A Bivariate Longitudinal Model for Psychometric Data

by

Matthew Berkowitz

B.Com., University of British Columbia, 2009

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Matthew Berkowitz**
Degree:	**Master of Science (Statistics)**
Title:	**A Bivariate Longitudinal Model for Psychometric Data**
Examining Committee:	**Chair:** Jinko Graham Professor
	Rachel Altman Senior Supervisor Associate Professor
	Joan Hu Professor Supervisor
	Thomas Loughin Professor External Examiner
Date Defended:	**April 30, 2020**

Abstract

Psychometric test data are useful for predicting a variety of important life outcomes and personality characteristics. The Cognitive Reflection Test (CRT) is a short, well-validated rationality test, designed to assess subjects' ability to override intuitively appealing but incorrect responses to a series of math- and logic-based questions. The CRT is predictive of many other cognitive abilities and tendencies, such as verbal intelligence, numeracy, and religiosity. Cognitive psychologists and psychometricians are concerned with whether subjects improve their scores on the test with repeated exposure, as this may threaten the test's predictive validity.

This project uses the first publicly available longitudinal dataset derived from subjects who took the CRT multiple times over a predefined period. The dataset includes a multitude of predictors, including number of previous exposures to the test (our variable of primary interest). Also included are two response variables measured with each test exposure: CRT score and time taken to complete the CRT. These responses serve as a proxy for underlying latent variables, "rationality" and "reflectiveness", respectively. We propose methods to describe the relationship between the responses and selected predictors. Specifically, we employ a bivariate longitudinal model to account for the presumed dependence between our two responses. Our model also allows for subpopulations ("clusters") of individuals whose responses exhibit similar patterns. We estimate the parameters of our one- and two-cluster models via adaptive Gaussian quadrature. We also develop an Expectation-Maximization algorithm for estimating models with greater numbers of clusters.

We use our fitted models to address a range of subject-specific questions in a formal way (building on earlier work relying on ad hoc methods). In particular, we find that test exposure has a greater estimated effect on test scores than previously reported and we find evidence of at least two subpopulations. Additionally, our work has generated numerous avenues for future investigation.

Keywords: Bivariate Longitudinal Model; Cluster Model, EM Algorithm, Gaussian Quadrature, Adaptive Quadrature; Mixed Model; Cognitive Reflection Test

Acknowledgements

I want to express my sincerest gratitude to Rachel Altman, a spectacular supervisor—and friend—who encouraged and challenged me every step of the way. I came away from our meetings feeling energized, motivated, and in positive spirits. Our conversations spanned much more than just statistics, traversing topics in philosophy, psychology, morality, religion, politics, and beer. We often agreed, but when we didn't, it was at least as enjoyable and perhaps even more fruitful—always in the common spirit of exploring ideas and pursuing truth. Rachel, thank you—I am extremely lucky and grateful to have you in my life.

Moreover, I want to thank all the professors I had the fortune to be instructed by during the MSc program: Derek Bingham, Joan Hu, Richard Lockhart, Tom Loughin, and of course, Rachel. A special thanks to Marie Loughin for doing an impeccable job managing us TAs—it was a pleasure working with you. To my fellow graduate students, especially those in my awesome cohort, thank you for making the program such an enriching, entertaining, and supportive experience. My sincerest appreciation goes to Megan Kurz for partnering on all class projects, never-ending support—especially the tech support—and for being a great friend.

Last but definitely not least, I want to thank my parents for their incredible support and encouragement throughout the program. To my wife, Rachel (a different Rachel!), thank you for putting up with my incessant stats talk and for being my rock.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Cognitive Reflection Test (CRT) (Frederick, 2005) was developed to assess a subject's "reflectiveness", operationalized in the cognitive psychology literature as the ability to override an incorrect but intuitively appealing response (a so-called "gut instinct"). The CRT is a short, three-question test that is predictive of many cognitive abilities and tendencies (Bialek and Pennycook, 2018). It was a precursor to the Comprehensive Assessment of Rational Thinking (CART), a more in-depth "rationality" test currently being developed (Stanovich et al., 2016). "Rationality" subsumes the construct of "reflectiveness" by referring to the ability to override intuitive responses *to obtain a correct answer*, as operationalized on the CART.

Part of this literature is concerned with disentangling the concepts of "intelligence" (as measured by Intelligence Quotient [IQ] tests) and "rationality" (as measured by the CRT or CART). Of particular interest to researchers is whether subjects tend to improve their scores over time (for example, via repeated exposure to the same test questions), in which case the tests may not retain their predictive validity. With respect to IQ, the literature provides no convincing evidence that IQ scores improve in the long-term (Haier, 2014). But, with respect to rationality scores, the literature is so far sparse. The first study to assess this question was Meyer et al. (2018), who administered the CRT to subjects multiple times over a predefined time period. We use the data from that longitudinal study in the present work.

Our project extends the work of Meyer et al. (2018), who used conventional linear regression modelling in an attempt to answer various questions about changes in subjects' CRT scores over time. These models did not sufficiently take into account the longitudinal nature of the data, the dependence among responses measured on the same individual, or the discreteness of the test scores. Though Meyer et al. (2018) intimates that the CRT dataset suggests

the presence of subpopulations, their models do not account for them. To address these limitations, we develop a bivariate longitudinal model to describe the relationship between various predictors (including measures of prior exposure to the test) and two dependent response variables: subjects' score and time spent completing the test. We conceive of the random effects in this model as representing reflectiveness and rationality. We also present an extension of this model that allows a different bivariate longitudinal model for different subpopulations of individuals via a latent cluster variable.

Our model extends the generalized linear mixed model (Agresti, 2013) to include a second response variable. Our approach is similar to that of Kondo et al. (2017), who proposed a multiple longitudinal outcome mixture model that incorporates random effects (REs) and clusters. We use adaptive Gaussian quadrature (AGQ) to estimate the parameters of our one- and two-cluster models. We also develop an Expectation-Maximization (EM) algorithm to estimate the parameters of our multi-cluster models.

The rest of this paper is organized as follows. In Chapter 2, we describe the CRT dataset. In Chapter 3, we present our models and estimation method. In Chapter 4, we use our models to address questions concerning the effect of prior exposure and compare our findings with those provided by Meyer et al. (2018). We conclude with a discussion of the analyses' limitations and possible future work in Chapter 5.

Chapter 2

Cognitive Reflection Test (CRT) Data

2.1 CRT Dataset Overview

The individuals in this study comprised over 14,000 subjects from Amazon Mechanical Turk (MTurk)—a crowdsourcing website where volunteers can participate in tasks—and over 28,000 observations across four separate series of surveys. (See Appendix A for a discussion of the reliability of MTurk samples.) The data were collected from November 2013 to April 2015. We chose the largest series, Fall 2014 (which included observations from Sept. 3, 2014 to Jan. 12, 2015), to be the focus of our present work. The raw dataset is available publicly from the Judgment and Decision Making journal's website (http://journal.sjdm.org/vol13.3.html).

After data wrangling (see Sections 2.2–2.4), the Fall 2014 series consisted of 6,228 observations on 2,920 unique subjects. The number of times that subjects took the test varied, ranging from 1 to 15 within this series. Figure 2.1 summarizes the distribution of this variable.

2.2 Responses of Interest

Meyer et al. (2018) treated CRT scores as the sole response variable in their analyses (using the time that subjects took to complete the test as a predictor in one). In contrast, we consider time to completion as another response variable, reasoning that it conveys information about the underlying latent variable ("reflectiveness") that we're interested in capturing.

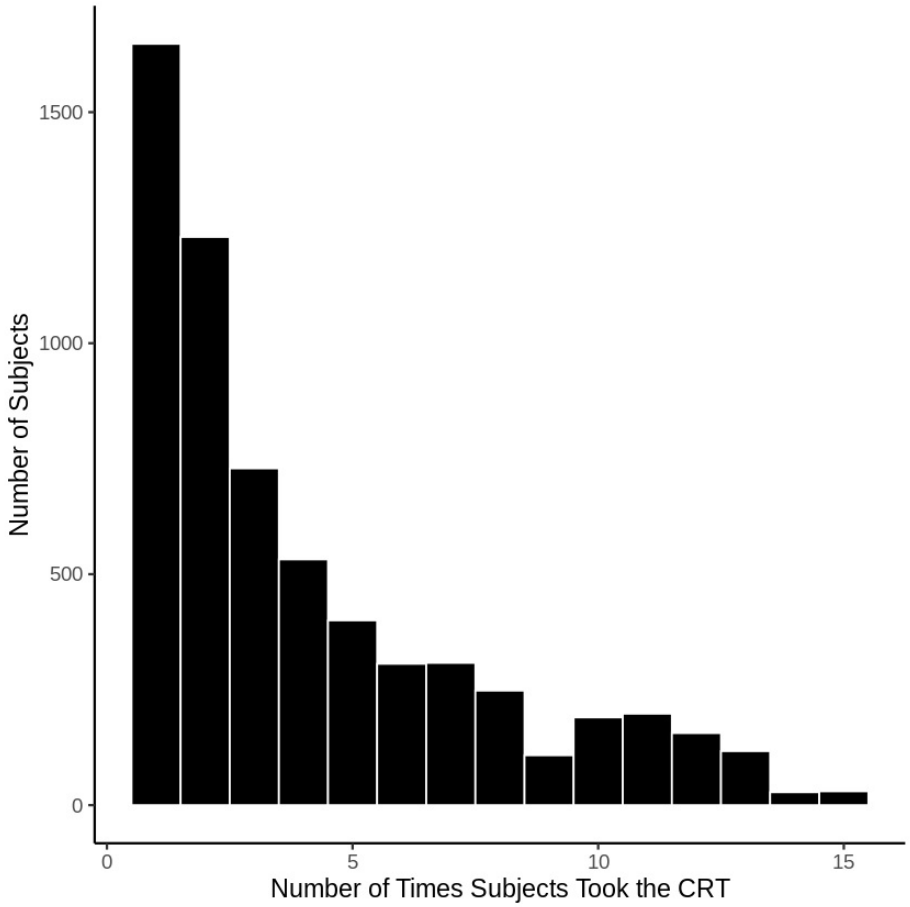

Figure 2.1: Distribution of subjects' exposures, i.e., number of times subjects took the CRT.

2.3 Predictors

Various predictor variables may influence the distribution of our two response variables. In this section we discuss our selection of these variables and our handling of idiosyncratic and missing values.

Our primary predictor of interest is the number of times a subject has taken the CRT *within* the series, including the current test. This variable is denoted by `nPrevS` and takes values from 1 to 15. It is a time-varying, numeric predictor. Subjects may have taken the CRT prior to these series, but we do not have access to this information.

Unlike `nPrevS`, the remaining predictors we selected were self-reported, and each presents challenges to address. First, subjects self-reported the number of questions they had seen from the CRT previously, denoted by `numSeen`. This variable is also numeric and time-varying. It takes values from 0 to 3. In theory, this predictor should be time-invariant after a subject's first test exposure, since all returning subjects would have seen all CRT items at

that time point. However, inconsistencies occur in practice: Subjects don't always report "3" after the first test exposure, and some even report decreasing values over time. Therefore, we had to determine whether to keep the values as reported or to implement a modification. As Meyer et al. (2018) noted, `numSeen` could be informative not only for its intended purpose (measuring CRT items seen), but also as a proxy for a subject's memory of the CRT and mathematical ability. That is, a subject's seeing the items but not remembering them is arguably equivalent to never having seen the items. Thus, this predictor potentially conveys useful information about the responses even though it doesn't accurately represent number of CRT items seen previously.

An additional concern is that `nPrevS` and `numSeen` could be highly correlated since they both measure familiarity with the CRT—albeit one objectively and the other subjectively. However, we think this concern is unwarranted for two reasons. First, as discussed, `numSeen` likely captures indirect information not reflected in `nPrevS`. Second, in a preliminary analysis based on separate models for each response variable, the estimated correlation of these two predictors was relatively low in absolute magnitude.

The predictor `aveSATS` refers to a subject's self-reported SAT score, averaged over the course of the Fall 2014 series. It is a standardized, continuous predictor.

The binary categorical predictor `male` denotes a subject's self-reported sex. However, `male` was not always constant throughout the series. In the case of only two observations per subject with different sex values, we exclude both observations; otherwise, we replace discrepant values with the most commonly used value reported by the subject.

Lastly, `age` denotes a subject's standardized, self-reported age, which we treat as continuous. Subjects had to be at least 18 years old to participate. Since subjects may have had a birthday between their first and final exposures to the CRT, their ages could have increased by a year across test exposures; however, we simply use their self-reported age at first exposure to avoid time-variance. If subjects had greater than one-year increases in `age` across observations, we replace the values with subjects' starting `age` values so that it is not time-varying. For the remaining discrepancies, our correctives involved some subjective judgment. If the values do not vary too erratically, we either replace the discrepant value(s) with the modal value or, in the case of no modal value, we use the median value. If the discrepancies are too great to make an educated modification, we simply exclude the observations.

We initially hoped to include a seven-level ordinal variable denoting subjects' reported level of education (the levels are undefined in the dataset, but higher values denote higher levels of education). However, we encountered issues relating to matrix sparsity when attempting to fit the models with this variable. Ultimately, we reasoned that much of the information

contained in this variable is likely contained within `aveSATS`, and thus decided to exclude it. Table 2.2 provides further support for this decision.

Table 2.1 summarizes the response and predictor variables.

Variable	Variable Type	Description
CRT score	Response (Discrete)	CRT score
CRT time	Response (Continuous)	Log of time spent on CRT
nPrevS	Explanatory (Discrete)	Exposure number within series (time-varying)
numSeen	Explanatory (Discrete)	# of CRT items seen before (time-varying)
aveSATS	Explanatory (Continuous)	SAT score (standardized)
male	Explanatory (Categorical)	Sex
age	Explanatory (Continuous)	Age (standardized)
identifier	Random factor	Subject ID

Table 2.1: CRT variables selected

2.4 Missing Data

A substantial number of observations had missing values for at least one predictor. The most common predictor with missing values is `aveSATS`, with over half of the original 14,500 observations in the Fall 2014 series missing subjects' SAT scores (and many of these also missing other predictor values). These values could represent *unreported* SAT scores or *non-existent* SAT scores. In particular, we expect most American MTurks with post-secondary education to have written the SAT (it is a mandatory test for admission into many American colleges and universities). This expectation is reflected in Table 2.2: Lower education levels are associated with higher percentages of missing values. Higher education levels are also associated with higher average SAT scores, adding further justification to our aforementioned assumption that much of the information in the education variable is contained in `aveSATS`. (Note that this table was constructed using only one observation per subject so as not to overestimate missingness.)

Education Level	# of Observations	# of Missing SAT Responses	Proportion Missing	Ave SAT Score
1	87	76	87%	1,124
2	743	599	81%	1,152
3	2,785	1,676	60%	1,208
4	2,463	1,110	45%	1,253
5	504	198	39%	1,286
6	108	42	39%	1,289
7	71	29	41%	1,330

Table 2.2: Percentage of `aveSATS` values missing by education level

However, other MTurks (including the roughly one-quarter of MTurks who are not American; see Appendix A) likely do not have SAT scores. In other words, we think that the missing data mechanism is likely related to other demographic characteristics about which we may not have information. That is, the missing data mechanism is likely either missing at random (MAR) or missing not at random (MNAR), but we cannot distinguish which. Since imputation could introduce unintended bias in the predictor values, we elect to exclude observations with missing SAT values from our analysis. We discuss possible implications of this decision in Chapter 5.

Once the observations with missing `aveSATS` values are removed, variables `numSeen`, `age`, and `male` each have a relatively small proportion of missing values (8%, 2%, and 3%, respectively). We omit all the observations with missing values of these predictors. Other than `aveSATS`, we treat these missing predictor values as MAR, as we can reasonably assume that a missing value is unrelated to the missing data but related to an observed variable or parameter of interest (e.g., subjects did not self-report this value due to an inability to recall, which may be related to `aveSATS`). The implications are likely minimal due to the small proportion of missing values.

Finally, about 1.5% of the total observations in the Fall 2014 series contained missing values for time to completion of the CRT, the second response variable. These missing values occurred because subjects did not submit their test. The time they spent on the test was not recorded. If this time had been recorded, we may have been able to include these (right-censored) responses in our analysis. But the missing values were misleadingly coded as "1", giving the illusion that those observations correspond to a very quick completion of the CRT. The missing values are clearly MNAR, and we have no reasonable way of imputing them. However, given that they comprise a small proportion of the observations and thus will have minimal impact, we discard them.

Our final dataset contains 6,228 observations from the Fall 2014 series. The target population is relatively well-educated American adults.

2.5 Data Visualization

Here we provide further visualizations of the dataset to explore and motivate our proposed models in the next chapter. First, we examine visualizations of the CRT score distribution. Histograms of CRT score for different values of `nPrevS` are shown in Figure 2.2 (we omit the cases where `nPrevS` ≥ 4 due to lack of data) and for different categories of `aveSATS` at `nPrevS` $= 1$ in Figure 2.3. The former reveals bathtub-shaped distributions for each value of `nPrevS`. The latter reveals bathtub-shaped distributions for each of the first two categories of `aveSATS` and skewed left distributions with peaks at the maximum CRT score for the final

two categories. Histograms of the distribution of CRT score conditional on other predictor variables reveal similar shapes (see Appendix C).

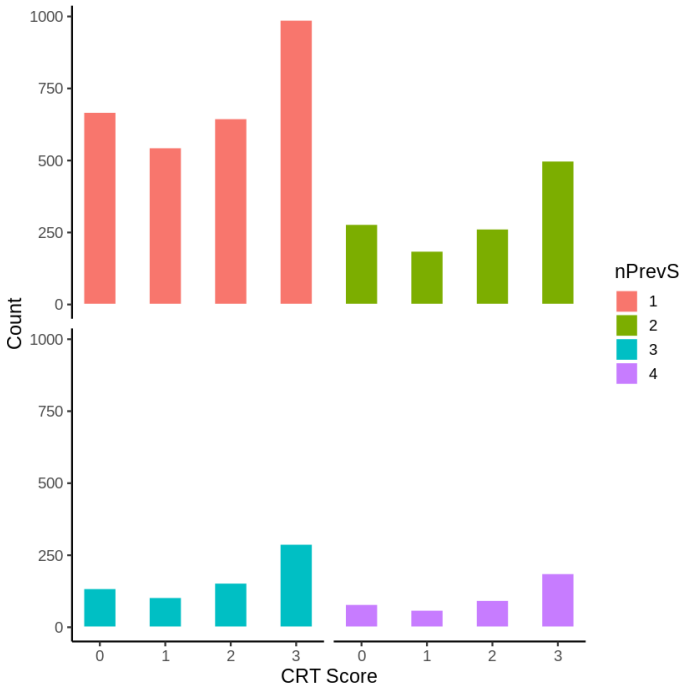

Figure 2.2: Distribution of CRT score by `nPrevS`

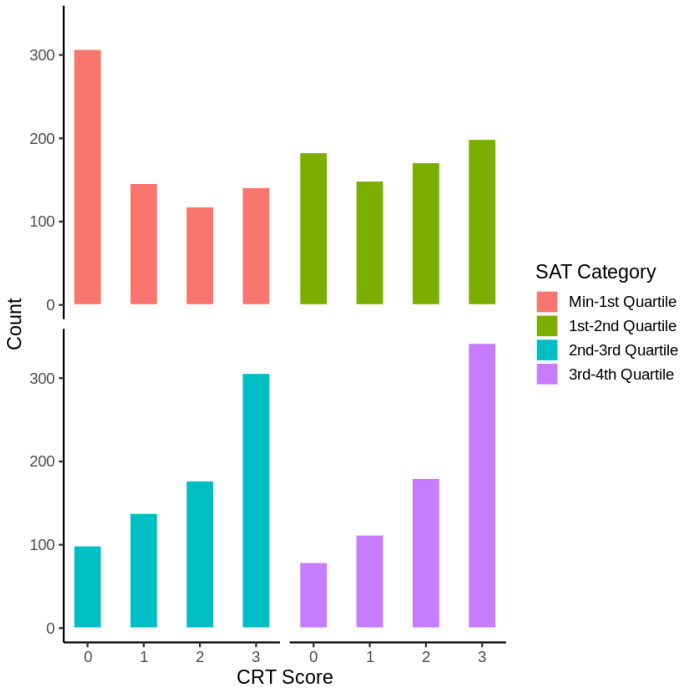

Figure 2.3: Distribution of CRT score by `aveSATS` (for `nPrevS`=1)

Figure 2.4 displays the distribution of the time response (on the logarithmic scale), broken down by `nPrevS` (left) and by `numSeen` at `nPrevS` = 1 (right). The former graph reveals an approximately normal distribution for each value of `nPrevS`. We also observe that additional test exposures are associated with lower times to completion. The latter graph likewise reveals an approximately normal distribution for each value of `numSeen` at subjects' first test exposure. The times to completion are markedly different for the lowest and highest values of `numSeen`. With values of `nPrevS` > 1 (see Appendix C), this difference is much less, implying that the effect of `numSeen` on CRT time to completion is most pronounced at the first test exposure. Similar graphs for the other predictors suggest little effect on time to completion (see Appendix C).

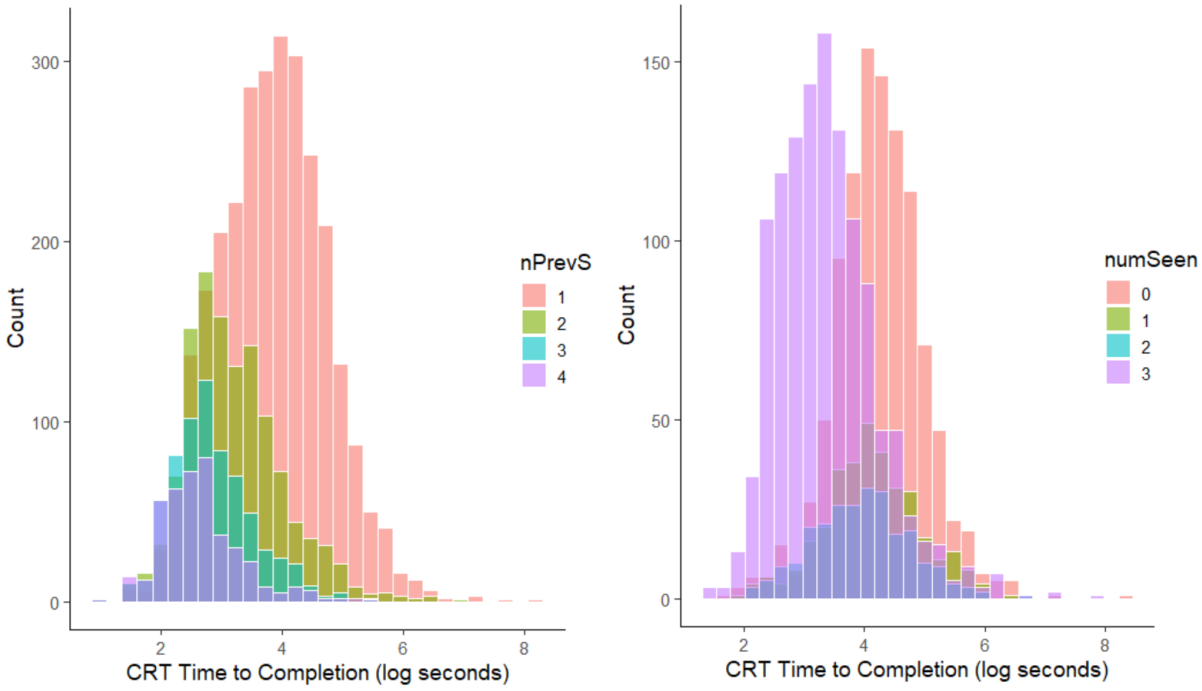

Figure 2.4: Distribution of the logarithm of time to completion for `nPrevS`≤ 4 (left) and for `numSeen` at `nPrevS`=1 (right)

Next, Figure 2.5 displays the ordinary least squares (OLS) estimates of the effects of `nPrevS` when CRT score (left) and CRT log time to completion (right) are regressed on the predictors separately for each subject (for subjects who completed the test more than once). We do not make formal inference based on these estimates; we use them simply for visualizing the trends in subjects' observed test scores and completion times. The plot for CRT score reveals a peak at 0, describing the vast majority of subjects whose scores remained constant over time. The majority of the remaining estimates are greater than 0, with a small proportion less than 0. The plot for time to completion reveals a peak at 0, with the majority of estimates being negative, implying that subjects generally took less time to complete the test with additional exposures. We also observe a small but non-negligible proportion

of subjects who spent an increasing amount of time to complete the test with additional exposures.

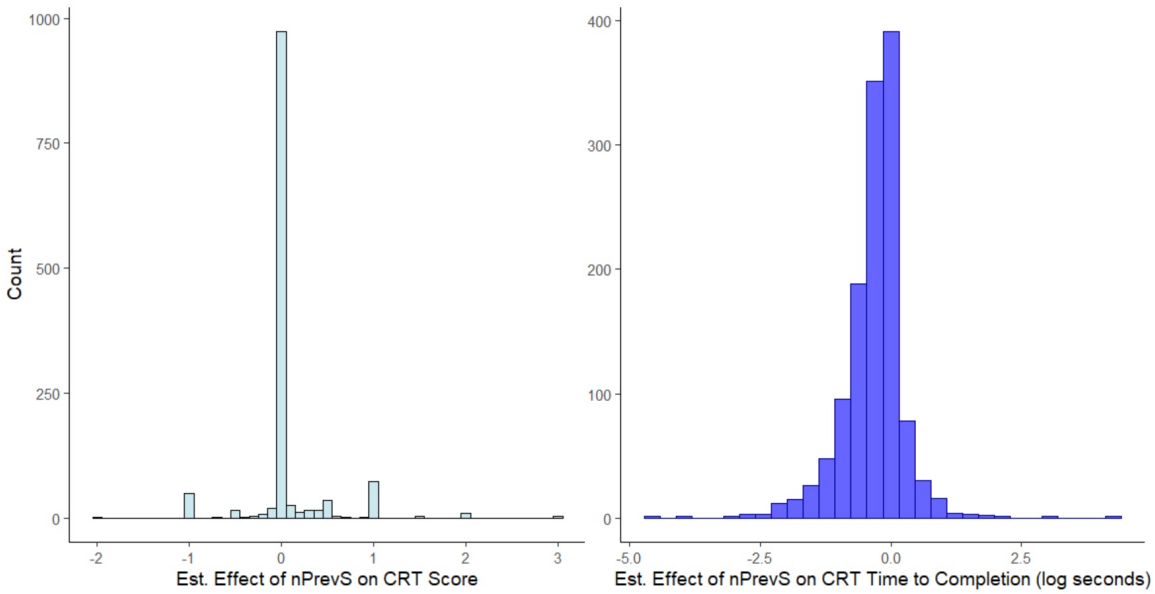

Figure 2.5: OLS estimates of the effects of `nPrevS` when CRT score is regressed on the predictors separately for each subject (left); and when CRT log time to completion is regressed on the predictors separately for each subject (right).

Lastly, we use the pair of scatterplots in Figure 2.6 to explore the changes in the two response variables over time: The left panel shows subjects' average time to completion vs. the OLS estimates of the effects of `nPrevS` when their *scores* are regressed on the predictors (separately for each subject); the right panel shows OLS estimates of the effects of `nPrevS` when subjects' log *times to completion* are regressed on the predictors vs. the OLS estimates of the effects of `nPrevS` when the *scores* are regressed on the predictors (both separately for each subject). Both sets of points are broken down by initial CRT scores. These plots use data only from subjects who appeared more than once in the series. The patterns in both plots are difficult to detect visually, but hint at very slight, positive correlations between the pairs of variables at each level of first score. The leftmost scatterplot exemplifies how subjects who improved their CRT scores on subsequent tests generally spent more time on the test than did subjects with constant scores; the rightmost scatterplot similarly exemplifies how subjects who improved their scores on subsequent tests generally spent more time on *each subsequent test* than did subjects with constant scores.

Moreover, of the 44% of subjects who appeared more than once in the series, 73% had constant CRT scores and their average decrease in time spent completing the test was 0.33 log seconds per additional test exposure; 18% had *increasing* scores, an average CRT score improvement of 0.70 per additional test exposure, and an average decrease in time spent of

0.27 log seconds; and 9% had *decreasing* CRT scores, an average CRT score decrease of 0.60, and an average decrease in time spent of 0.42 log seconds. In other words, the small subset of subjects who improved their test scores over time reflected longer than did subjects who exhibited constant scores. These statistics and the scatterplots in Figure 2.6 are consistent with the observation by Meyer et al. (2018) that a small proportion of subjects "continue to spend time on the test".

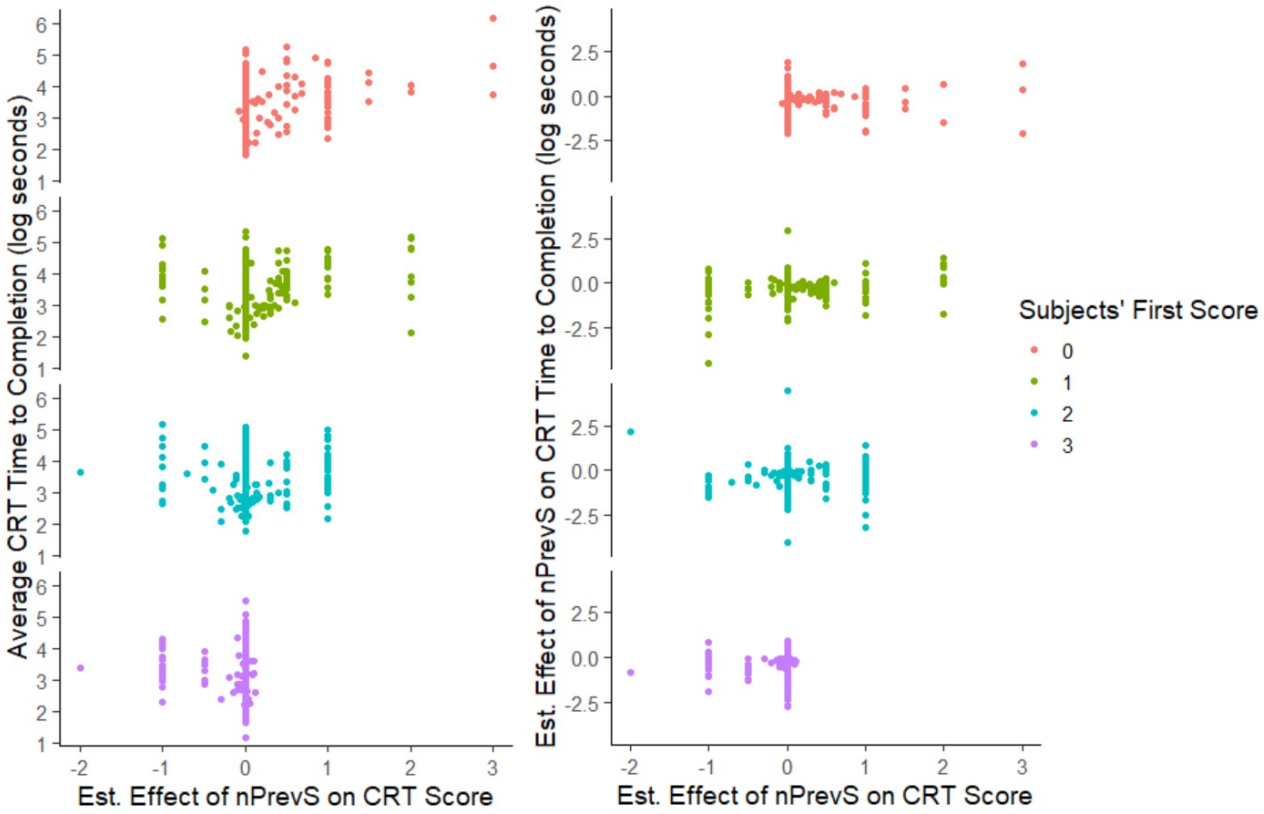

Figure 2.6: Average time to completion (log scale) vs. OLS estimates of the effects of `nPrevS` on CRT score by subjects' first test score (left); OLS estimates of the effects of `nPrevS` on log time to completion vs. OLS estimates of the effects of `nPrevS` on CRT score by subjects' first test score (right)

Chapter 3

Statistical Methods

To model our unbalanced longitudinal data and explore the relationship between our predictors and bivariate response, we consider extensions of traditional generalized linear mixed models. In the following sections, we describe bivariate longitudinal models that can be applied to the CRT data and, in particular, the estimation and computational challenges that can arise in maximizing the likelihoods. Ultimately, we propose three models; the first serves as our foundational model, and the second and third extend the first to allow for subpopulations ("clusters") of individuals with similar levels of rationality and reflectiveness.

3.1 Models

Let Y_{ij} and T_{ij} denote subject i's CRT score and response time (on the logarithmic scale), respectively, on the j^{th} attempt of the CRT in the Fall 2014 series, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Since a subject is awarded one point for each correct answer on the CRT, $Y_{ij} \in \{0, 1, 2, 3\}$. In contrast, T_{ij} takes values on the real line.

3.1.1 Bivariate Longitudinal Model

To deal with the repeated measures, we use a random intercept in the model for each of our bivariate responses. Among other implications, these random effects allow for correlation among scores or times to completion observed on the same individual.

Let $\mathbf{x}_{ij}$ denote the vector of predictor variables associated with subject i on the j^{th} attempt of the CRT. We model the test scores as

$$Y_{ij} \mid U_i \sim \text{Bin}(3, \theta_{ij}),$$

where

$$\text{logit}\theta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + U_i$$

and where the random effects, U_i, are independent and distributed as $N(0, \sigma_u^2)$. We conceive of U_i as a latent variable representing "rationality". Likewise, we model the logarithm of the time to completion as

$$T_{ij} \mid V_i \sim \text{N}(\mu_{ij}, \sigma_t^2),$$

where

$$\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\alpha} + V_i$$

and where the random effects, V_i, are independent and distributed as $N(0, \sigma_v^2)$. We conceive of V_i as a latent variable representing "reflectiveness".

We assume that $Y_{ij} \mid U_i$ is independent of $Y_{ij'}$, $j' \neq j$, all T_{ij}'s, and V_i. We also assume that $T_{ij} \mid V_i$ is independent of $T_{ij'}$, $j' \neq j$, all Y_{ij}'s, and U_i. Finally, we assume that the joint distribution of the random effects is bivariate normal, that is,

$$(U_i, V_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}.$$

Figure 2.4 motivates the model for $T_{ij} \mid V_i$. Histograms of the logarithm of time to completion given combinations of predictor variables reveal that the marginal distribution of T_{ij} is approximately normal. From this perspective, the proposed models for $T_{ij} \mid V_i$ and V_i (which imply that T_{ij} is normally distributed) are reasonable.

With these assumptions, we can write the likelihood as a product of the conditional distributions:

$$\begin{aligned}
\mathcal{L}^{[1]}(\boldsymbol{\psi}) &= \prod_i \int\int \Big(\prod_j f_{Y_{ij}|U_i}(y_{ij}|u_i) f_{T_{ij}|V_i}(t_{ij}|v_i) \Big) \cdot f_{U_i,V_i}(u_i, v_i) du_i dv_i \\
&= \prod_i \int\int \Big(\prod_j \theta_{ij}^{y_{ij}} (1-\theta_{ij})^{3-y_{ij}} \cdot \frac{1}{\sigma_t} \exp\Big(-\frac{(t_{ij}-\mu_{ij})^2}{2\sigma_t^2} \Big) \Big) \cdot \\
&\quad \frac{1}{\sigma_u\sigma_v\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2(1-\rho^2)} \Big(\frac{u_i^2}{\sigma_u^2} + \frac{v_i^2}{\sigma_v^2} - \frac{2\rho u_i v_i}{\sigma_u\sigma_v} \Big) \Big] \, du_i dv_i, \qquad (3.1)
\end{aligned}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_t, \sigma_u, \sigma_v, \rho)$ is the vector of parameters to be estimated. We omit terms that are constant with respect to the unknown parameters throughout this report. In ad-

dition, we use superscripts with square brackets to denote the number of clusters in the model.

Based on our chosen model, we can find a closed form for the marginal distribution of the time to completion. In particular, the vector of times to completion of the i^{th} subject, $\mathbf{T}_i$, is distributed as multivariate normal with $\mathrm{E}[T_{ij}] = \mu_{ij}$, $\mathrm{Var}[T_{ij}] = \sigma_v^2 + \sigma_t^2$, and

$$\mathrm{Cov}[T_{ij}, T_{ik}] = \sigma_v^2$$

for $j \neq k$. Since times observed on different subjects are assumed independent, $\mathrm{Cov}[T_{ij}, T_{hk}] = 0$ for $i \neq h$.

The marginal distribution of CRT score is

$$f_{Y_{ij}}(y_{ij}) = \int f_{Y_{ij}|U_i}(y_{ij}|u_i) f_{U_i}(u_i) du_i,$$

which does not have a closed form. Likewise, the marginal mean, variance, and covariance of CRT scores do not have a closed form. We can say, however, that they depend on the predictor variables in a complicated way.

The assumption that test scores are conditionally binomial distributed may, at first, seem suspect because the outcomes (correct/incorrect) of the three questions posed to each individual at each exposure are not necessarily independent with common probability of success. However, Figures 2.2 and 2.3 help to justify the model for $Y_{ij} \mid U_i$. In particular, the histograms of the CRT score responses for given combinations of predictor variables reveal that the marginal distribution of Y_{ij} has a "bathtub" shape. This shape can be captured by a mixture of binomial distributions where the mixture distribution is a normal distribution, i.e., our specified distribution of $Y_{ij} \mid U_i$. We thus use this model for the overall test scores but do *not* interpret the scores as arising from a series of three independent trials (questions) with a common probability of success (correctness).

3.1.2 Bivariate Longitudinal Model with Two Clusters

Our second proposed model extends the first model by postulating that test subjects comprise distinct clusters. This model is based on an alternative interpretation of the marginal distribution of the CRT scores depicted in Figures 2.2 and 2.3. In particular, we surmise that this (bimodal) distribution arises due to two distinct subpopulations or clusters of individuals. In this new model, we hypothesize that the first cluster corresponds to subjects whose CRT scores remain relatively stable over time, while the second corresponds to subjects whose CRT scores improve over time. The proportion of subjects whose scores decrease

over time is expected to be negligible. Our original model can be considered a special case of this extended model where the probability associated with one cluster is 0.

Let $\bar{\mathbf{x}}_{ij}$ be the vector of all predictor variables except `nPrevS` observed on subject i at time j. Let s_{ij} be the value of `nPrevS` observed on subject i at time j. Let $C_i \in \{1,2\}$ be a latent cluster indicator, where clusters correspond to the two subpopulations described above. We assume that the C_i's are independent and distributed as $P(C_i = c_i) = \gamma_{c_i}$. As per our original model, we assume that (U_i, V_i) are independent, bivariate normal distributed random effects. We then assume that $Y_{ij} \mid U_i, C_i$ is distributed as $\text{Bin}(3, \theta_{ij})$, where

$$\text{logit}\theta_{ij} = \beta_{c_i 0} + \beta_{c_i 1} s_{ij} + \bar{\mathbf{x}}'_{ij}\boldsymbol{\beta} + u_i.$$

We further assume that $T_{ij} \mid V_i, C_i$ is distributed as $N(\mu_{ij}, \sigma_t^2)$, where

$$\mu_{ij} = \alpha_{c_i 0} + \alpha_{c_i 1} s_{ij} + \bar{\mathbf{x}}'_{ij}\boldsymbol{\alpha} + v_i.$$

The intercepts and effects of `nPrevS` are allowed to differ by cluster but, for parsimony, we assume that the other regression coefficients are common across clusters.

We expect that one cluster will correspond to the subpopulation of individuals whose CRT scores remained relatively stable over time and that the other cluster will correspond to the subpopulation of whose scores improved over time.

The likelihood is

$$\begin{aligned}\mathcal{L}^{[2]}(\boldsymbol{\psi}) &= \prod_i \int\int \sum_{c_i} \left(\prod_j f_{Y_{ij}|U_i,C_i}(y_{ij}|u_i,c_i) f_{T_{ij}|V_i,C_i}(t_{ij}|v_i,c_i) \right) \cdot f_{C_i}(c_i) \cdot f_{U_i,V_i}(u_i,v_i) du_i dv_i \\ &= \prod_i \int\int \sum_{c_i} \left(\prod_j \theta_{ij}^{y_{ij}} (1-\theta_{ij})^{3-y_{ij}} \cdot \frac{1}{\sigma_t} \exp\left(-\frac{(t_{ij}-\mu_{ij})^2}{2\sigma_t^2} \right) \right) \cdot \gamma_{c_i} \cdot f_{U_i,V_i}(u_i,v_i) du_i dv_i, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_t, \sigma_u, \sigma_v, \rho, \gamma_2)$ is the vector of parameters to be estimated. (We exclude γ_1 from $\boldsymbol{\psi}$ since it can be computed as $\gamma_1 = 1 - \gamma_2$ and hence is not a free parameter.)

3.1.3 Bivariate Longitudinal Model with Four Clusters

Our third proposed model increases the number of latent cluster labels to four. This four-cluster model uses theoretical understandings derived from the literature, namely the slightly differing operationalizations of reflectiveness and rationality that we outlined in Chapter 1. Specifically, we hypothesize "high" and "low" categories for each latent variable, resulting in four combinations of the two categories. We now let $C_i \in \{1,2,3,4\}$, where clusters correspond to the four combinations of rational/not rational and reflective/not

reflective. As in the two-cluster model, we assume that the C_i's are independent and distributed as $P(C_i = c_i) = \gamma_{c_i}$. We define $\boldsymbol{\gamma} = (\gamma_2, \gamma_3, \gamma_4)$. As in the prior two models, we assume that the tuples (U_i, V_i) are independent and distributed as bivariate normal. We further assume that $Y_{ij} \mid U_i, C_i$ is distributed as $\text{Bin}(3, \theta_{ij})$, where

$$\text{logit}\theta_{ij} = \beta_{c_i 0} + \beta_{c_i 1} s_{ij} + \bar{\mathbf{x}}'_{ij}\boldsymbol{\beta} + u_i.$$

We further assume that $T_{ij} \mid V_i, C_i$ is distributed as $N(\mu_{ij}, \sigma_t^2)$, where

$$\mu_{ij} = \alpha_{c_i 0} + \alpha_{c_i 1} s_{ij} + \bar{\mathbf{x}}'_{ij}\boldsymbol{\alpha} + v_i.$$

The purpose of this model is to allow a coarse categorization (via the clusters) of individuals as rational/not rational and reflective/not reflective. The random effects U_i and V_i account for the remaining variation in the underlying levels of these characteristics. We envision that cluster 1 would correspond to the subpopulation of individuals who are neither rational nor reflective. We would expect $\beta_{11} = 0$, as we expect that subjects who aren't reflective do not improve their CRT scores with repeated test exposure. Cluster 2 would correspond to the subpopulation of individuals who are not rational but are reflective. Like in cluster 1, we would expect $\beta_{21} = 0$ and β_{20} to be relatively low, but α_{20} to be relatively high. Cluster 3 would correspond to the subpopulation of individuals who are rational and reflective. Here we would expect β_{30} and α_{30} to be relatively high, and expect β_{31} to be positive and α_{31} to be 0 or negative. Cluster 4 would correspond to the subpopulation of individuals who are rational but either aren't reflective or provide no information about their reflectiveness because they quickly chose the correct answers. We therefore expect $\beta_{41} = 0$. We further expect β_{40} to be high and α_{40} to be low.

The likelihood is

$$\begin{aligned}
\mathcal{L}^{[4]}(\boldsymbol{\psi}) &= \prod_i \int\int \sum_{c_i} \Big(\prod_j f_{Y_{ij}|U_i,C_i}(y_{ij}|u_i,c_i) f_{T_{ij}|V_i,C_i}(t_{ij}|v_i,c_i) \Big) \cdot f_{C_i}(c_i) \cdot f_{U_i,V_i}(u_i,v_i) du_i dv_i \\
&= \prod_i \int\int \sum_{c_i} \Big(\prod_j \theta_{ij}^{y_{ij}} (1-\theta_{ij})^{3-y_{ij}} \cdot \frac{1}{\sigma_t} \exp\Big(-\frac{(t_{ij}-\mu_{ij})^2}{2\sigma_t^2} \Big) \Big) \cdot \gamma_{c_i} \cdot f_{U_i,V_i}(u_i,v_i) du_i dv_i,
\end{aligned} \tag{3.3}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_t, \sigma_u, \sigma_v, \rho, \boldsymbol{\gamma})$ is the vector of parameters to be estimated.

3.2 Estimation

Direct maximization of the likelihoods (3.1)–(3.3) requires integrating complex functions with respect to u_i and v_i. These integrals do not have closed form solutions. Instead, we

use adaptive Gaussian quadrature (AGQ) (Pinheiro and Chao, 2006) to find the maximum likelihood estimates (MLEs) of the one- and two-cluster model parameters. To estimate the four-cluster model, which contains many more parameters, we develop an Expectation-Maximization (EM) algorithm (Dempster et al., 1977). We elect to use AGQ rather than EM for the former models due to improved efficiency.

3.2.1 Adaptive Gaussian Quadrature

Gaussian quadrature is a numerical procedure for approximating integrals that involves a finite, weighted sum of the output of a function evaluated at particular inputs. A Gaussian quadrature rule has the form

$$\int_a^b w(z)h(z)dz \approx \sum_{k=1}^{Q} w_k h(z_k), \quad z_1 < z_2 < \ldots < z_Q,$$

where w_k are weights, z_k are quadrature points or abscissae, and Q is the chosen number of quadrature points. If $h(z)$ is a polynomial of degree $2Q-1$ or less, the approximation is exact. When the weight function is $w(z) = e^{-z^2}$, the Gauss-Hermite quadrature (GHQ) rule is commonly used to determine the weights and abscissae. However, large numbers of quadrature points are often needed to obtain accurate approximations of likelihoods, which can become prohibitively expensive computationally (see Section 4.6).

Adaptive Gaussian Quadrature (AGQ) solves some of the issues of GHQ by adapting the quadrature points and weights to the function we wish to integrate. With too few quadrature points used, the peak of the integrand may be located between adjacent quadrature points so that a substantial portion of the likelihood contribution may be lost. AGQ has been described as an attempt to shift and rescale the quadrature points to lie under the peak of the integrand (Rabe-Hesketh and Skrondal, 2002).

The iterative steps for maximum likelihood estimation using AGQ can be summarized, in the context of our proposed one-cluster model, as

1. Predict $(\hat{u}_i, \hat{v}_i)$ using starting values of our model parameters, $\boldsymbol{\psi}$.

2. Use $(\hat{u}_i, \hat{v}_i)$ and Gauss-Hermite quadrature to form an approximate log-likelihood.

3. Maximize this approximate log-likelihood to find an updated estimate of $\boldsymbol{\psi}$.

4. Repeat steps 1–3 until convergence (defined as the event that the difference between consecutive estimates is less than a chosen value, δ) is achieved.

The details of this procedure are as follows. We first define the logarithm of the joint density of our response variables and random effects as

$$g_{U_i,V_i}(u_i, v_i) = \log\Big(f_{\mathbf{Y_i},\mathbf{T_i}|U_i,V_i}(\mathbf{y}_i, \mathbf{t}_i \mid u_i, v_i) f_{U_i,V_i}(u_i, v_i) \Big).$$

We then maximize $g_{U_i,V_i}(u_i, v_i)$ by computing $(\hat{u}_i, \hat{v}_i)$ such that $g'_{U_i,V_i}(\hat{u}_i, \hat{v}_i) = 0$. Using a Laplace approximation of $g_{U_i,V_i}(u_i, v_i)$ around $(\hat{u}_i, \hat{v}_i)$, we can show that $\exp\{g_{U_i,V_i}(u_i, v_i)\}$—and hence the posterior distribution of (U_i, V_i)—is approximately proportional to a normal density with mean $\boldsymbol{\mu}_A = (\hat{u}_i, \hat{v}_i)$ and variance $\boldsymbol{\Sigma}_A = \Big[g''_{U_i,V_i}(\hat{u}_i, \hat{v}_i) \Big]^{-1}$.

Let $\phi(\cdot; \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ be the density of a bivariate normal random variable with mean $\boldsymbol{\mu}_A$ and variance-covariance matrix $\boldsymbol{\Sigma}_A$. Defining $\mathbf{B}_i = (U_i, V_i)$, we next rewrite the marginal density of $(\mathbf{Y}_i, \mathbf{T}_i)$ as

$$\begin{aligned}
& f_{\mathbf{Y}_i,\mathbf{T}_i}(\mathbf{y}_i, \mathbf{t}_i) \\
= & \iint f_{\mathbf{Y}_i,\mathbf{T}_i|\mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i \mid \mathbf{b}_i) f_{\mathbf{B}_i}(\mathbf{b}_i)\, d\mathbf{b}_i \\
= & \iint \left\{ \frac{f_{\mathbf{Y}_i,\mathbf{T}_i|\mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i \mid \mathbf{b}_i) f_{\mathbf{B}_i}(\mathbf{b}_i)}{\phi(\mathbf{b}_i; \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)} \right\} \phi(\mathbf{b}_i; \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)\, d\mathbf{b}_i.
\end{aligned}$$

The Laplace approximation result implies that the term $\{\cdot\}$ is approximately constant with respect to $\mathbf{b}_i$. Consequently, we can use Gauss-Hermite quadrature with relatively few quadrature points to evaluate this integral with high accuracy. In particular, substituting $\mathbf{z}_i = \dfrac{\mathbf{b}_i - \boldsymbol{\mu}_A}{\boldsymbol{\Sigma}_A}$, we can write

$$\begin{aligned}
& f_{\mathbf{Y}_i,\mathbf{T}_i}(\mathbf{y}_i, \mathbf{t}_i) \\
= & \iint \frac{f_{\mathbf{Y}_i,\mathbf{T}_i|\mathbf{B}_i}\left(\mathbf{y}_i, \mathbf{t}_i \mid \boldsymbol{\Sigma}_A^{-\frac{1}{2}}\mathbf{z}_i + \boldsymbol{\mu}_A\right) f_{\mathbf{B}_i}\left(\boldsymbol{\Sigma}_A^{-\frac{1}{2}}\mathbf{z}_i + \boldsymbol{\mu}_A\right)}{\exp(-||\mathbf{z}_i||^2)} \boldsymbol{\Sigma}_A^{-\frac{1}{2}}\ \exp(-||\mathbf{z}_i||^2)\, d\mathbf{z}_i \\
\approx & (2\pi)^{q/2} |\boldsymbol{R}_i|^{-1} \sum_{\mathbf{k}} f_{\mathbf{Y}_i,\mathbf{T}_i|\mathbf{B}_i}\left(\mathbf{y}_i, \mathbf{t}_i \mid \boldsymbol{\Sigma}_A^{-\frac{1}{2}}\mathbf{z}_\mathbf{k} + \boldsymbol{\mu}_A\right) f_{\mathbf{B}_i}\left(\boldsymbol{\Sigma}_A^{-\frac{1}{2}}\mathbf{z}_\mathbf{k} + \boldsymbol{\mu}_A\right) W_\mathbf{k},
\end{aligned}$$

where $\mathbf{k} = (k_1, k_2)$, $k_1, k_2 = 1, \ldots, Q$, indexes the $Q \times 2$ grid of abscissae, $W_\mathbf{k} = \exp(||\boldsymbol{z}_\mathbf{k}||^2) w_{k_1} w_{k_2}$, and $\boldsymbol{R} = \boldsymbol{\Sigma}_A^{-1}$.

The AGQ approximation to the log-likelihood function is

$$\ell^{[AGQ]}(\boldsymbol{\psi}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{n_i} \left(-\log|\boldsymbol{R}_i| + \log\left\{ \sum_{\boldsymbol{j}}^{Q} f_{\mathbf{Y}_i,\mathbf{T}_i|\mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i \mid \mathbf{b}_i) f_{\mathbf{B}_i}(\mathbf{b}_i) W_{\boldsymbol{k}} \right\} \right).$$

When $Q = 1$, this approximation is the Laplace approximation. Higher values of Q lead to greater accuracy, however, and are thus preferable. Pinheiro and Chao (2006) argue that $Q \leq 7$ is generally sufficient. In our case, $Q = 15$ quadrature points seemed sufficient to evaluate the integrals in our log-likelihood accurately.

The computational efficiency is thus generally much greater for AGQ compared to GHQ. Figure 3.1, adapted from Rabe-Hesketh and Skrondal (2002), illustrates the difference between GHQ and AGQ using normal prior and posterior densities. (Note that the prior refers to the assumed marginal density of the random effect, and the posterior refers to the conditional density of the random effect given the data.)

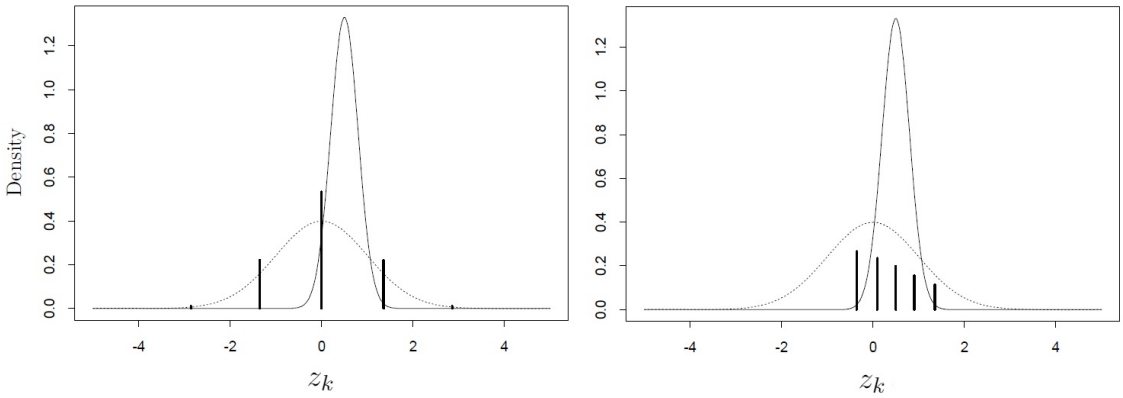

Figure 3.1: Prior (dotted curves) and posterior (solid curves) densities and quadrature points (bars) for standard GHQ (left) and AGQ (right). The heights of the bars represent the weights assigned to the quadrature points.

3.2.2 EM Algorithm

Direct maximization of the likelihood (after approximating the integrals using AGQ) is infeasible for models with large numbers of clusters. Thus, for our four-cluster model, we take a different approach to estimation: the EM algorithm. This algorithm is an iterative algorithm consisting of an E-step and an M-step at each iteration. At the E-step, an objective function is defined using the parameter estimates at the current iteration. At the M-step, this function is then maximized to obtain updated parameter estimates. This procedure is repeated until convergence, as defined in Section 3.2.1.

The foundation of the EM algorithm is the complete log-likelihood, which is the likelihood that would arise if all latent variables were, in fact, observed. The complete log-likelihood

associated with the cluster model with K clusters is

$$\begin{aligned}
\ell_c^{[K]}(\boldsymbol{\psi}) &= \log\left[\prod_i \left(\prod_j f_{Y_{ij}|U_i,C_i}(y_{ij}|u_i,c_i) f_{T_{ij}|V_i,C_i}(t_{ij}|v_i,c_i)\right) \cdot f_{U_i,V_i}(u_i,v_i) \cdot f_{C_i}(c_i)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\log\left[f_{Y_{ij}|U_i,C_i}(y_{ij}|u_i,c_i)\right] + \log\left[f_{T_{ij}|V_i,C_i}(t_{ij}|v_i,c_i)\right]\right) \\
&\quad + \sum_{i=1}^{n}\log\left[f_{V_i|U_i}(v_i|u_i)f(u_i)\right] + \sum_{i=1}^{n}\log\left[f_{C_i}(c_i)\right].
\end{aligned}$$

The E-Step

In the E-step, we define the objective function as

$$Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) = \mathrm{E}^{(p)}[\ell_c^{[K]}(\boldsymbol{\psi}) \mid \mathbf{Y}, \mathbf{T}],$$

i.e., the expectation of the complete log-likelihood with respect to the current conditional distribution of the random effects given our observed data and the current estimates of our parameters, $\boldsymbol{\psi}^{(p)}$. In the M-step we find the maximizer of $Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$, namely

$$\boldsymbol{\psi}^{(p+1)} = \arg\max_{\boldsymbol{\psi}} Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}).$$

Repeating these steps guarantees that the value of the log-likelihood (3.1) will increase (or at least not decrease) with increasing p (Wu, 1983).

To provide more detail, for the multi-cluster model, $Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$ takes the form

$$
\begin{aligned}
& Q^{(K)}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) \\
= & \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathrm{E}^{(p)}\left\{\log\left[\theta_{ij}^{y_{ij}}(1-\theta_{ij})^{3-y_{ij}}\right]\middle|\mathbf{Y}_i, \mathbf{T}_i\right\} \\
& + \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathrm{E}^{(p)}\left\{\log\left[\frac{1}{\sigma_t}\exp\Big(-\frac{(t_{ij}-\mu_{ij})^2}{2\sigma_t^2}\Big)\right]\middle|\mathbf{Y}_i, \mathbf{T}_i\right\} \\
& + \sum_{i=1}^{n} \mathrm{E}^{(p)}\left\{\log\left[f_{V_i|U_i}(v_i|u_i)f_{U_i}(u_i)\right]\middle|\mathbf{Y}_i, \mathbf{T}_i\right\} \\
& + \sum_{i=1}^{n} \mathrm{E}^{(p)}\left\{\log(\gamma_{c_i})\middle|\mathbf{Y}_i, \mathbf{T}_i\right\} \\
= & \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{c_i=1}^{K} \gamma_{c_i}^{(p)} \iint \left\{\log\left[\theta_{ij}^{y_{ij}}(1-\theta_{ij})^{3-y_{ij}}\right]\right\} f^{(p)}_{U_i,V_i|Y_i,T_i,C_i}(u_i, v_i|y_i, t_i, c_i)du_i dv_i \\
& + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\sum_{c_i=1}^{K} \gamma_{c_i}^{(p)} \iint \left\{\log\left[\frac{1}{\sigma_t}\exp\left(-\frac{(t_{ij}-\mu_{ij})^2}{2\sigma_t^2}\right)\right]\right\} f^{(p)}_{U_i,V_i|Y_i,T_i,C_i}(u_i, v_i|y_i, t_i, c_i)du_i dv_i \\
& + \sum_{i=1}^{n}\sum_{c_i=1}^{K} \gamma_{c_i}^{(p)} \iint \left\{\log\left[f_{V_i|U_i}(v_i|u_i)f_{U_i}(u_i)\right]\right\} f^{(p)}_{U_i,V_i|Y_i,T_i,C_i}(u_i, v_i|y_i, t_i, c_i)du_i dv_i \\
& + \sum_{i=1}^{n}\sum_{c_i=1}^{K} \gamma_{c_i}^{(p)} \iint \log\left(\gamma_{c_i}\right) f^{(p)}_{U_i,V_i|Y_i,T_i,C_i}(u_i, v_i|y_i, t_i, c_i)du_i dv_i \\
\equiv & \; Q_1^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) + Q_2^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) + Q_3^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) + Q_4^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}),
\end{aligned}
$$

where the functions $Q_1^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$, $Q_2^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$, $Q_3^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$, and $Q_4^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$ correspond to the four sets of summations.

These expressions require evaluating double integrals that do not have a closed form solution. Therefore, to complete the E-step, we propose evaluating the integrals in the objective function empirically by sampling methods such as Markov chain Monte Carlo (MCMC) or importance sampling, in which case the estimation method is called a Monte Carlo Expectation-Maximization (MCEM) algorithm (Neath, 2013). Indeed, Kondo et al. (2017) took this approach. However, this approach is much less computationally efficient than AGQ, hence our decision to use the latter for our one- and two-cluster models (Pinheiro and Chao, 2006). Given enough computing power, standard Gauss-Hermite quadrature could be a feasible alternative (see Appendix E for details on this approach).

The M-Step

Now we turn to the M-step, maximization. The four sets of summations in $Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$ are functions of disjoint sets of unknown parameters. We thus split it into the sum of four

different objective functions:

$$Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) \equiv Q_1^{[K]}(\boldsymbol{\beta}, \boldsymbol{\psi}^{(p)}) + Q_2^{[K]}(\boldsymbol{\alpha}, \sigma_t^2, \boldsymbol{\psi}^{(p)}) + Q_3^{[K]}(\sigma_u^2, \sigma_v^2, \rho, \boldsymbol{\psi}^{(p)}) + Q_4^{[K]}(\boldsymbol{\gamma}, \boldsymbol{\psi}^{(p)}).$$

These functions can be approximated using Monte Carlo sampling or possibly Gauss-Hermite quadrature (see Appendix E) and then maximized separately.

We maximize these functions for the current estimates of the parameters, $\boldsymbol{\psi}^{(p)}$. We then iterate the E- and M-steps until the distance between consecutive estimates is less than a specified (small) value, δ.

3.2.3 Starting Values

To obtain starting values for the parameter estimates in the one-cluster model, we first fit separate (generalized) linear mixed models to the CRT scores and completion times, treating these responses as independent. That is, we maximized

$$\mathcal{L}^{[Y]}(\boldsymbol{\psi}) = \prod_i \int \left(\prod_j f_{Y_{ij}|U_i}(y_{ij}|u_i) \right) \cdot f_{U_i}(u_i) du_i$$

and

$$\mathcal{L}^{[T]}(\boldsymbol{\psi}) = \prod_i \int \left(\prod_j f_{T_{ij}|V_i}(t_{ij}|v_i) \right) \cdot f_{V_i}(v_i) dv_i.$$

For our correlation parameter, we used a starting value of 0.

For our two-cluster model, to obtain starting values for the fixed and random effect parameters common to each cluster, we first fit the two-cluster model with no random effects. We used the MLEs of the parameters in this model—along with small values for σ_u and σ_v and 0 for ρ—as starting values for estimating the full two-cluster model.

3.3 Predicting Random Effects

Predicting random effects is often not of interest, especially when they may not have any physical meaning. However, in our case, we construe them as representing subjects' rationality and reflectiveness, which are fundamental characteristics of interest.

We are interested in predicting U_i and V_i given $\mathbf{Y}$ and $\mathbf{T}$. To this end, after computing the MLEs of the model parameters, $\hat{\boldsymbol{\psi}}$, we can return to step 1 in the iterative estimation procedure discussed in Section 3.2.2. The prediction $(\hat{u}_i, \hat{v}_i)$ is the posterior mode of the distribution of (U_i, V_i) given the observed data. It can be interpreted as the level of rationality and reflectiveness of the i^{th} subject. Values of zero correspond to subjects with

average levels of rationality and reflectiveness, while values less than and greater than zero indicate below and above average levels, respectively. The magnitude of the values should be interpreted relative to the estimated standard deviations of U_i and V_i.

3.4 Implementation

We implemented the aforementioned methods (with the exception of Monte Carlo sampling) in `R`. We used the function `GLMMadaptive::mixed_model` to fit the binomial generalized linear mixed model to the score data and the `lme4::lmer` function to fit the linear mixed model to the completion time data (as described in Section 3.2.3). We also used the `nlm` function for maximizing objective functions and the package `gaussquad` to obtain the Gauss-Hermite quadrature points and weights. Otherwise, we wrote our own code.

Chapter 4

Results

Having described the statistical methods we used to analyze our data, we now discuss the fitted models and use them to answer a variety of field-related questions.

4.1 One-Cluster Model: Fit and Interpretation

For our one-cluster model, the parameter estimates and associated standard errors are displayed in Table 4.1.

Parameter	β_0	β_1	β_2	β_3	β_4	β_5
Estimate	−0.688	0.064	0.305	1.105	0.963	0.231
SE	0.109	0.016	0.031	0.060	0.119	0.057

Parameter	α_0	α_1	α_2	α_3	α_4	α_5
Estimate	4.324	−0.115	−0.275	−0.052	−0.044	0.016
SE	0.028	0.004	0.009	0.013	0.028	0.013

Parameter	$\log(\sigma_t)$	$\log(\sigma_u)$	$\log(\sigma_v)$	$\log[(1+\rho)/(1-\rho)]$
Estimate	−0.549	0.928	−0.632	0.080
SE	0.012	0.027	0.025	0.058

Table 4.1: One-cluster model parameter estimates and standard errors

Our primary question of interest—whether repeat exposures are associated with increases in CRT scores—can now be addressed. The 95% confidence interval (CI) for β_1 (the coefficient of `nPrevS`) is [0.033, 0.095], suggesting that the effect of repeat exposures on test scores is indeed positive. The estimated effect of the subjective metric of CRT item exposure, `numSeen`, is also positive, but stronger in magnitude (95% CI [0.245, 0.365]). These estimates

are difficult to interpret concisely because they have a complicated relationship with the mean test score. However, we can estimate and compare mean test scores for different combinations of the predictor variables. Table 4.1 presents estimated mean CRT scores ($\hat{\mathrm{E}}[Y]$) for different values of `nPrevS` and `numSeen` based on our one-cluster model. The standardized predictors `aveSATS` and `age` are set to 0, and `male` is set to 1 (i.e., these estimates are for male subjects with average SAT score and age).

		$\hat{E}[Y]$				$\hat{E}[Y]$	
nPrevS	**numSeen**	$u_i = 0$	Population	**nPrevS**	**numSeen**	$u_i = 0$	Population
1	0	1.209	1.356	1	2	1.662	1.595
2	0	1.255	1.381	2	2	1.709	1.620
3	0	1.302	1.406	3	2	1.756	1.645
4	0	1.350	1.431	4	2	1.803	1.669
5	0	1.397	1.456	5	2	1.848	1.694
1	1	1.434	1.475	1	3	1.883	1.713
2	1	1.482	1.501	2	3	1.927	1.737
3	1	1.530	1.526	3	3	1.971	1.761
4	1	1.578	1.551	4	3	2.014	1.786
5	1	1.626	1.576	5	3	2.056	1.810

Table 4.2: Estimated mean CRT scores, $\hat{E}[Y]$ for the average subject ($u_i = 0$) and the population of subjects, for different values of `nPrevS` and `numSeen`

The first column of estimated mean CRT scores is for an average subject, that is, for a subjects with $u_i = 0$. The second column is for the population of subjects, which we obtained by simulating u_i from the $N(0, \hat{\sigma}_u^2)$ distribution. For a fixed value of `numSeen`, the estimated mean CRT score for an average subject increases by about 0.046 for each unit increase in `nPrevS`, and the estimated mean score for all subjects increases by about 0.025 for each unit increase in `nPrevS`. For a fixed value of `nPrevS`, the estimated mean CRT score for an average subject increases by about 0.223 for each unit increase in `numSeen`, and the estimated mean score for all subjects increases by about 0.118 for each unit increase in `numSeen`. These results call into question the CRT's predictive validity, i.e., test exposure has a non-trivial effect on test scores.

The estimated (approximate) per exposure increase in mean CRT score for an average subject (0.046) contrasts with the estimated 0.024 increase reported by Meyer et al. (2018), who used OLS to estimate this effect by regressing CRT score on number of test exposures. However, the 0.024 estimate is based on the entirety of the Fall 2014 dataset; the estimate would be 0.011 if based on the subset of these data that we use.

As additional confirmation of the effect of `nPrevS` on CRT score, we conduct a likelihood ratio test of $\beta_1 = 0$. The p-value ≤ 0.001, suggesting very strong evidence that score changes with increased exposure.

We estimate that time to completion decreases by 0.115 (95% CI [0.106, 0.124]) log seconds for each additional test exposure. Likewise, we estimate a decrease of 0.275 (95% CI [0.258, 0.292]) log seconds spent on the test with each unit increase in `numSeen`.

Meyer et al. (2018) reported slightly negative correlations between CRT score and time to completion *within* subject (for those who took the test at least twice). We take a different but related approach by estimating the correlation of our random effects, ρ, which describes the correlation between subjects' rationality and reflectiveness *after* accounting for within subject variation due to the predictors. The very weak estimated correlation of 0.040 (95% CI [−0.017, 0.096]) between latent variables is consistent with our descriptive data analysis in Chapter 2 (see Figure 2.4). This estimate suggests that our two response variables are nearly uncorrelated.

4.2 Two-Cluster Model: Fit and Interpretation

The estimates of the parameters of our two-cluster model are displayed in Table 4.3.

Parameter	β_{10}	β_{11}	β_{20}	β_{21}	β_2	β_3	β_4	β_5
Estimate	−0.420	0.024	−0.825	0.070	0.310	1.104	0.962	0.236
SE	0.221	0.047	0.142	0.017	0.031	0.060	0.119	0.057

Parameter	α_{10}	α_{11}	α_{20}	α_{21}	α_2	α_3	α_4	α_5
Estimate	5.134	−0.368	3.958	−0.081	−0.246	−0.058	−0.043	0.024
SE	0.116	0.032	0.053	0.007	0.009	0.013	0.027	0.013

Parameter	$\log(\sigma_t)$	$\log\left(\frac{\gamma_2}{1-\gamma_2}\right)$	$\log(\sigma_u)$	$\log(\sigma_v)$	$\log[(1+\rho)/(1\text{-}\rho)]$
Estimate	−0.612	0.630	0.926	−0.820	−0.033
SE	0.013	0.244	0.027	0.040	0.086

Table 4.3: Two-cluster model parameter estimates and standard errors

We now have two additional fixed effects (an intercept and coefficient of `nPrevS` for the second cluster) for each response variable and a cluster probability parameter. For the CRT score response, the CIs for the intercepts in each cluster [−0.853, 0.013] and [−1.103, −0.547], overlap and the CIs for the effects of `nPrevS`, [−0.068, 0.116] and [−0.037, 0.103],

also overlap, preventing us from drawing any firm conclusions about the differences between the clusters.

On the other hand, subjects in cluster 1 appear to spend more time on their first test than subjects in cluster 2, with estimated intercepts of 5.134 (95% CI [4.907, 5.361]) and 3.958 (95% CI [3.854, 4.062]), respectively. Additionally, subjects in cluster 2 appear to reflect for longer on subsequent tests compared to subjects in cluster 1; the estimated effects of `nPrevS` on time to completion are -0.368 (95% CI [-0.431, -0.305]) and -0.081 (95% CI [-0.095, -0.067]) for clusters 1 and 2, respectively.

The estimates for the cluster probability parameters are $\hat{\gamma}_1 = 0.348$ (95% CI [0.248, 0.462]) and $\hat{\gamma}_2 = 0.652$ (95% CI [0.538, 0.752]).

Though the estimates of the cluster-specific parameters of the model for CRT scores do not point to distinct subpopulations in terms of rationality, the two-cluster model fits substantially better than the one-cluster model. In particular, the maximized value of the log-likelihood is $-12{,}426$, whereas the maximized value of the log-likelihood of the one-cluster model is $-12{,}545$.

The estimates of σ_u based on the one- and two-cluster models are very close: 2.524 (95% CI [2.394, 2.662]) in the one-cluster model vs. 2.489 (95% CI [2.174, 2.850]) in the two-cluster model. The estimates of σ_v are similarly close: 0.440 (95% CI [0.407, 0.476]) in the one-cluster model vs. 0.480 (95% CI [0.406, 0.569]) in the two-cluster model. Using the two-cluster model, the estimate of ρ is still indistinguishable from 0: $\hat{\rho} = -0.017$ (95% CI [-0.100, 0.068]). In other words, after adjusting for the predictors and clusters, we have no evidence of correlation between a subject's rationality and reflectiveness.

4.3 Additional Results

In both models, the single best predictor of CRT score is `aveSATS`. This result reaffirms the commonly reported finding in the literature that CRT and SAT scores are positively correlated and useful predictors of one another, e.g., Frederick (2005). While this association exists, Stanovich et al. (2016) has shown that the underlying latent variables (rationality on the CRT and CART, and intelligence on IQ tests, for which SAT is a significant predictor) are distinct. That is, they are characterized by different cognitive processes and predict different outcomes and traits.

Moreover, whether using the one- or two-cluster model, we estimate a large, positive effect of `male` on CRT score, a moderate-to-weak effect of `age` on CRT score, and negligible effects of `male` and `age` on time to completion. These findings confirm the effects of sex and age on CRT test scores that have been reported in the literature.

Unfortunately, while we attempted to fit the four-cluster model using both AGQ and the EM algorithm with GHQ, we were not able to obtain reliable results in time for this report.

4.4 Model Assessment

As an informal check of the fit of our one-cluster model, we compare the distributions of observed CRT scores and times to completion at `nPrevS=1` to the estimated distributions of the score and time responses using parameter estimates from our fitted model. See Appendix D for the relevant plots and further details on how the distributions were estimated. The estimated distribution of CRT scores corresponds reasonably well to the real data. The estimated distribution of completion times corresponds very closely to the real completion times.

4.5 Random Effects Predictions

Figure 4.1 depicts histograms of the predicted latent variables, $\hat{u}_i$ and $\hat{v}_i$, based on the final parameter estimates of our one-cluster model and step 1 of the iterative estimation procedure discussed in Section 3.2.1. They represent the deviations in rationality and reflectiveness from that of an average subject (i.e., 0), on the scale of each latent variable's estimated standard deviation. For example, since $\hat{\sigma}_u = 2.530$, a value of $\hat{u} = 5.06$ corresponds to a subject with rationality lying two standard deviations above the mean. The apparent bimodal distribution of rationality provides further evidence of two or more clusters.

4.6 Computational Challenges

Fitting our proposed models provided notable computational challenges. Given the two-dimensional integral, the large sample size, and the large number of parameters to be estimated, especially in the cluster models, estimation was a computationally arduous process. Using Google Compute (8 vCPUs, 52 GB memory), we initially used GHQ and the EM algorithm to fit the one-cluster model. Using $Q = 5$ quadrature points, each iteration of the EM algorithm took about 1.5 hours; with $Q = 15$, each iteration took over 8 hours. For the two-cluster model, the average run times were about 2.5 and 20 hours, respectively. Using the large number of quadrature points that would have been necessary to find the MLEs would have been prohibitive. On the other hand, using AGQ, the algorithm for fitting the one-cluster model converged in roughly 2 hours.

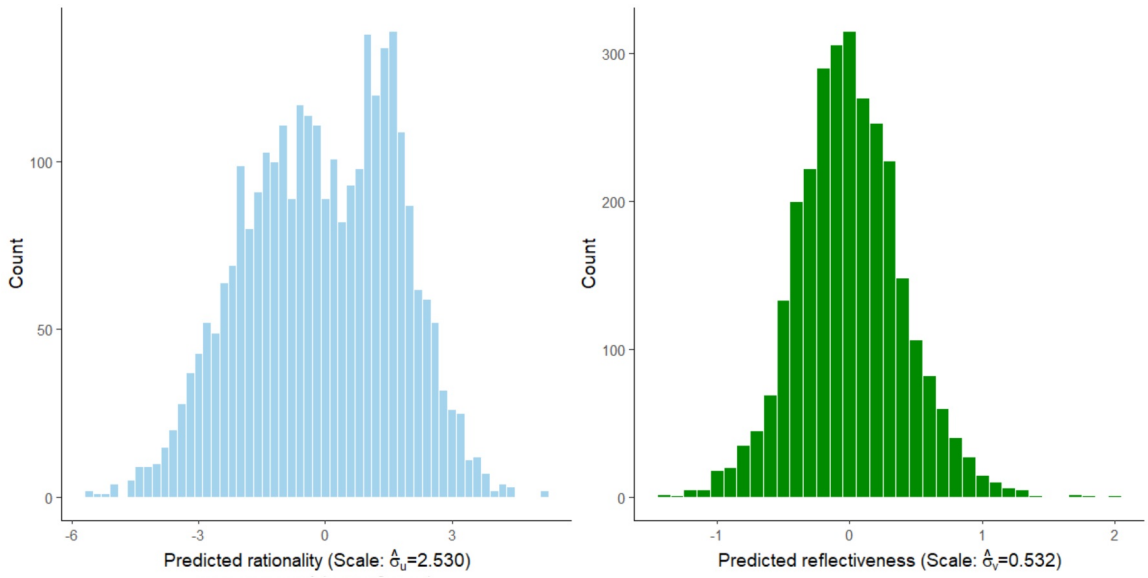

Figure 4.1: Distributions of predicted latent variables

Chapter 5

Discussion and Future Work

We expect that our results provide more trustworthy estimates and associated standard errors of the effect of test exposure (and four other self-reported covariates) on CRT score and time to completion than the original models used by Meyer et al. (2018). In particular, 1) our models more appropriately account for the repeated measures within individual, using information from all exposures rather than simply the difference between final and initial scores; 2) we consider the two response variables jointly, thus using the information in all the available data to estimate the model parameters; and 3) we make more defensible distributional assumptions—namely, we treat the CRT score response as conditionally binomial rather than marginally normal. These differences translate into the larger estimated effect of `nPrevS` on CRT scores that we found compared to Meyer et al. (2018). Our findings suggest that, at least in some contexts, the CRT's predictive validity may not be retained upon repeat exposure.

Though we expected that our two-cluster model would separate subjects according to their levels of rationality and reflectiveness, our results lend support only to distinct subpopulations in terms of reflectiveness. The wide, overlapping CIs of the cluster-specific parameters in the model for CRT scores exemplify the difficulty in estimating this model. The restricted range of the CRT test responses likely contributes to the difficulty in partitioning the sources of variance into within-cluster and between-cluster, which makes separating subjects into distinct clusters a challenge. Moreover, when we fit the two-cluster model to just the CRT scores, the parameter estimates were consistent with our original predictions, i.e., the majority of subjects were classified as belonging to a cluster with low, relatively stable scores and the remainder were classified as belonging to a cluster with higher, increasing scores. When we fit the two-cluster model to just the CRT completion times, the parameter estimates were similar to those based on the joint model, except for that of γ_2, whose sign was reversed. These findings imply that, when fitting the joint model, the completion times

dominate and the scores are forced into inappropriate clusters. In other words, two clusters are insufficient.

Our four-cluster model addresses these issues—and has a nice interpretation, as described in Section 3.1.3. Preliminary results suggest that this model fits substantially better than our other models; its maximized log-likelihood is greater than −12,317, which is far greater than that of our one-cluster model (−12,545) and two-cluster model (−12,426). In addition, preliminary parameter estimates are sensible and consistent with our expectations.

Caution is required in terms of the generalizability of our results. Though MTurk participants are generally regarded as reasonably representative of the population (see Appendix A), our decision to include only observations with self-reported SAT scores presumably biases our sample towards educated American adults. However, as discussed, the inclusion of `numSeen` likely provides important information regarding cognitive proficiency (like mathematical ability and memory), thereby acting in part as a differentiator in reasoning ability.

Though we did not perform a formal simulation study, we fit several versions of our models to simulated data to ensure accuracy of the parameter estimates. We began with a simple version where U_i and V_i are independent, which is equivalent to a standard linear mixed model for the completion time response and a binomial generalized linear mixed model for the score response. The parameter estimates were very accurate, which was reassuring given the highly unbalanced nature of the CRT dataset. We proceeded to test our one-cluster bivariate mixed model, tweaking the sample size, number of repeat measures, included predictors, and number of quadrature points. Reasonable estimates were obtained in each scenario, especially with larger sample sizes. These simulations provide assurance that our proposed methods yield trustworthy results.

Regarding our choice to assume that the random effects are normally distributed, our review of the relevant literature provides some alleviation of concerns about the ramifications of misspecifying these distributions. In the linear mixed model setting, both Butler and Louis (1992) and Verbeke and Lesaffre (1997) demonstrated that incorrectly specifying the distribution of the random effects has a negligible effect on the fixed-effect estimates, while Verbeke and Lesaffre (1997) also demonstrated that the variance estimates are consistent, but that to obtain the correct asymptotic covariance matrix, the Fisher information matrix requires a "sandwich"-type correction. In the generalized linear mixed model setting, Agresti et al. (2004) found that misspecification of the distribution of the random effects when the variance was large could lead to efficiency loss in the prediction of probability outcomes and parameter estimates, suggesting a nonparametric distribution in these cases. Litière et al. (2008) corroborated this finding, additionally discovering that the fixed effect estimates can

become biased as the variance of the random effects becomes high. Given that our starting values for the variance parameters (see end of Section 3.2.3) are not particularly large, we are not too concerned about the aforementioned scenario. However, Litière et al. (2008) caution that, because the estimate of the variance "is the only tool to study the variability of the true random-effects distribution", it is also possible that bias in our starting values could in turn bias the estimates of the fixed effects. We have also made the (perhaps strong) assumption that the random effects distribution does not depend on the predictors, an issue for which Heagerty and Zeger (2000) provide an alternative approach. In the end, we justified our choice of distributions for the random effects by assessing the appropriateness of the implied marginal distributions of the responses, and by relying on the conclusion of McCulloch and Neuhaus (2011) that "most aspects of statistical inference are highly robust to [assuming a normal distribution for the random effects]".

We have numerous ideas for further work in this area. One involves extending our bivariate longitudinal model by treating CRT score as multinomial rather than binomial. This approach was used by Campitelli and Gerrans (2013), who expanded the categories of incorrect CRT responses to distinguish between wrong "intuitive" answers (for example, the "\$0.10" answer on the Bat & Ball problem, or "24 days" on the Lilypads problem) and wrong "idiosyncratic" answers (wrong answers other than the "intuitive" ones). Adopting this approach in the bivariate longitudinal model context may prove informative, though would be even more computationally burdensome.

In each of our proposed models, we elected to treat CRT time to completion as a response variable in our proposed models, whereas Meyer et al. (2018) treated it as a predictor of CRT score. We could take a similar approach by expressing the relationship between the two variables as $f_{\mathbf{Y}_i,\mathbf{T}_i}(\mathbf{y}_i, \mathbf{t}_i) = f_{\mathbf{Y}_i|\mathbf{T}_i}(\mathbf{y}_i|\mathbf{t}_i) f_{\mathbf{T}_i}(\mathbf{t}_i)$, which would allow us to create separate longitudinal models for $\mathbf{T}_i$ and $\mathbf{Y}_i|\mathbf{T}_i$. The latter would treat $\mathbf{T}_i$ as a predictor variable. Exploring (competing) models in this class would be worthwhile.

Future models could also incorporate additional terms, such as the seven-level categorical education variable that we elected to exclude, or an interaction term for `nPrevS` and `numSeen` to address our observation that the influence of `numSeen` was more pronounced at subjects' first test exposures. In addition, we could extend our models to include information that may be contained in the number of tests taken by each subject (the n_i's).

Lastly, because of the interest of cognitive psychologists and psychometricians in understanding the relationship between IQ and metrics of rationality, we could use a nonparametric model (e.g., a random forest) to predict SAT (a strong predictor of IQ) using CRT scores.

Overall, our novel approach in modelling the CRT data allows us to rigorously answer key questions of interest in the cognitive psychology and psychometric literature. We hope that our methods and analysis have contributed meaningfully to this area of inquiry and will motivate future research.

Bibliography

Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. Wiley.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, 47(3):639–653.

Bialek, M. and Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavioural Research Methods*, 50(5):1953–1959.

Butler, S. and Louis, T. (1992). Random effects models with non-parametric priors. *Statistics in Medicine*, 11(14-15):1981–2000.

Campitelli, G. and Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, 42(3):434–447.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42.

Haier, R. (2014). Increased intelligence is a myth (so far). *Frontiers in Systems Neuroscience*, 8(34).

Heagerty, P. and Zeger, S. (2000). Marginalized Multilevel Models and Likelihood Inference. *Statistical Science*, 15(1):1–26.

Kondo, Y., Zhao, Y., and Petkau, J. (2017). Identification of treatment responders based on multiple longitudinal outcomes with applications to multiple sclerosis patients. *Statistics in Medicine*, 36(12):1862–1883.

Litière, S., Alonso, A., and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27(16):3125–3144.

McCulloch, C. and Neuhaus, J. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science*, 26(3):388–402.

Meyer, A., Frederick, S., and Zhou, E. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making*, 13(3):246–259.

Neath, R. (2013). On Convergence Properties of the Monte Carlo EM Algorithm. *The Institute of Mathematical Statistics*, 10:43–62.

Paolacci, G., Chandler, J., and Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

Pinheiro, J. and Chao, E. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 15(1):58–81.

Rabe-Hesketh, S. and Skrondal, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Stanovich, K., West, R., and Toplak, M. (2016). *The Rationality Quotient: Toward a Test of Rational Thinking*. The MIT Press.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, 23(4):541–546.

Wu, J. (1983). On the convergence properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103.

Appendix A

MTurk Reliability

Paolacci et al. (2010) assess the quality of Mechanical Turk (MTurk) participant samples by comparing MTurk samples to university/college student laboratory samples and Internet samples. They proposed various criteria with which to judge the representativeness of MTurk samples, as well as the overall quality of the data the samples produce. This work involved looking at demographic factors (e.g., age, sex, race, and education) and statistical properties of the samples (e.g., coverage error, non-response error, subject motivation, and experimenter effects). Past surveys found that 70-80% of MTurks were from the U.S. More women than men participated (65% vs. 35%). The sample mean and median ages were 36 and 33, respectively, which are slightly lower than those of both the U.S. population and typical Internet users. All MTurk participants must have a bank account in the U.S. Paolacci et al. (2010) summarize, "Our demographic data suggests that Mechanical Turk workers are at least as representative of the U.S. population as traditional subject pools, with sex, race, age and education of Internet samples all matching the population more closely than college undergraduate samples and internet samples in general. . . ".

MTurks are thus thought to be an inexpensive, relatively high quality source of data for psychological experiments. For this reason, we are comfortable with treating our MTurk sample as representative of a relatively well-educated American population for the purpose of our analyses.

Appendix B

CRT Original Questions

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

_____ cents

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

_____ minutes

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

_____ days

Note that modified versions of these questions were given in the other series that we excluded in our analysis.

Appendix C

Further Data Visualization

In Section 2.5, we presented graphs of the distributions of each of our response variables for various values of the predictors. Specifically, in Figures 2.2 and 2.3, we presented histograms of our CRT score response variable for different levels of `nPrevS` and for different levels of `aveSATS` for `nPrevs` = 1. Below are histograms of CRT score for different levels of the other predictors: `numSeen` (Figure C.1), `age` (Figure C.2), and `male` (Figure C.3), each at `nPrevs` = 1. While the CRT score histograms for `age` and `male` do not reveal any substantial effects of these variables, the histograms for `numSeen` reveal a similar pattern to the histograms for `aveSATS`, i.e., right-skewed distributions at the lowest level (in this case, `numSeen` = 0), and bathtub-shaped distributions for higher levels.

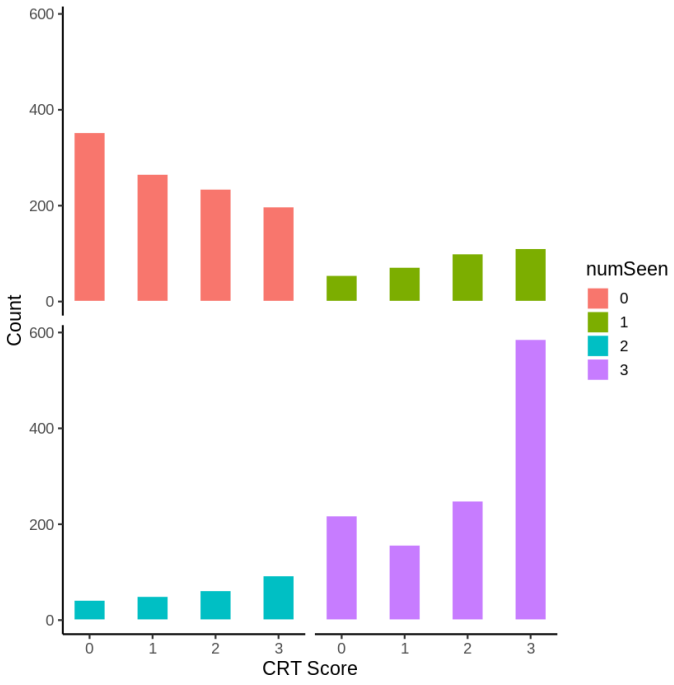

Figure C.1: Distribution of CRT score for `numSeen` at `nPrevS`=1

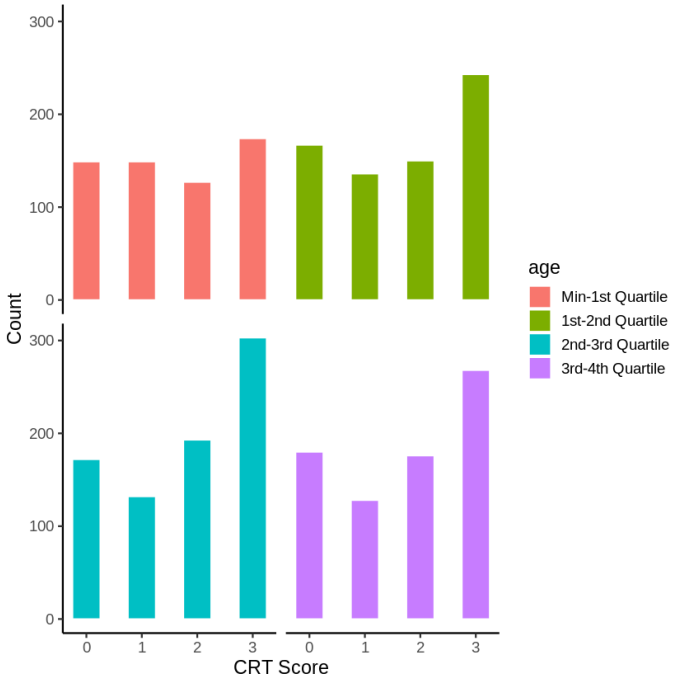

Figure C.2: Distribution of CRT score for `age` at `nPrevS`=1

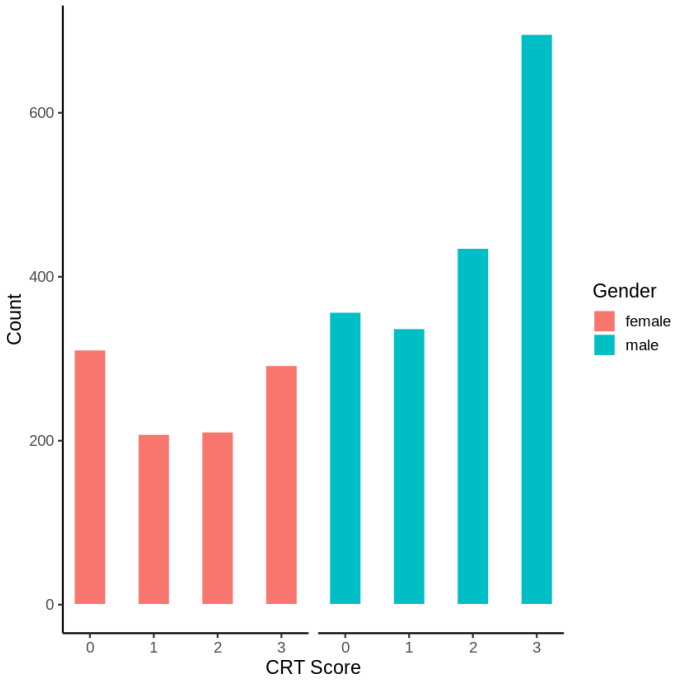

Figure C.3: Distribution of CRT score for `male` at `nPrevS`=1

We also presented histograms of CRT time to completion for different levels of `nPrevS` and for different levels of `numSeen` at `nPrevs` = 1 (see Figure 2.3). Below are histograms of CRT time to completion for different levels of `aveSATS` (Figure C.4), `age` (Figure C.5), and `male`

(Figure C.6) each at `nPrevs` = 1. None of these figures reveals any obvious distributional differences across levels.

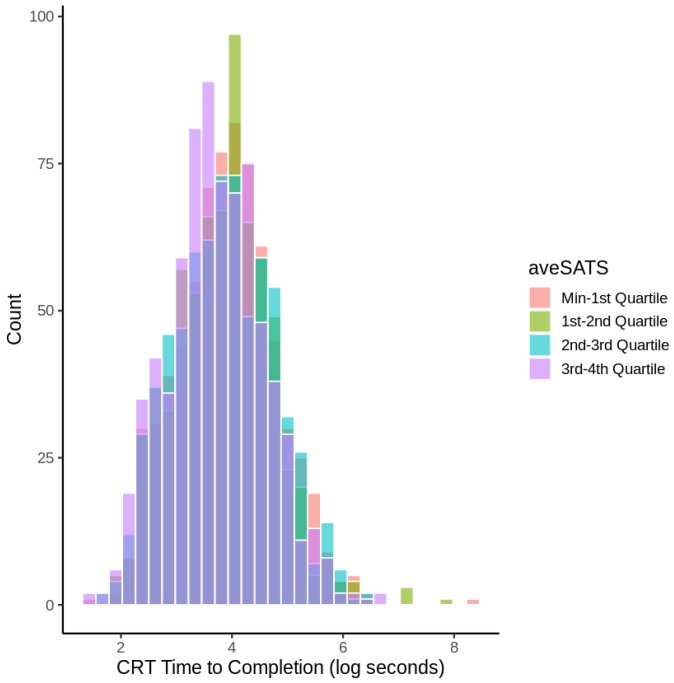

Figure C.4: Distribution of the logarithm of time to completion for `aveSATS` at `nPrevS`=1

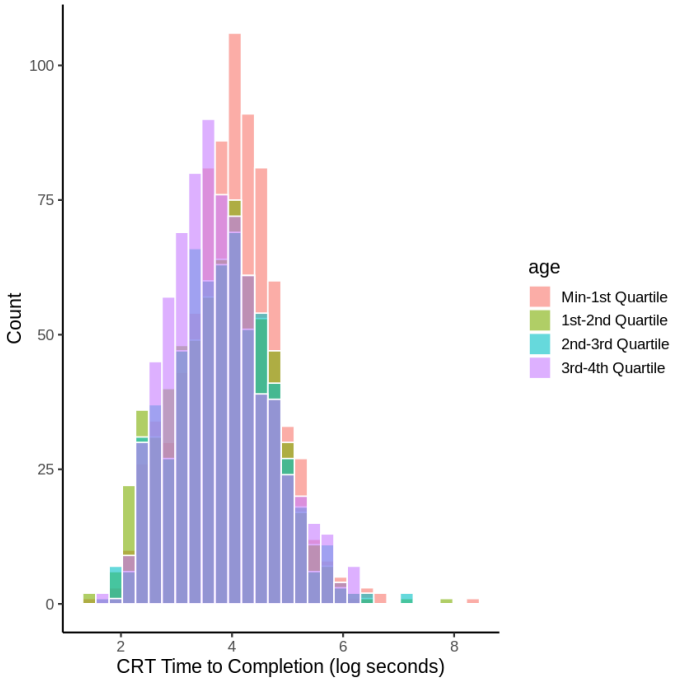

Figure C.5: Distribution of the logarithm of time to completion for `age` at `nPrevS`=1

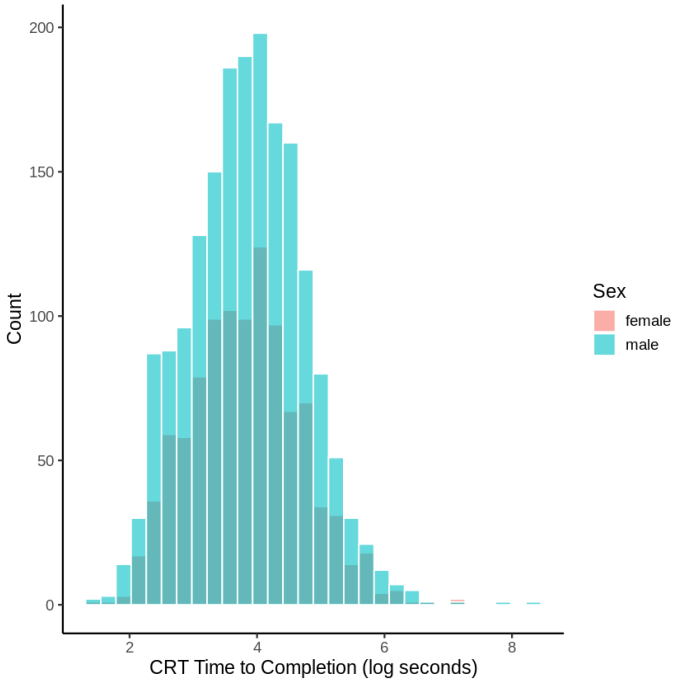

Figure C.6: Distribution of the logarithm of time to completion for `male` at `nPrevS`=1

Additionally, Figure C.7 displays histograms of CRT time to completion for different levels of `numSeen` for `nPrevS` = 2 as a contrast to the histogram on the right side of Figure 2.4 (where `nPrevS` = 1). We can observe that, at subsequent test exposures, the distribution of `numSeen` is slightly right-skewed.

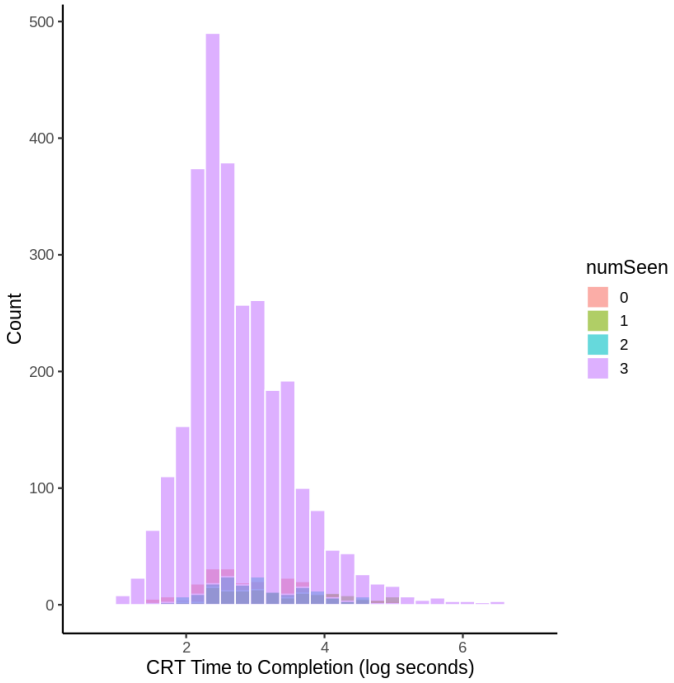

Figure C.7: Distribution of the logarithm of time to completion for `numSeen` at `nPrevS`=2

Appendix D

Further Model Assessment

To provide an informal check of our one-cluster model fit, Figure D.1 displays both the real CRT score and time to completion responses, along with their respective estimated marginal distributions.

For the score response, we estimate the probabilities of each CRT score using the estimated parameters and the observed predictor values, restricted to `nPrevS`=1. Since the marginal distribution of Y_{ij} does not have a closed form, we use Gauss-Hermite quadrature with 100 quadrature points to approximate the four probabilities. The bars on the leftmost plot correspond to the empirical probabilities of success for each CRT score, while the red horizontal lines correspond to the estimated probabilities.

For the time to completion, the marginal distribution has a closed form, namely

$$T_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma_v^2 + \sigma_t^2),$$

where

$$\mu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\alpha}.$$

The histogram on the right reflects the empirical distribution of time to completion, while the curve reflects the estimated distribution.

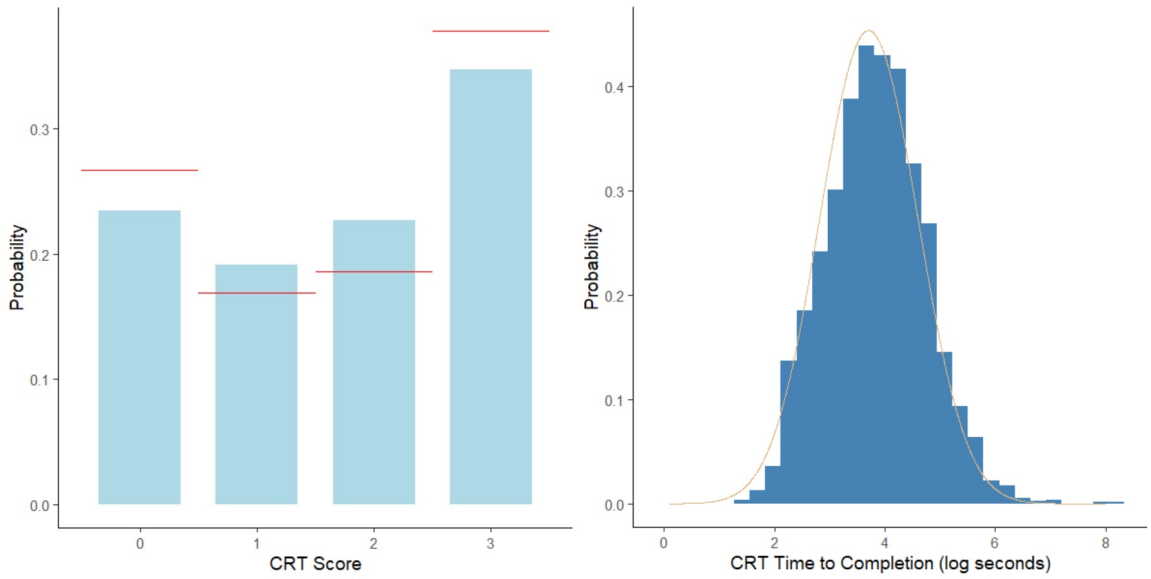

Figure D.1: Observed and estimated distributions of CRT score (left) and time to completion (right) at `nPrevS`=1

Appendix E

Gauss-Hermite Quadrature

As discussed in Section 3.2.3, given sufficient computing resources, standard Gaussian quadrature could be used to evaluate the integrals in our multi-cluster model's objective function, $Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)})$.

Recall that when the weight function is $w(z) = e^{-z^2}$, the GHQ rule is commonly used to determine the weights and abscissae. By performing some variable transformations, we will show that our objective functions are of this form.

We rewrite the joint density of U_i and V_i as

$$f^{(p)}_{U_i,V_i}(u_i, v_i) = f^{(p)}_{V_i|U_i}(v_i|u_i) \cdot f^{(p)}_{U_i}(u_i),$$

where

$$V_i \mid U_i \sim N\left\{\frac{\sigma_v^{(p)}}{\sigma_u^{(p)}}\rho^{(p)}u_i, \left[1 - \left(\rho^{(p)}\right)^2\right]\left(\sigma_v^{(p)}\right)^2\right\},$$

and

$$U_i \sim N\left(0, \left(\sigma_u^{(p)}\right)^2\right).$$

The densities of U_i and $V_i|U_i$ can now be transformed to be amenable to Gauss-Hermite quadrature approximation. Specifically, for $f^{(p)}_{U_i}(u_i)$, let $z_i^2 = \dfrac{u_i^2}{2\left(\sigma_u^{(p)}\right)^2}$. Then, $u_i = \sqrt{2}\sigma_u^{(p)} z_i$ and $du_i = \sqrt{2}\sigma_u^{(p)} dz_i$. For $f^{(p)}_{V_i|U_i}(v_i|u_i)$, let $z_i^{*2} = \dfrac{1}{2\left(s^{(p)}\right)^2}\left(v_i - \dfrac{\sigma_v^{(p)}}{\sigma_u^{(p)}}\rho^{(p)} \cdot u_i\right)^2$, where $(s^{(p)})^2 = \left(1 - \left(\rho^{(p)}\right)^2\right)\left(\sigma_v^{(p)}\right)^2$. Then, $v_i = \sqrt{2}s^{(p)} z_i^* + \dfrac{\sigma_v^{(p)}}{\sigma_u^{(p)}}\rho^{(p)} u_i$ and $dv_i = \sqrt{2}s^{(p)} dz_i^*$.

With these transformations, we can rewrite our objective function as

$$
\begin{aligned}
& Q^{[K]}(\boldsymbol{\psi}, \boldsymbol{\psi}^{(p)}) \\
= \; & \sum_{i=1}^{n} \left[D_i^{[K](p)} \right]^{-1} \sum_{j=1}^{n_i} \iint h_1^{[K](p)}(\boldsymbol{\beta}, z_i, z_i^*) \; e^{-z_i^2} dz_i \; e^{-z_i^{*2}} dz_i^* \\
& + \sum_{i=1}^{n} \left[D_i^{[K](p)} \right]^{-1} \sum_{j=1}^{n_i} \iint h_2^{[K](p)}(\boldsymbol{\alpha}, \sigma_t^2, z_i, z_i^*) \; e^{-z_i^2} dz_i \; e^{-z_i^{*2}} dz_i^* \\
& + \sum_{i=1}^{n} \left[D_i^{[K](p)} \right]^{-1} \iint h_3^{[K](p)}(\sigma_u^2, \sigma_v^2, \rho, z_i, z_i^*) \; e^{-z_i^2} dz_i \; e^{-z_i^{*2}} dz_i^* \\
& + \sum_{i=1}^{n} \left[D_i^{[K](p)} \right]^{-1} \iint h_4^{[K](p)}(\boldsymbol{\gamma}, z_i, z_i^*) \; e^{-z_i^2} dz_i \; e^{-z_i^{*2}} dz_i^*, \qquad \text{(E.1)}
\end{aligned}
$$

where

$$
\begin{aligned}
h_1^{[K](p)}(\boldsymbol{\beta}, z_i, z_i^*) &= \log \left(\left[\frac{(e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i})}{1+e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i}} \right]^{y_{ij}} \left[1 - \frac{(e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i})}{1+e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i}} \right]^{3-y_{ij}} \right) \cdot f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|u_i, v_i, c_i), \\
h_2^{[K](p)}(\boldsymbol{\alpha}, \sigma_t^2, z_i, z_i^*) &= \log \left(\frac{1}{\sigma_t} \exp \left(-\frac{(t_{ij} - (\mathbf{x}'_{ij}\boldsymbol{\alpha} + v_i))^2}{2\sigma_t^2} \right) \right) \cdot f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|u_i, v_i, c_i), \\
h_3^{[K](p)}(\sigma_u^2, \sigma_v^2, \rho, z_i, z_i^*) &= \log \left(f_{V_i|U_i}(v_i|u_i) f_{U_i}(u_i) \right) \cdot f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|u_i, v_i, c_i) \\
h_4^{[K](p)}(\boldsymbol{\gamma}, z_i, z_i^*) &= \log (\gamma_{c_i}) \cdot f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|u_i, v_i, c_i).
\end{aligned}
$$

Evaluating $D_i^{[K](p)}$ requires an additional quadrature step:

$$
\begin{aligned}
D_i^{[K](p)} &\equiv \iint f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|k_i, l_i, c_i) \gamma_{c_i}^{(p)} f^{(p)}_{U_i,V_i}(u_i, v_i) du_i dv_i \\
&\equiv \iint h_0^{[K](p)}(z_i, z_i^*) \; e^{-z_i^2} dz_i \; e^{-z_i^{*2}} dz_i^*,
\end{aligned}
$$

where

$$
h_0^{[K](p)}(z_i, z_i^*) = f^{(p)}_{Y_i,T_i|U_i,V_i,C_i}(y_i, t_i|u_i, v_i, c_i).
$$

With these variable transformations and equation manipulations, we can now approximate the integrals.

However, large numbers of quadrature points are often needed to find the MLEs, which can become prohibitively expensive computationally, and hence we ultimately recommend MCEM as the preferred alternative.