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

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.

Chapter 2

Cognitive Re ection 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 (re ectiveness) that we're interested in capturing.

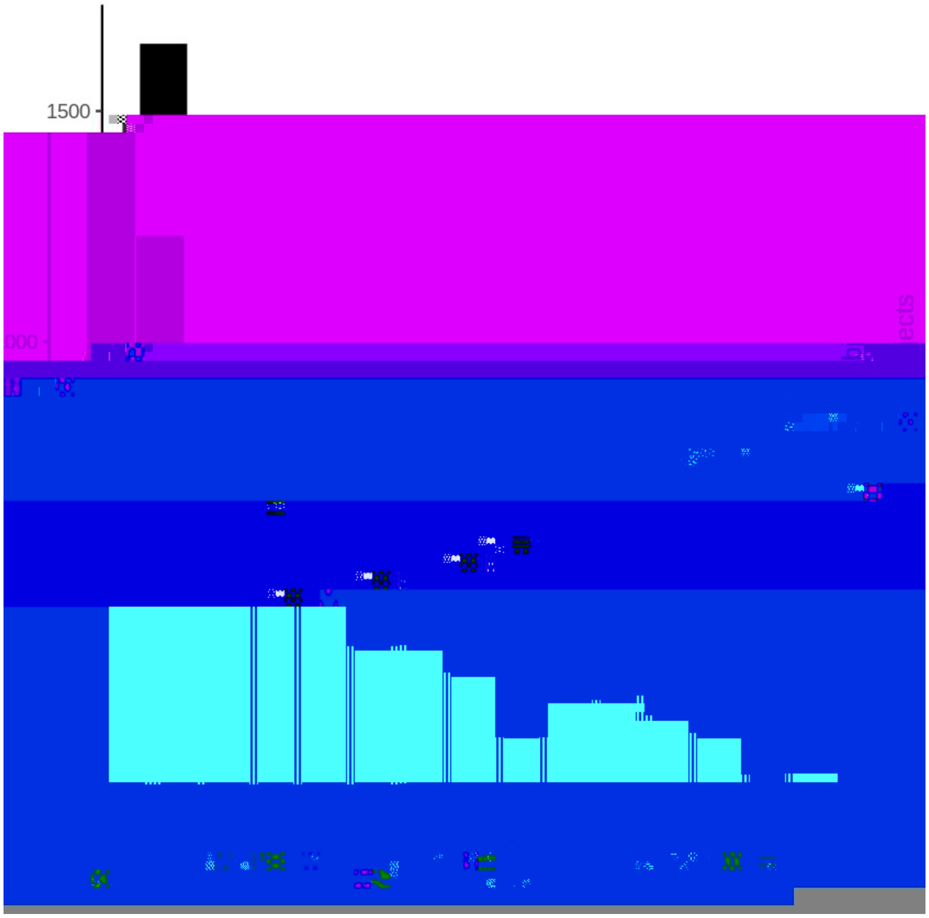

Figure 2.1: Distribution of subjects' exposures, i.e., number of times subjects took the CRT.

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 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

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

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

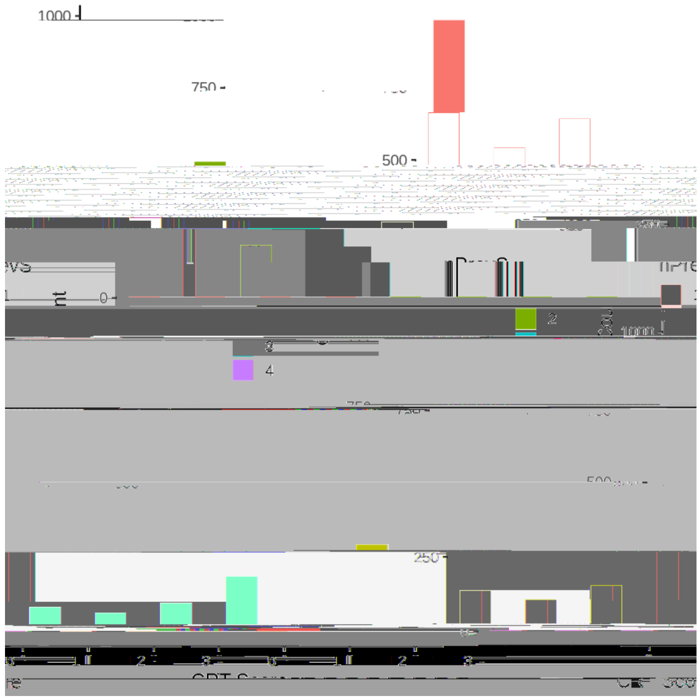

Figure 2.4 displays the distribution of the time response (on the logarithmic scale), broken down by nPrevS (left) and by numSeen at nPrevS = 1 (right). The former graph reveals an approximately normal distribution for each value of nPrevS. We also observe that additional test exposures are associated 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).

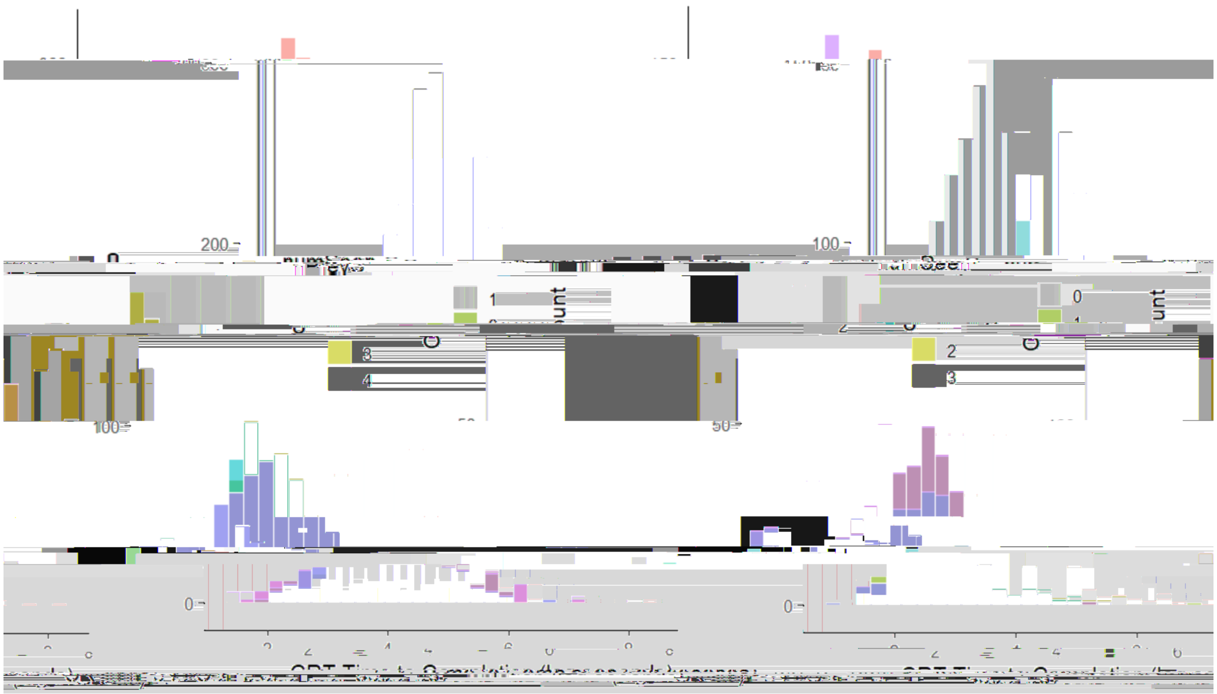

Figure 2.4: Distribution of the logarithm of time to completion for nPrevS $\leq$ 4 (left) and for numSeen at nPrevS=1 (right)

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

0.27 log seconds; and 9% had decreasingCRT 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”.

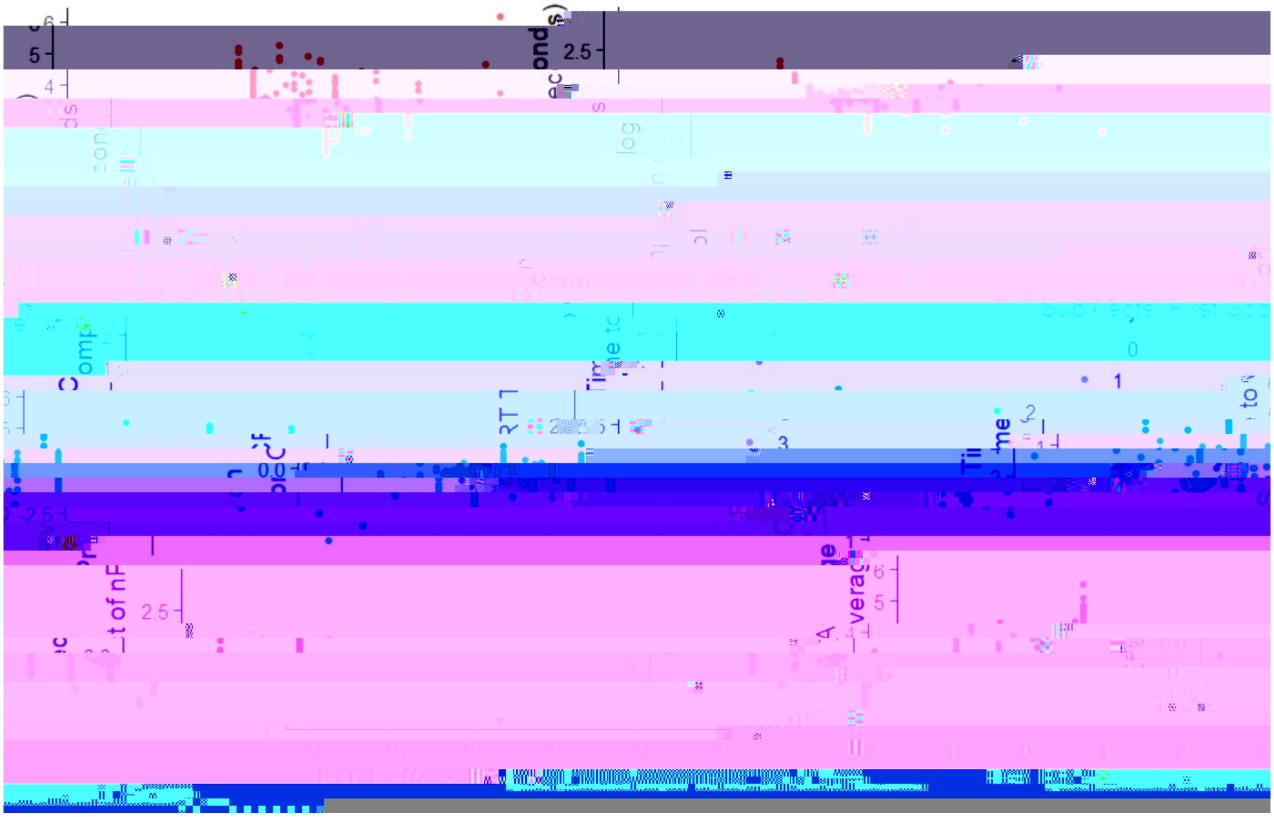

Figure 2.6: Average time to completion (log scale) vs. OLS estimates of the e□ects of nPrevS on CRT score by subjects’ first test score (left); OLS estimates of the e□ects of nPrevS on log time to completion vs. OLS estimates of the e□ects of nPrevS on CRT score by subjects’ first test score (right)

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 inter(the),i1r39:.c3o(63henltimately)83(,)-315(w)28(e)-3

$$\text{logit}\,\theta_{ij} = x_{ij}' \quad + U_i$$

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

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

where

$$\mu_{ij} = x_{ij}' \quad + V_i$$

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,

$$(U_i, V_i) \quad N \quad (0, \quad),$$

where

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

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:

$$L^{[1]}(\quad) = \prod_i \int\int \prod_j f_{Y_{ij}|U_i}(y_{ij}|u_i) f_{T_{ij}|V_i}(t_{ij}|v_i) \quad f_{U_i,V_i}(u_i \quad (\quad) = \prod_i$$

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, T_i, 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 $\bar{x}_{ij}$ 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 $\mathrm{Bin}(3, \theta_{ij})$, where

$$\operatorname{logit} \theta_{ij} = \beta_{c_i}$$

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 $\quad = (\gamma_2, \gamma_3, \gamma_4)$. 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 $\mathrm{Bin}(3, \theta_{ij})$, where

$$\operatorname{logit}\theta_{ij} = \beta_{c_i 0} + \beta_{c_i 1} s_{ij} + \overline{x}_{ij}^{0} \quad + u_i .$$

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

$$\mu_{ij} = \alpha_{c_i 0} + \alpha_{c_i 1} s_{ij} + \overline{x}_{ij}^{0} \quad + v_i .$$

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

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

where $\quad = (\quad, \quad, \sigma_t, \sigma_u, \sigma_v, \rho, \quad)$ 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

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 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 be-

$_{ij}$

associated with the cluster model with K clusters is

$$\ell_c^{[K]}(\;) = \log \left[\prod_i \left(\prod_j f_{Y_{ij} \mid U_i ;C} \right. \right.$$

1f 5.729 -1.216 Td 1.382 08.216 Td [(i)]TJ/F29 70.9091 Tf 8.682 59682436 [(()]TJ/F38 20.9091 Tf 4.2

different objective functions:

$$Q^{[K]}(\ ,\ ^{(p)}) \quad Q_1^{[K]}(\ ,\ ^{(p)}) + Q_2^{[K]}(\ ,\sigma_t^2,\ ^{(p)}) + Q_3^{[K]}(\sigma_u^2,\sigma_v^2,\rho,\ ^{(p)}) + Q_4^{[K]}(\ ,\ ^{(p)}).$$

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, $^{(p)}$. 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

$$L^{[Y]}(\ ) = \prod_i \int \prod_j f_{Y_{ij}|U_i}(y_{ij}|u_i) \quad f_{U_i}(u_i)du_i$$

and

$$L^{[T]}(\ ) = \prod_i \int \prod_j f_{T_{ij}|V_i}(t_{ij}|v_i) \quad f_{V_i}(v_i)dv_i.$$

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 Y and T. To this end, after computing the MLEs of the model parameters, $\hat{}$, 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

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

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.

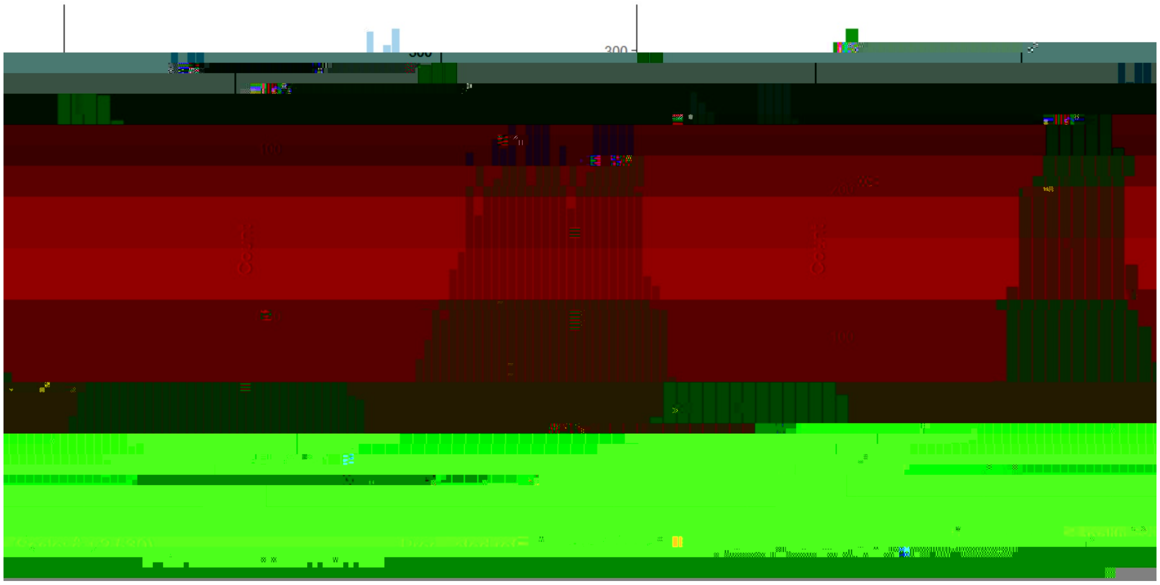

Figure 4.1: Distributions of predicted latent variables

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 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

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., Ca□o, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random e□ects distribution reduces e□ciency, 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.

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 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

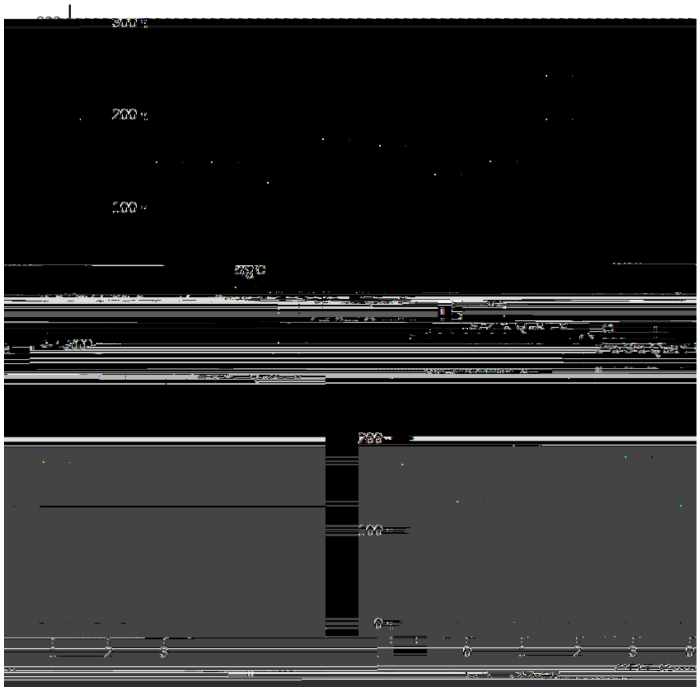

Figure C.2: Distribution of CRT score for age at nPrevS=1

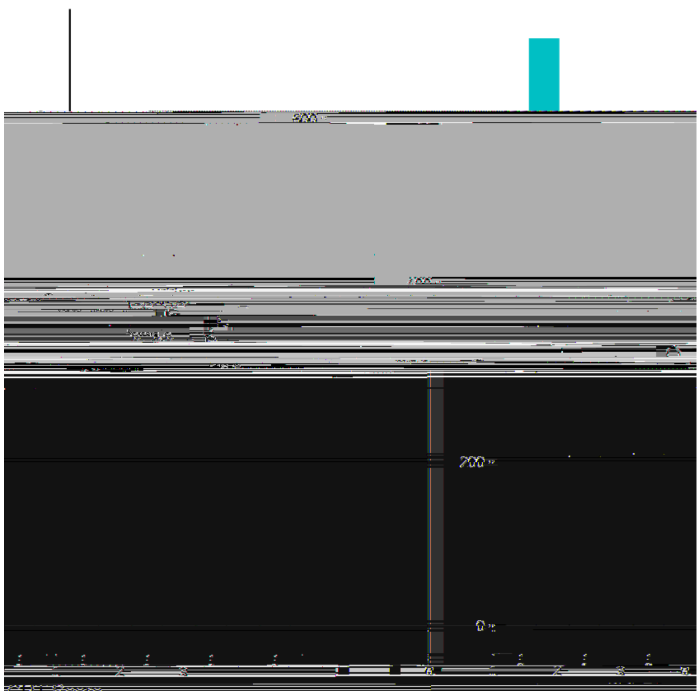

Figure C.3: Distribution of CRT score for male at nPrevS=1

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.

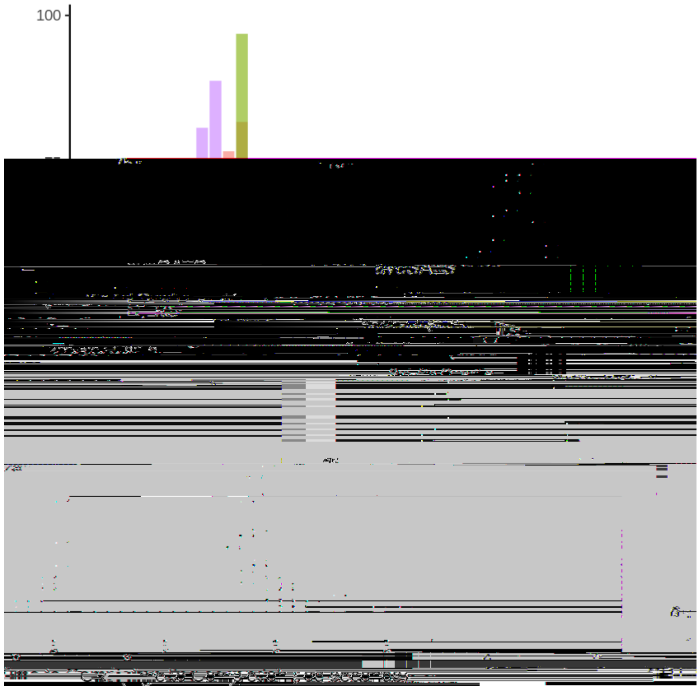

Figure C.4: Distribution of the logarithm of time to completion for aveSATS at nPrevS=1

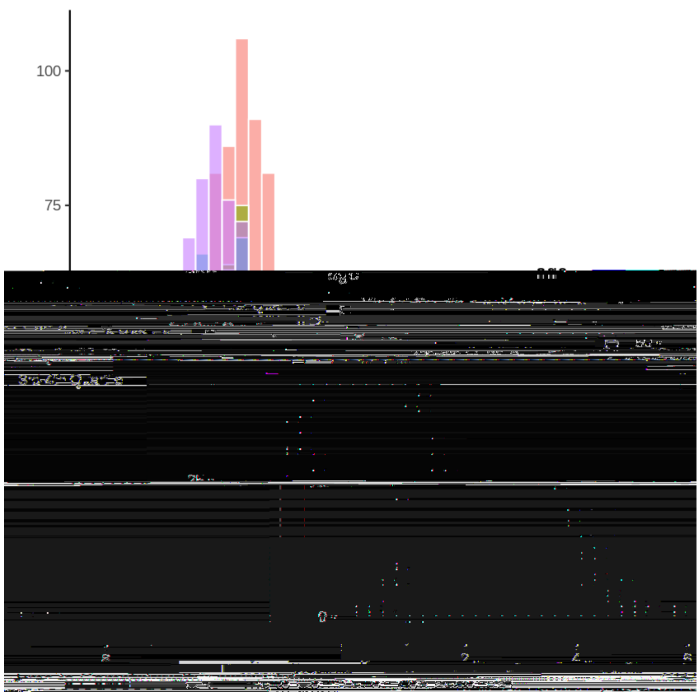

Figure C.5: Distribution of the logarithm of time to completion for age at nPrevS=1

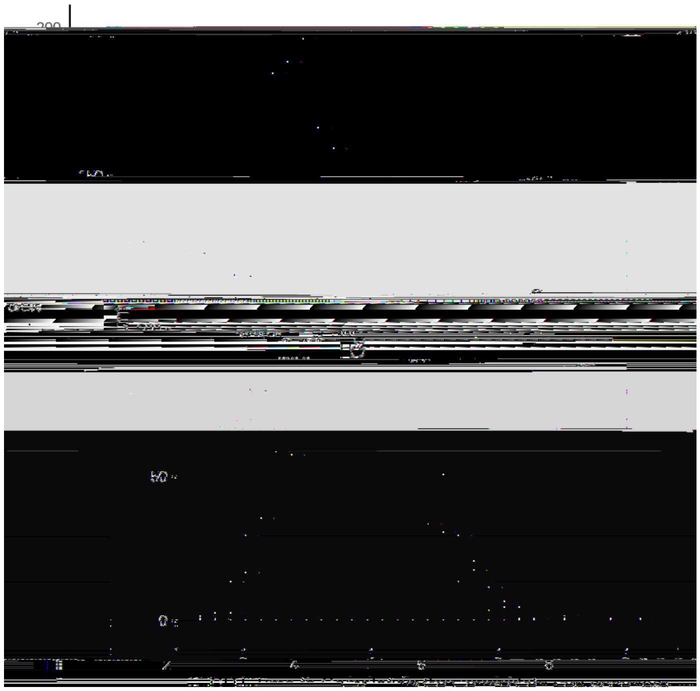

Figure C.6: Distribution of the logarithm of time to completion for male at nPrevS=1

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.

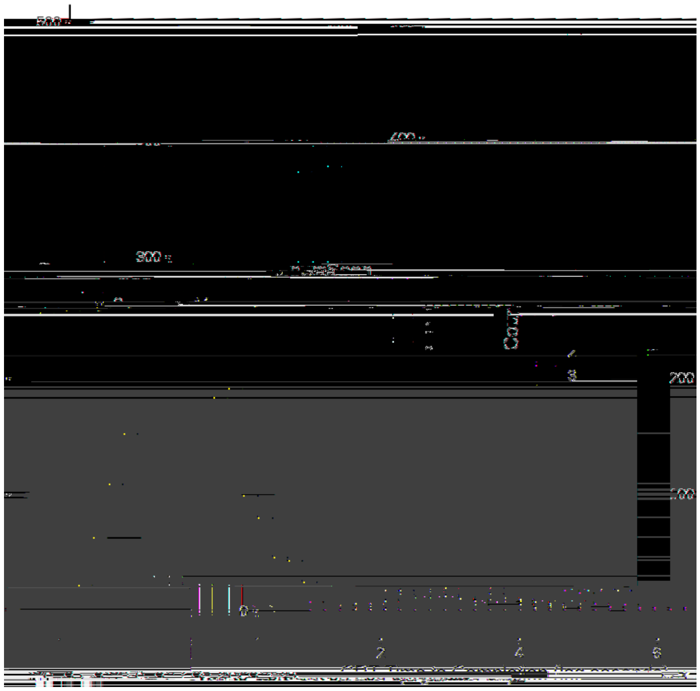

Figure C.7: Distribution of the logarithm of time to completion for numSeen at nPrevS=2

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

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

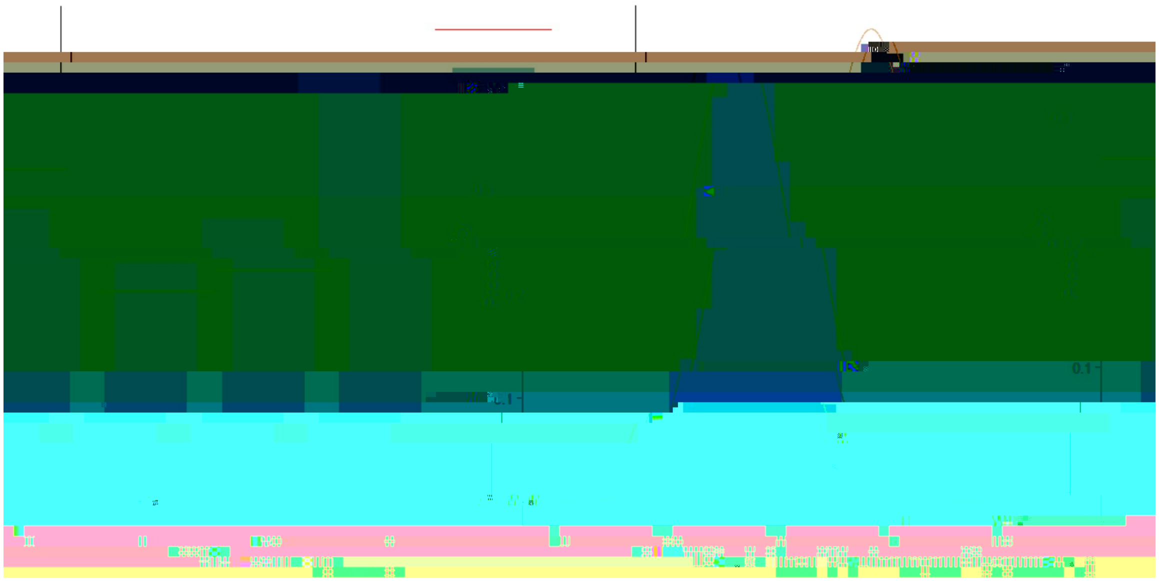

Figure D.1: Observed and estimated distributions of CRT score (left) and time to completion (right) at nPrevS=1

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, $Q^{[K]}(\; , \; ^{(p)})$.

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

$$f^{(p)}_{U_i;V_i}(u_i, v_i) = f^{(p)}_{V_i}$$

With these transformations, we can rewrite our objective function as

$$
\begin{aligned}
& Q^{[K]}(\ , \ ^{(p)}) \\
= & \sum_{i=1}^{n} \Big[D_i^{[K](p)} \Big]^{i} {}_{1} \sum_{j=1}^{n_i} \iint h_1^{[K](p)}(\ , z_i, z_i) \, e^{\ z_i^2} dz_i \ e^{\ z_i \ 2} dz_i \\
& + \sum_{i=1}^{n} \Big[D_i^{[K](p)} \Big]^{i} {}_{1} \sum_{j=1}^{n_i} \iint h_2^{[K](p)}(\ , \sigma_t^2, z_i, z_i) \, e^{\ z_i^2} dz_i \ e^{\ z_i \ 2} dz_i \\
& + \sum_{i=1}^{n} \Big[D_i^{[K](p)} \Big]^{i} {}_{1} \iint h_3^{[K](p)}(\sigma_u^2, \sigma_v^2, \rho, z_i, z_i) \, e^{\ z_i^2} dz_i \ e^{\ z_i \ 2} dz_i \\
& + \sum_{i=1}^{n} \Big[D_i^{[K](p)} \Big]^{i} {}_{1} \iint h_4^{[K](p)}(\ , z_i, z_i) \, e^{\ z_i^2} dz_i \ e^{\ z_i \ 2} dz_i \, ,
\end{aligned}
\tag{E.1}
$$

where

$$
h_1^{[K](p)}(\ , z_i, z_i) \ =
$$