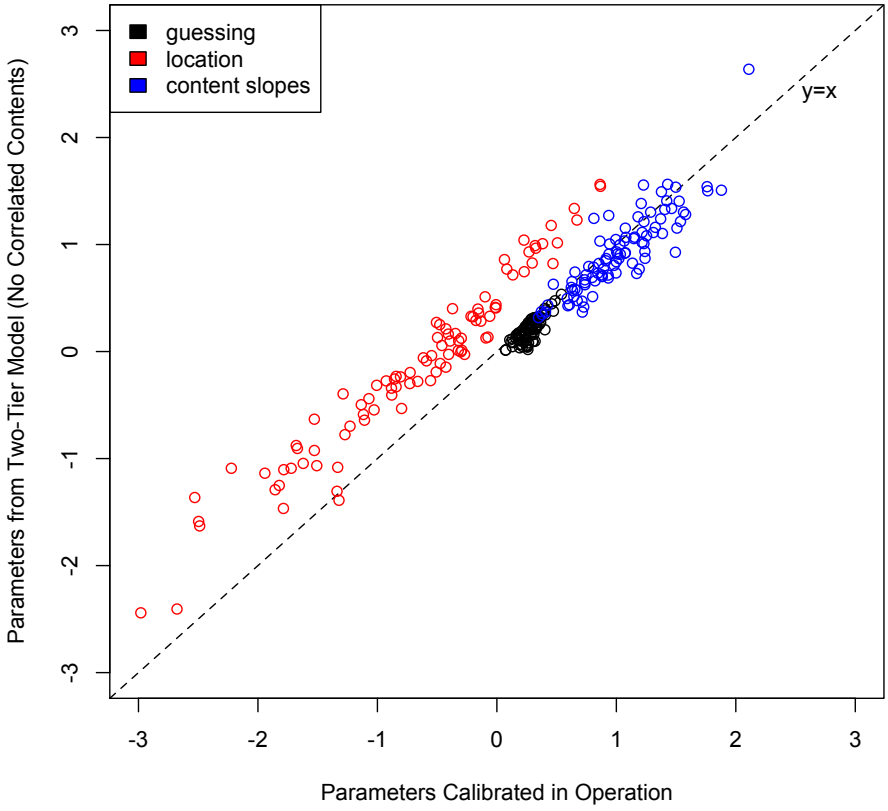

Figure 3.13: Item parameters for 3PL Items - given operational item parameters versus two-tier covariate model without correlated content domains

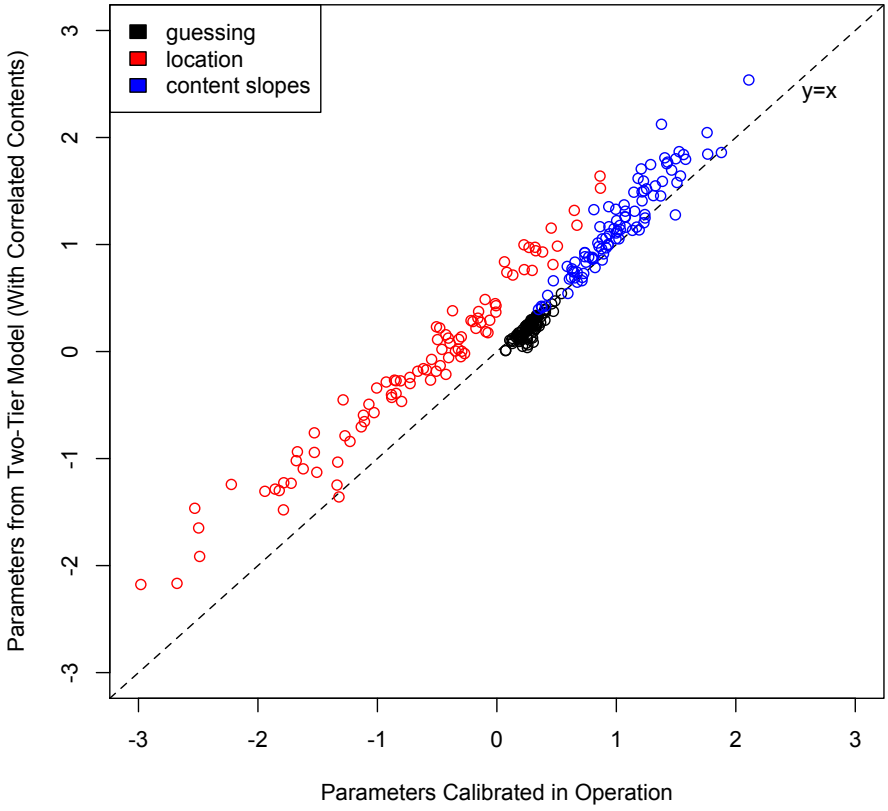

Figure 3.14: Item parameters for 3PL Items - given operational item parameters versus two-tier covariate model with correlated content domains

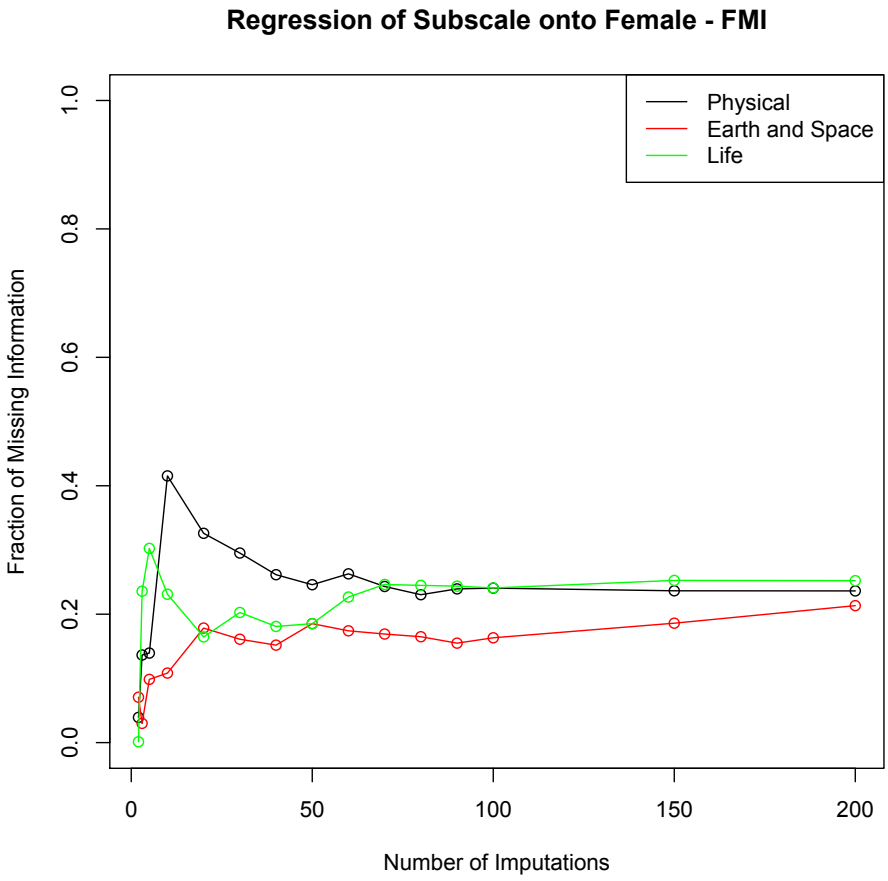

Figure 3.15: Fraction of missing information of regression of content subscale onto female by the number of imputations

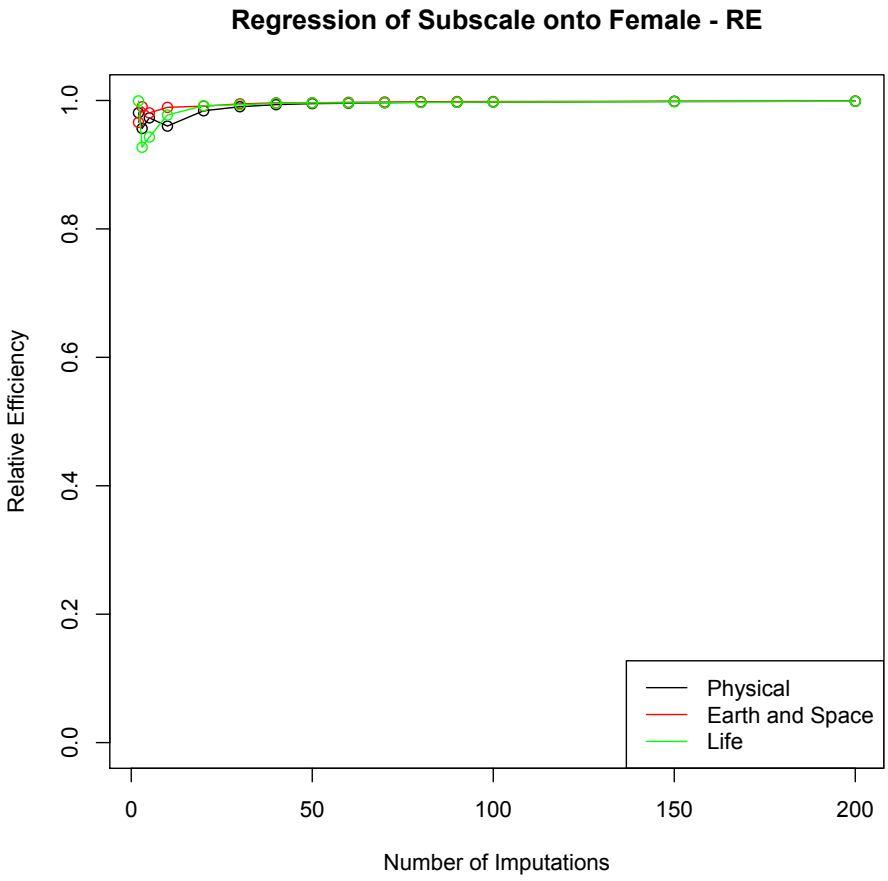

Figure 3.16: Relative efficiency (to theoretical infinite imputations) of regression of content subscale onto female by the number of imputations

3.7 Tables

Table 3.1: The structure of slopes for the simulation of the NAEP Science Assessment

Item	Content 1	Content 2	Content 3	Practice 1	Practice 2	Practice 3
1	$\lambda_{1,1}$	0	0	$\lambda_{1,4}$	0	0
2	$\lambda_{2,1}$	0	0	0	$\lambda_{2,5}$	0
3	$\lambda_{3,1}$	0	0	0	0	$\lambda_{3,6}$
4	0	$\lambda_{4,2}$	0	$\lambda_{4,4}$	0	0
5	0	$\lambda_{5,2}$	0	0	$\lambda_{5,5}$	0
6	0	$\lambda_{6,2}$	0	0	0	$\lambda_{6,6}$
7	0	0	$\lambda_{7,3}$	$\lambda_{7,4}$	0	0
8	0	0	$\lambda_{8,3}$	0	$\lambda_{8,5}$	0
9	0	0	$\lambda_{9,3}$	0	0	$\lambda_{9,6}$

Table 3.2: The structure of slopes for the calibration of the NAEP Science Assessment

Item	Content 1	Content 2	Content 3	Practice 1	Practice 2	Practice 3	Practice 4
1	$\lambda_{1,1}$	0	0	$\lambda_{1,4}$	0	0	0
2	$\lambda_{2,1}$	0	0	0	$\lambda_{2,5}$	0	0
3	$\lambda_{3,1}$	0	0	0	0	$\lambda_{3,6}$	0
4	$\lambda_{4,1}$	0	0	0	0	0	$\lambda_{4,7}$
5	0	$\lambda_{5,2}$	0	$\lambda_{5,4}$	0	0	0
6	0	$\lambda_{6,2}$	0	0	$\lambda_{6,5}$	0	0
7	0	$\lambda_{7,2}$	0	0	0	$\lambda_{7,6}$	0
8	0	$\lambda_{8,2}$	0	0	0	0	$\lambda_{8,7}$
9	0	0	$\lambda_{9,3}$	$\lambda_{9,4}$	0	0	0
10	0	0	$\lambda_{10,3}$	0	$\lambda_{10,5}$	0	0
11	0	0	$\lambda_{11,3}$	0	0	$\lambda_{11,6}$	0
12	0	0	$\lambda_{12,3}$	0	0	0	$\lambda_{12,7}$

Table 3.3: Average fit indexes across 50 simulations of BIB data for Models 2, 4, 5, & 6

Model	Content Domains	Practice Domains	Covariates	BIB data	-2LL	AIC	BIC
Model 2	3	3	Yes	Yes	90916.1318	92002.1318	95860.1361
Model 4	3	3	No	Yes	130999.5066	131905.5066	135124.0629
Model 5	3	0	Yes	Yes	91836.1777	92622.1777	95414.4347
Model 6	3	0	No	Yes	131646.1988	132252.1988	134405.0077

Table 3.4: Fitted models to simulated data: overall bias (standardized) & RMSE

	Content Domains	Practice Domains	Covariates	BIB data	Location Parameters		Slope Parameters		Regression	
					Mean Rel Bias	CV	Mean Rel Bias	CV	Mean Rel Bias	CV
Model 1	3	3	Yes	No	0.067	0.073	0.80	0.81	-0.47	0.53
Model 2	3	3	Yes	Yes	0.158	0.297	1.53	1.61	-0.54	0.71
Model 3	3	3	Yes	Yes	0.135	0.267	1.44	1.50	-0.54	0.71
Model 4	3	3	No	Yes	0.039	0.342	6.87	7.05	NA	NA
Model 5	3	0	Yes	Yes	0.008	0.021	1.27	1.30	-0.55	0.72
Model 6	3	0	No	Yes	-0.073	0.266	5.98	6.04	NA	NA

Table 3.5: Bias and RMSEA of correlation estimates between simulated scientific content domains across 50 simulations

Corr	True Value	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1 & 2	0.7	0.089	0.089	-0.027	0.032	-0.024	0.029	-0.034	0.035	-0.027	0.032	-0.032	0.033
1 & 3	0.7	0.100	0.100	0.001	0.015	0.003	0.015	0.063	0.063	0.003	0.015	0.063	0.064
2 & 3	0.7	0.095	0.096	0.005	0.017	0.007	0.017	0.055	0.056	0.005	0.017	0.055	0.056

Table 3.6: Model Fit Statistics for 2011 NAEP Science Assessment

Model	Two Tier	Covariates	Correlated Primary Domains	Runtime (in hours)	Free Parameters	Deviance	AIC	BIC	Passed Second Order
1	No	No	No	11.3	465	5489169.4	5490099.4	5494623.6	Yes
2	No	Yes	No	17.1	501	5714670.5	5715672.5	5720537.3	No
3	No	Yes	Yes	19.1	498	5298138.5	5299134.5	5303970.2	Yes
4	Yes	No	No	23.7	604	5466576.4	5467784.4	5473660.9	Yes
5	Yes	Yes	No	40.46	637	5281953.6	5283227.6	5289412.9	Yes
6	Yes	Yes	Yes	43.3	640	5691043.0	5692323.0	5698537.5	No

CHAPTER 4

Periodontal Disease Classification and Issues with Partial Mouth Recording

4.1 Underestimation of partial mouth periodontal examinations

Periodontal examinations are conducted in dental epidemiological studies for surveillance and detection and to estimate the prevalence of periodontal disease in a given population. The typical exam involves probing the teeth at up to six sites per tooth, measuring attachment loss, pocket depth, and recession (in mm) in addition to indicators of bleeding on probing. Partial-mouth recording protocols can refer to a number of possible study designs. A full-mouth periodontal examination typically is conducted on six sites per tooth on all teeth excluding the third-molars. Half-mouth periodontal examinations refer to sampling only half of the teeth, commonly by diagonal quadrants. Partial-mouth examinations may also incorporate fewer measurements per tooth, with some studies such as NHANES 1999-2000 probing only two sites per tooth, with other examinations recording 3 or 4.

Half-mouth periodontal examinations have been implemented in a number of studies as a means of reducing costs and patient burden, but there is conflicting research on the underestimation of prevalence of periodontal disease that occurs when only half-mouth data are used. Some studies show little differences in the estimates of prevalence from data on two quadrants [Hun87, DEK02], while other

studies have shown that sensitivity is decreased when using partial recording protocols [SKA05]. In an unpublished thesis, Maitra [Mai12] showed that there is a high degree of association between sites using circular statistics and argued that the findings suggest a sub-sample of teeth should be sufficient for estimating periodontal disease status. A literature review of 12 studies with 32 partial mouth protocols found that half-mouth six-site protocols or full-mouth three-site (mesiobuccal, midbuccal, and distolingual) sampling had the greatest sensitivity of disease prevalence and lowest relative bias for severity [TGD13].

The National Health and Nutrition Examination Study (NHANES) currently uses the full-mouth, six-site periodontal examination after previously using the half-mouth data [EDW12]. A convenience sample of 454 adults was given a full-mouth periodontal examination to measure the "true" periodontal prevalence rate and the results were compared to the partial-mouth recording protocols of NHANES III and NHANES 2001-2004. The partial-mouth recording protocols were found to underestimate of the prevalence of periodontal disease by as much as 50% or greater, depending on how periodontitis was defined [ETW10].

Periodontal disease is often defined as having at least one site where measured attachment loss or pocket depth is greater than or equal to 4mm. However, the definitions of periodontal disease vary in the literature and has been defined on ranges of attachment loss between ≥ 2 to 6 mm or on pocket depths above 3 to 6 mm [SEM09]. In the study of the oral consequences of methamphetamine use, the stages of periodontal disease are classified by the CDC-AAP criteria [EDW12]:

- Severe periodontitis is classified by the presence of two or more interproximal sites (sites that neighbor an adjacent tooth) on different teeth with attachment loss ≥ 6 mm and one or more interproximal sites with pocket depth ≥ 5.

- Moderate periodontitis is characterized by two or more interproximal sites on different teeth with attachment loss ≥ 4 or two or more interproximal

sites (on different teeth) with pocket depth ≥ 5.

- Mild periodontitis is categorized by two or more interproximal sites with attachment loss ≥ 3 and two or interproximal sites with pocket depth ≥ 4 or one site with pocket depth ≥ 5 on different teeth.

These definitions/categorizations of periodontal disease are made without respect to the sampling design, but may be sensitive to designs in which the full-mouth data are not observed. Some alternative approaches that have been proposed by others will be discussed Section 4.2.

4.2 Existing methods for dealing with partial mouth data

A recent data analysis was conducted using the full-mouth, six-site periodontal examination from NHANES 2009-2010 by Tran et al. [TGD14]. In this analysis, the authors applied various partial mouth recording protocols and estimated the prevalence of periodontal disease and associated measures and found that partial mouth recording protocols, particularly half-mouth designs, underestimated prevalence as compared with the full mouth data. The authors proposed a "half-reduced" estimate of disease status, which essentially reduces the required number of sites with extreme pocket depth or attachment loss by half for each threshold diagnosis of periodontal disease. The half-reduced criteria is as follows:

- Severe periodontitis is classified by the presence of one or more interproximal sites with attachment loss ≥ 6 mm and one or more interproximal sites with pocket depth ≥ 5 on a different tooth or ≥ 6 on the same site.

- Moderate periodontitis is characterized by one or more interproximal sites on different teeth with attachment loss 4 or one or more interproximal sites (on different teeth) with pocket depth ≥ 5.

- Mild periodontitis is categorized by one or more interproximal sites with attachment loss ≥ 3 and one or more interproximal sites with pocket depth ≥ 4 or one site with pocket depth ≥ 5.

The half-reduced disease status is akin to doubling the number of observed sites with extreme pocket depth or attachment loss or imputing what is observed on one half of the mouth for the other half. Thus this method assumes that each half of the mouth is exchangeable with the other, which might not be true given the evidence from the χ^2 tests, and can possibly overestimate the prevalence of disease.

Reich et al. [RB10] modeled the missing data from a full-mouth periodontal examination (i.e. teeth not present in the subject) using a latent spatial process model modeling binary and continuous outcomes on site level data. Their model found that incorporating information about the missingness of the teeth produced more reliable estimates of the periodontal disease status. However, these methods were designed to account for missing periodontal observations due to teeth being unrecordable or not present. In this study, we do not seek to predict observations on teeth that do not exist: rather the goal is to account for what would have been observed under a full-mouth examination.

4.3 Unadjusted prevalence of periodontal disease among methamphetamine users and a matched cohort of non-using subjects from NHANES

The proportion of observed teeth meeting attachment loss thresholds for both the MA users and the matched 2011 NHANES subjects can be seen in Figure 4.3, while the proportions for each pocket depth threshold can be seen in Figure 4.4, and alternative representations can be seen in Figures 4.1 and 4.2. Almost all of

the observed and present teeth from MA users meet attachment loss thresholds of ≥ 2, even on anterior teeth (teeth 6-11 and 22-27). The most severe attachment loss threshold of ≥ 6 happens more frequently among MA users than NHANES subjects, particularly on the maxillary first and second molars (teeth 2, 3, 14, and 15), occurring in approximately 20% of each maxillary molar observed among MA users as opposed to less than 10% of maxillary molars in NHANES subjects. In addition, attachment loss between 4 mm and 6mm is also more prevalent in MA users maxillary molars, occurring in approximately 40% of first and second maxillary molars, as compared with 20% among NHANES observed teeth. Mandibular molars (teeth 18, 19, 30, and 31) have lower rates of attachment loss ≥ 6 mm among observed teeth for MA users, but have equally high rates of attachment loss between 4 and 6 mm. The mandibular anterior teeth (teeth 23-28) appear to have relatively high rates of both the mid and extreme levels of attachment loss for MA users.

While instances of the maximum pocket depth meeting the highest threshold of ≥ 7 are less common, they occur slightly more frequently on the right second molars (teeth 2 and 18). The lower anterior teeth (22-27), particularly mandibular canines (teeth 22 and 27), are more likely to have pocket depths ≥ 3. Methamphetamine users are have higher frequencies of pocket depth for each tooth than the observed NHANES subjects.

4.3.1 Prevalence among methamphetamine users

Using the CDC-AAP definitions, of the 546 methamphetamine users with periodontal data, 116 (22.3%) are classified as having severe periodontitis and 292 (53.5%) have moderate periodontitis. Only 35 (6.2%) have mild periodontitis, and 104 (19.1%) have no periodontal disease.

When comparing outcomes from subjects randomized to half-mouth examinations to those randomized full-mouth examinations, there is a higher prevalence

of severe periodontal disease within the full-mouth group compared to all in the half-mouth group. Table 4.1 shows the frequency of each periodontal disease classification by randomization to either full or half-mouth examinations. Over 27% of the subjects full-mouth were classified as having severe periodontitis compared to 17% of those randomized to the half-mouth exam. Subjects randomized to the half-mouth examination had higher rates of classification to mild disease (7.9%) or no disease (25.6%) compared with those who received the full-mouth exam (3.9% with mild disease and 10.0% no disease). Half-mouth examinees also had lower rates of moderate periodontal disease classification (49.5%) than the full-mouth participants (59.0%).

Two half-mouth sampling designs were conducted: upper right and lower left (teeth 1-8 and 17-24) and upper left and lower right (teeth 9-16 and 25-32). Table 4.2 displays the frequencies of periodontal disease status further stratified by half-mouth sampling design. It can be seen that subjects randomized to the upper right/lower left quadrants have slightly higher rates of classification to both severe and moderate periodontal disease (18.8% and 52.8% respectively) than the subjects randomized to the upper left/lower right exam (15.1% and 46.1%). When a logistic regression analysis is conducted to predict the probability of having severe periodontal disease by randomization to full mouth, upper left/lower right, or upper right/lower left, the full-mouth exam had slightly higher rates of classification of servere periodontitis as compared to those randomized to the upper left/lower right (parameter estimate: -0.33, SE: 0.17, $p(\chi^2) = 0.05$). However, those randomized to the upper right/lower left examination did not have significantly lower rates of classification of severe periodontitis as compared to those who received the full-mouth exam.

These results indicate that there may be a selection effect in the sampling of the sites for the periodontal examination. This would make logical sense when

considering it only takes two sites of a given threshold (on separate teeth) to qualify as having periodontal disease. If a subject only has two sites with attachment loss greater than 6 mm, and those two particular sites are located on the same quadrant (or on the quadrants diagonal from each other), these would be observed under the full-mouth examination and one of the half-mouth examinations, but not the other half-mouth examination. These results confirm what has been seen in other studies in which disease prevalence is underestimated when using partial mouth protocols. However, because there exist data on the full-mouth examination for some subjects, it may be possible to model the data on the unobserved quadrants.

4.3.2 Prevalence among NHANES 2011-2012 subjects

Using the procedure detailed in the background section, 1090 subjects were selected from the 2011-2012 NHANES study. All subjects received full mouth, six-site periodontal examinations. We estimate the prevalence of periodontal disease under the same four-site procedure used on the methamphetamine users. In addition, we also apply each of the four-site, half-mouth sampling designs to determine the bias of each procedure, knowing the full sample frequencies of periodontal disease.

The number and percentage of the 1090 NHANES subjects classified as having periodontal disease are displayed in Table 4.3. It can be seen that the UR-LL design would underestimate the overall prevalence of periodontal disease by an absolute 12.6% (relative bias of 28.0%), and the UL-LR design would underestimate the prevalence by 18.7% (relative bias of 41.3%). The pattern of underestimation is similar to that found in the methamphetamine-using sample: although both half-mouth sampling designs underestimate the prevalence of periodontal disease, detection of periodontal disease is more likely in upper-right/lower-left

examinations than in upper-left, lower right exams.

4.4 A note on missingness mechanisms

In this study, there are two types of missing data. Using the terminology defined by Rubin [Rub76], data from the half-mouth examinations on the quadrants of teeth unobserved can be considered missing completely at random (MCAR), and data on teeth that are not present in an individual's mouth, which may be missing not at random (MNAR) or missing at random (MAR). For the purposes of this analysis, we assume that the data from the teeth missing in the subject are MAR. While it is important not to be too cavalier about this assumption, which can be violated in practice, it is also valuable to view the MAR assumption as a starting point for analysis that often would be reasonable to consider as a candidate assumption.

4.5 Multiple Imputation of Periodontal Examination Data

An alternative way of conducting an analysis on partial mouth data is to borrow information from what is known from full mouth data. Given one half mouth of the data, and if some information is known about the relationship across the population between the observed and the unobserved teeth, we might be able to better predict what would have been observed on the other half of the mouth.

There is some precedent for imputing oral health data for improved estimates of population prevalence of dental diseases related to caries. Pahel et al. [PPS11] imputed the number of carious teeth in children using a zero-inflated poisson model. Schuller and van Buuren [SB14] examined multiple imputation methods under different assumptions of the missing data mechanism for Decayed, Missing,

or Filled (DMF) surface and teeth indexes incorporating socioeconomic status (SES) information. In that study, DMF scores were treated as continuous variables and were imputed using chained equations (MICE) [VBG06].

The individual site measurements of the periodontal examination can be thought of having a hierarchical structure, in which sites are located within teeth, and teeth are located within an individual subject. The assumption of nesting within teeth may be relaxed with alternative coding of the data to incorporate tooth effects as well as site measurement type effects.

The imputations were generated using the R package PAN, which is designed to impute multivariate repeated measures data using a Gibbs sampler as described in Schafer (1997). The underlying model used by PAN is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{4.1}$$

for $i = 1, ..., m$, where

- i indexes the study participant,
- m is the total number of subjects,
- n_i is the number of measurements (observed and unobserved) on the ith subject
- y_i represents the $(n_i \times r)$ matrix of incomplete multivariate outcome data for subject i,
- X_i is the $(n_i \times p)$ matrix of p fixed covariates,
- Z_i is the $(n_i \times q)$ matrix of q random covariates,
- β is the $(p \times r)$ matrix of coefficients common to the population,

- b_i is the ($q \times r$ matrix of subject specific coefficients, and
- ϵ_i is the $(n_i \times r)$ matrix of residual errors.

The model index i refers to the study participant, and the outcome y_i is a matrix of outcomes for r different types of measurements on n_i possible locations in the mouth. The random covariates for this study, Z_i, will just be a column of 1's to indicate a random intercept model. For modeling the periodontal disease status, we include dummy variables to uniquely identify each site and tooth as fixed effects. Thus in the sample of methamphetamine users where four sites per tooth were recorded, there are three dummy variables to indicate site type (B, M, or DL versus D), six dummy variables to indicate tooth type (with incisors as reference), one dummy variable indicating mandibular tooth, and one variable to indicate the right side of the mouth. In addition, we also include effects such as methamphetamine used, age, and other variables that may be relevant to periodontal disease status and subsequent analyses.

4.5.1 Post-Imputation Processing and Analysis

First, the imputed values of attachment loss and pocket depth were rounded to the nearest integer. Next, to avoid imputing nonsensical values where there would have been no measurement (i.e. the tooth is not present or cannot be measured), imputed values on missing teeth were set to missing again. The purpose of the multiple imputation is to capture the disease status for partial-mouth data as compared to what would have been observed had the full mouth been recorded. Thus imputing values for teeth that would not have been observed even under the full-mouth sampling scheme is to be avoided.

The CDC-AAP definitions of mild, moderate, and severe periodontitis were then applied to the imputed datasets. Using the computed status variables, the SAS functions PROC UNIVARIATE and PROC MIANALYZE were used to gen-

erate overall estimates and standard errors for prevalence of each periodontal disease classification.

4.5.2 Results from Analysis for MA Users

The resulting prevalence estimates, standard errors, fraction of missing information (FMI), and relative efficiency (of using 5 imputations relative to an infinite number of imputations) are presented in Table 4.4. The percentage of subjects classified as having no periodontal disease decreases from an original estimate of 19.1% to 5.5% after imputing values for the unobserved (but present) teeth. The rates of mild periodontal disease also decrease from 6.2% to 2.0%. Prevalence of moderate periodontal disease increases to 65.8% from 53.5%, and the estimated percent of subjects with severe periodontal disease goes from 21.3% to 26.7%. With each of these prevalence estimates, the relative efficiency is over 95%, while the fraction of missing information varies between 0.068 to 0.21 (higher for no or mild periodontal disease). With the highest fraction of missing information at 0.21, five imputations appears to be sufficient.

4.6 Multiple Imputation Analysis for NHANES subjects

Because we observe the full-mouth examination on the NHANES subjects, an opportunity is presented in which we can test our imputation procedure to see if we can replicate what would is observed for the full-mouth examination. In this case, we randomly assigned the 1090 NHANES subjects to a full, upper-left and lower-right, or upper-right and lower-left examination with the same probabilities as observed in the proportion of MA subjects. The first application of this randomization resulted in 473 subjects with the full mouth examination, 328 with upper-right and lower-left, and 289 with upper-left, lower-right. If assigned to a

half-mouth examination, all observations on the "unobserved" teeth were set to missing. The multiple imputations were generated and analyzed using the same procedure as with the MA subjects, with the exception of the background characteristics on the individual participants. Two imputation models were considered: one without any background variables, and one which included age, gender, ethnicity, and smoking status.

4.7 Results: Multiple Imputation Analysis for NHANES subjects

Table 4.5 displays the original (no missing-by-design), missing-by-design, and multiply-imputed prevalence estimates for the NHANES subjects where some subjects have been randomized to the half-mouth examinations under the model without personal covariates. It can be seen that when the planned-missingness is imposed on the NHANES sample in the proportions assigned to the MA subjects, the prevalence of any periodontal disease decreases by 7.8%. The rates of mild periodontal disease remain approximately the same, although 0.8% lower, and the rates of both moderate and severe periodontal disease status are reduced by 4.2% and 2.7% respectively.

While the point estimate obtained for prevalence any disease status using the multiple imputation without covariates procedure is greater than the original value by 3.5%, the 95% interval does contain the target value. The estimates of percentages with mild periodontal disease and severe periodontal disease are still slightly lower (by 0.7% and 2.0% respectively), but the estimate of prevalence of moderate periodontal disease is an additional 5.3% larger. These results indicate that the procedure increases the likelihood of detecting moderate periodontal disease.

The randomization and multiple imputation procedure was repeated on the

NHANES data five times to show what would happen when different subjects are randomized. The resulting estimates from the randomization and imputation (with background characteristics) for each of these five trials can be seen in Table 4.6. We still see a slight overestimation of moderate periodontal disease in addition to lower estimates of severe periodontal disease prevalence. However, the 95% confidence interval for the estimated proportion of the sample having no periodontal disease does cover the true value in each trial, indicating that the rate of any periodontal disease (mild, moderate, or severe) is only slightly overestimated.

4.8 Conclusions

The rates of periodontal disease using the CDC definitions can be underestimated when planned-missing-data designs are used without any adjustment for the design. Presented here is an alternative to simply changing the definition; we show how multiple imputation can be used to impute values for the unobserved (but present) measurements. The procedure enables us to replace the missing-by-design values with plausible values given a hierarchical normal model for the data, and the analysis also allows for the standard errors to be adjusted for the missing observations.

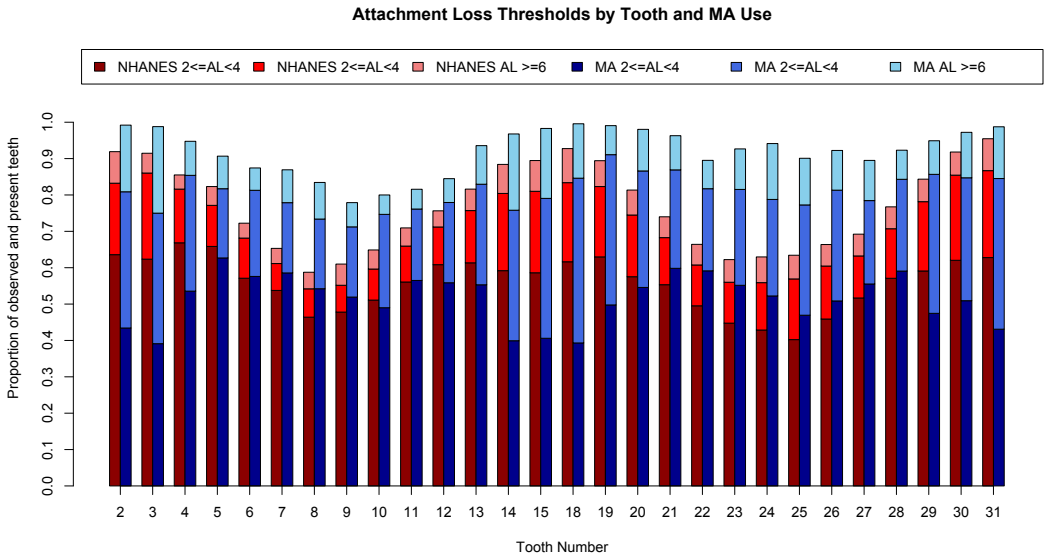

Figure 4.1: Proportion of each observed tooth meeting given attachment loss thresholds

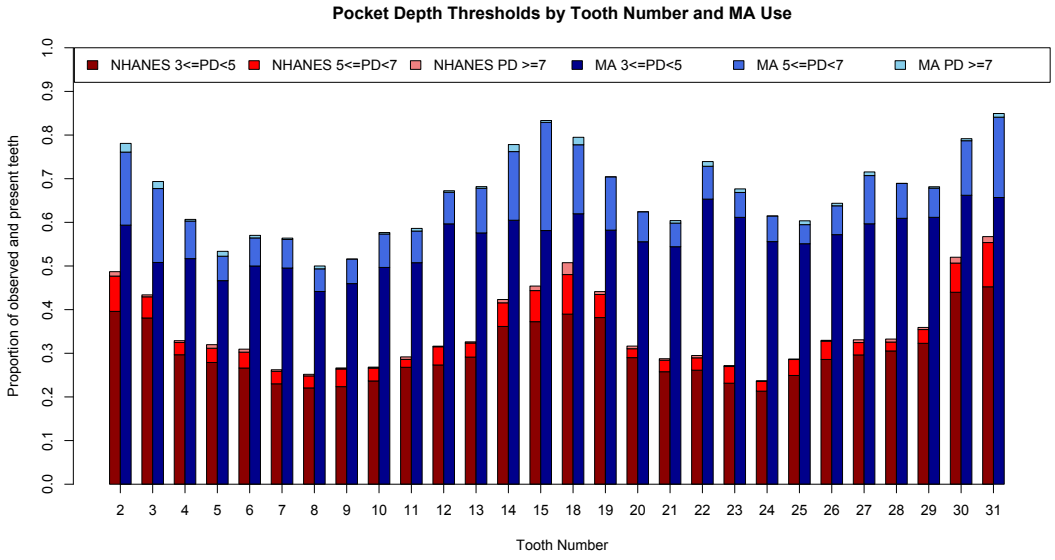

Figure 4.2: Proportion of each observed tooth meeting given pocket depth thresholds

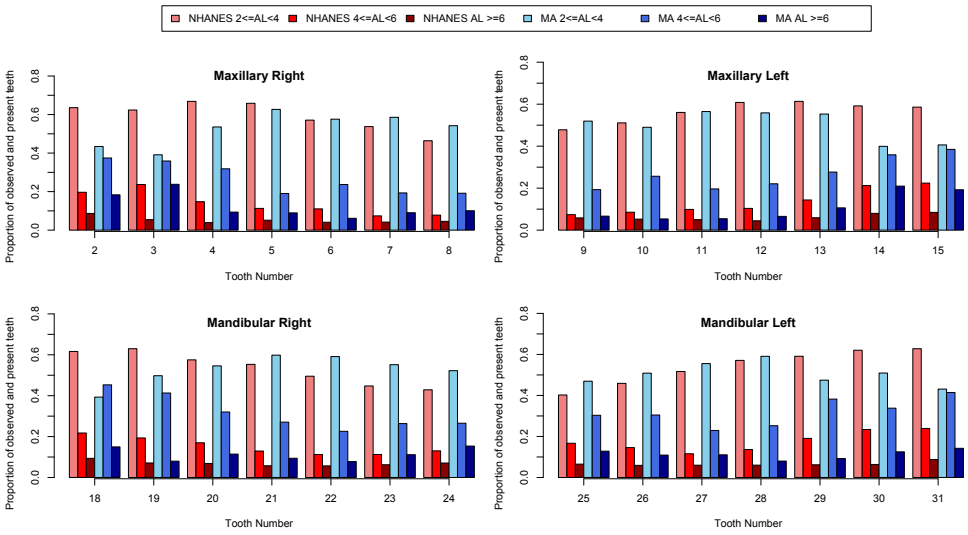

Figure 4.3: Proportion of each observed tooth meeting given attachment loss thresholds

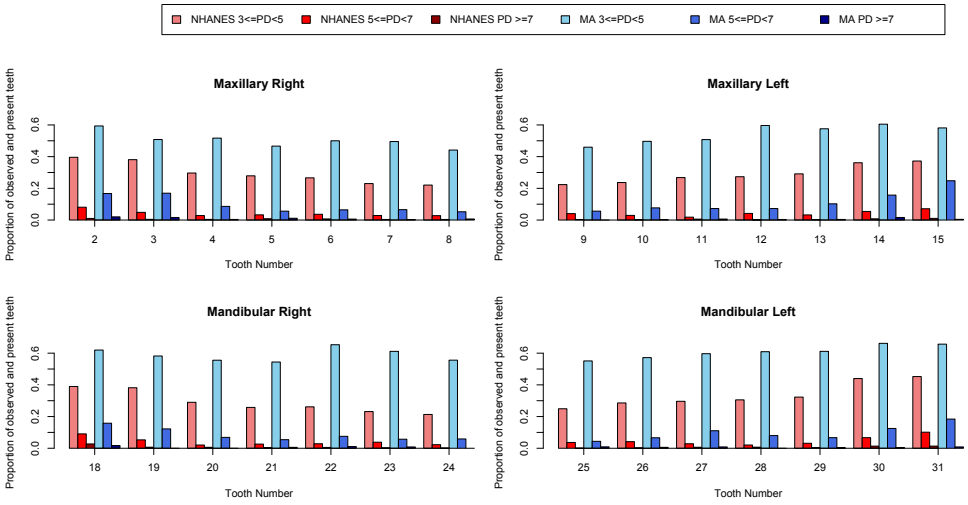

Figure 4.4: Proportion of each observed tooth meeting given pocket depth thresholds

Periodontal disease status	Full-Mouth Exam	Half-Mouth Exam
No periodontal disease	23 (10.4%)	81 (25.6%)
Mild periodontal disease	9 (3.9%)	25 (7.8%)
Moderate periodontal disease	135 (59.9%)	157 (49.5%)
Severe periodontal disease	62 (27.1%)	54 (17.0%)

Table 4.1: Periodontal disease classifications by half-mouth versus full-mouth exams. $\chi^2 = 28.6$ on 3 df, $p < 0.0001$

Periodontal disease status	Full-Mouth Exam	Half-Mouth Exam Upper Left/Lower Right	Half-Mouth Exam Upper Right/Lower Left
No periodontal disease	23 (10.0%)	46 (30.3%)	35 (21.2%)
Mild periodontal disease	9 (3.9%)	13 (8.6%)	12 (7.3%)
Moderate periodontal disease	135 (59.0%)	70 (46.1%)	87 (52.7%)
Severe periodontal disease	62 (27.1%)	23 (15.1%)	31 (18.8%)

Table 4.2: Periodontal disease classifications by randomization. $\chi^2 = 33.4$ on 6 df, $p < 0.0001$

Periodontal disease status	Full-Mouth Exam	Half-Mouth Exam UR-LL	Half-Mouth Exam UL-LR
No periodontal disease	494 (45.3%)	632 (58.0%)	698 (64.0%)
Mild periodontal disease	32 (2.9%)	25 (2.3%)	29 (2.7%)
Moderate periodontal disease	409 (37.5%)	331 (30.4%)	293 (26.9%)
Severe periodontal disease	155 (14.2%)	102 (9.4%)	70 (6.4%)

Table 4.3: Periodontal disease classifications of NHANES 2011-2012 subjects under four-site, full-mouth and half-mouth criteria, McNemar $\chi^2 < 0.0001$ for all 3 two-way comparisons

Disease status	Original Prevalence Estimate	Multiple Imputation Analysis Prevalence Estimate (SE)	95% Interval	FMI	RE
None	19.1%	5.5% (1.1%)	(3.4,7.7)	0.21	0.96
Mild	6.2%	2.0% (0.66%)	(0.74, 3.3)	0.17	0.97
Moderate	53.5%	65.8% (2.2%)	(61.4,70.1)	0.16	0.97
Severe	21.3%	26.7% (1.9%)	(22.8,30.5)	0.068	0.99

Table 4.4: Periodontal disease classifications after imputation for MA Users

Disease status	Original Prevalence Estimate	Missing by Design Prevalence Estimate	Multiple Imputation Analysis			
			Prevalence Estimate (SE)	95% Interval	FMI	RE
None	45.3%	53.1%	41.8% (1.9%)	(37.8%,45.8%)	0.45	0.92
Mild	2.9%	2.1%	2.2 % (0.45%)	(1.3%,3.0%)	0.041	0.99
Moderate	37.5%	33.3%	43.8% (1.9%)	(39.8%,47.8%)	0.45	0.92
Severe	14.2%	11.5%	12.2% (1.0%)	(10.2%,14.2%)	0.032	0.99

Table 4.5: NHANES Periodontal disease classifications after imputation - without background characteristics

Disease status	Original Prevalence Estimate	Missing by Design Prevalence Estimate	Multiple Imputation Analysis			
			Prevalence Estimate (SE)	95% Interval	FMI	RE
Trial 1						
None	45.3%	53.1%	41.5% (1.5%)	(38.5%,44.5%)	0.06	0.99
Mild	2.9%	2.1%	2.4 % (0.63%)	(1.0%,3.6%)	0.52	0.91
Moderate	37.5%	33.3%	43.7% (1.6%)	(40.4%,47.0%)	0.18	0.97
Severe	14.2%	11.5%	12.4% (1.0%)	(10.4%,14.5%)	0.05	0.99
Trial 2						
None	45.3%	53.1%	42.4 (1.9%)	(38.4%,46.4%)	0.45	0.92
Mild	2.9%	2.2%	2.4% (0.48%)	(1.5%,3.4%)	0.08	0.98
Moderate	37.5%	33.7%	43.6% (1.9%)	(39.5%,47.6%)	0.45	0.92
Severe	14.2%	11.0%	11.7% (0.98%)	(9.7%,13.6%)	0.02	0.99
Trial 3						
None	45.3%	52.4%	42.8 (1.7%)	(39.6%,46.1%)	0.20	0.96
Mild	2.9%	2.8%	2.6% (0.53%)	(1.5%,3.6%)	0.20	0.96
Moderate	37.5%	33.9%	43.1% (1.7%)	(39.6%,46.6%)	0.28	0.95
Severe	14.2%	10.8%	11.4% (0.99%)	(9.5%,13.4%)	0.05	0.99
Trial 4						
None	45.3%	52.5%	43.3 (1.6%)	(40.2%,46.4%)	0.09	0.98
Mild	2.9%	2.8%	2.6% (0.56%)	(1.5%,3.7%)	0.29	0.95
Moderate	37.5%	33.9%	42.2% (1.6%)	(39.2%,45.3%)	0.08	0.98
Severe	14.2%	10.8%	11.8% (1.0%)	(9.8%,13.8%)	0.06	0.99
Trial 5						
None	45.3%	52.5%	42.4 (1.6%)	(39.1%,45.7%)	0.19	0.96
Mild	2.9%	2.5%	2.6% (0.52%)	(1.6%,3.6%)	0.13	0.97
Moderate	37.5%	34.1%	43.2% (1.7%)	(39.9%,46.5%)	0.13	0.97
Severe	14.2%	11.0%	11.8% (1.0%)	(9.8%,13.8%)	0.058	0.99

Table 4.6: NHANES Periodontal disease classifications after imputation with characteristics - Five Trials

Disease status	Original Prevalence Estimate	Missing by Design Prevalence Estimate	Multiple Imputation Analysis Prevalence Estimate (SE)	95% Interval	FMI	RE
Trial 1						
None	45.3%	53.1%	39.5% (1.6%)	(36.4%,42.5%)	0.10	0.98
Mild	2.9%	2.1%	3.1 % (0.61%)	(2.0%,4.4%)	0.25	0.95
Moderate	37.5%	33.3%	41.8% (1.7%)	(38.5%,45.1%)	0.20	0.96
Severe	14.2%	11.5%	15.5% (1.1%)	(13.2%,17.7%)	0.09	0.98

Table 4.7: NHANES Periodontal disease classifications after imputation with characteristics on Trial 1

Table 4.8: Coding of each site in the periodontal examination

Site	B	M	DL	T2	T3	T4	T5	T6	T7	Maxillary	Right
2D	0	0	0	1	0	0	0	0	0	1	1
2B	1	0	0	1	0	0	0	0	0	1	1
2M	0	1	0	1	0	0	0	0	0	1	1
2DL	0	0	1	1	0	0	0	0	0	1	1
3D	0	0	0	0	1	0	0	0	0	1	1
3B	1	0	0	0	1	0	0	0	0	1	1
3M	0	1	0	0	1	0	0	0	0	1	1
3DL	0	0	1	0	1	0	0	0	0	1	1
4D	0	0	0	0	0	1	0	0	0	1	1
4B	1	0	0	0	0	1	0	0	0	1	1
4M	0	1	0	0	0	1	0	0	0	1	1
4DL	0	0	1	0	0	0	0	0	0	1	1
5D	0	0	0	0	0	0	1	0	0	1	1
5B	1	0	0	0	0	0	1	0	0	1	1
5M	0	1	0	0	0	0	1	0	0	1	1
5DL	0	0	1	0	0	0	1	0	0	1	1
6D	0	0	0	0	0	0	0	1	0	1	1
6B	1	0	0	0	0	0	0	1	0	1	1
6M	0	1	0	0	0	0	0	1	0	1	1
6DL	0	0	1	0	0	0	0	1	0	1	1
7D	0	0	0	0	0	0	0	0	1	1	1
7B	1	0	0	0	0	0	0	0	1	1	1
7M	0	1	0	0	0	0	0	0	1	1	1
7DL	0	0	1	0	0	0	0	0	1	1	1
8D	0	0	0	0	0	0	0	0	0	1	1
8B	1	0	0	0	0	0	0	0	0	1	1
8M	0	1	0	0	0	0	0	0	0	1	1
8DL	0	0	1	0	0	0	0	0	0	1	1

CHAPTER 5

Item Response Theory Modeling of the Decayed, Missing, and Filled Index of Oral Health

5.1 The Use of the DMF Index in Epidemiology of Oral Health

In this section we discuss IRT models of the Decayed, Missing, and Filled index. The Decayed, Missing, and Filled (DMF) index is one of the most commonly used metrics for evaluating extent of caries disease in oral health epidemiology. There are two most common forms of DMF indices, the DMFT, which sums the number of decayed, missing, or filled teeth in the mouth, and the DMFS, which counts the number of decayed, missing, or filled surfaces (where there are 4-5 surfaces per tooth). There has been some debate as to how to count the surfaces and the contribution of each type.

While the DMF status of each surface and tooth is observed in the methamphetamine study as well as NHANES, we propose using these models to indicate which teeth and statuses may be making the greatest contribution to the total scores. In addition, the item response functions for each tooth may be useful in selecting teeth that are most indicative of methamphetamine use. The resulting IRT calibrations may be used as a basis for future use of planned-missing-data-designs. The DMFT and DMFS as basic summed scores fail to account for the patterns within the data. Decayed, missing, and filled teeth provide equal contributions to the overall measure, even though it can be argued that decayed or

missing teeth are more extreme manifestations of caries than filled teeth. In addition, one decayed, missing, or filled anterior tooth provides the same contribution to the DMFT as a similar condition on a molar, but anterior caries are far less common.

The goal of this research is to provide an alternative means of scoring the caries examination to account for differences in caries type and teeth and determine which patterns produce the highest DMFT scores. Item response theory models are proposed to analyze the data and account for information from different teeth and caries type.

5.2 Motivating Study

5.3 Methods

The motivation behind the use of item response theory (IRT) models stems from viewing the DMFT measurements as individual categorical measurements of an overall disease status. Each tooth can be considered rated on a nominal scale, where the categories are normal, decayed, missing, or filled. While normal would be the lowest category on an ordinal scale, followed by filled, determining the order for missing versus decayed may be more complicated. The summed score of DMFT is one way of measuring the overall caries extent, but in an IRT context, we would consider the disease status to be a latent trait, denoted θ.

5.3.1 Item Response Theory

5.3.1.1 Two Parameter Logistic Model

The two parameter logistic (2PL) item response model is common model used for binary measurements. This model is represented by

$$P_j(\theta) \equiv P(x_j = 1|\theta, a_j, c_j) = 1/(1 + exp[-a_j(\theta - c_j)]),$$

where

- a_j is the slope parameter for item j, which characterizes the relationship between the latent domain and the probability of a correct response, and
- c_j is the location parameter that is generally indicative of the difficulty of the item.

5.3.1.2 Graded Response Model

The Graded Response Model was developed by Samejima (1969) for ordinal or scale items. Let item j have K graded categories. The cumulative probabilities of response are

$$\begin{aligned}
P(x_j \geq 0|\theta) &= 1.0 \\
P(x_j \geq 1|\theta) &= \frac{1}{1 + exp[-c_{j,1} + a_j\theta]} \\
&\ldots \\
P(x_j \geq K - 1|\theta) &= \frac{1}{1 + exp[-c_{j,K-1} + a_j\theta]} \\
P(x_j \geq K|\theta) &= 0,
\end{aligned}$$

where a_j is the slope parameter for item j and $c_{j,k}$ are the item intercepts for the $k = 0, ..., K - 1$ levels.

5.3.1.3 Nominal Model

The Nominal Response Model (Bock 1972) is used when there is no ordering between the categories of an item.

$$P(x_j = k|\theta) = \frac{e^{a_{jk}(\theta - b_{jk})}}{\sum_{l=1}^{L} e^{a_{jl}(\theta - b_{jl})}}$$

5.3.2 Modeling the DMFT

The individual items in the DMFT each show the status of the tooth as decayed, missing, filled, or normal. The different item response models considered for the tooth status along with the coding of each response are displayed in Table 5.1. The most simple model for the data would be the two parameter logistic model, in which decayed, missing, and filled statuses are treated equally as compared to being a sound tooth. The coding of the tooth status into binary data is equivalent to how the summed DMFT treats the possible tooth statuses, and thus the scores resulting from this model will likely be the closest to the DMFT obtained by summing the number of decayed, missing, or filled teeth.

In the most complex item response model, we consider the nominal model, in which no ordering between decayed, missing, or filled is assumed. The nominal model enables us to look at how the probability of being either decayed, missing, or filled on a given tooth is ranked given the overall disease status. We would expect that for a given tooth, lower disease statuses would have a higher probability of being filled and those with higher disease statuses would be categorized into either decayed or missing. However the ranking between decayed and missing would give us insights as to which categorization is more indicative of extreme caries experience for a given tooth.

Model	Number of Categories	Coding of Teeth			
		Normal	Filled	Decayed	Missing
2PL	2	0	1	1	1
Graded	4	0	1	2	3
Graded	3	0	1	2	2
Nominal	4	0	1	2	3

Table 5.1: Possible mechanisms for coding decayed, missing, or filled teeth

The other two models for the data considered are graded response models. First, decayed and missing teeth are treated as equal responses, for a graded response model with three categories. A graded response model with four categories is also fit to the data in which the missing tooth status is considered the most extreme manifestation of caries experience, followed by decayed, and then filled.

5.3.3 IRT Calibration

The process by which item parameters are estimated is called calibration. A number of methods have been developed to estimate the item parameters. Most frequently applied is the Bock-Aitkin EM algorithm [BA81], which evaluates the likelihood at specified quadrature nodes across the distribution of θ. While this approach also allows for more comprehensive model fit indexes to be produced, it can be quite slow for multidimensional IRT models. An alternative approach when considering multidimensional IRT models would be to use the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm [Cai08], which uses stochastic simulation to approximate the posterior distribution. Item response models are calibrated to the combined set of MA users and NHANES subjects, as well as separately to detect differences in item parameter estimates.

5.3.4 Scoring based on IRT

There are a number of potential methods for generating scores based on the item response models. First, *expected a posteriori* (EAP) scores can be computed by

taking the expected value of the posterior distribution of the latent trait for each individual given the item responses [BM82]. Alternatively, *maximum a posteriori* (MAP) scores can be computed from the maximum value of the person-posterior. The approach set forth by [MJM92b] is to impute multiple plausible values from the posterior distribution and treat the latent abilities as missing data. The plausible values approach should only be considered when individual scores are not of interest, rather the interest of inference is on population-level contrasts, such as the difference in average scores between two groups.

5.3.5 Differential item functioning

Differential item functioning (DIF) is the phenomenon where the item parameters (such as the slope or location parameter in a 2PL model) take on different values for different groups [TSW93]. As an example in education research, girls may respond differently to a given question regarding a passage about football than boys, resulting in different item slopes to be estimated between the two groups (akin to an interaction in a regression model). Differential item functioning occurs when, for the same value of the latent variable or vector, there are different expected probabilities of response between groups. One specific aim of the oral health study is to characterize the patterns and relationships of oral health outcomes by whether or not a person uses methamphetamine. It may be possible to identify specific sites in the mouth on which methamphetamine users are more likely to experience bleeding or high attachment loss as compared with non-methamphetamine users with the same overall disease status by assessing differential item functioning.

In practice, to test for DIF, we can specify groups during the model calibration step. Item parameters are then estimated for each group and compared for differences. To determine which teeth have different behaviors between the MA users and NHANES participants for the same overall caries disease status, DIF testing can be done by defining the groups as either MA or NHANES. The drawback

of using only two groups, however, is that there may be other confounding variables related to DIF, as the NHANES and MA samples have different population characteristics. As described previously, we have used a propensity score model to select subjects from the NHANES study that are demographically similar to those from the MA study. Five propensity score subgroups were defined where covariates are balanced within each subgroup. Instead of using only two groups, we can then define 10 groups: MA and NHANES within each propensity subgroup. We can then test, within a propensity score subgroup, for DIF between the MA and NHANES, which may indicate which sites are more indicative of periodontal disease between the two groups. The results of the DIF analysis may not only allow us to identify sites that are indicative of MA use, but also to identify sites in which a partial mouth recording protocol should not ignore.

While testing for differential item functioning under various constraints is an active area of research in the literature in educational measurement, we adopt a straightforward approach by Thissen et al. [TSW93]. Items are calibrated separately for each group, and then Wald χ^2 statistics are used to test for differences in item parameters between the two groups.

5.4 Results

Attempts at doing separate calibrations by propensity-score subgroup failed to converge to a maximum likelihood solution, likely due to the small number of participants from the MA sample in the lower-propensity-score groups. However, many models calibrated MA users only, NHANES subjects only, and all combined converged to maximum likelihood solutions.

5.4.1 Comparing EAP Scores

Figure 5.2 displays the plot of the 2PL EAP scores versus the DMFT summed score. The relationship between the two scores is very close, with curvature at the tails indicating that the 2PL scores are more distinct at the tails as opposed to the summed scores. The Nominal EAP scores are plotted against the summed scores in Figure 5.3. The dispersion of Nominal EAP scores is much greater at each summed score than the 2PL model. Logically, this makes sense. The 2PL model for the data assumes that decayed, missing, and filled are scored equally on a binary scale, so the 2PL EAPs at each summed score will likely be more closely related. In the nominal model, decayed, missing, and filled categorizations are give different contributions to the EAP score, and the information and discrimination of each category will vary by tooth. Thus it is not surprising that the Nominal EAP scores have greater dispersion as the summed DMFT index increases.

To compare the EAP scores between the methamphetamine users and NHANES subjects, there had to be constraints placed on the model. Using the NHANES subjects as a reference with a prior on the scaled score of a standard normal distribution, the mean of the MA subjects when calibrated with the NHANES subjects was allowed to vary relative to the NHANES mean. The boxplots of the 2PL EAP scores are displayed in Figure 5.4. The NHANES subjects have a mean EAP of 0, while the MA users have a slightly higher mean score. Similarly, the MA subjects have a slightly higher mean Nominal EAP as compared with the NHANES subjects (as seen in Figure 5.5).

5.4.2 Differential Item Functioning

The test statistics for DIF in the 2PL model are shown in Figure 5.7. For all teeth, there is some significant degree of overall differential item functioning (Total X2), where p-values are all below 0.003. When examining DIF specifically in the slopes,

not all teeth had different discrimination parameters between MA and NHANES. The teeth with significantly different (at $\alpha = 0.05$) slopes were 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 18, 20, 21, 23, 24, 25, 26, 27, and 30. The teeth without significantly different slope parameters were 2, 3, 7, 10, 19, 22, 28, 29, and 31. There does not appear to be a clear pattern other than the majority of the teeth without different slopes appear to be on the right side. All of the teeth can be seen to have significantly different location parameters for MA users as compared with NHANES subjects. The difference in the location parameters can be thought of as the difference in the score at which 50% of the population would have a decayed, missing, or filled tooth for that particular tooth. The MA users have significantly lower location parameters for each tooth, and thus the DMF score at which 50% of the MA users would have a particular tooth be decayed, missing, or filled is significantly lower than the NHANES sample.

The DIF statistics for the Nominal model can be found in Figure 5.8. Similarly to the 2PL models, all teeth have significantly different overall item response models between MA and NHANES subjects. The location parameters appear to be the prominent source of distinctions between the study arms. While slope parameters and category parameters have some significant differences, the location parameters are all significantly different.

5.4.3 Item Information and Selection for Planned-Missingness

The item information of each tooth at specific values (under a 2PL model) of the latent domain are printed in Figure 5.9 for the combined MA-NHANES sample. At the tail ends of the distribution, but particularly for the lower end at $\theta = -2.8$ or -2.4, none of the tooth provide much information. The anterior teeth are the least likely to be informative at the lower end of the distribution, but the most likely to be informative at the higher disease statuses. Conversely decayed, missing, or filled molars are the most informative teeth at the lower tails of the distribution,

but are the least informative for higher disease status. These trends indicate that for little to no progression of disease, anterior teeth are the least likely to have any caries, while molars are more likely to be decayed missing or filled for those who have very little disease progression. However, for determining extreme caries disease, the presence of decayed, missing, or filled status on anterior teeth is less likely unless the disease progression is advanced. Similar item information patterns are found when using the nominal model (Figure 5.10). Anterior teeth provide little information at the lower ends of the distribution of latent disease status.

Selecting sets of teeth for planned-missing-data study designs would follow a similar process using either the nominal or 2PL model. To adequately cover the distribution of the latent disease, a mixture of both anterior and posterior teeth would need to be included. The half-mouth sampling by diagonal quadrants (as done in the periodontal examination) would satisfy this criteria.

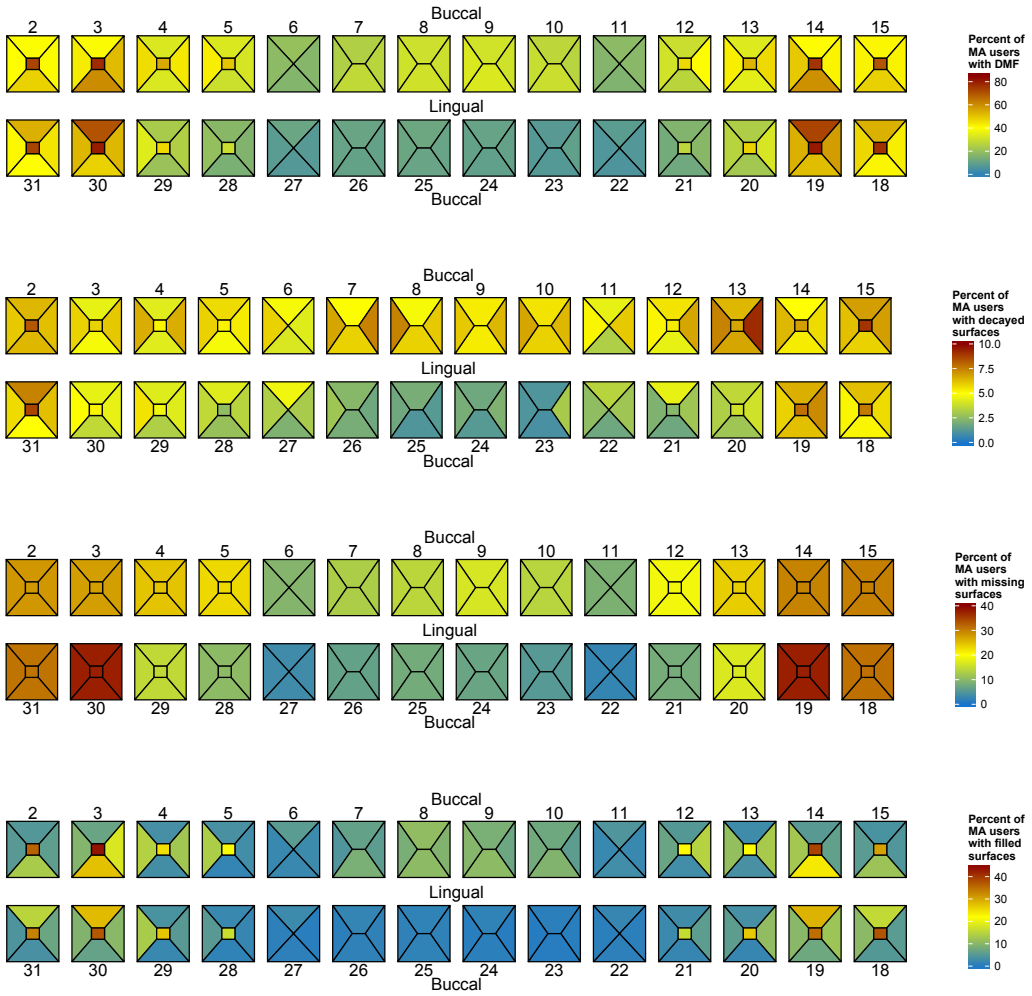

Figure 5.1: Percent of Decayed, Missing, and Filled surfaces among methamphetamine users

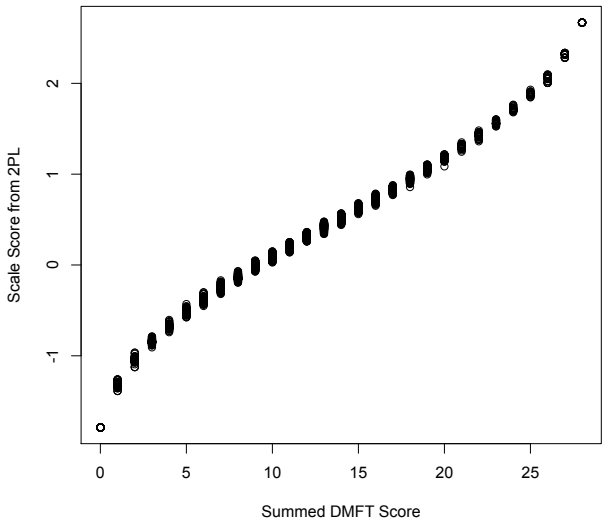

Figure 5.2: 2PL EAP Scale Scores vs. DMFT Summed Score

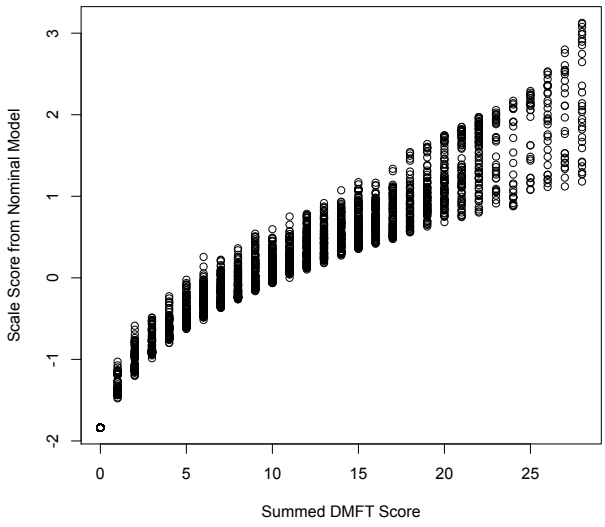

Figure 5.3: Nominal Scale Scores vs. DMFT Summed Score

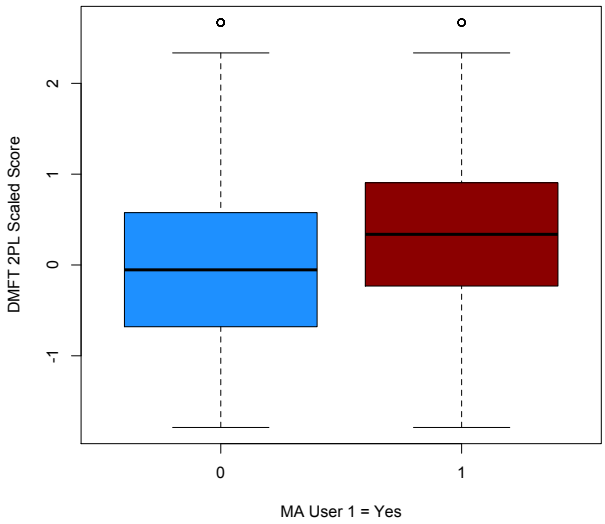

Figure 5.4: Boxplot of 2PL Scale Scores

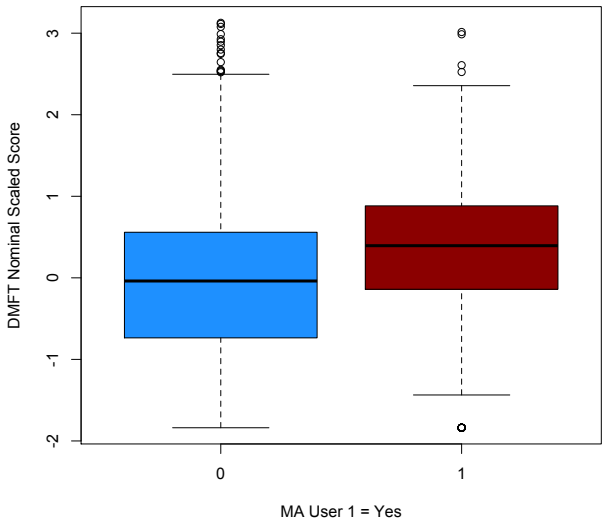

Figure 5.5: Boxplot of Nominal Scale Scores

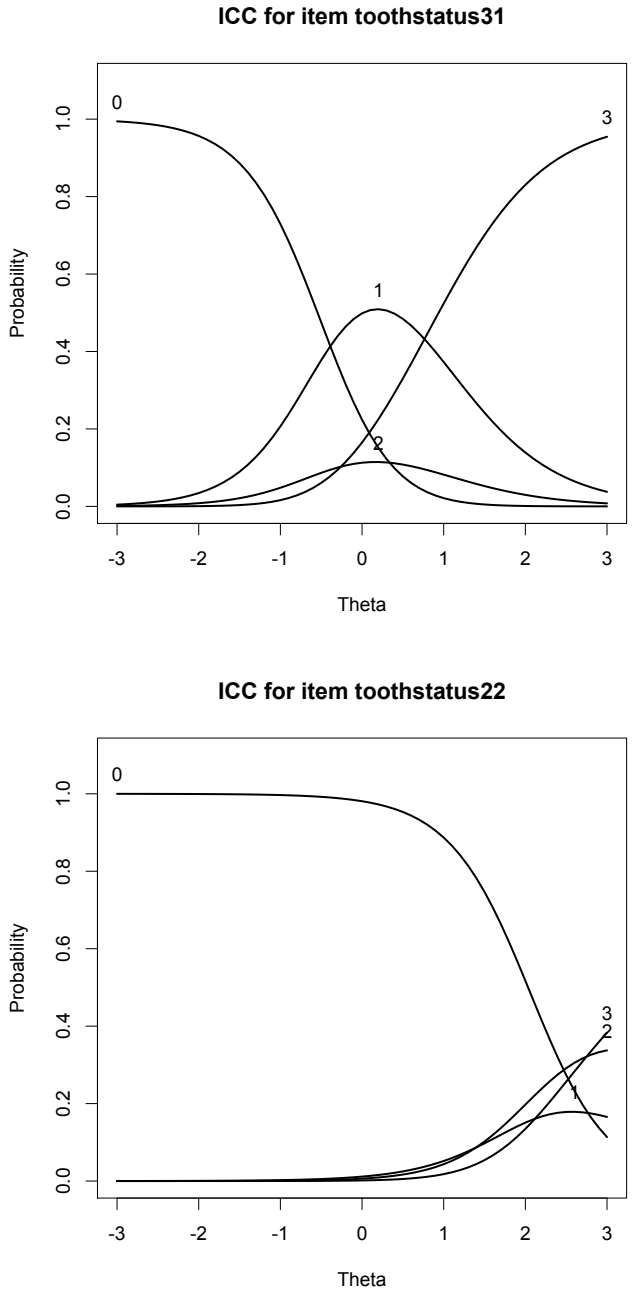

Figure 5.6: Example Item Characteristic Curves for Nominal Model

DIF Statistics for Graded Items
Item numbers in:

Grp1	Grp2	Total X2	d.f.	p	X2a	d.f.	p	X2c|a	d.f.	p
1	1	84.4	2	0.0001	1.4	1	0.2308	83.0	1	0.0001
2	2	62.2	2	0.0001	1.1	1	0.2950	61.1	1	0.0001
3	3	91.6	2	0.0001	5.6	1	0.0180	86.0	1	0.0001
4	4	83.2	2	0.0001	4.1	1	0.0431	79.1	1	0.0001
5	5	53.5	2	0.0001	21.9	1	0.0001	31.6	1	0.0001
6	6	27.2	2	0.0001	0.2	1	0.6702	27.1	1	0.0001
7	7	67.0	2	0.0001	13.9	1	0.0002	53.2	1	0.0001
8	8	79.6	2	0.0001	11.4	1	0.0007	68.2	1	0.0001
9	9	44.2	2	0.0001	3.0	1	0.0823	41.2	1	0.0001
10	10	14.6	2	0.0007	5.1	1	0.0236	9.5	1	0.0021
11	11	104.5	2	0.0001	20.0	1	0.0001	84.5	1	0.0001
12	12	113.5	2	0.0001	27.1	1	0.0001	86.5	1	0.0001
13	13	68.4	2	0.0001	7.0	1	0.0084	61.5	1	0.0001
14	14	93.0	2	0.0001	19.2	1	0.0001	73.8	1	0.0001
15	15	89.4	2	0.0001	8.7	1	0.0032	80.7	1	0.0001
16	16	73.9	2	0.0001	0.8	1	0.3623	73.1	1	0.0001
17	17	88.1	2	0.0001	10.3	1	0.0014	77.9	1	0.0001
18	18	65.0	2	0.0001	12.1	1	0.0005	52.9	1	0.0001
19	19	12.5	2	0.0020	1.7	1	0.1954	10.8	1	0.0010
20	20	14.3	2	0.0008	4.1	1	0.0441	10.2	1	0.0014
21	21	18.3	2	0.0001	5.4	1	0.0196	12.9	1	0.0003
22	22	27.0	2	0.0001	7.5	1	0.0061	19.5	1	0.0001
23	23	48.4	2	0.0001	19.9	1	0.0001	28.5	1	0.0001
24	24	28.9	2	0.0001	4.1	1	0.0429	24.9	1	0.0001
25	25	56.0	2	0.0001	1.2	1	0.2655	54.8	1	0.0001
26	26	67.5	2	0.0001	1.6	1	0.2019	65.9	1	0.0001
27	27	67.6	2	0.0001	3.5	1	0.0607	64.1	1	0.0001
28	28	54.2	2	0.0001	1.0	1	0.3134	53.2	1	0.0001

Figure 5.7: DIF Testing in 2PL Model

DIF Statistics for Nominal Items
Item numbers in:

Grp1	Grp2	Total X2	d.f.	p	X2s	d.f.	p	X2a|s	d.f.	p	X2c|a,s	d.f.	p
1	1	189.4	6	0.0001	8.0	2	0.0183	0.4	1	0.5122	180.9	3	0.0001
2	2	153.9	6	0.0001	7.8	2	0.0201	1.3	1	0.2500	144.8	3	0.0001
3	3	222.9	6	0.0001	7.5	2	0.0234	1.0	1	0.3163	214.4	3	0.0001
4	4	246.3	6	0.0001	11.7	2	0.0029	0.3	1	0.5772	234.3	3	0.0001
5	5	173.0	6	0.0001	10.8	2	0.0044	16.8	1	0.0001	145.4	3	0.0001
6	6	157.7	6	0.0001	16.4	2	0.0003	0.7	1	0.3936	140.6	3	0.0001
7	7	206.6	6	0.0001	10.4	2	0.0056	12.2	1	0.0005	184.0	3	0.0001
8	8	218.1	6	0.0001	9.7	2	0.0079	7.9	1	0.0050	200.5	3	0.0001
9	9	201.6	6	0.0001	12.1	2	0.0024	2.3	1	0.1265	187.2	3	0.0001
10	10	120.7	6	0.0001	5.7	2	0.0575	1.3	1	0.2542	113.7	3	0.0001
11	11	264.3	6	0.0001	5.4	2	0.0683	10.0	1	0.0015	248.9	3	0.0001
12	12	276.4	6	0.0001	1.2	2	0.5574	8.7	1	0.0031	266.5	3	0.0001
13	13	160.4	6	0.0001	3.6	2	0.1673	0.0	1	0.9426	156.8	3	0.0001
14	14	222.4	6	0.0001	9.6	2	0.0084	2.9	1	0.0878	209.9	3	0.0001
15	15	182.4	6	0.0001	14.2	2	0.0008	0.1	1	0.7979	168.1	3	0.0001
16	16	139.1	6	0.0001	5.4	2	0.0662	0.8	1	0.3848	132.9	3	0.0001
17	17	179.1	6	0.0001	3.9	2	0.1428	0.0	1	0.9764	175.2	3	0.0001
18	18	128.1	6	0.0001	1.9	2	0.3946	0.0	1	0.8564	126.2	3	0.0001
19	19	43.7	6	0.0001	1.9	2	0.3974	0.8	1	0.3683	41.0	3	0.0001
20	20	44.3	6	0.0001	2.2	2	0.3286	0.2	1	0.6343	41.9	3	0.0001
21	21	46.4	6	0.0001	0.8	2	0.6574	0.0	1	0.8640	45.6	3	0.0001
22	22	68.1	6	0.0001	1.8	2	0.4064	0.4	1	0.5307	65.9	3	0.0001
23	23	70.2	6	0.0001	2.3	2	0.3125	1.2	1	0.2726	66.7	3	0.0001
24	24	60.2	6	0.0001	0.1	2	0.9350	1.0	1	0.3104	59.1	3	0.0001
25	25	130.1	6	0.0001	5.4	2	0.0674	2.4	1	0.1246	122.3	3	0.0001
26	26	150.4	6	0.0001	5.2	2	0.0758	2.0	1	0.1578	143.3	3	0.0001
27	27	172.8	6	0.0001	12.7	2	0.0018	0.1	1	0.7480	160.0	3	0.0001
28	28	143.9	6	0.0001	11.8	2	0.0028	1.4	1	0.2410	130.7	3	0.0001

Figure 5.8: DIF Testing in Nominal Model

Item Information Function Values at 15 Values of theta from -2.8 to 2.8 for Group 1: MANHANES

Item	Label	-2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	toothstatus2	0.01	0.04	0.10	0.24	0.55	1.05	1.45	1.28	0.75	0.35	0.14	0.06	0.02	0.01	0.00
2	toothstatus3	0.03	0.07	0.19	0.46	0.95	1.47	1.44	0.90	0.42	0.17	0.07	0.02	0.01	0.00	0.00
3	toothstatus4	0.00	0.00	0.01	0.05	0.15	0.44	1.13	2.02	1.94	1.03	0.39	0.13	0.04	0.01	0.00
4	toothstatus5	0.00	0.00	0.01	0.03	0.09	0.28	0.81	1.78	2.23	1.42	0.57	0.19	0.06	0.02	0.01
5	toothstatus6	0.00	0.00	0.00	0.00	0.01	0.04	0.13	0.36	0.89	1.65	1.84	1.17	0.51	0.19	0.06
6	toothstatus7	0.00	0.00	0.01	0.02	0.05	0.12	0.30	0.66	1.16	1.43	1.13	0.63	0.28	0.12	0.05
7	toothstatus8	0.00	0.01	0.01	0.03	0.06	0.14	0.29	0.55	0.88	1.07	0.93	0.61	0.32	0.16	0.07
8	toothstatus9	0.00	0.00	0.01	0.03	0.06	0.13	0.29	0.58	0.95	1.16	0.98	0.61	0.31	0.14	0.06
9	toothstatus10	0.00	0.00	0.01	0.02	0.04	0.12	0.31	0.73	1.36	1.66	1.22	0.61	0.25	0.09	0.03
10	toothstatus11	0.00	0.00	0.00	0.01	0.02	0.05	0.14	0.38	0.94	1.72	1.83	1.12	0.47	0.17	0.06
11	toothstatus12	0.00	0.00	0.01	0.03	0.08	0.27	0.78	1.75	2.26	1.48	0.60	0.20	0.06	0.02	0.01
12	toothstatus13	0.00	0.01	0.02	0.05	0.15	0.43	1.04	1.79	1.78	1.02	0.42	0.15	0.05	0.02	0.01
13	toothstatus14	0.03	0.07	0.18	0.44	0.92	1.43	1.43	0.92	0.44	0.18	0.07	0.03	0.01	0.00	0.00
14	toothstatus15	0.01	0.03	0.09	0.23	0.55	1.09	1.54	1.35	0.78	0.35	0.14	0.05	0.02	0.01	0.00
15	toothstatus18	0.02	0.06	0.17	0.41	0.87	1.40	1.46	0.97	0.47	0.20	0.08	0.03	0.01	0.00	0.00
16	toothstatus19	0.04	0.10	0.27	0.64	1.26	1.66	1.31	0.69	0.29	0.11	0.04	0.01	0.01	0.00	0.00
17	toothstatus20	0.00	0.01	0.02	0.04	0.11	0.28	0.65	1.23	1.57	1.23	0.66	0.28	0.11	0.04	0.02
18	toothstatus21	0.00	0.00	0.01	0.01	0.03	0.08	0.20	0.46	0.90	1.30	1.27	0.83	0.42	0.18	0.07
19	toothstatus22	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.15	0.37	0.79	1.31	1.44	1.02	0.52
20	toothstatus23	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.05	0.15	0.42	1.04	1.83	1.83	1.04	0.42
21	toothstatus24	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.18	0.42	0.87	1.36	1.39	0.92	0.46
22	toothstatus25	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.17	0.39	0.78	1.20	1.28	0.91	0.49
23	toothstatus26	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.17	0.47	1.07	1.75	1.66	0.95	0.40
24	toothstatus27	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.18	0.43	0.86	1.31	1.33	0.89	0.45
25	toothstatus28	0.00	0.00	0.01	0.01	0.04	0.09	0.20	0.45	0.86	1.24	1.22	0.82	0.42	0.19	0.08
26	toothstatus29	0.00	0.01	0.02	0.04	0.11	0.29	0.67	1.24	1.55	1.20	0.64	0.28	0.11	0.04	0.02
27	toothstatus30	0.05	0.12	0.31	0.69	1.22	1.48	1.14	0.61	0.27	0.11	0.04	0.02	0.01	0.00	0.00
28	toothstatus31	0.02	0.06	0.16	0.40	0.87	1.41	1.49	0.99	0.48	0.20	0.08	0.03	0.01	0.00	0.00
	Test Information:	1.23	1.61	2.59	4.87	9.22	14.83	19.35	23.05	25.17	23.77	20.77	17.62	13.44	8.16	4.29
	Expected s.e.:	0.90	0.79	0.62	0.45	0.33	0.26	0.23	0.21	0.20	0.21	0.22	0.24	0.27	0.35	0.48

Marginal reliability for response pattern scores: 0.92

Figure 5.9: Item Information for 2PL Model

DMFT Graded(4) MA & NHANES Combined Full Fit Statistics
Calibration

Item Information Function Values at 15 Values of theta from -2.8 to 2.8 for Group 1: MANHANES

Item	Label	-2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	-0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	toothstatus2	0.02	0.05	0.11	0.24	0.50	0.90	1.28	1.40	1.35	1.33	1.18	0.81	0.44	0.21	0.09
2	toothstatus3	0.05	0.10	0.22	0.44	0.77	1.09	1.22	1.15	1.10	1.13	1.03	0.73	0.41	0.20	0.10
3	toothstatus4	0.00	0.00	0.01	0.04	0.14	0.41	1.04	1.97	2.42	2.37	2.06	1.20	0.49	0.17	0.06
4	toothstatus5	0.00	0.00	0.01	0.02	0.08	0.26	0.77	1.77	2.62	2.67	2.45	1.52	0.62	0.20	0.06
5	toothstatus6	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.39	1.18	2.52	3.24	2.96	1.75	0.66	0.20
6	toothstatus7	0.00	0.00	0.00	0.01	0.03	0.10	0.30	0.79	1.67	2.37	2.46	2.08	1.22	0.51	0.18
7	toothstatus8	0.00	0.00	0.00	0.01	0.04	0.11	0.30	0.75	1.47	2.04	2.14	1.88	1.20	0.56	0.22
8	toothstatus9	0.00	0.00	0.00	0.01	0.04	0.11	0.31	0.76	1.49	2.08	2.17	1.86	1.14	0.52	0.20
9	toothstatus10	0.00	0.00	0.00	0.01	0.03	0.09	0.29	0.87	1.97	2.83	2.87	2.31	1.18	0.43	0.13
10	toothstatus11	0.00	0.00	0.00	0.00	0.01	0.03	0.12	0.42	1.31	2.78	3.46	3.06	1.67	0.58	0.16
11	toothstatus12	0.00	0.00	0.01	0.02	0.07	0.24	0.73	1.74	2.65	2.73	2.50	1.55	0.62	0.20	0.06
12	toothstatus13	0.00	0.01	0.02	0.05	0.14	0.39	0.95	1.73	2.16	2.15	1.88	1.16	0.51	0.19	0.07
13	toothstatus14	0.04	0.10	0.21	0.43	0.77	1.11	1.26	1.19	1.14	1.17	1.04	0.71	0.39	0.19	0.09
14	toothstatus15	0.02	0.04	0.10	0.23	0.51	0.94	1.36	1.51	1.46	1.42	1.20	0.76	0.38	0.17	0.07
15	toothstatus18	0.05	0.10	0.21	0.41	0.72	1.04	1.20	1.18	1.13	1.12	0.95	0.65	0.36	0.18	0.08
16	toothstatus19	0.08	0.16	0.32	0.58	0.89	1.10	1.13	1.08	1.07	1.00	0.77	0.48	0.26	0.13	0.06
17	toothstatus20	0.00	0.01	0.02	0.05	0.12	0.28	0.58	1.01	1.37	1.46	1.43	1.30	0.92	0.50	0.24
18	toothstatus21	0.00	0.00	0.01	0.02	0.05	0.10	0.20	0.39	0.68	0.97	1.13	1.15	1.10	0.92	0.63
19	toothstatus22	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.07	0.14	0.26	0.45	0.70	0.92	1.02	1.02
20	toothstatus23	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.15	0.34	0.67	1.10	1.37	1.28	0.90
21	toothstatus24	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.08	0.17	0.35	0.64	0.99	1.18	1.06	0.74
22	toothstatus25	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.08	0.16	0.33	0.59	0.90	1.09	1.02	0.75
23	toothstatus26	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.08	0.17	0.35	0.65	1.02	1.25	1.20	0.89
24	toothstatus27	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.08	0.17	0.32	0.57	0.85	1.06	1.12	1.04
25	toothstatus28	0.00	0.00	0.01	0.02	0.05	0.10	0.20	0.39	0.68	0.98	1.14	1.16	1.11	0.94	0.64
26	toothstatus29	0.00	0.01	0.02	0.06	0.13	0.29	0.57	0.97	1.29	1.37	1.34	1.22	0.88	0.50	0.25
27	toothstatus30	0.09	0.18	0.35	0.60	0.89	1.06	1.07	1.01	1.01	0.97	0.77	0.49	0.27	0.14	0.07
28	toothstatus31	0.05	0.10	0.20	0.40	0.70	1.00	1.16	1.14	1.11	1.09	0.93	0.63	0.36	0.18	0.09
	Test Information:	1.40	1.88	2.85	4.70	7.71	11.88	17.35	25.06	34.27	41.50	42.75	36.20	25.16	15.98	10.06
	Expected s.e.:	0.85	0.73	0.59	0.46	0.36	0.29	0.24	0.20	0.17	0.16	0.15	0.17	0.20	0.25	0.32

Marginal reliability for response pattern scores: 0.93

Figure 5.10: Item Information for Nominal Model

CHAPTER 6

IRT models for Periodontal Examination Data

6.1 Motivation

When addressing issues of estimation arising when data are missing by design, defining what we are estimating is a crucial piece in the analysis strategy. Periodontal disease categorizations have been defined through CDC definitions which count the number of sites with certain thresholds. However, periodontal disease status or progression can be thought of as an unobserved, or latent, trait that is measured and defined by taking a number of measurements on a specific number of locations on the mouth. Another way of solving the underestimation problem is to find alternative ways of defining the measurement model.

In this research, item response theory models are proposed as a method for estimating the underlying disease status given the repeated measurements of the periodontal examination. In the next chapter, we will also discuss the extension of these models for scoring the caries examination and identifying teeth with surface conditions that may indicate methamphetamine use. In item response theory estimation, the model parameters are estimated given the observed data, even when number of items observed on each individual may vary. Using the estimated item parameters and observed data, we can approximate the posterior distributions of the individual disease status. Scores for the latent domain of interest, or disease status, can be estimated using some function of the posterior, commonly

the expected a posteriori or the maximum a posteriori. With planned-missing-data designs in which the inference is on a population rather than the individual, the individual posteriors may have wide variances. Thus instead of single point scores, multiple imputations from the posterior, or plausible values, can be used instead of point estimates.

6.2 Methods

Clinical attachment loss measurements were classified using the graded response model with five categories: 0 mm to 1 mm = 0, 2mm to 3 mm = 1, 4 mm to 5 mm = 2, 6+ mm = 3. For models to investigate the specific relationship between sites within teeth and the overall disease status, site level measurements were used. Other models defined items as the maximum attachment loss for the tooth, motivated by the traditional definitions of periodontal disease that require two or more sites at a given threshold from different teeth (thereby only incorporating information about the maximum attachment loss per tooth).

6.3 Calibration of IRT Models for MA Users Only

Item calibration and scoring was conducted using the software flexMIRT 2.0. The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was used for estimation of item parameters as well as generation of multiple imputations of the latent traits. The bifactor models were compared to the unidimensional models using likelihood ratio tests (the unidimensional models are nested within a given bifactor model, with the additional factor parameters constrained to zero). Residual dependence between items was tested in unidimensional models using LD χ^2

statistics with quadrature-based item calibration; however, current versions of the software using MH-RM for high dimensional models do not provide LD χ^2 statistics.

IRT models assume that given the value of the latent domains or vectors θ, that items are conditionally dependent. In this study that assumption may not be valid. While the periodontal disease status represents the primary latent domain, other factorizations are considered to account for residual dependence due to mouth location. In this analysis, bifactor models were calibrated for:

- Quadrants
- Clusters of teeth
- Tooth type
- Tooth type and location (mandibular/maxillary, left/right)

One specific aim of the study of methamphetamine users is to determine the relationship between methamphetamine use and oral health outcomes on a population scale. As such, individual disease status is not the target of inference, rather the estimation of regression coefficients and contrasts. The distribution of the latent disease status was assumed to follow a standard normal distribution for some models. However, a latent regression model to incorporate potential covariates in generating multiple imputations of the latent traits may be necessary to reduce bias in the population contrasts. Following the methodology developed by Mislevy et al. (1992) for large-scale educational assessments, the distribution of disease status can be assumed to have a conditionally normal distribution with $E(\theta_i) = \gamma x_i$, where γ represents the regression slopes onto background characteristics, denoted by x. The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was used for estimation of item parameters as well as generation of

multiple imputations of the latent traits. The MH-RM algorithm was adapted to incorporate covariates in the imputation step, thereby simultaneously estimating the latent regression parameters with the item parameters (Cai and Harrell, 2014, in progress).

6.3.1 Results for Methamphetamine Users

Unidimensional IRT models on either the maximum attachment loss of the four sites on each of 28 teeth or on all of the 112 site measurements showed strong associations between each measurement and the latent disease progression (Figures 3 & 5). The item slopes, or factor loadings, onto the general dimension for periodontal disease progression were positive and significantly different than zero in each of the models calibrated, including the bifactor models. However, tests of local dependence indicate that a single latent domain fails to account for all covariance between measurements. Bifactor models appeared to account for additional dependence, but more data would be necessary to address more complex structures. Table 1 includes summary information for several of the models calibrated on the data.

The unidimensional model using site measurements provides the highest marginal reliability for pattern scores, but the reliability is still high for the unidimensional model using maximum attachment loss per tooth. Reliability slightly decreases when regression covariates are included. When covariates were introduced in the conditioning model, the estimates for the slope parameters were consistently lower (Figure 4). Of the 8 background characteristics included, only two had significant associations with periodontal disease status.

6.4 Calibration of IRT models for both MA users and NHANES subjects

When calibrating models for two different groups, considerations should be made to as to whether the model is expected to behave differently for each group. When trying to arrive at comparable estimates of the latent periodontal disease to compare MA users with the non-using cohort selected from NHANES, ideally we would estimate the item parameters on the combined sample from both groups. However, if we want to examine the differences in the relationship between a tooth's attachment loss and periodontal disease between the two samples, we calibrate the IRT models separately and test for differential item functioning.

6.4.1 Differential Item Functioning between MA users and NHANES subjects

Differential item functioning was first examined by having both groups have equal means and dispersion of the latent disease status. Similar to the results seen in Chapter 5, all 28 teeth have significant differential item functioning. As seen in Figure 6.7, the major source of variation in the item parameters tends to be the location parameters. DIF between slope parameters vary; for molars, the slope parameters are not significantly different. However, for anterior teeth, the discrimination is significantly different between MA users and NHANES subjects. In fact, the slope parameters tend to be higher on the anterior teeth for the NHANES subjects than the MA users, signifying that the anterior teeth are more discriminating among the non-MA users. The location parameters are significantly different for all 28 teeth, and upon examination, it appears that the locations for each category of the graded model are lower for the MA users. This occurs because a higher proportion of MA users have more severe attachment loss at each tooth.

We can allow the means to differ by freeing the mean parameter for one group,

in this case NHANES. We recalibrate the model to compare the item parameters between the two groups with different means to check again for differential item functioning. The estimate of the mean periodontal disease status for NHANES subjects is -0.27, which is lower than the zero-valued fixed mean for the MA users. Not all teeth exhibited differential item functioning when the NHANES mean was freely estimated, as seen in Figure 6.8. The item response functions for tooth 5 and tooth 12 are not significantly different between the two groups. The remaining teeth are all significantly different between users and non-users, but it is not always the location parameter driving the difference in the item response function (although it usually is). For tooth number 11, the slope parameters are significantly different, but not the location parameters.

6.5 Concluding remarks

Item response theory models can be used as an alternative metric for inference about periodontal disease, particularly when partial-mouth-recording protocols are used and not all sites are observed. There do appear to be significant differences in the latent disease and the relationship between each site and the disease status between MA-using subjects and a matched non-using cohort.

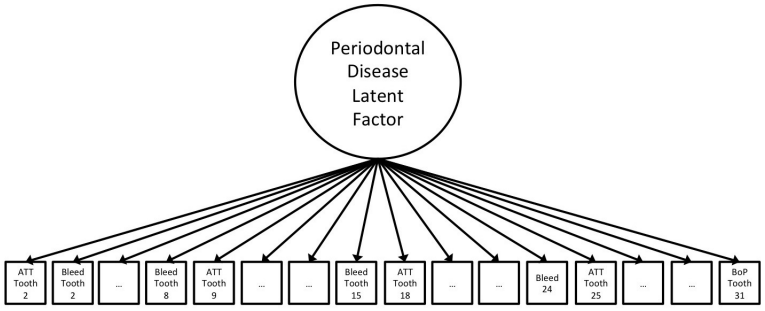

Figure 6.1: A unidimensional IRT model for tooth level data.

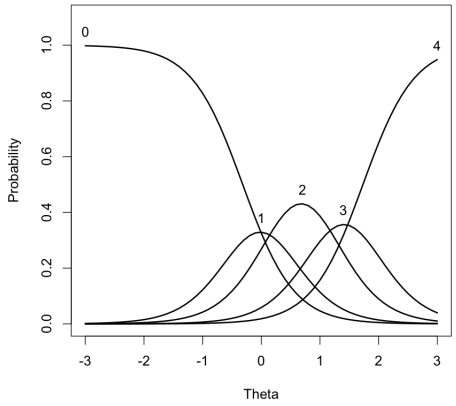

Figure 6.2: Item characteristic curve for the unidimensional graded response model for $K = 5$ levels counting the number of sites on Tooth 30 which have attachment loss $\geq 4mm$.

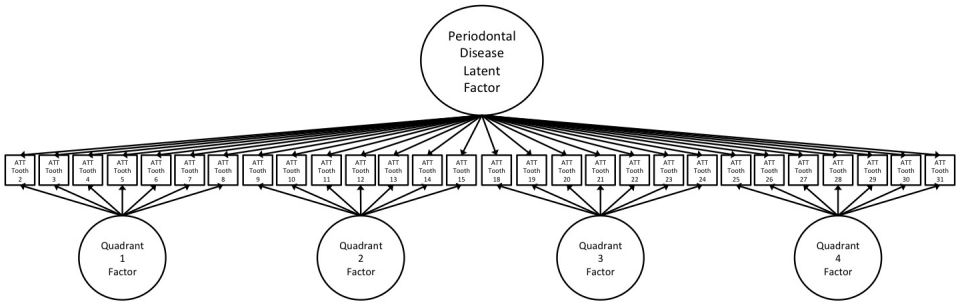

Figure 6.3: A bifactor model with four secondary factors (one factor for each quadrant) and one general factor

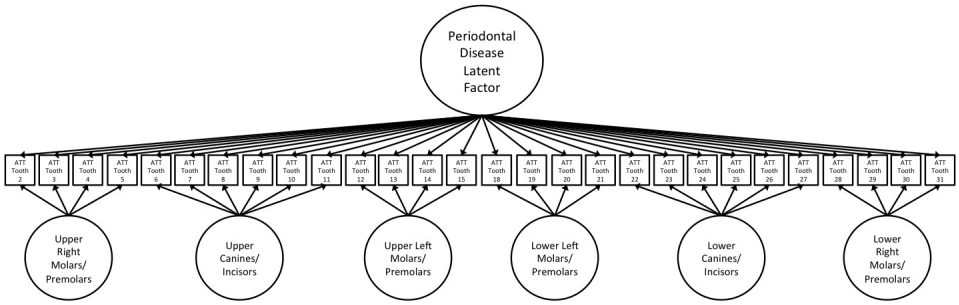

Figure 6.4: A bifactor model with seven secondary factors (one factor for each tooth type) and one general factor

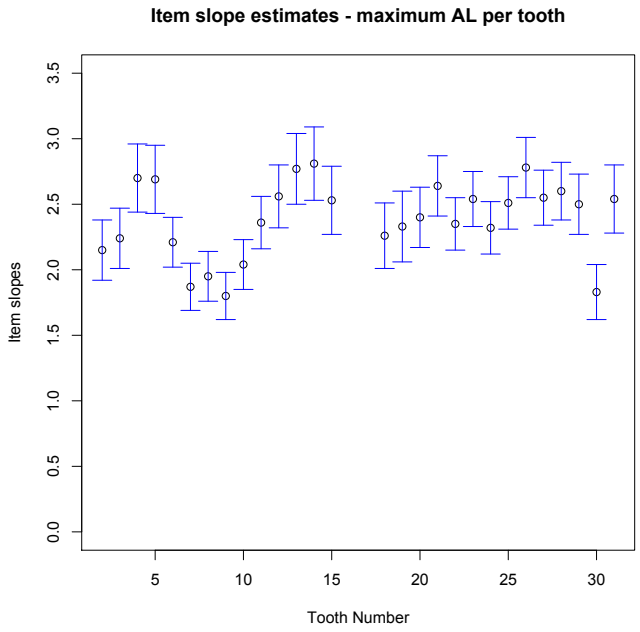

Figure 6.5: Item slopes for the unidimensional model with 28 items (28 maximum attachment loss per tooth)

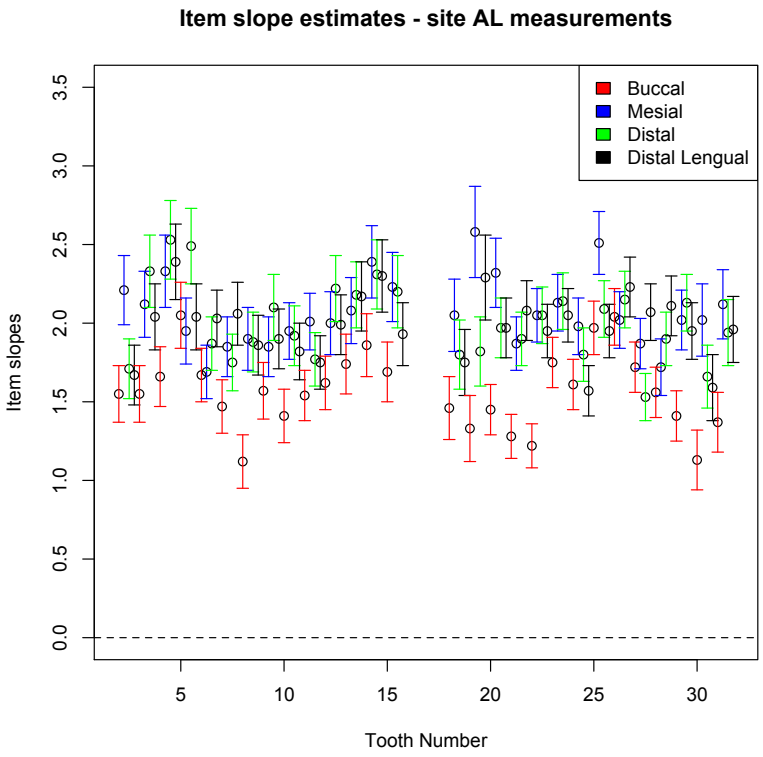

Figure 6.6: Item slopes for the unidimensional model with 112 items (categorized attachment loss on each site)

Attachment Loss Graded MA & NHANES Combined Full Fit Statistics
Calibration

DIF Statistics for Graded Items
Item numbers in:

Grp1	Grp2	Total X2	d.f.	p	X2a	d.f.	p	X2c|a	d.f.	p
1	1	81.5	4	0.0001	0.2	1	0.6798	81.3	3	0.0001
2	2	114.0	4	0.0001	0.1	1	0.7047	113.9	3	0.0001
3	3	84.3	4	0.0001	0.2	1	0.6731	84.1	3	0.0001
4	4	36.2	4	0.0001	0.5	1	0.4875	35.7	3	0.0001
5	5	75.2	4	0.0001	4.9	1	0.0268	70.3	3	0.0001
6	6	105.4	4	0.0001	5.6	1	0.0183	99.8	3	0.0001
7	7	139.5	4	0.0001	18.7	1	0.0001	120.8	3	0.0001
8	8	108.3	4	0.0001	41.1	1	0.0001	67.2	3	0.0001
9	9	101.6	4	0.0001	21.9	1	0.0001	79.7	3	0.0001
10	10	51.5	4	0.0001	13.6	1	0.0002	37.9	3	0.0001
11	11	46.3	4	0.0001	3.9	1	0.0483	42.4	3	0.0001
12	12	68.7	4	0.0001	0.8	1	0.3728	67.9	3	0.0001
13	13	86.1	4	0.0001	0.1	1	0.7667	86.1	3	0.0001
14	14	87.7	4	0.0001	0.8	1	0.3839	86.9	3	0.0001
15	15	86.6	4	0.0001	0.4	1	0.5260	86.2	3	0.0001
16	16	63.2	4	0.0001	1.6	1	0.2028	61.6	3	0.0001
17	17	108.5	4	0.0001	7.3	1	0.0069	101.2	3	0.0001
18	18	142.2	4	0.0001	5.7	1	0.0169	136.5	3	0.0001
19	19	125.5	4	0.0001	5.3	1	0.0215	120.2	3	0.0001
20	20	194.3	4	0.0001	12.0	1	0.0005	182.3	3	0.0001
21	21	187.7	4	0.0001	4.5	1	0.0345	183.2	3	0.0001
22	22	162.6	4	0.0001	8.1	1	0.0043	154.5	3	0.0001
23	23	182.8	4	0.0001	8.5	1	0.0036	174.3	3	0.0001
24	24	159.2	4	0.0001	16.5	1	0.0001	142.7	3	0.0001
25	25	91.2	4	0.0001	1.3	1	0.2628	89.9	3	0.0001
26	26	87.9	4	0.0001	0.1	1	0.8184	87.9	3	0.0001
27	27	38.4	4	0.0001	1.1	1	0.2847	37.3	3	0.0001
28	28	64.5	4	0.0001	0.0	1	0.8813	64.5	3	0.0001

Figure 6.7: Test statistics for differential item functioning for each tooth when latent means for both groups are equal

Group Parameter Estimates:

Group	Label	P#	mu	s.e.	P#	s2	s.e.	sd	s.e.
1	MA		0.00	----		1.00	----	1.00	----
2	NHANES		-0.27	----		1.00	----	1.00	----

Attachment Loss Graded MA & NHANES Combined Full Fit Statistics
Calibration

DIF Statistics for Graded Items
Item numbers in:

Grp1	Grp2	Total X2	d.f.	p	X2a	d.f.	p	X2c|a	d.f.	p
1	1	36.5	4	0.0001	0.2	1	0.6666	36.3	3	0.0001
2	2	54.0	4	0.0001	0.1	1	0.7038	53.8	3	0.0001
3	3	26.9	4	0.0001	0.2	1	0.6710	26.7	3	0.0001
4	4	4.0	4	0.4131	0.5	1	0.4828	3.5	3	0.3268
5	5	19.5	4	0.0006	5.0	1	0.0253	14.5	3	0.0023
6	6	38.6	4	0.0001	5.6	1	0.0176	33.0	3	0.0001
7	7	57.9	4	0.0001	18.5	1	0.0001	39.4	3	0.0001
8	8	53.9	4	0.0001	39.2	1	0.0001	14.7	3	0.0021
9	9	45.3	4	0.0001	21.6	1	0.0001	23.8	3	0.0001
10	10	17.9	4	0.0013	13.7	1	0.0002	4.1	3	0.2477
11	11	9.0	4	0.0608	4.1	1	0.0434	4.9	3	0.1784
12	12	18.3	4	0.0011	0.9	1	0.3570	17.4	3	0.0006
13	13	31.9	4	0.0001	0.1	1	0.7462	31.8	3	0.0001
14	14	38.3	4	0.0001	0.9	1	0.3466	37.5	3	0.0001
15	15	42.9	4	0.0001	0.4	1	0.5184	42.5	3	0.0001
16	16	29.8	4	0.0001	1.8	1	0.1833	28.0	3	0.0001
17	17	46.1	4	0.0001	8.1	1	0.0045	38.0	3	0.0001
18	18	60.8	4	0.0001	6.1	1	0.0136	54.7	3	0.0001
19	19	41.7	4	0.0001	5.7	1	0.0168	35.9	3	0.0001
20	20	87.5	4	0.0001	12.5	1	0.0004	75.0	3	0.0001
21	21	87.4	4	0.0001	4.8	1	0.0281	82.6	3	0.0001
22	22	63.7	4	0.0001	8.6	1	0.0034	55.1	3	0.0001
23	23	75.3	4	0.0001	9.0	1	0.0027	66.3	3	0.0001
24	24	58.9	4	0.0001	16.7	1	0.0001	42.2	3	0.0001
25	25	29.4	4	0.0001	1.5	1	0.2290	27.9	3	0.0001
26	26	31.0	4	0.0001	0.1	1	0.8111	30.9	3	0.0001
27	27	10.1	4	0.0389	1.3	1	0.2505	8.8	3	0.0326
28	28	26.4	4	0.0001	0.0	1	0.8647	26.4	3	0.0001

Figure 6.8: Test statistics for differential item functioning for each tooth when latent mean for NHANES subjects is freely estimated

Tooth	Overall Perio Disease	Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
2	λ_{12}	λ_{22}	0	0	0
3	λ_{13}	λ_{23}	0	0	0
4	λ_{14}	λ_{24}	0	0	0
5	λ_{15}	λ_{25}	0	0	0
6	λ_{16}	λ_{26}	0	0	0
7	λ_{17}	λ_{27}	0	0	0
8	λ_{18}	λ_{28}	0	0	0
9	λ_{19}	0	λ_{39}	0	0
10	λ_{110}	0	λ_{310}	0	0
11	λ_{111}	0	λ_{311}	0	0
12	λ_{112}	0	λ_{312}	0	0
13	λ_{113}	0	λ_{313}	0	0
14	λ_{114}	0	λ_{314}	0	0
15	λ_{115}	0	λ_{315}	0	0
18	λ_{118}	0	0	λ_{418}	0
19	λ_{119}	0	0	λ_{419}	0
20	λ_{120}	0	0	λ_{420}	0
21	λ_{121}	0	0	λ_{421}	0
22	λ_{122}	0	0	λ_{422}	0
23	λ_{123}	0	0	λ_{423}	0
24	λ_{124}	0	0	λ_{424}	0
25	λ_{125}	0	0	0	λ_{525}
26	λ_{126}	0	0	0	λ_{526}
27	λ_{127}	0	0	0	λ_{527}
28	λ_{128}	0	0	0	λ_{528}
29	λ_{129}	0	0	0	λ_{529}
30	λ_{130}	0	0	0	λ_{530}
31	λ_{131}	0	0	0	λ_{531}

Table 6.1: A bifactor model on tooth-level data with one general periodontal disease domain and four quadrant subdomains

CHAPTER 7

Future Research & Discussion

7.1 Future Research in Planned Missing Data in Oral Health

The work presented here is only the beginning in terms of analyzing planned missing data in oral health research. This section discusses ideas for how to expand this research.

7.1.1 Multiple Imputation from Hierarchical Spatial Models

In Chapter 4, multiple imputation analyses were performed using two-level hierarchical normal models for attachment loss and pocket depth, where measurement sites were nested within people. However, this hierarchical model did not account for the spatial structure inherent in the data. Reich et. al [RB10] pursued work characterizing the spatial processes involved in the periodontal examination, but these models could potentially be used to impute values of attachment loss and pocket depth for the unobserved teeth in planned-missing-data designs.

7.1.2 Spatial Models for the DMFT Index

In Chapter 6, we present IRT models for the DMF index. The DMF observations are categorical, but may also have a similar hierarchical/spatial pattern as the periodontal examination. An idea for future research could involve characterizing the spatial processes between surfaces in the DMF index.

7.2 Future Research involving IRT and Education Research

7.2.1 Approximating the Missing Information Matrix in MH-RM using Multiple Imputation

Failure of complex IRT models to converge to a maximum likelihood estimate was one issue found when estimating based on data with a high number of missing observations. In many instances, models failed to pass the second order test (negative second derivatives of the likelihood). This is likely the result of how the complete data information matrix is estimated in practice, which is by taking the sum of the observed and missing information matrices. However, in the case with a high number of missing observations, the estimated missing data information matrix can be negative. Here, we propose using the between-imputation variance to approximate the missing information matrix.

In the MH-RM algorithm, we impute values for the latent domain in the first step of each cycle. In practice, generally only one value of θ is imputed, but multiple imputations of θ can be drawn during this step. If we take multiple imputations of θ, we can approximate the matrix of missing data information by utilizing the between-imputation matrix and the total variance matrix.

7.2.2 Cognitive Diagnostic Models

Cognitive diagnostic models are similar to item response models except the latent proficiency domain is a binary variable rather than continuous. The use of cognitive diagnostic models and their application to the periodontal examination data as well as the NAEP framework may be explored in the future.

7.2.3 Incorporation of weighting for complex sample design from NAEP

It should be noted that the complex sampling design from NAEP will not be addressed in this research with the design weights. Typically, the sampling weights are incorporated in secondary analyses in estimation of population contrasts, and the level two weights for clusters are used for jack-knife variance estimates.

7.2.4 Prediction of statistical proficiency from NAEP Data

The statistical proficiency domain estimated in this disseration research could be quite valuable in surveying the existing skills of students in the United States and what factors are related to higher proficiency of data analysis. Using data from the High School Transcript Study, the association between particular courses, such as AP Biology, Statistics, or Calculus, and achievement levels could be useful towards crafting education policy for improvement of statistical learning. Contrasts on subpopulations may highlight where policy efforts should be targeted.

7.2.5 Sample size for planned-missing-data designs

Not addressed here are the sample sizes needed for a given planned-missing-data design. Future research could address through simulation the sample sizes necessary for an examination with a given number of items and proposed missing design under specific analysis models.

7.2.6 Longitudinal planned missingness

The models in this research are geared towards designs with a one-time observation of a study participant. However, in studies involving interventions with follow-ups, it may be desirable to limit the information collected at a given timepoint. Future extensions of this work would be to adapt the procedures to handle measurements

missing by design across time points.

CHAPTER 8

Appendix A - Simulated IRT and Regression Parameter Bias Tables

Table 8.1: Regression Parameters and Bias

Item	True Value	Model 1			Model 2			Model 3			Model 5		
		Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE
1	0.52	-0.22	-0.41	0.22	-0.28	-0.54	0.28	-0.28	-0.54	0.28	-0.29	-0.55	0.29
2	-0.47	0.20	-0.42	0.20	0.26	-0.55	0.26	0.26	-0.55	0.26	0.26	-0.56	0.26
3	-1.63	0.71	-0.44	0.71	0.94	-0.58	0.94	0.93	-0.57	0.93	0.96	-0.58	0.96
4	-2.31	1.02	-0.44	1.02	1.34	-0.58	1.34	1.33	-0.58	1.33	1.36	-0.59	1.36
5	0.97	-0.42	-0.43	0.42	-0.55	-0.57	0.55	-0.55	-0.56	0.55	-0.56	-0.58	0.56
6	-0.36	0.15	-0.41	0.15	0.19	-0.54	0.19	0.19	-0.53	0.19	0.20	-0.54	0.20
7	0.19	-0.09	-0.48	0.09	-0.12	-0.63	0.12	-0.12	-0.63	0.12	-0.12	-0.64	0.12
8	1.59	-0.69	-0.43	0.69	-0.90	-0.57	0.90	-0.89	-0.56	0.89	-0.91	-0.57	0.91
9	0.85	-0.39	-0.46	0.39	-0.52	-0.61	0.52	-0.52	-0.61	0.52	-0.53	-0.62	0.53
10	-0.96	0.42	-0.43	0.42	0.55	-0.57	0.55	0.54	-0.56	0.54	0.55	-0.58	0.55
11	-0.30	0.12	-0.41	0.12	0.16	-0.54	0.16	0.16	-0.53	0.16	0.16	-0.54	0.16
12	-1.74	0.76	-0.44	0.76	1.00	-0.58	1.00	1.00	-0.57	1.00	1.02	-0.59	1.02
13	-0.20	0.09	-0.44	0.09	0.11	-0.56	0.11	0.11	-0.56	0.11	0.11	-0.57	0.12
14	0.26	-0.10	-0.38	0.10	-0.13	-0.50	0.13	-0.12	-0.49	0.13	-0.13	-0.50	0.13
15	2.40	-1.06	-0.44	1.06	-1.40	-0.58	1.40	-1.38	-0.58	1.39	-1.42	-0.59	1.42
16	2.13	-0.94	-0.44	0.94	-1.23	-0.58	1.23	-1.22	-0.57	1.22	-1.25	-0.59	1.25
17	0.07	-0.01	-0.23	0.02	-0.02	-0.24	0.02	-0.02	-0.24	0.02	-0.01	-0.23	0.02
18	0.09	-0.04	-0.47	0.04	-0.05	-0.61	0.05	-0.05	-0.60	0.05	-0.05	-0.62	0.06
19	0.24	-0.10	-0.41	0.10	-0.13	-0.54	0.13	-0.13	-0.54	0.13	-0.13	-0.55	0.14
20	2.24	-0.96	-0.43	0.96	-1.26	-0.56	1.26	-1.25	-0.56	1.25	-1.28	-0.57	1.28
21	0.54	-0.23	-0.42	0.23	-0.30	-0.56	0.30	-0.30	-0.56	0.30	-0.31	-0.57	0.31
22	0.62	-0.26	-0.43	0.26	-0.35	-0.56	0.35	-0.34	-0.56	0.34	-0.35	-0.57	0.35
23	-1.04	0.46	-0.44	0.46	0.61	-0.59	0.61	0.60	-0.58	0.60	0.62	-0.60	0.62
24	0.92	-0.41	-0.44	0.41	-0.54	-0.59	0.54	-0.53	-0.58	0.53	-0.55	-0.60	0.55
25	-1.54	0.68	-0.44	0.68	0.89	-0.58	0.89	0.89	-0.57	0.89	0.91	-0.59	0.91
26	0.53	-0.23	-0.44	0.23	-0.30	-0.58	0.30	-0.30	-0.57	0.30	-0.31	-0.58	0.31
27	-2.73	1.18	-0.43	1.18	1.55	-0.57	1.55	1.54	-0.56	1.54	1.57	-0.58	1.57
28	1.12	-0.49	-0.44	0.49	-0.65	-0.58	0.65	-0.64	-0.58	0.64	-0.66	-0.59	0.66
29	2.16	-0.93	-0.43	0.93	-1.22	-0.56	1.22	-1.21	-0.56	1.21	-1.24	-0.57	1.24
30	0.38	-0.17	-0.45	0.17	-0.23	-0.60	0.23	-0.23	-0.59	0.23	-0.23	-0.61	0.23
31	1.29	-0.55	-0.43	0.55	-0.74	-0.57	0.74	-0.73	-0.57	0.73	-0.75	-0.58	0.75
32	-0.72	0.31	-0.43	0.31	0.40	-0.56	0.41	0.40	-0.56	0.40	0.41	-0.57	0.41
33	-1.62	0.69	-0.43	0.69	0.93	-0.57	0.93	0.92	-0.57	0.92	0.94	-0.58	0.94
34	-0.75	0.32	-0.42	0.32	0.42	-0.56	0.42	0.42	-0.55	0.42	0.42	-0.56	0.42
35	1.00	-0.43	-0.43	0.43	-0.56	-0.56	0.56	-0.56	-0.56	0.56	-0.57	-0.57	0.57
36	-1.12	0.48	-0.43	0.48	0.64	-0.57	0.64	0.63	-0.57	0.63	0.65	-0.58	0.65
37	-1.43	0.63	-0.44	0.63	0.84	-0.59	0.84	0.83	-0.58	0.83	0.85	-0.60	0.85
38	0.50	-0.19	-0.39	0.19	-0.25	-0.50	0.25	-0.25	-0.49	0.25	-0.25	-0.50	0.25
39	-1.53	0.67	-0.44	0.67	0.89	-0.59	0.90	0.89	-0.58	0.89	0.91	-0.60	0.91
40	-0.83	0.35	-0.43	0.35	0.47	-0.56	0.47	0.46	-0.56	0.46	0.47	-0.57	0.47
41	-1.27	0.55	-0.43	0.55	0.73	-0.58	0.73	0.73	-0.57	0.73	0.74	-0.59	0.74
42	0.01	-0.02	-2.95	0.03	-0.04	-4.50	0.04	-0.04	-4.56	0.04	-0.04	-4.80	0.04
43	-0.21	0.09	-0.43	0.09	0.12	-0.58	0.12	0.12	-0.58	0.12	0.12	-0.59	0.12
44	1.90	-0.82	-0.43	0.82	-1.10	-0.58	1.10	-1.09	-0.58	1.09	-1.12	-0.59	1.12
45	0.53	-0.21	-0.40	0.21	-0.29	-0.54	0.29	-0.28	-0.53	0.28	-0.29	-0.54	0.29

Table 8.2: Regression Parameters and Bias (Continued)

Item	True Value	Model 1			Model 2			Model 3			Model 5		
		Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE	Bias	Rel Bias	RMSE
46	1.33	-0.57	-0.43	0.57	-0.76	-0.57	0.76	-0.75	-0.56	0.75	-0.77	-0.58	0.77
47	1.69	-0.74	-0.43	0.74	-0.98	-0.58	0.98	-0.98	-0.58	0.98	-1.00	-0.59	1.00
48	-0.25	0.11	-0.44	0.11	0.14	-0.59	0.15	0.14	-0.59	0.14	0.15	-0.60	0.15
49	0.42	-0.18	-0.42	0.18	-0.24	-0.56	0.24	-0.23	-0.56	0.23	-0.24	-0.57	0.24
50	3.32	-1.42	-0.43	1.42	-1.90	-0.57	1.90	-1.88	-0.57	1.88	-1.93	-0.58	1.93
51	1.61	-0.70	-0.43	0.70	-0.93	-0.58	0.93	-0.93	-0.58	0.93	-0.95	-0.59	0.95
52	0.57	-0.24	-0.42	0.24	-0.31	-0.56	0.31	-0.31	-0.55	0.31	-0.32	-0.56	0.32
53	-0.09	0.03	-0.37	0.04	0.05	-0.49	0.05	0.04	-0.49	0.05	0.05	-0.50	0.05
54	-0.24	0.11	-0.46	0.11	0.15	-0.63	0.15	0.15	-0.63	0.15	0.16	-0.64	0.16
55	-0.71	0.30	-0.42	0.30	0.40	-0.56	0.40	0.39	-0.55	0.39	0.40	-0.57	0.40
56	0.76	-0.33	-0.43	0.33	-0.44	-0.58	0.44	-0.43	-0.58	0.43	-0.45	-0.59	0.45
57	-3.03	1.29	-0.43	1.29	1.72	-0.57	1.72	1.71	-0.56	1.71	1.75	-0.58	1.75
58	0.06	-0.02	-0.28	0.02	-0.02	-0.34	0.02	-0.02	-0.33	0.02	-0.02	-0.34	0.02
59	2.77	-1.18	-0.43	1.18	-1.57	-0.57	1.57	-1.56	-0.56	1.56	-1.59	-0.58	1.59
60	-0.26	0.12	-0.45	0.12	0.15	-0.58	0.15	0.15	-0.58	0.15	0.15	-0.59	0.15
61	1.55	-0.75	-0.49	0.75	-0.95	-0.62	0.95	-0.95	-0.61	0.95	-0.97	-0.62	0.97
62	-1.18	0.57	-0.48	0.57	0.73	-0.61	0.73	0.72	-0.61	0.72	0.74	-0.62	0.74
63	-0.81	0.38	-0.47	0.38	0.48	-0.59	0.48	0.48	-0.58	0.48	0.48	-0.59	0.48
64	-0.71	0.33	-0.47	0.33	0.42	-0.59	0.42	0.41	-0.58	0.41	0.42	-0.59	0.42
65	1.52	-0.74	-0.48	0.74	-0.93	-0.61	0.93	-0.93	-0.61	0.93	-0.95	-0.62	0.95
66	-1.10	0.53	-0.48	0.53	0.67	-0.61	0.67	0.67	-0.61	0.67	0.68	-0.62	0.68
67	0.38	-0.20	-0.51	0.20	-0.25	-0.66	0.25	-0.25	-0.66	0.25	-0.26	-0.68	0.26
68	3.56	-1.74	-0.49	1.74	-2.21	-0.62	2.21	-2.19	-0.62	2.19	-2.24	-0.63	2.24
69	-0.78	0.38	-0.49	0.38	0.48	-0.62	0.48	0.47	-0.61	0.47	0.49	-0.63	0.49
70	-1.07	0.52	-0.48	0.52	0.65	-0.61	0.65	0.65	-0.61	0.65	0.66	-0.62	0.66
71	-0.65	0.31	-0.48	0.31	0.39	-0.61	0.39	0.39	-0.60	0.39	0.39	-0.61	0.40
72	-1.93	0.93	-0.49	0.93	1.19	-0.62	1.19	1.18	-0.61	1.18	1.21	-0.63	1.21
73	-0.23	0.11	-0.49	0.11	0.14	-0.62	0.14	0.14	-0.62	0.14	0.14	-0.63	0.14
74	0.46	-0.21	-0.46	0.21	-0.26	-0.57	0.26	-0.26	-0.57	0.26	-0.26	-0.58	0.26
75	0.57	-0.26	-0.46	0.26	-0.33	-0.58	0.33	-0.33	-0.57	0.33	-0.33	-0.58	0.33
76	0.73	-0.34	-0.47	0.34	-0.42	-0.58	0.42	-0.42	-0.58	0.42	-0.43	-0.59	0.43
77	0.93	-0.44	-0.48	0.44	-0.56	-0.61	0.56	-0.56	-0.60	0.56	-0.57	-0.61	0.57
78	0.10	-0.05	-0.50	0.05	-0.06	-0.64	0.06	-0.06	-0.63	0.06	-0.06	-0.65	0.06
79	0.67	-0.32	-0.48	0.32	-0.41	-0.61	0.41	-0.40	-0.61	0.40	-0.41	-0.62	0.41
80	3.18	-1.53	-0.48	1.53	-1.94	-0.61	1.94	-1.92	-0.61	1.92	-1.96	-0.62	1.96
81	-0.18	0.10	-0.57	0.10	0.13	-0.75	0.13	0.13	-0.75	0.13	0.14	-0.77	0.14
82	1.34	-0.66	-0.49	0.66	-0.83	-0.62	0.83	-0.83	-0.62	0.83	-0.84	-0.63	0.85
83	-0.08	0.03	-0.39	0.03	0.03	-0.46	0.04	0.03	-0.46	0.04	0.03	-0.45	0.04
84	0.31	-0.15	-0.48	0.15	-0.19	-0.60	0.19	-0.19	-0.60	0.19	-0.19	-0.61	0.19
85	-0.73	0.34	-0.47	0.34	0.43	-0.60	0.43	0.43	-0.59	0.43	0.44	-0.60	0.44
86	0.00	0.01	3.49	0.01	0.01	6.44	0.02	0.01	6.52	0.02	0.01	6.97	0.02
87	-2.83	1.36	-0.48	1.36	1.71	-0.61	1.71	1.70	-0.60	1.70	1.74	-0.61	1.74
88	0.63	-0.30	-0.48	0.30	-0.38	-0.61	0.38	-0.38	-0.60	0.38	-0.38	-0.61	0.39
89	3.75	-1.81	-0.48	1.81	-2.30	-0.61	2.30	-2.28	-0.61	2.28	-2.33	-0.62	2.33
90	-0.22	0.11	-0.50	0.11	0.14	-0.63	0.14	0.14	-0.63	0.14	0.14	-0.64	0.14

Table 8.3: 2PL Content Slope Parameters and Bias

Item	True Value	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias
1	1.078	0.82	0.76	1.54	1.43	1.49	1.38	6.86	6.36	1.18	1.09	5.73	5.31
2	1.317	0.97	0.74	1.74	1.32	1.66	1.26	8.65	6.57	1.65	1.25	8.10	6.15
3	1.993	1.81	0.91	3.48	1.75	3.19	1.60	15.71	7.88	3.02	1.52	13.42	6.73
4	1.039	0.80	0.77	1.54	1.48	1.45	1.39	6.68	6.43	1.14	1.10	5.60	5.39
5	1.418	1.07	0.75	2.43	1.71	2.13	1.50	10.75	7.58	1.62	1.14	8.18	5.77
6	1.758	1.60	0.91	3.13	1.78	2.92	1.66	13.46	7.66	2.55	1.45	11.47	6.52
7	1.388	1.06	0.77	2.01	1.45	1.89	1.36	8.98	6.47	1.74	1.25	8.00	5.76
8	2.003	1.51	0.76	3.04	1.52	2.81	1.40	14.13	7.05	2.86	1.43	13.47	6.72
9	1.334	1.23	0.93	2.24	1.68	2.11	1.58	9.74	7.30	1.81	1.36	8.42	6.31
10	1.589	1.18	0.74	2.28	1.43	2.06	1.30	10.71	6.74	1.94	1.22	9.60	6.04
11	1.533	1.20	0.78	2.32	1.52	2.14	1.39	10.90	7.11	1.82	1.19	8.58	5.60
12	2.159	1.64	0.76	3.40	1.57	3.04	1.41	15.90	7.36	2.55	1.18	12.69	5.88
13	1.35	1.24	0.92	2.78	2.06	2.43	1.80	11.04	8.18	1.70	1.26	7.93	5.87
14	2.105	1.63	0.78	3.43	1.63	3.13	1.49	14.96	7.11	2.84	1.35	12.81	6.08
15	1.431	1.07	0.74	1.93	1.35	1.79	1.25	9.36	6.54	1.65	1.15	8.39	5.86
16	1.45	1.31	0.91	2.27	1.56	2.20	1.52	10.34	7.13	2.02	1.39	9.32	6.42
17	1.561	1.21	0.77	2.05	1.31	1.97	1.26	9.49	6.08	1.75	1.12	8.40	5.38
18	0.813	0.60	0.73	1.05	1.29	1.02	1.26	5.19	6.38	1.01	1.24	4.98	6.13
19	1.681	1.56	0.93	2.59	1.54	2.46	1.47	11.86	7.06	2.17	1.29	10.21	6.07
20	1.215	0.90	0.74	1.59	1.31	1.54	1.26	7.82	6.44	1.46	1.20	7.34	6.04
21	1.365	1.06	0.78	1.77	1.30	1.72	1.26	8.40	6.16	1.61	1.18	7.62	5.58
22	1.334	0.99	0.74	1.71	1.28	1.65	1.24	8.51	6.38	1.62	1.22	8.01	6.00
23	1.647	1.53	0.93	2.94	1.78	2.77	1.68	11.92	7.24	2.01	1.22	9.50	5.77
24	1.768	1.36	0.77	2.46	1.39	2.35	1.33	11.33	6.41	1.97	1.12	9.54	5.40
25	1.619	1.22	0.76	2.11	1.30	2.03	1.26	10.48	6.47	1.92	1.19	9.53	5.89
26	1.395	1.28	0.92	2.06	1.47	1.98	1.42	9.58	6.87	1.75	1.26	8.33	5.97
27	1.35	1.04	0.77	1.89	1.40	1.79	1.32	8.30	6.15	1.57	1.16	7.35	5.44
28	1.887	1.41	0.75	2.40	1.27	2.28	1.21	11.94	6.33	2.20	1.16	11.09	5.88
29	1.414	1.29	0.91	2.12	1.50	2.05	1.45	9.80	6.93	1.95	1.38	9.02	6.38
30	2.067	1.57	0.76	2.86	1.38	2.74	1.32	13.62	6.59	2.84	1.37	13.37	6.47
31	1.463	1.11	0.76	1.96	1.34	1.89	1.29	8.94	6.11	1.85	1.27	8.43	5.77
32	1.356	1.01	0.75	1.96	1.44	1.85	1.36	9.17	6.76	1.45	1.07	7.52	5.54
33	2.081	1.96	0.94	3.60	1.73	3.40	1.64	16.02	7.70	3.26	1.57	14.33	6.88
34	1.185	0.89	0.75	1.57	1.33	1.52	1.28	7.30	6.16	1.44	1.22	6.78	5.72
35	0.94	0.71	0.75	1.27	1.35	1.23	1.31	6.16	6.55	1.09	1.16	5.49	5.84
36	0.848	0.77	0.91	1.36	1.60	1.31	1.55	6.05	7.14	1.21	1.43	5.51	6.50
37	1.215	0.92	0.76	1.62	1.34	1.58	1.30	7.55	6.21	1.43	1.18	6.77	5.57
38	1.647	1.21	0.73	2.20	1.34	2.11	1.28	10.82	6.57	1.97	1.20	9.85	5.98
39	1.527	1.39	0.91	2.28	1.49	2.19	1.44	10.45	6.84	2.19	1.43	9.87	6.47
40	1.58	1.18	0.75	2.20	1.39	2.11	1.33	10.48	6.64	2.03	1.29	9.76	6.18
41	1.576	1.19	0.75	2.21	1.40	2.12	1.35	9.78	6.20	1.78	1.13	8.49	5.39
42	1.318	0.98	0.75	1.72	1.30	1.67	1.27	8.54	6.48	1.48	1.12	7.51	5.70
43	1.747	1.62	0.92	2.75	1.58	2.68	1.53	12.35	7.07	2.43	1.39	10.97	6.28
44	1.79	1.38	0.77	2.48	1.39	2.38	1.33	11.36	6.35	2.17	1.21	10.17	5.68
45	1.12	0.83	0.74	1.56	1.39	1.50	1.34	7.46	6.66	1.27	1.14	6.34	5.66
46	1.702	1.55	0.91	2.78	1.63	2.64	1.55	12.42	7.30	2.52	1.48	11.17	6.57
47	1.484	1.16	0.78	2.00	1.35	1.94	1.31	9.19	6.19	1.52	1.02	7.60	5.12
48	1.462	1.10	0.75	2.04	1.39	1.97	1.35	9.64	6.59	2.01	1.38	9.48	6.49
49	1.212	1.09	0.90	1.86	1.53	1.80	1.49	8.41	6.94	1.72	1.42	7.75	6.39
50	1.589	1.17	0.74	2.14	1.35	2.05	1.29	10.39	6.54	2.04	1.28	9.90	6.23

Table 8.4: 2PL Content Slope Parameters and Bias (continued)

Item	True Value	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias
51	1.773	1.35	0.76	2.46	1.39	2.34	1.32	11.18	6.31	2.12	1.20	9.75	5.50
52	1.177	0.88	0.75	1.58	1.34	1.55	1.31	7.48	6.36	1.35	1.14	6.54	5.55
53	1.677	1.53	0.91	2.64	1.57	2.48	1.48	11.61	6.92	2.05	1.22	9.74	5.81
54	1.48	1.13	0.77	1.87	1.26	1.81	1.22	8.94	6.04	1.77	1.20	8.33	5.63
55	1.419	1.06	0.75	2.00	1.41	1.94	1.37	9.65	6.80	1.53	1.08	7.77	5.48
56	1.035	0.94	0.91	1.67	1.62	1.61	1.55	7.31	7.06	1.35	1.31	6.27	6.06
57	1.712	1.32	0.77	2.43	1.42	2.33	1.36	10.89	6.36	2.37	1.38	10.55	6.16
58	1.599	1.20	0.75	2.20	1.38	2.12	1.32	10.61	6.63	2.12	1.32	10.08	6.31
59	1.581	1.49	0.94	2.90	1.84	2.71	1.71	12.05	7.62	2.14	1.35	9.78	6.19
60	1.107	0.82	0.74	1.44	1.30	1.40	1.26	7.11	6.43	1.26	1.14	6.43	5.81
61	0.859	0.65	0.76	1.20	1.40	1.16	1.35	5.40	6.29	0.90	1.05	4.39	5.11
62	0.949	0.71	0.75	1.27	1.33	1.23	1.29	6.15	6.48	1.21	1.28	5.78	6.09
63	1.312	1.22	0.93	2.10	1.60	2.03	1.55	9.36	7.13	1.97	1.51	8.64	6.58
64	1.564	1.20	0.77	2.28	1.46	2.15	1.37	10.31	6.59	1.78	1.14	8.44	5.39
65	1.727	1.28	0.74	2.41	1.39	2.25	1.30	11.31	6.55	1.78	1.03	9.15	5.30
66	1.144	1.05	0.92	1.79	1.56	1.73	1.51	8.04	7.03	1.62	1.42	7.29	6.37
67	1.562	1.18	0.75	2.23	1.43	2.14	1.37	9.83	6.30	1.54	0.98	7.69	4.92
68	0.878	0.65	0.74	1.17	1.33	1.13	1.29	5.71	6.51	1.07	1.22	5.30	6.04
69	1.434	1.30	0.91	2.35	1.64	2.26	1.57	10.20	7.11	1.84	1.28	8.47	5.91
70	1.135	0.85	0.75	1.59	1.40	1.53	1.35	7.71	6.80	1.35	1.19	6.74	5.94
71	1.615	1.24	0.77	2.14	1.33	2.06	1.27	9.91	6.13	1.87	1.16	8.72	5.40
72	1.111	0.82	0.74	1.54	1.39	1.48	1.34	7.37	6.64	1.27	1.14	6.38	5.74
73	1.889	1.75	0.92	3.44	1.82	3.21	1.70	14.34	7.59	2.60	1.38	11.62	6.15
74	1.441	1.11	0.77	2.08	1.44	1.98	1.38	9.37	6.50	1.67	1.16	7.93	5.50
75	1.079	0.80	0.74	1.40	1.30	1.36	1.26	6.95	6.44	1.30	1.20	6.39	5.92
76	1.349	1.24	0.92	2.29	1.70	2.18	1.61	10.21	7.57	1.99	1.47	8.90	6.60
77	1.473	1.12	0.76	2.20	1.49	2.11	1.43	9.51	6.46	1.95	1.32	8.53	5.79
78	1.192	0.88	0.74	1.59	1.33	1.54	1.29	7.86	6.59	1.49	1.25	7.40	6.21
79	1.336	1.24	0.93	2.27	1.70	2.10	1.57	9.49	7.11	1.73	1.30	8.10	6.07
80	1.375	1.01	0.74	1.91	1.39	1.82	1.32	9.38	6.82	1.72	1.25	8.53	6.20
81	0.75	0.58	0.77	1.02	1.37	0.98	1.31	4.51	6.02	0.71	0.95	3.61	4.81
82	1.399	1.05	0.75	1.95	1.39	1.86	1.33	9.46	6.76	1.58	1.13	7.99	5.71
83	1.444	1.31	0.91	2.36	1.63	2.23	1.54	10.48	7.26	2.06	1.43	9.30	6.44
84	1.098	0.84	0.76	1.77	1.62	1.69	1.54	7.32	6.67	1.27	1.15	5.68	5.17
85	1.61	1.17	0.73	2.39	1.48	2.27	1.41	11.07	6.88	1.77	1.10	8.97	5.57
86	1.742	1.61	0.92	2.81	1.61	2.65	1.52	12.33	7.08	2.37	1.36	10.83	6.22
87	1.565	1.20	0.77	2.34	1.50	2.24	1.43	9.82	6.27	2.03	1.29	8.73	5.58
88	1.678	1.24	0.74	2.26	1.34	2.16	1.29	10.81	6.44	2.24	1.34	10.69	6.37
89	1.527	1.40	0.92	2.50	1.64	2.41	1.58	11.27	7.38	2.32	1.52	10.27	6.73
90	1.969	1.48	0.75	3.03	1.54	2.89	1.47	13.91	7.07	2.81	1.43	13.10	6.65
91	1.201	0.92	0.77	1.62	1.34	1.57	1.31	7.39	6.16	1.49	1.24	6.78	5.64
92	1.254	0.93	0.74	1.84	1.47	1.74	1.38	8.51	6.79	1.60	1.27	7.64	6.09
93	1.15	1.05	0.91	1.93	1.68	1.86	1.62	8.57	7.45	1.66	1.45	7.57	6.59
94	1.527	1.17	0.77	2.13	1.39	2.04	1.34	9.51	6.23	1.97	1.29	8.69	5.69
95	1.229	0.91	0.74	1.63	1.33	1.54	1.26	7.89	6.42	1.20	0.98	6.41	5.21
96	1.561	1.46	0.93	2.60	1.66	2.48	1.59	11.55	7.40	2.39	1.53	10.53	6.74
97	1.141	0.86	0.75	1.68	1.47	1.58	1.39	7.07	6.20	1.29	1.13	6.02	5.27
98	1.064	0.82	0.77	1.56	1.46	1.48	1.39	7.29	6.85	1.48	1.39	6.92	6.51
99	1.698	1.54	0.91	2.94	1.73	2.74	1.62	12.61	7.43	2.51	1.48	11.10	6.54
100	1.592	1.19	0.75	2.43	1.52	2.27	1.42	11.11	6.98	2.18	1.37	10.17	6.39

Table 8.5: 2PL Content Slope Parameters and Bias (continued)

Item	True Value	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias	Bias	Rel Bias
101	0.882	0.68	0.77	1.25	1.41	1.20	1.36	5.67	6.43	1.00	1.14	4.74	5.37
102	0.991	0.73	0.73	1.40	1.41	1.35	1.36	6.62	6.68	1.25	1.26	6.04	6.10
103	1.813	1.61	0.89	2.77	1.53	2.62	1.45	12.89	7.11	2.53	1.40	11.44	6.31
104	1.297	1.00	0.77	1.88	1.45	1.83	1.41	8.38	6.46	1.74	1.34	7.68	5.93
105	0.88	0.65	0.74	1.22	1.39	1.18	1.34	5.65	6.42	0.90	1.03	4.72	5.36
106	2.337	2.17	0.93	4.46	1.91	4.02	1.72	18.94	8.11	3.87	1.66	16.46	7.05
107	1.854	1.46	0.79	2.77	1.49	2.49	1.34	12.08	6.52	2.33	1.25	10.55	5.69
108	1.663	1.24	0.75	2.45	1.48	2.35	1.41	12.11	7.28	2.19	1.32	10.74	6.46
109	1.296	1.18	0.91	2.11	1.62	2.02	1.56	9.56	7.38	1.92	1.48	8.69	6.70
110	1.385	1.05	0.76	1.97	1.42	1.88	1.36	9.38	6.77	1.86	1.35	9.02	6.51
111	1.238	0.95	0.77	1.84	1.49	1.70	1.37	7.86	6.35	1.28	1.03	6.29	5.08
112	1.987	1.47	0.74	2.88	1.45	2.72	1.37	13.99	7.04	2.25	1.13	11.30	5.68
113	1.793	1.62	0.90	3.54	1.98	3.20	1.79	14.68	8.19	2.51	1.40	11.41	6.36
114	1.597	1.21	0.76	2.79	1.75	2.52	1.58	11.30	7.08	1.80	1.13	8.24	5.16
115	1.966	1.46	0.74	2.82	1.44	2.62	1.33	13.74	6.99	2.13	1.08	10.99	5.59
116	1.577	1.40	0.89	2.77	1.76	2.53	1.61	12.34	7.83	2.19	1.39	10.15	6.44
117	1.724	1.34	0.78	2.73	1.59	2.52	1.46	11.46	6.65	2.33	1.35	10.31	5.98
118	1.343	1.01	0.76	1.95	1.45	1.85	1.38	9.28	6.91	1.63	1.21	8.23	6.12
119	1.416	1.31	0.92	2.44	1.72	2.27	1.60	10.27	7.25	1.61	1.14	7.88	5.56
120	1.558	1.17	0.75	2.13	1.36	2.02	1.30	10.28	6.60	2.04	1.31	9.96	6.39
121	1.109	0.85	0.76	1.67	1.51	1.60	1.44	7.41	6.68	1.38	1.24	6.28	5.66
122	1.417	1.05	0.74	2.17	1.53	1.98	1.40	10.17	7.18	1.71	1.21	8.51	6.00
123	1.544	1.44	0.93	3.07	1.99	2.86	1.85	12.87	8.33	2.43	1.57	10.37	6.72
124	1.037	0.79	0.76	1.60	1.54	1.55	1.49	6.86	6.62	1.37	1.32	6.19	5.97
125	1.249	0.94	0.75	2.05	1.64	1.93	1.54	9.07	7.26	1.49	1.19	7.42	5.94
126	1.66	1.51	0.91	3.37	2.03	2.91	1.75	13.30	8.01	2.38	1.43	10.33	6.22
127	1.067	0.81	0.76	1.65	1.54	1.57	1.48	6.96	6.52	1.22	1.14	5.74	5.38
128	1.774	1.34	0.75	2.78	1.57	2.58	1.45	12.51	7.05	2.43	1.37	11.45	6.46
129	1.641	1.48	0.90	3.13	1.91	2.82	1.72	13.17	8.03	2.26	1.37	10.02	6.11
130	1.762	1.33	0.75	2.54	1.44	2.40	1.36	11.99	6.81	1.92	1.09	9.85	5.59
131	1.066	0.82	0.77	1.56	1.46	1.50	1.40	6.85	6.42	1.42	1.33	6.30	5.91
132	1.132	0.85	0.75	1.68	1.48	1.57	1.39	7.51	6.63	1.28	1.13	6.43	5.68
133	1.493	1.35	0.90	2.82	1.89	2.60	1.74	11.46	7.67	2.22	1.49	9.65	6.46
134	1.614	1.25	0.78	2.76	1.71	2.65	1.64	11.22	6.95	2.34	1.45	9.65	5.98
135	1.546	1.17	0.76	2.29	1.48	2.05	1.33	10.74	6.95	1.92	1.24	9.34	6.04
136	1.649	1.48	0.90	3.27	1.98	2.72	1.65	13.21	8.01	2.36	1.43	10.67	6.47
137	1.714	1.34	0.78	3.51	2.05	3.10	1.81	12.03	7.02	2.61	1.52	10.72	6.25
138	1.394	1.05	0.75	2.08	1.49	1.96	1.41	9.84	7.06	1.75	1.25	8.45	6.06
139	1.438	1.31	0.91	2.53	1.76	2.37	1.65	10.45	7.27	2.09	1.46	9.26	6.44
140	1.199	0.89	0.75	1.88	1.57	1.77	1.47	7.86	6.55	1.68	1.40	7.80	6.50
141	1.072	0.81	0.75	1.76	1.64	1.64	1.53	7.46	6.96	1.25	1.17	5.81	5.42
142	0.902	0.66	0.74	1.33	1.47	1.26	1.40	5.89	6.53	1.10	1.21	5.10	5.65
143	1.611	1.47	0.91	3.05	1.89	2.61	1.62	12.16	7.55	1.94	1.21	9.08	5.64
144	1.071	0.82	0.77	1.85	1.72	1.66	1.55	7.14	6.67	1.29	1.21	5.89	5.50
145	1.395	1.03	0.74	2.50	1.79	2.18	1.56	12.21	8.76	1.64	1.18	8.07	5.78
146	1.9	1.69	0.89	3.30	1.74	2.97	1.56	15.27	8.04	2.59	1.36	11.80	6.21
147	1.279	0.95	0.75	2.00	1.56	1.90	1.49	8.32	6.51	1.89	1.47	8.14	6.36
148	1.726	1.28	0.74	2.74	1.59	2.53	1.47	11.83	6.85	2.31	1.34	11.02	6.38
149	1.487	1.38	0.93	3.20	2.15	2.86	1.92	11.37	7.65	2.29	1.54	9.80	6.59
150	2.609	2.00	0.77	5.65	2.17	4.51	1.73	19.11	7.33	4.66	1.79	18.69	7.16

References

[BA81] R Darrell Bock and Murray Aitkin. "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm." *Psychometrika*, **46**(4):443–459, 1981.

[BHB13] Daniel J Bauer, Andrea L Howard, Ruth E Baldasaro, Patrick J Curran, Andrea M Hussong, Laurie Chassin, and Robert A Zucker. "A trifactor model for integrating ratings across multiple informants." *Psychological methods*, **18**(4):475, 2013.

[BM82] R Darrell Bock and Robert J Mislevy. "Adaptive EAP estimation of ability in a microcomputer environment." *Applied psychological measurement*, **6**(4):431–444, 1982.

[Cai08] Li Cai. *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. ProQuest, 2008.

[Cai10] Li Cai. "A two-tier full-information item factor analysis model with applications." *Psychometrika*, **75**(4):581–612, 2010.

[CJ99] Jon D Cohen and Tao Jiang. "Comparison of partially measured latent traits across nominal subgroups." *Journal of the American Statistical Association*, **94**(448):1035–1044, 1999.

[CSH11] Li Cai, Ji Seung Yang, and Mark Hansen. "Generalized full-information item bifactor analysis." *Psychological methods*, **16**(3):221, 2011.

[DEK02] Sherie A Dowsett, George J Eckert, and Michael J Kowolik. "The applicability of half-mouth examination to periodontal disease assessment in untreated adult populations." *Journal of Periodontology*, **73**(9):975–981, 2002.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[DS10] Matthias von Davier and Sandip Sinharay. "Stochastic approximation methods for latent regression item response models." *Journal of Educational and Behavioral Statistics*, **35**(2):174–193, 2010.

[EDW12] PI Eke, BA Dye, L Wei, GO Thornton-Evans, and RJ Genco. "Prevalence of Periodontitis in Adults in the United States: 2009 and 2010." *Journal of Dental Research*, **91**(10):914–920, 2012.

[ERC02] Phil Edwards, Ian Roberts, Mike Clarke, Carolyn DiGuiseppi, Sarah Pratap, Reinhard Wentz, and Irene Kwan. "Increasing response rates to postal questionnaires: systematic review." *BMJ*, **324**(7347):1183, 2002.

[ETW10] PI Eke, GO Thornton-Evans, L Wei, WS Borgnakke, and BA Dye. "Accuracy of NHANES periodontal examination protocols." *Journal of dental research*, **89**(11):1208–1213, 2010.

[GTC01] John W Graham, Bonnie J Taylor, and Patricio E Cumsille. "Planned missing-data designs in analysis of change." 2001.

[HC13] Carrie R Houts and Li Cai. "flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring Users Manual Version 2.0." 2013.

[Hun87] RJ Hunt. "The efficiency of half-mouth examinations in estimating the prevalence of periodontal disease." *Journal of Dental Research*, **66**(5):1044–1048, 1987.

[LR02] R Little and D Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002.

[Mai12] Samopriyo Maitra. *Applications of Circular Distributions and Spatial Point Processes to the Analysis of Periodontal Data*. PhD thesis, The University of Michigan, 2012.

[Mis91] Robert J Mislevy. "Randomization-based inference about latent variables from complex samples." *Psychometrika*, **56**(2):177–196, 1991.

[MJM92a] RJ Milsevy, E Johnson, and E Muraki. "Scaling Procedures in NAEP." *Journal of Educational and Behavioral Statistics*, **17**, 1992.

[MJM92b] Robert J Mislevy, Eugene G Johnson, and Eiji Muraki. "Scaling procedures in NAEP." *Journal of Educational and Behavioral Statistics*, **17**(2):131–154, 1992.

[MKB79] Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic press, 1979.

[Mur92] Eiji Muraki. "A generalized partial credit model: Application of an EM algorithm." *Applied psychological measurement*, **16**(2):159–176, 1992.

[NAE11] "NAEP Technical Documentation.", 2011.

[NAE12] "The Nation's Report Card: Science 2011.", 2012.

[PPS11] Bhavna T Pahel, John S Preisser, Sally C Stearns, and R Gary Rozier. "Multiple imputation of dental caries data using a zero-inflated Poisson regression model." *Journal of public health dentistry*, **71**(1):71–78, 2011.

[PR10] Raymond F Palmer and Donald R Royall. "Missing data? Plan on it!" *Journal of the American Geriatrics Society*, **58**(s2):S343–S348, 2010.

[RB10] Brian J Reich and Dipankar Bandyopadhyay. "A latent factor model for spatial data with informative missingness." *The Annals of Applied Statistics*, **4**(1):439, 2010.

[RG95] TE Raghunathan and JE Grizzle. "A Split Questionnaire Survey Design." *Journal of the American Statistical Association*, **90**, 1995.

[RJD14a] F. Rijmen, M. Jeon, M. von Davier, and S. Rabe-Hesketh. *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, chapter A General Psychometric Approach for Educational Survey Assessments: Flexible Statistical Models and Efficient Estimation. CRC Press, Boca Raton, FL, 2014.

[RJD14b] Frank Rijmen, Minjeong Jeon, Matthias von Davier, and Sophia Rabe-Hesketh. "A Third-Order Item Response Theory Model for Modeling the Effects of Domains and Subdomains in Large-Scale Educational Assessment Surveys." *Journal of Educational and Behavioral Statistics*, p. 1076998614531045, 2014.

[RR83] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, **70**(1):41–55, 1983.

[RR84] Paul R Rosenbaum and Donald B Rubin. "Reducing bias in observational studies using subclassification on the propensity score." *Journal of the American Statistical Association*, **79**(387):516–524, 1984.

[RR85] Paul R Rosenbaum and Donald B Rubin. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician*, **39**(1):33–38, 1985.

[Rub76] Donald B Rubin. "Inference and missing data." *Biometrika*, **63**(3):581–592, 1976.

[Sam69] Fumiko Samejima. "Estimation of latent ability using a response pattern of graded scores." *Psychometrika Monograph Supplement*, 1969.

[SB14] AA Schuller and S van Buuren. "Estimation of caries experience by multiple imputation and direct standardization." *Caries research*, **48**(2):91–+, 2014.

[SD05] Sandip Sinharay and Matthias von Davier. "Extension of the NAEP BGROUP program to higher dimensions." *Research Report-Educational Testing Service*, **5**, 2005.

[SEM09] Amir Savage, Kenneth A Eaton, David R Moles, and Ian Needleman. "A systematic review of definitions of periodontitis and methods that have been used to identify this disease." *Journal of clinical periodontology*, **36**(6):458–467, 2009.

[SI02] Steven L Scott and Edward H Ip. "Empirical Bayes and item-clustering effects in a latent variable hierarchical model: a case study from the National Assessment of Educational Progress." *Journal of the American Statistical Association*, **97**(458):409–419, 2002.

[SJT14] Lynne Steuerle Schofield, Brian Junker, Lowell J Taylor, and Dan A Black. "Predictive inference using latent variables with covariates." *Psychometrika*, pp. 1–21, 2014.

[SKA05] Cristiano Susin, Albert Kingman, and Jasim M Albandar. "Effect of partial recording protocols on estimates of prevalence of periodontal disease." *Journal of Periodontology*, **76**(2):262–267, 2005.

[TG97] N Thomas and N Gan. "Generating Multiple Imputations for Matrix Sampling Data Analyzed with Item Response Models." *Journal of Educational and Behavioral Statistics*, **22**, 1997.

[TGD13] Duong T Tran, Isabel Gay, Xianglin L Du, Yunxin Fu, Richard D Bebermeyer, Ana S Neumann, Charles Streckfus, Wenyaw Chan, and Muhammad F Walji. "Assessing periodontitis in populations: a systematic review of the validity of partial-mouth examination protocols." *Journal of clinical periodontology*, **40**(12):1064–1071, 2013.

[TGD14] Duong T Tran, Isabel Gay, Xianglin L Du, Yunxin Fu, Richard D Bebermeyer, Ana S Neumann, Charles Streckfus, Wenyaw Chan, and Muhammad F Walji. "Assessment of partial-mouth periodontal examination protocols for periodontitis surveillance." *Journal of clinical periodontology*, **41**(9):846–852, 2014.

[Tho00] N Thomas. "Assessing Model Sensitivity of the Imputation Methods Used in the National Assessment of Educational Progress." *Journal of Educational and Behavioral Statistics*, **25**, 2000.

[TSW93] David Thissen, Lynne Steinberg, and Howard Wainer. "Detection of differential item functioning using the parameters of item response models." 1993.

[VBG06] Stef Van Buuren, Jaap PL Brand, CGM Groothuis-Oudshoorn, and Donald B Rubin. "Fully conditional specification in multivariate imputation." *Journal of statistical computation and simulation*, **76**(12):1049–1064, 2006.

[Yen84] Wendy M Yen. "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model." *Applied Psychological Measurement*, **8**(2):125–145, 1984.