






# A bivariate longitudinal cluster model with application to the Cognitive Reflection Test

Matthew Berkowitz<sup>a</sup>   and Rachel MacKay Altman<sup>a</sup> 

<sup>a</sup>Simon Fraser University, Burnaby, British Columbia, Canada.

**Abstract** ■ The Cognitive Reflection Test (CRT) is a test designed to assess subjects' ability to override intuitively appealing but incorrect responses. Psychologists are concerned with whether subjects improve their scores on the test with repeated exposure, in which case, the test's predictive validity may be threatened. In this paper, we take a novel approach to modelling data recorded on subjects who took the CRT multiple times. We develop bivariate, longitudinal models to describe the responses, CRT score and time taken to complete the CRT. These responses serve as a proxy for the underlying latent variables "numeracy" and "reflectiveness", respectively—two components of "rationality". Our models allow for subpopulations of individuals whose responses exhibit similar patterns. We assess the reasonableness of our models via new visualizations of the data. We estimate their parameters by modifying the method of adaptive Gaussian quadrature. We then use our fitted models to address a range of subject-specific questions in a formal way. We find evidence of at least three subpopulations, which we interpret as representing individuals with differing combinations of numeracy and reflectiveness, and determine that, in some subpopulations, test exposure has a greater estimated effect on test scores than previously reported.

**Keywords** ■ Bivariate Longitudinal Model; Cluster Model, Gaussian Quadrature, Adaptive Quadrature; Mixed Model; Cognitive Reflection Test.

 [mberkowi@sfu.ca](mailto:mberkowi@sfu.ca)

 [10.20982/tqmp.18.1.p021](https://doi.org/10.20982/tqmp.18.1.p021)

**Acting Editor** ■ Denis Cousineau (Université d'Ottawa)

**Reviewers**

■ One anonymous reviewer

## Introduction

The Cognitive Reflection Test (CRT) (Frederick, 2005) was developed to assess a subject's ability to override an incorrect but intuitively appealing response (a so-called "gut instinct"), a key component of *rationality*. The CRT is a short, three-question test that is predictive of many cognitive abilities and tendencies (Bialek & Pennycook, 2018). It was a precursor to the Comprehensive Assessment of Rational Thinking (CART), a more in-depth test of rationality currently being developed (Stanovich, West, & Toplak, 2016). Both tests can provide information about subjects' rationality. The CRT focuses specifically on *numeracy*, an aspect of rationality concerned with the ability to reason and apply concepts involving numbers (Attali & Bar-Hillel, 2020; Erceg, Galic, & Ružojčić, 2020). Numeracy in this context can be operationalized as the number of correct re-

sponses to the test questions. The tests can also yield insight into subjects' *reflectiveness*, which we define as the quality of considering a question carefully rather than reporting the first response that springs to mind—using so-called "System 2" thinking, as per Kahneman (2013). Reflectiveness can be operationalized as time spent completing a test of rationality.

A key question in the literature is whether subjects tend to improve their test scores over time (for example, via repeated exposure to the same test questions), in which case the tests may not retain their predictive validity.

In the case of intelligence, an aspect of cognitive ability that is related to (though distinct from) rationality, the literature provides no convincing evidence that intelligence quotient (IQ) test scores improve over time (Haier, 2014). But, with respect to rationality (as measured by scores on the CRT or CART), the literature is so far sparse. The first



authors to assess this question were Meyer, Frederick, and Zhou (2018), who administered the CRT to subjects multiple times over a predefined time period. They affirmed that test scores remain approximately stable over time, a conclusion reached in other investigations as well (Stagnaro, Pennycook, & Rand, 2018).

Our research extends the work of Meyer et al. (2018), who attempt to answer various questions about changes in subjects' CRT scores over time. They use conventional linear regression modelling—regressing CRT score on the current test exposure number and the total number of test exposures. Furthermore, via various tables and graphs, they present descriptive statistics that explore informally how test scores change over time, e.g., they compare average CRT scores for different values of other variables. These models and expositions do not fully account for the longitudinal nature of the data, the dependence among responses measured on the same individual, the discreteness of the test scores, or the role of other predictor variables. Though Meyer et al. (2018) intimates that the CRT dataset suggests the presence of subpopulations, their models do not incorporate them explicitly. To address these limitations, we develop a bivariate longitudinal model for these authors' data. Our model describes 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 also present an extension that allows a different bivariate longitudinal model for different subpopulations of individuals via a latent cluster variable.

Our model is a special case of the multiple longitudinal outcome mixture model (MLOMM) developed by Kondo, Zhao, and Petkau (2017), which is an extension of the generalized linear mixed model (GLMM) (Agresti, 2013) to include additional response variables and multiple latent clusters. However, for estimating the parameters of these models, we take a different approach from Kondo et al. (2017), who use a Monte Carlo expectation-maximization (MCEM) algorithm to estimate the parameters of a two-cluster MLOMM. Specifically, we modify the adaptive Gaussian quadrature (AGQ) approach proposed by Pinheiro and Chao (2006) for estimating the parameters of GLMMs. Our method allows for parameter estimation even in models with more than two clusters.

The rest of this paper is organized as follows. In the next section, we describe the CRT dataset and provide novel visualizations of its features. We then present our models and new estimation method. Subsequently, 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.

## Cognitive Reflection Test Data

### CRT data 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 cleaning (see the next two sections), 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 1 summarizes the distribution of this variable.

### Responses of interest

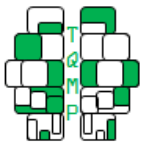
Meyer et al. (2018) treated CRT scores as the sole response variable in their analyses (using the time that subjects took to complete the test as a predictor in one). In contrast, we consider time to completion as a second response variable. We view test score and time to completion as the operationalizations of numeracy and reflectiveness, respectively—two components of rationality. Individual test scores range from 0 to 3 and completion times range from 2 to 4002 seconds (or 0.69 to 8.29 log seconds).

### Predictors

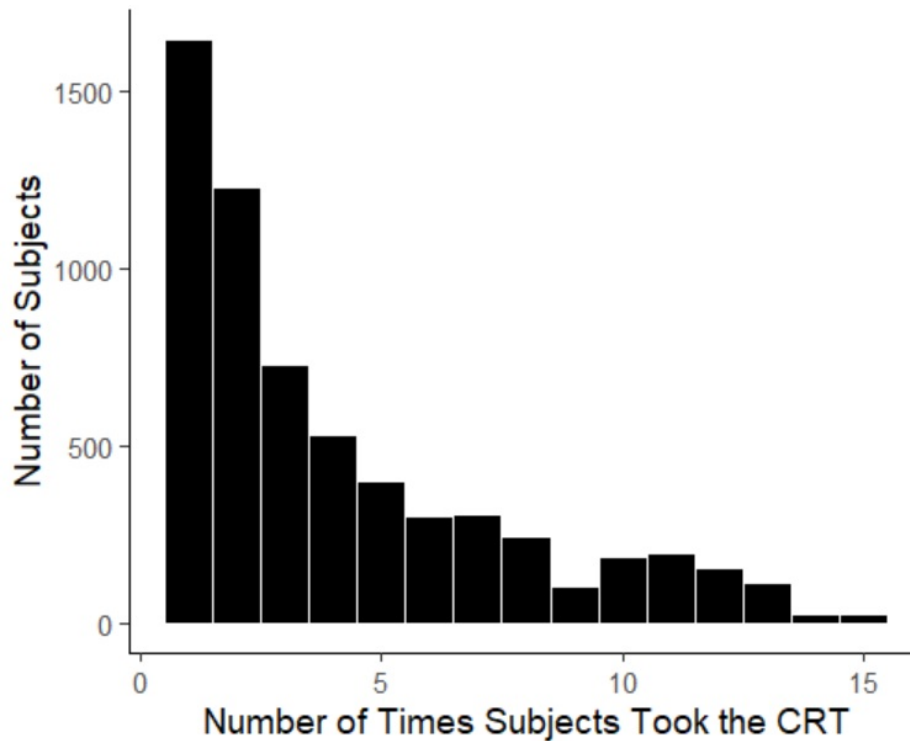
Various predictor variables may influence the distribution of our two response variables. In this section, we discuss our selection of these variables and our handling of idiosyncratic and missing values.

Our primary predictor of interest (the exposure variable, in the language of causal inference) is the number of times a subject has taken the CRT *within* the series, including the current test. This variable is denoted by  $n_{PrevS}$  and takes values from 1 to 15. It is a time-varying, numeric predictor. We acknowledge that subjects may have seen the CRT questions prior to participating in this study, but  $n_{PrevS}$  remains our best objective measure of exposure.

Subjects self-reported the number of CRT questions they had seen previously ( $numSeen$ ), a numeric variable taking 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 three CRT items. However, subjects don't always report "3" after the first



**Figure 1** ■ Distribution of subjects' exposures, i.e., number of times subjects took the CRT.



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 an appropriate transformation. As Meyer et al. (2018) noted, the original variable could be informative not only for its intended purpose (measuring number of CRT items seen) but also as a proxy for a subject's memory of the CRT and mathematical ability. In this spirit, we convert it to a categorical variable (denoted by *memory*), defined as "0" when *nTotal*=1 (i.e., the case where *numSeen* gives no information about the true number of items seen by the subject), "1" when *numSeen* provides evidence of a faulty memory (i.e., *numSeen* < 3 for any test exposure other than the first), and "2" when *numSeen* = 3 for *nPrevS* > 1 (i.e., the case where a subject may be reporting *numSeen* correctly). Both *numSeen* and *nPrevS* measure familiarity with the CRT—albeit one subjectively and the other objectively. However, *memory* presumably captures indirect information about memory not reflected in *nPrevS*.

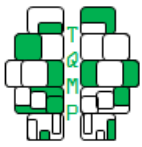
The predictor *aveSATs* refers to a subject's standardized average self-reported SAT score over the course of the Fall 2014 series.

The binary categorical predictor *male* denotes a subject's self-reported sex (with values of "0" and "1" cor-

responding to female and male, respectively). However, *male* was not always constant throughout the series. In the case of only two observations per subject with different values of *male*, we exclude both observations; otherwise, we replace discrepant values with the most commonly used value reported by the subject.

Next, *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. In most cases, we use their self-reported age at first exposure to avoid time variance due to birthdays during the study period. However, some subjects' reported ages vary by more than 1 year, indicating errors. In these cases, 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.

Lastly, we also include *nTotal*, the total number of times that a subject takes the test, as a predictor in our models. This choice mirrors that of Meyer et al. (2018) in one of their analyses. Although Meyer et al. (2018) do not provide an explanation for their choice, adjusting for *nTotal* is sensible. Our rationale is that less motivated

**Table 1** ■ CRT variables selected

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)
memory	Explanatory (Categorical)	Faulty memory
aveSATS	Explanatory (Continuous)	SAT score (standardized)
male	Explanatory (Categorical)	Sex
age	Explanatory (Continuous)	Age (standardized)
nTotal	Explanatory (Discrete)	Total number of test exposures
identifier	Random factor	Subject ID

subjects may be more likely to have both lower scores and lower values of `nTotal`, i.e., we can view the differing values of `nTotal` as arising due to a missing test score problem. Including `nTotal` in the model is a way of adjusting for subjects' motivation; after this adjustment, we view the missing test scores as missing at random.

Table 1 summarizes the response and predictor variables.

A substantial number of observations have missing values for at least one predictor. See Supplementary Information for details on how we handled missing data.

Our final dataset is intended to be a sample from the population of relatively well-educated American adults. It contains 6,228 observations from the Fall 2014 series.

### Data visualization

We now provide further visualizations of the dataset to explore and motivate our proposed models in the next section.

First, we examine the CRT score distribution. Histograms of CRT score for different values of `nPrevS` are shown in Figure 2 (we omit the cases where `nPrevS`  $\geq 4$  due to lack of data) and for different categories of `aveSATS` at `nPrevS` = 1 in Figure 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.

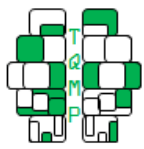
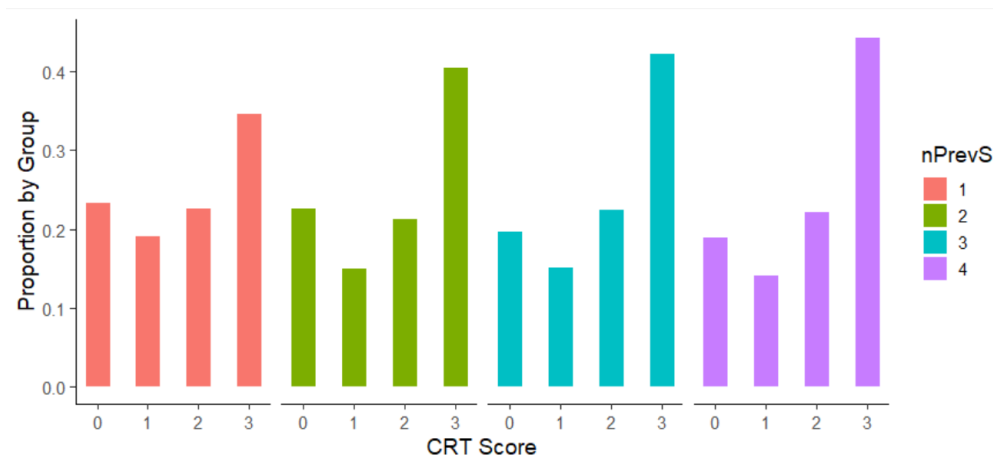
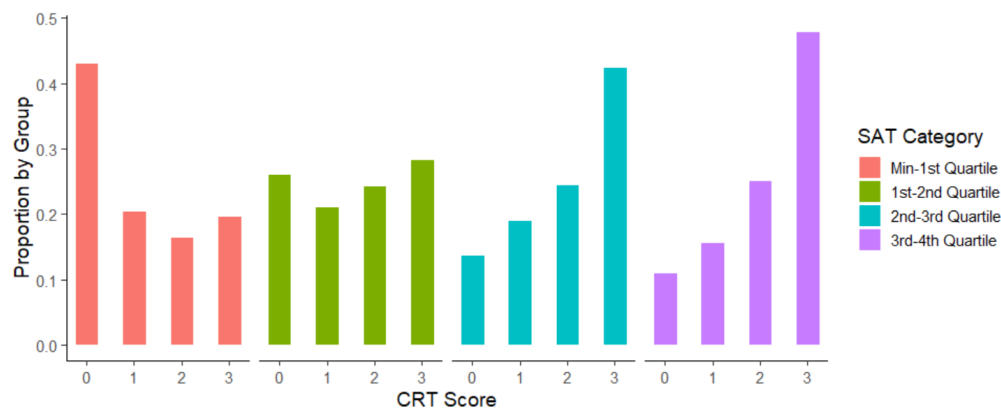
Figure 4 displays the distribution of the time response (on the logarithmic scale), broken down by whether subjects took the test once or multiple times. The graph reveals an approximately normal distribution for both groups, i.e., for `nPrevS` = 1 and `nPrevS` > 1. We also observe that additional test exposures are associated with lower times to completion. Histograms of completion time at different

levels of the other predictors (not shown) similarly reveal approximately normal distributions but with no indication of distributional differences among the different levels.

Next, Figure 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 `nPrevS` 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 has a peak near 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 proportion of subjects who spent an increasing amount of time on additional exposures.

Lastly, we explore the changes in the two response variables over time. 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, with 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, with an average CRT score decrease of 0.60 and an average decrease in time spent of 0.42 log seconds. Although all subjects decreased their time spent on subsequent tests, on average, in the aforementioned groups (constant, decreasing, and increasing scores), this decrease was least for the small subset who improved their test scores over time.

The scatterplots in Figure 6 also illustrate the relationship between changes in scores and completion times for subjects who took the test multiple times. The left panel

**Figure 2** ■ Distribution of CRT score by  $nPrevS$ **Figure 3** ■ Distribution of CRT score by  $aveSATs$  (for  $nPrevS=1$ )

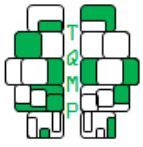
shows subjects' average log time to completion vs. changes in test scores (as estimated by OLS, separately for each subject). This scatterplot suggests a positive correlation between change in CRT score and average time spent on the test, i.e., subjects who improved their scores tended to spend more time on the test, and vice versa. The right panel shows changes in log times to completion (also estimated by OLS, separately for each subject) vs. changes in test scores. Similarly, the second scatterplot suggests a positive correlation between change in CRT score and change in completion time. In this case, subjects who improved their scores (positive estimated effect of  $nPrevS$  on score) tended to spend an approximately constant amount of time on each test (estimated effect of  $nPrevS$  on completion time of approximately 0). Subjects who did not improve

their scores (non-positive estimated effect of  $nPrevS$  on score) tended to spend a decreasing amount of time on each test (negative estimated effect of  $nPrevS$  on completion time).

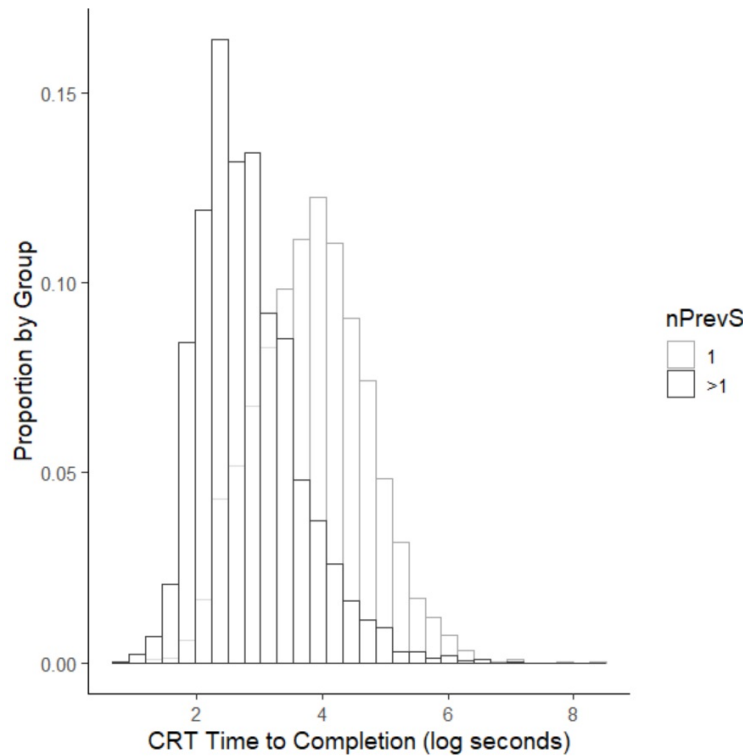
These plots use data only from subjects who appeared more than once in the series and whose first test scores were 0 or 1 (since only those subjects had the possibility of improving their scores substantially). These statistics and scatterplots are consistent with the observation by Meyer et al. (2018) that a small proportion of subjects “continue to spend time on the test”.

### Statistical Methods

To model our bivariate longitudinal data and explore the existence of subpopulations, we consider the class of mod-



**Figure 4 ■** Distribution of the logarithm of time to completion by nPrevS



els proposed by Kondo et al. (2017). In the following sections, we describe models that are appropriate for the CRT data and, in particular, the computational challenges that can arise in estimating their parameters. Ultimately, we propose two models. The first serves as our foundational model for testing the hypotheses specified by Meyer et al. (2018). The second extends the first to allow for the description of subpopulations (“clusters”) of individuals whose responses exhibit similar patterns (after adjusting for predictor variables).

### 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,  $i = 1, \dots, n$ ,  $j = 1, \dots, n_i$ . 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. Throughout, we use the notation  $f_G$  to denote the probability mass (or density) function of a random variable  $G$ .

#### Bivariate Longitudinal Model

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

$$Y_{ij} \mid U_i = u_i \sim \text{Binomial}(3, \theta_{ij}),$$

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 numeracy. Likewise, we model the logarithm of the time to completion as

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

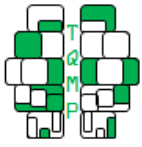
where

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

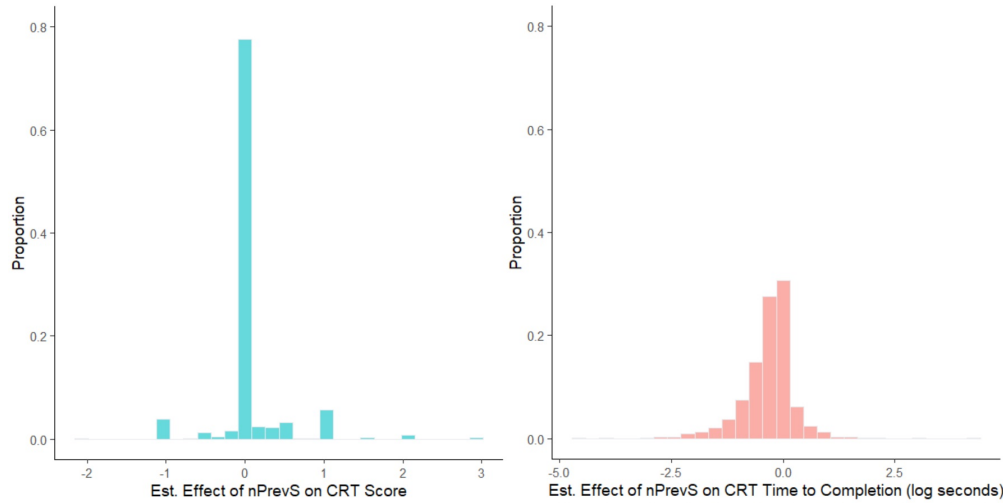
and where the random effects,  $V_i$ , are independent and distributed as  $N(0, \sigma_v^2)$ . We conceive of  $V_i$  as a latent variable representing reflectiveness. Among other implications, the random effects allow for correlation among scores and times to completion observed on the same individual.

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





**Figure 5 ■** OLS estimates of the effects of nPrevS when CRT score is regressed on nPrevS separately for each subject (left); and when CRT log time to completion is regressed on nPrevS separately for each subject (right).



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

where

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

The purpose of allowing  $U_i$  and  $V_i$  to be correlated (via the parameter  $\rho$ ) is to allow dependence between any score and any completion time observed on the same subject.

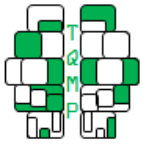
The assumption that test scores are conditionally binomial distributed may, at first, seem suspect because the outcomes (correct/incorrect) of the three questions posed to each individual at each exposure are not necessarily independent with common probability of success. However, Figures 2 and 3 help to justify the model for  $Y_{ij} | U_i$ . 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 mixing distribution is a normal distribution, i.e., our specified distribution of  $Y_{ij} | U_i$ . 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).

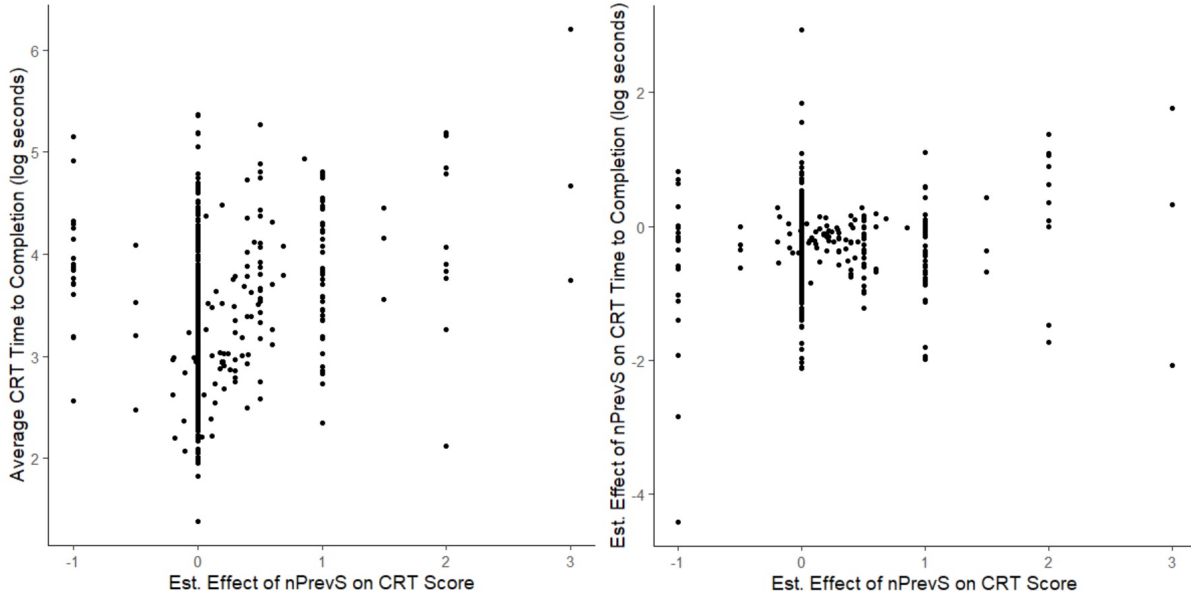
Figure 4 motivates the model for  $T_{ij} | V_i$ ; the histogram of the logarithm of time to completion given combinations of predictor variables reveal that the marginal distribution of  $T_{ij}$  is approximately normal. From this perspective, the proposed models for  $T_{ij} | V_i$  and  $V_i$  (which imply that  $T_{ij}$  is normally distributed) are reasonable.

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

$$\begin{aligned} \mathcal{L}^{[1]}(\psi) &= \prod_{i=1}^n \int_{v_i=-\infty}^{\infty} \int_{u_i=-\infty}^{\infty} \left( \prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_i) f_{T_{ij}|V_i}(t_{ij}|v_i) \right) \cdot f_{U_i, V_i}(u_i, v_i) du_i dv_i \\ &= \prod_{i=1}^n \int_{v_i=-\infty}^{\infty} \int_{u_i=-\infty}^{\infty} \left\{ \prod_{j=1}^{n_i} \theta_{ij}^{y_{ij}} (1 - \theta_{ij})^{3-y_{ij}} \cdot \frac{1}{\sigma_t} \exp\left(-\frac{(t_{ij} - \mu_{ij})^2}{2\sigma_t^2}\right) \right\} \\ &\quad \cdot \frac{1}{\sigma_u \sigma_v \sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)} \left(\frac{u_i^2}{\sigma_u^2} + \frac{v_i^2}{\sigma_v^2} - \frac{2\rho u_i v_i}{\sigma_u \sigma_v}\right)\right\} du_i dv_i, \end{aligned} \quad (1)$$



**Figure 6 ■** Average time to completion (log scale) vs. OLS estimates of the effects of nPrevS on CRT 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 (right). Each plot is restricted to subjects who took the test more than once and whose first test scores were 0 or 1.



where  $\psi = (\beta, \alpha, \sigma_t, \sigma_u, \sigma_v, \rho)$  is the 20-dimensional vector of parameters to be estimated. We omit terms that are constant with respect to the unknown parameters. In addition, we use superscripts with square brackets to denote the number of clusters in the model.

#### Bivariate Longitudinal Model with Latent Clusters

Our second proposed model extends the first model by postulating that test subjects comprise distinct clusters. We justify this model on two grounds. First, Meyer et al. (2018), as part of their analysis, imply that the presence of subpopulations might drive some of their findings. Second, the multi-cluster model can be seen as an alternative way to represent the marginal distribution of the CRT scores depicted in Figures 2 and 3. In particular, this (bimodal) distribution could arise due to at least two distinct clusters of individuals. Our original model can be considered a special case of this extended model.

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  $C_i \in \{1, 2, \dots, K\}$  be a latent cluster indicator. We assume that the  $C_i$ 's are independent and distributed as  $P(C_i = c) = \gamma_c$ . As per our original model, we take

$U_i$  and  $V_i$  to be random effects with normal distributions, independent across subjects, representing numeracy and reflectiveness, respectively (i.e., we now consider a model with three latent variables rather than two). However, in this case, the distributions of  $U_i$  and  $V_i$  describe variation in numeracy and reflectiveness among subjects *within a given cluster* rather than among subjects in the overall population. For our final analysis, we ultimately make the simplifying assumption that  $U_i$  and  $V_i$  are also independent within subject (i.e., that  $\rho = 0$ ). We discuss this choice in more detail in the section titled “Fitted multi-cluster model”. We take  $Y_{ij} \mid U_i = u_i, C_i = c_i$  to be distributed as  $\text{Binomial}(3, \theta_{ij})$ , where

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

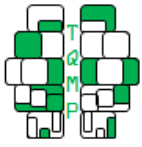
where  $\boldsymbol{\beta} = (\beta_2, \dots, \beta_7)$ . We further assume that  $T_{ij} \mid V_i = v_i, C_i = c_i$  is distributed as  $N(\mu_{ij}, \sigma_t^2)$ , where

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

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.

The likelihood is





$$\begin{aligned}
 \mathcal{L}^{[K]}(\psi) &= \prod_{i=1}^n \int_{v_i=-\infty}^{\infty} \int_{u_i=-\infty}^{\infty} \sum_{c_i=1}^K \left( \prod_{j=1}^{n_i} f_{Y_{ij}|U_i, C_i}(y_{ij}|u_i, c_i) f_{T_{ij}|V_i, C_i}(t_{ij}|v_i, c_i) \right) \\
 &\quad f_{C_i}(c_i) \cdot f_{U_i, V_i}(u_i, v_i) du_i dv_i \\
 &= \prod_{i=1}^n \int_{v_i=-\infty}^{\infty} \int_{u_i=-\infty}^{\infty} \sum_{c_i=1}^K \left\{ \prod_{j=1}^{n_i} \theta_{ij}^{y_{ij}} (1 - \theta_{ij})^{3-y_{ij}} \cdot \frac{1}{\sigma_t} \exp\left(-\frac{(t_{ij} - \mu_{ij})^2}{2\sigma_t^2}\right) \right\} \\
 &\quad \gamma_{c_i} \cdot f_{U_i, V_i}(u_i, v_i) du_i dv_i,
 \end{aligned} \tag{2}$$

where  $\psi = (\beta, \alpha, \sigma_t, \sigma_u, \sigma_v, \gamma_2, \dots, \gamma_K)$  is the vector of parameters to be estimated. (We exclude  $\gamma_1$  from  $\psi$  since it can be computed as  $\gamma_1 = 1 - \sum_{c=2}^K \gamma_c$  and hence is not a free parameter.) Altogether, this model has  $5K + 14$  parameters to be estimated, so, for example, the 4-cluster model has 34 parameters to be estimated.

### Estimation

Direct maximization of the likelihoods (1) and (2) requires integrating complex functions with respect to  $u_i$  and  $v_i$ . These integrals do not have closed forms. We therefore modify the adaptive Gaussian quadrature (AGQ) procedure of Pinheiro and Chao (2006) to find approximations to the likelihoods. We then maximize these approximations to obtain the (approximate) maximum likelihood estimates (MLEs).

AGQ was originally proposed as an alternative to Gauss-Hermite quadrature for approximating the likelihood of a GLMM. As explained by Rabe-Hesketh and Skrondal (2002), the key result underlying this method is that the integrand is proportional to the posterior distribution of the random effects, which, in turn, is approximately proportional to a certain normal density. The integrand can then be rewritten as a product of this density and an approximately constant function. Using this form of the integral and Gauss-Hermite quadrature (which allows exact evaluation of integrals of low-degree polynomials with respect to a normal density), leads to accurate and efficient evaluation of the likelihood. Although Pinheiro and Chao (2006) developed the method for a univariate response and a single cluster, we show in this section that the same arguments can be used to justify its use when evaluating likelihoods based on multivariate responses and multiple clusters.

Our approach differs from that of Kondo et al. (2017), who used MCEM to estimate the parameters in their multi-cluster model. We chose our approach based on the expectation of greater efficiency (given the results of Pinheiro and Chao (2006) for the 1-cluster, univariate response case) and because of challenges in achieving convergence when applying the MCEM method when  $K > 2$  (Y. Kondo, personal communication, September 12, 2020).

Let  $Q$  be the chosen number of Gauss-Hermite quadrature points for evaluating one of the *one-dimensional* integrals in (2), and let  $z_k$  and  $w_k$  represent the  $k^{th}$  abscissa and weight, respectively,  $k = 1, \dots, Q$ . Evaluating the two-dimensional integral will thus, in general, require  $Q^2$  quadrature points. We use  $S = \{(k_1, k_2) : k_1, k_2 \in \{1, \dots, Q\}\}$  to index these points. We then define  $\mathbf{k} = (k_1, k_2)$  and  $\mathbf{z}_{\mathbf{k}} = (z_{k_1}, z_{k_2})$ .

We summarize the iterative steps for maximum likelihood estimation using AGQ as follows:

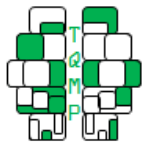
- Choose starting values of the model parameters,  $\psi$ .
- Using the current estimate of  $\psi$ , predict  $(U_i, V_i)$  by numerically maximizing

$$g_{U_i, V_i}(u_i, v_i) = \log\{f_{\mathbf{Y}_i, \mathbf{T}_i|U_i, V_i}(\mathbf{y}_i, \mathbf{t}_i | u_i, v_i) \times f_{U_i, V_i}(u_i, v_i)\}.$$

Define  $\mu_{Ai} = (\hat{u}_i, \hat{v}_i)$  as the maximizer of  $g(\cdot)$  and let  $\Sigma_{Ai} = \{\nabla^2 g_{U_i, V_i}(\hat{u}_i, \hat{v}_i)\}^{-1}$ .

- Let  $\mathbf{B}_i = (U_i, V_i)$ . Define  $\mathbf{b}_{i\mathbf{k}} = \Sigma_{Ai}^{1/2} \mathbf{z}_{\mathbf{k}} + \mu_{Ai}$  and  $W_{\mathbf{k}} = \exp(-\|\mathbf{z}_{\mathbf{k}}\|^2) w_{k_1} w_{k_2}$ . Then numerically maximize the approximate log-likelihood (Eq. 3 below) to obtain an updated estimate of  $\psi$ .
- Repeat steps 2–3 until convergence (defined as the event that the maximum difference across consecutive parameter estimates is less than some chosen value).

$$\ell^{[AGQ]}(\psi) = \sum_{i=1}^n \left( \frac{1}{2} \log |\Sigma_{Ai}| + \log \left\{ \sum_{\mathbf{k} \in S} f_{\mathbf{Y}_i, \mathbf{T}_i|\mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i | \mathbf{b}_{i\mathbf{k}}) f_{\mathbf{B}_i}(\mathbf{b}_{i\mathbf{k}}) W_{\mathbf{k}} \right\} \right) \tag{3}$$



The details of this procedure appear in Appendix C.

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 the univariate, 1-cluster case. In our case, we started with  $Q = 15$  and then gradually increased  $Q$  (to a maximum of 50) to ensure stability of the parameter estimates. In general,  $Q = 35$  seemed to be sufficient.

Similarly, the convergence criterion chosen in step 3 can be weak for the initial iterations of the algorithm and more stringent for later iterations.

To obtain starting values for the parameters in the 1-cluster model, we first fit separate (generalized) linear mixed models to the CRT scores and completion times, treating these responses as independent. This approach is equivalent to fitting our 1-cluster model with  $\rho = 0$ . We then used these parameter estimates—along with 0 for  $\rho$ —as starting values for fitting the general 1-cluster model. For our multi-cluster models, we used the MLEs of the parameters of the 1-cluster model (with some added jitter) as starting values, treating the parameters associated with each cluster (including the cluster probabilities) as close to identical.

### Implementation

We implemented our proposed methods 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 the “Estimation” section). We 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 available on the journal’s web site.

### Results

We now present our fitted models and use them to answer a variety of field-related questions.

#### Fitted 1-cluster model

Recall that the primary question of Meyer et al. (2018) is whether CRT scores increase with number of test exposures, adjusting for the other predictor variables. In this section, we address this question using our 1-cluster model (fit with  $Q = 35$ ), which builds on the work of these authors by describing the test scores and completion times simultaneously and longitudinally.

The 95% confidence interval (CI) for  $\beta_1$  (the coefficient of  $n_{\text{PrevS}}$ ) is [0.055, 0.117]. This estimate is difficult to interpret concisely because it has a complicated relationship

with the mean test score. However, we can estimate and compare mean test scores for different values of  $n_{\text{PrevS}}$  and the other predictor variables. Table 3 presents estimated mean CRT scores (and their standard errors) for different values of  $n_{\text{PrevS}}$  and  $n_{\text{Total}}$  using our fitted 1-cluster model. The standardized predictors `aveSATs` and `age` are set to 0, `male` is set to 1, and `memory` is set to 0 (i.e., these estimates are for male subjects with average SAT score, average age, and for whom we have no information about memory).

The first row of estimated mean CRT scores is for an average subject, that is, for a subject with  $U = 0$ . The second row is for the population of subjects, obtained using AGQ to determine the required marginal distribution of CRT score. When holding `aveSATs`, `age`, `male`, and `memory` fixed at the values specified, for each additional test exposure, the estimated mean CRT score for an average subject increases by about 0.046. More specifically, considering only the values of  $n_{\text{Total}}$  and consecutive values  $n_{\text{PrevS}}$  represented in the table, the estimated increase ranges from 0.039 to 0.053 (each with a standard error of approximately 0.003). On the other hand, the estimated mean score in the population of subjects with these predictor values increases by about 0.031, with estimated increases (for the chosen range of values of  $n_{\text{Total}}$  and  $n_{\text{PrevS}}$ ) lying between 0.030 and 0.032 (with standard errors lying between 0.008 and 0.010).

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  $n_{\text{PrevS}}$  and  $n_{\text{Total}}$ . However, the 0.024 estimate is based on the entire dataset and does not adjust for the other predictors. Using just the Fall 2014 data and all the predictors, the OLS-estimated effect of  $n_{\text{PrevS}}$  is 0.017. Overall, we agree with Meyer et al. (2018) that repeated test exposure has a small but non-trivial effect on test scores. But, because we are able to compute standard errors for the estimated mean score increases (for specified values of the predictors), our conclusion is more strongly justified.

We estimate that mean time to completion decreases by 0.114 (95% CI [0.104, 0.124]) log seconds for each additional test exposure (for both an average subject and in the population).

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,  $\rho$ . The weak, negative estimated correlation of  $-0.066$  (95% CI [ $-0.123$ ,  $-0.010$ ]) is consistent with the findings of Meyer et al. (2018). This estimate



**Table 2 ■** 1-cluster model parameter estimates and standard errors. The value of the maximized log-likelihood is  $-12,764$ .

Parameter	Est (SE)	Parameter	Est (SE)	Parameter	Est (SE)
$\beta_0$ (Intercept)	-0.383 (0.112)	$\alpha_0$ (Intercept)	4.137 (0.027)	$\log(\sigma_t)$	-0.505 (0.012)
$\beta_1$ (nPrevS)	0.086 (0.016)	$\alpha_1$ (nPrevS)	-0.114 (0.005)	$\log(\sigma_u)$	0.944 (0.027)
$\beta_2$ (memory1)	-1.315 (0.202)	$\alpha_2$ (memory1)	-0.228 (0.046)	$\log(\sigma_v)$	-0.621 (0.026)
$\beta_3$ (memory2)	0.384 (0.155)	$\alpha_3$ (memory2)	-0.481 (0.035)	$\log\left(\frac{1+\rho}{1-\rho}\right)$	-0.133 (0.058)
$\beta_4$ (aveSATS)	1.110 (0.061)	$\alpha_4$ (aveSATS)	-0.074 (0.013)		
$\beta_5$ (male)	0.915 (0.121)	$\alpha_5$ (male)	0.024 (0.028)		
$\beta_6$ (age)	0.259 (0.058)	$\alpha_6$ (age)	-0.012 (0.013)		
$\beta_7$ (nTotal)	0.116 (0.035)	$\alpha_7$ (nTotal)	-0.060 (0.007)		

**Table 3 ■** Estimated mean CRT scores for the average subject,  $\hat{E}[Y|U = 0]$ , and the population of subjects,  $\hat{E}[Y]$ , for different values of nPrevS and nTotal

nPrevS	nTotal	$\hat{E}[Y U = 0]$ (SE)	$\hat{E}[Y]$ (SE)
1	2	2.10 (0.036)	1.84 (0.061)
2	2	2.15 (0.034)	1.87 (0.059)
1	3	2.17 (0.040)	1.88 (0.068)
2	3	2.22 (0.038)	1.91 (0.065)
3	3	2.27 (0.036)	1.94 (0.063)
1	4	2.24 (0.045)	1.92 (0.078)
2	4	2.29 (0.043)	1.95 (0.073)
3	4	2.33 (0.041)	1.99 (0.070)
4	4	2.38 (0.039)	2.02 (0.068)
1	5	2.30 (0.051)	1.97 (0.088)
2	5	2.35 (0.048)	2.00 (0.083)
3	5	2.39 (0.045)	2.03 (0.078)
4	5	2.43 (0.043)	2.06 (0.075)
5	5	2.47 (0.042)	2.09 (0.072)

suggests that our two response variables are weakly negatively correlated after accounting for variation due to the predictors.

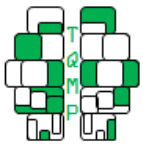
### Fitted multi-cluster model

We now present the results of fitting our multi-cluster model, an alternative way to describe the observed patterns in the data.

As mentioned in the “Bivariate Longitudinal Model with Latent Clusters” section, we chose to specify  $\rho = 0$  in our final model. We justify this choice on the basis that the latent clusters allow for correlation between  $Y_{ij}$  and  $T_{ij}$ , therefore reducing the need to allow for additional correlation via  $\rho$ . This argument is supported by the data:  $\rho$  did not differ significantly from 0 in the 3-cluster model with nPrevS as the only predictor variable and  $Q = 15$  (results not shown). Given that the latter took weeks to fit (compared to less than two days to fit the 3-cluster model with

all the predictors and  $\rho = 0$ ), massive reduction in computational effort also motivates our decision.

To choose  $K$ , the number of clusters, we fit the model with  $\rho = 0$  for  $K \in \{1, 2, 3, 4\}$  (using  $Q = 50$ ) and obtained the maximum values of the log-likelihoods (see Table 4). We do not formally compare the fits of these models because the likelihood ratio test statistics do not have the usual chi-squared distributions. However, informally, we can compare the AIC values (defined as  $-2$  times the maximized log-likelihood plus twice the number of model parameters) associated with the models. The AIC is lowest for the 4-cluster model, suggesting that that model provides the best description of the data of the four models considered. Moreover, as we discuss below, the 4-cluster model is highly interpretable with respect to the psychometric concerns that we laid out in the introduction. Therefore, while Table 4 displays parameter estimates and standard errors for all four models, from this point on, we focus our atten-



**Table 4 ■** Cluster model parameter estimates, standard errors, and maximum values of the log-likelihoods. Note that  $\gamma_1^* = 0$  and  $\gamma_c = \exp(\gamma_c^*) / \sum_c \{\exp(\gamma_c^*) + 1\}$  for  $c = 1, 2, 3, 4$ .

Parameter	1-cluster	2-cluster	3-cluster	4-cluster
	Est (SE)	Est (SE)	Est (SE)	Est (SE)
$\beta_{10}$ (Intercept)	−0.386 (0.112)	−0.616 (0.156)	−1.143 (0.193)	−1.076 (0.195)
$\beta_{11}$ (nPrevS)	0.087 (0.016)	0.131 (0.038)	0.099 (0.032)	0.074 (0.031)
$\beta_{20}$ (Intercept)	-	−0.219 (0.148)	0.362 (0.217)	0.539 (0.221)
$\beta_{21}$ (nPrevS)	-	0.074 (0.018)	0.070 (0.021)	0.036 (0.021)
$\beta_{30}$ (Intercept)	-	-	0.034 (0.268)	0.112 (0.277)
$\beta_{31}$ (nPrevS)	-	-	0.234 (0.101)	0.219 (0.105)
$\beta_{40}$ (Intercept)	-	-	-	−1.546 (0.492)
$\beta_{41}$ (nPrevS)	-	-	-	0.805 (0.122)
$\beta_2$ (memory1)	−1.316 (0.202)	−1.283 (0.205)	−1.235 (0.205)	−1.249 (0.205)
$\beta_3$ (memory2)	0.382 (0.155)	0.401 (0.157)	0.430 (0.157)	0.427 (0.157)
$\beta_4$ (aveSATS)	1.111 (0.061)	1.112 (0.061)	1.117 (0.061)	1.124 (0.061)
$\beta_5$ (male)	0.916 (0.121)	0.920 (0.121)	0.922 (0.121)	0.929 (0.122)
$\beta_6$ (age)	0.258 (0.058)	0.250 (0.058)	0.240 (0.058)	0.230 (0.058)
$\beta_7$ (nTotal)	0.116 (0.035)	0.105 (0.036)	0.077 (0.037)	0.072 (0.036)
$\alpha_{10}$ (Intercept)	4.137 (0.027)	4.957 (0.066)	4.406 (0.070)	4.366 (0.072)
$\alpha_{11}$ (nPrevS)	−0.114 (0.005)	−0.371 (0.020)	−0.240 (0.016)	−0.248 (0.017)
$\alpha_{20}$ (Intercept)	-	3.640 (0.048)	3.465 (0.056)	3.368 (0.052)
$\alpha_{21}$ (nPrevS)	-	−0.068 (0.006)	−0.052 (0.006)	−0.053 (0.006)
$\alpha_{30}$ (Intercept)	-	-	5.659 (0.121)	5.752 (0.139)
$\alpha_{31}$ (nPrevS)	-	-	−0.637 (0.041)	−0.652 (0.042)
$\alpha_{40}$ (Intercept)	-	-	-	4.555 (0.141)
$\alpha_{41}$ (nPrevS)	-	-	-	−0.051 (0.020)
$\alpha_2$ (memory1)	−0.227 (0.046)	−0.262 (0.046)	−0.238 (0.045)	−0.252 (0.044)
$\alpha_3$ (memory2)	−0.481 (0.035)	−0.479 (0.036)	−0.457 (0.036)	−0.457 (0.035)
$\alpha_4$ (aveSATS)	−0.074 (0.013)	−0.078 (0.013)	−0.078 (0.013)	−0.076 (0.013)
$\alpha_5$ (male)	0.024 (0.028)	0.011 (0.028)	0.008 (0.027)	0.018 (0.027)
$\alpha_6$ (age)	−0.012 (0.013)	0.011 (0.013)	0.017 (0.013)	0.023 (0.013)
$\alpha_7$ (nTotal)	−0.060 (0.007)	−0.029 (0.007)	−0.023 (0.007)	−0.021 (0.006)
$\log(\sigma_t)$	−0.505 (0.012)	−0.605 (0.013)	−0.639 (0.013)	−0.637 (0.013)
$\log(\sigma_u)$	0.944 (0.027)	0.943 (0.027)	0.908 (0.030)	0.910 (0.030)
$\log(\sigma_v)$	−0.620 (0.026)	−0.880 (0.043)	−0.931 (0.048)	−1.116 (0.064)
$\gamma_2^*$	-	0.307 (0.138)	−0.090 (0.167)	−0.211 (0.147)
$\gamma_3^*$	-	-	−1.063 (0.258)	−1.204 (0.277)
$\gamma_4^*$	-	-	-	−2.084 (0.268)
Max. log-lik.	−12,767	−12,596	−12,547	−12,510
AIC	25,496	25,144	25,036	24,952
# of Params	19	24	29	34

tion on the fitted 4-cluster model.

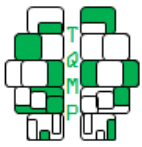
We first consider our primary question of interest, i.e., the effect of repeat exposures on test score. The 95% CIs for  $\beta_{11}$ ,  $\beta_{21}$ ,  $\beta_{31}$ , and  $\beta_{41}$  (the effects of nPrevS for clusters 1–4) are [0.013, 0.135], [−0.005, 0.077], [0.013, 0.425] and [0.566, 1.044], respectively. The CI for effect of nPrevS in cluster 4 does not overlap with the corresponding CIs for clusters 1–3. These findings suggest that the effect of repeat exposures on test scores is positive in each subpopulation and that these effects differ across some subpopulations.

The 95% CIs for many of the estimated intercepts do not overlap. Specifically, the CI for  $\beta_{10}$ , [−1.458, −0.694], does not overlap with the CI for  $\beta_{20}$  or  $\beta_{30}$  ([0.106, 0.972] and [−0.431, 0.655], respectively). And the CI for  $\beta_{40}$  ([−2.510, −0.582]) does not overlap with that for  $\beta_{20}$  or  $\beta_{30}$ . These results suggest distinct mean initial CRT scores among the clusters.

Focusing now on our second response variable, time to completion, the 95% CIs for the estimated effects of nPrevS are [−0.281, −0.215], [−0.065, −0.041], [−0.734, −0.570], and [−0.090, −0.012] in clusters 1–4, respectively. These CIs are mostly non-overlapping, suggesting that effect of nPrevS on completion time differs across all clusters.

None of the 95% CIs for the estimated intercepts overlaps except for those in the first and fourth clusters, suggesting distinct mean initial completion times among the clusters. The CIs for the estimated intercepts are [4.23, 4.51], [3.27, 3.47], [5.48, 6.02], and [4.28, 4.83] for clusters 1–4, respectively.

We now turn to the interpretation of the clusters. We ground our interpretations within the mindware instantiation continuum; see Stanovich, Toplak, and West (2008). Mindware refers to the ability to deploy various mental tools to improve problem-solving and decision-making.



Problems with mindware can impair one's performance. Specifically, a mindware gap refers to lacking the aforementioned tools, while contaminated mindware refers to having faulty such tools. Accordingly, all other factors being equal, we view low and high initial scores as indicative of low and high levels of numeracy, respectively. Likewise, we view low and high initial completion times as indicative of low and high levels of reflectiveness, respectively. Mindware levels can relate to both latent variables: low mindware levels can correspond to lower numeracy (producing lower test scores) and can prevent subjects from reflection (leading to lower completion times). In principle, we expect the scores of individuals who are highly reflective to increase over time. However, we cannot always observe such increases due to the serious ceiling effect (the maximum score achievable was 3) and the fact that subjects may have seen the test answers prior to participating in the study. With these considerations in mind, we interpret the subpopulations represented in the 4-cluster model as follows:

**Cluster 1.** The first cluster consists of subjects who are less numerate and moderately reflective. In other words, these subjects have low mindware levels, which prevents them from scoring highly and limits their ability to improve with time. They score relatively low initially (low estimated intercept,  $\hat{\beta}_{10}$ ) and only slightly higher over time. They spend an average amount of time on their first test (middling estimated intercept,  $\hat{\alpha}_{10}$ ) and spend less and less time on each subsequent test (middling estimated effect of  $nPrevS$ ,  $\hat{\alpha}_{11}$ ). Cluster 1 has an estimated cluster probability of 0.448 (95% CI [0.374, 0.521]).

**Cluster 2.** The second cluster corresponds to people who are more numerate than average *or* who have seen the test before and remember the answers. In other words, these subjects likely possess higher mindware levels, which allows them to score highly and respond quickly. They score moderately high initially (moderately high estimated intercept,  $\hat{\beta}_{20}$ ). Since their initial scores are so close to the maximum of 3, they are unable to increase their scores substantially over time (the estimated effect of  $nPrevS$ ,  $\hat{\beta}_{21}$ , is not significantly different from 0). Subjects in this cluster spend a relatively low amount of time on their first test (low estimated intercept,  $\hat{\alpha}_{20}$ ) and slightly less time on subsequent tests (slightly negative estimated effect of  $nPrevS$ ,  $\hat{\alpha}_{21}$ ). However, we *cannot* infer that subjects in this group are less reflective than average; to make such a statement, we would need to administer a new test without such a severe ceiling effect—one that would challenge the highly numerate subjects and allow the subset of highly reflective subjects (i.e., subjects with higher completion times) the possibility of achieving higher scores, on average, than their less reflective counterparts. Cluster 2

has an estimated cluster probability of 0.362 (95% CI [0.276, 0.449]), implying that clusters 1 and 2 are the largest subpopulations.

**Cluster 3.** The third cluster represents subjects who are less numerate but more reflective than average. In other words, these subjects possess moderately developed mindware, as seen by their initially moderately high scores (moderately high estimated intercept,  $\hat{\beta}_{30}$ ) and only slight score improvements over time (low estimated effect of  $nPrevS$ ,  $\hat{\beta}_{31}$ ). They spend a relatively high amount of time on their first test (high estimated intercept,  $\hat{\alpha}_{30}$ ) but much less time on each subsequent test (low estimated effect of  $nPrevS$ ,  $\hat{\alpha}_{31}$ ). With an estimated cluster probability of 0.134 (95% CI [0.062, 0.206]), this subpopulation is the smallest.

**Cluster 4.** The fourth cluster consists of subjects who are the least numerate but the most reflective. Thus, like subjects in cluster 3, they possess moderately developed mindware, expressed in this case by higher levels of reflectiveness that counteract the lower levels of numeracy. Specifically, they score very low initially (very low estimated intercept,  $\hat{\beta}_{40}$ ) and increase their scores over time much more dramatically than other subjects (very high estimated effect of  $nPrevS$ ,  $\hat{\beta}_{41}$ ). They spend a relatively high amount of time on their first test (high estimated intercept,  $\hat{\alpha}_{40}$ ) and slightly less time on each subsequent test (near-zero estimated effect of  $nPrevS$ ,  $\hat{\alpha}_{41}$ ). With an estimated cluster probability of 0.056 (95% CI [0.024, 0.087]), this subpopulation is one of the smallest.

The interpretation of these four clusters is consistent with the patterns evident in Figure 6 and the observation by Meyer et al. (2018) that the improvement in test scores over time is driven by a small minority of subjects. In contrast, our other models appear to blend heterogeneous response patterns, resulting in substantially higher AIC values.

### Additional results

In this section, we report on the effects of predictors other than  $nPrevS$  on CRT score and time to completion. We focus our discussion on the interpretation of the effects in the 1-cluster model. However, our overall conclusions apply to the effects in the multi-cluster models as well.

The estimated effect of the faulty memory indicator, *memory*, is strongly negative (95% CI [−1.711, −0.919]). That is, subjects who *incorrectly* self-reported their prior number of exposures were more likely to score *lower* on the CRT. The estimated effect of the good memory indicator was also significant (95% CI [0.080, 0.688]), similarly suggesting that subjects who *correctly* self-reported their prior number of exposures were more likely to score *higher* on the CRT. These results are consistent with the observation





of Meyer et al. (2018) that, among subjects who are known to have seen all three CRT questions previously (because they have taken the test at least once before), average CRT score increases with the number of CRT questions they report having seen.

We found *aveSATS* to have a very influential, positive effect on CRT score (95% CI [0.990, 1.230] for  $\beta_4$ ), reaffirming 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 rationality and intelligence are distinct. That is, they are characterized by different cognitive processes and predict different outcomes and traits. We also found *aveSATS* to have a weak, negative effect on time to completion (95% CI [−0.010, −0.049] for  $\alpha_4$ ).

Moreover, we found that *nTotal* was an important predictor for both response variables (95% CI [0.047, 0.185] for  $\beta_7$ , and 95% CI [−0.074, −0.046] for  $\alpha_7$  in the 1-cluster model). Those who choose to take the test many times may be more motivated (both in terms of answering the questions correctly and in terms of earning money), which may lead to higher test scores and less time spent taken to complete the test.

Finally, we find a large, positive effect of *male* on CRT score (95% CI [0.678, 1.152] for  $\beta_5$ ), a moderate-to-weak positive effect of *age* on CRT score (95% CI [0.145, 0.373] for  $\beta_6$ ), and insignificant effects of *male* (95% CI [−0.031, 0.079] for  $\alpha_5$ ) and *age* (95% CI [−0.037, 0.013] for  $\alpha_6$ ) on time to completion. These findings are consistent with the effects of sex and age on CRT test scores that have been reported in the literature (Zhang, Highhouse, & Rada, 2016).

### Model assessment

As an informal check of the fit of our 1- and 4-cluster models, we compare the empirical distributions of 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 models (listed in Table 4). See Appendix D for the relevant plots and further details. The empirical and estimated distributions of CRT scores correspond reasonably well, and the empirical and estimated distributions of completion times correspond very closely.

### Discussion

We expect that our models provide more trustworthy estimates and associated standard errors of the effect of test exposure (and the other covariates) on CRT score and time to completion than the original models used by Meyer et al. (2018). Our rationale is that 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; 3) we make more defensible distributional assumptions with the aid of our novel visualizations (see the data visualization section)—namely, we treat the CRT score response as conditionally binomial rather than marginally normal; and 4) our approach allows for the presence of subpopulations. Our findings suggest that, contrary to the conclusions of Meyer et al. (2018), the CRT's predictive validity is sometimes weakened upon repeated exposure, e.g., in the case of individuals in the subpopulations represented by clusters 3 and 4 in our 3- and 4-cluster models.

Our work is also a contribution to the understanding of two important components of rationality: numeracy and reflectiveness. In particular, our models have nice interpretations in terms of these constructs. Our fitted 1-cluster model shows that, in the overall population, numeracy and reflectiveness are negatively correlated. However, when considering subpopulations (via our fitted 3-cluster model), we find no evidence of correlation between these constructs.

From a statistical perspective, our proposed method for parameter estimation based on AGQ is another contribution. Based on preliminary simulation studies, our method is efficient and provides accurate estimates of the parameters in both the 1- and multi-cluster models.

An alternative estimation method that we considered was the EM algorithm, using Gaussian quadrature methods—rather than Monte Carlo methods, as proposed by Kondo et al. (2017)—in the E-step. We expected that this algorithm, which transforms the problem of maximizing the log-likelihood into a series of smaller maximization problems, would result in fewer convergence issues and be less sensitive to starting values. However, to obtain accurate approximations to the integrals in the E-step using Gauss-Hermite quadrature, a prohibitively large number of quadrature points were required. AGQ, suggested by Hall and Wang (2005), was similarly computationally burdensome since the integrands in the E-step are not proportional to the posterior distribution of the random effects—the key requirement for the efficiency of this method. In the end, direct maximization of the log-likelihood (which does involve an integrand that is proportional to the posterior distribution of the random effects) using AGQ and carefully chosen starting values was the most efficient and effective estimation method we examined.

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, 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. Likewise, in the generalized linear mixed model setting, McCulloch and Neuhaus (2011) conclude that “most aspects of statistical inference are highly robust to [assuming a normal distribution for the random effects]”. An exception is the case where the true random effects distribution depends on the predictors—see Heagerty and Zeger (2000). In the end, we justify our choice of distributions for the random effects by assessing the appropriateness of the implied marginal distributions of the responses.

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 (see Supporting Information) is presumably representative of more educated American adults.

We have numerous ideas for further work in this area. One involves extending our bivariate longitudinal model by treating CRT score as conditionally 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.

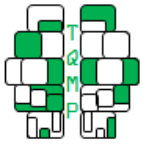
Overall, our novel approach in modelling the CRT data allows us to rigorously address key questions of interest in the cognitive psychology and psychometric literature. In addition, our explicit incorporation of numeracy and reflectiveness in our models paves the way for future research in this area.

#### Authors' note

We thank Dr. Andrew Meyer for sharing details about the data. This work was supported, in part, by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN/04304-2018).

#### References

- Agresti, A. (2013). *Categorical data analysis, third edition*. Wiley.
- Attali, Y., & Bar-Hillel, M. (2020). The false allure of fast lures. *Judgment and Decision Making*, 15(1), 93–111.
- Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavioural Research Methods*, 50(5), 1953–1959. doi:10.3758/s13428-017-0963-x
- Butler, S., & Louis, T. (1992). Random effects models with non-parametric priors. *Statistics in Medicine*, 11(14–15), 1981–2000. doi:10.1002/sim.4780111416
- Campitelli, G., & Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? a mathematical modeling approach. *Memory & Cognition*, 42(3), 434–447. doi:10.3758/s13421-013-0367-9
- Erceg, N., Galic, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection – testing convergent/divergent validity of two measures of cognitive reflection. *Judgment and Decision Making*, 15(5), 741–755.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42. doi:10.1257/089533005775196732
- Haier, R. (2014). Increased intelligence is a myth (so far). *Frontiers in Systems Neuroscience*, 8(34). doi:10.3389/fnsys.2014.00034
- Hall, D., & Wang, L. (2005). Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling*, 5, 21–37. doi:10.1191/1471082X05st0900a
- Heagerty, P., & Zeger, S. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1), 1–26. doi:10.1214/ss/1009212671
- Kahneman, D. (2013). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kondo, Y., Zhao, Y., & 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. doi:10.1002/sim.7230
- McCulloch, C., & Neuhaus, J. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26(3), 388–402. doi:10.1214/11-STS361
- Meyer, A., Frederick, S., & Zhou, E. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making*, 13(3), 246–259.
- Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5), 411–419.
- Pinheiro, J., & 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. doi:10.1198/106186006X96962
- Rabe-Hesketh, S., & Skrondal, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1), 1–21. doi:10.1177/1536867X0200200101



- Stagnaro, M., Pennycook, G., & Rand, D. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision Making*, 13(3), 260–267. doi:[10.2139/ssrn.3115809](https://doi.org/10.2139/ssrn.3115809)
- Stanovich, K., Toplak, M., & West, R. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior*, 36, 251–285. doi:[10.1016/S0065-2407\(08\)00006-2](https://doi.org/10.1016/S0065-2407(08)00006-2)
- Stanovich, K., West, R., & Toplak, M. (2016). *The rationality quotient: Toward a test of rational thinking*. The MIT Press.
- Verbeke, G., & 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. doi:[10.1016/S0167-9473\(96\)00047-3](https://doi.org/10.1016/S0167-9473(96)00047-3)
- Zhang, D., Highhouse, S., & Rada, T. (2016). Explaining sex differences on the cognitive reflection test. *Personality and Individual Differences*, 101(1), 425–427. doi:[10.1016/j.paid.2016.06.034](https://doi.org/10.1016/j.paid.2016.06.034)

## Appendix A: MTurk Reliability

Paolacci, Chandler, and Ipeirotis (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 U.S. bank account. 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 Questions

The following questions comprise the CRT:

- (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
- Modified versions of these questions were given in the other series that we excluded in our analysis.

## Appendix C: Details of the AGQ Algorithm

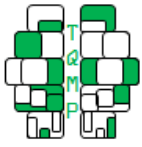
We first define the logarithm of the joint density of our response variables and random effects as

$$g_{U_i, V_i}(u_i, v_i) = \log \{ f_{\mathbf{Y}_i, \mathbf{T}_i | U_i, V_i}(\mathbf{y}_i, \mathbf{t}_i | u_i, v_i) f_{U_i, V_i}(u_i, v_i) \}.$$

We then maximize  $g_{U_i, V_i}(u_i, v_i)$  by computing  $(\hat{u}_i, \hat{v}_i)$  such that  $\nabla g_{U_i, V_i}(\hat{u}_i, \hat{v}_i) = 0$ . 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  $\boldsymbol{\mu}_{Ai} = (\hat{u}_i, \hat{v}_i)$  and variance  $\boldsymbol{\Sigma}_{Ai} = \{ \nabla^2 g_{U_i, V_i}(\hat{u}_i, \hat{v}_i) \}^{-1}$ .

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

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



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  $\mathbf{b}_i = \Sigma_{Ai}^{1/2} \mathbf{z}_i^* + \mu_{Ai}$ , we can write

$$\begin{aligned} & f_{\mathbf{Y}_i, \mathbf{T}_i}(\mathbf{y}_i, \mathbf{t}_i) \\ &= \iint \frac{f_{\mathbf{Y}_i, \mathbf{T}_i | \mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i | \mathbf{b}_i) f_{\mathbf{B}_i}(\mathbf{b}_i)}{\exp(-\|\mathbf{z}_i^*\|^2)} |\Sigma_{Ai}|^{1/2} \exp(-\|\mathbf{z}_i^*\|^2) d\mathbf{z}_i^*. \end{aligned}$$

Defining  $S$ ,  $\mathbf{k}$ ,  $\mathbf{b}_{ik}$ , and  $W_{\mathbf{k}}$  as in the “Estimation” section and letting  $\mathbf{R}_i = \Sigma_{Ai}^{-1/2}$ ,

$$\begin{aligned} & f_{\mathbf{Y}_i, \mathbf{T}_i}(\mathbf{y}_i, \mathbf{t}_i) \\ & \approx |\mathbf{R}_i|^{-1} \sum_{\mathbf{k} \in S} f_{\mathbf{Y}_i, \mathbf{T}_i | \mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i | \mathbf{b}_{ik}) f_{\mathbf{B}_i}(\mathbf{b}_{ik}) W_{\mathbf{k}}. \end{aligned}$$

The AGQ approximation to the log-likelihood function is then

$$\ell^{[AGQ]}(\psi) = \sum_{i=1}^{n_i} \left[ -\log |\mathbf{R}_i| + \log \left\{ \sum_{\mathbf{k} \in S} f_{\mathbf{Y}_i, \mathbf{T}_i | \mathbf{B}_i}(\mathbf{y}_i, \mathbf{t}_i | \mathbf{b}_{ik}) f_{\mathbf{B}_i}(\mathbf{b}_{ik}) W_{\mathbf{k}} \right\} \right].$$

## Appendix D: Model Assessment

To provide an informal check of our 1- and 4-cluster model fit, Figure 7 displays both the empirical CRT score and time to completion responses, along with their respective estimated marginal distributions under the two models. Specifically, we computed the estimated distributions for each individual (using each model’s parameter estimates) and then averaged these distributions over all individuals.

For the score response, we estimate the probabilities of each CRT score using the estimated parameters and the observed predictor values, restricted to  $n_{\text{PrevS}} = 1$ . 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 left-most plot correspond to the empirical probability of each CRT score, while the purple and red horizontal lines correspond to the probabilities estimated using our 1- and 4-cluster models, respectively.

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

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

where

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

The histogram on the right reflects the empirical distribution of time to completion. The purple and red curves are the estimated marginal distributions of completion time based on our 1- and 4-cluster models, respectively.

## Open practices

- The *Open Data* badge was earned because the data of the experiment(s) are available on [the journal’s web site](#).
- The *Open Material* badge was earned because supplementary material(s) are available on [the journal’s web site](#).

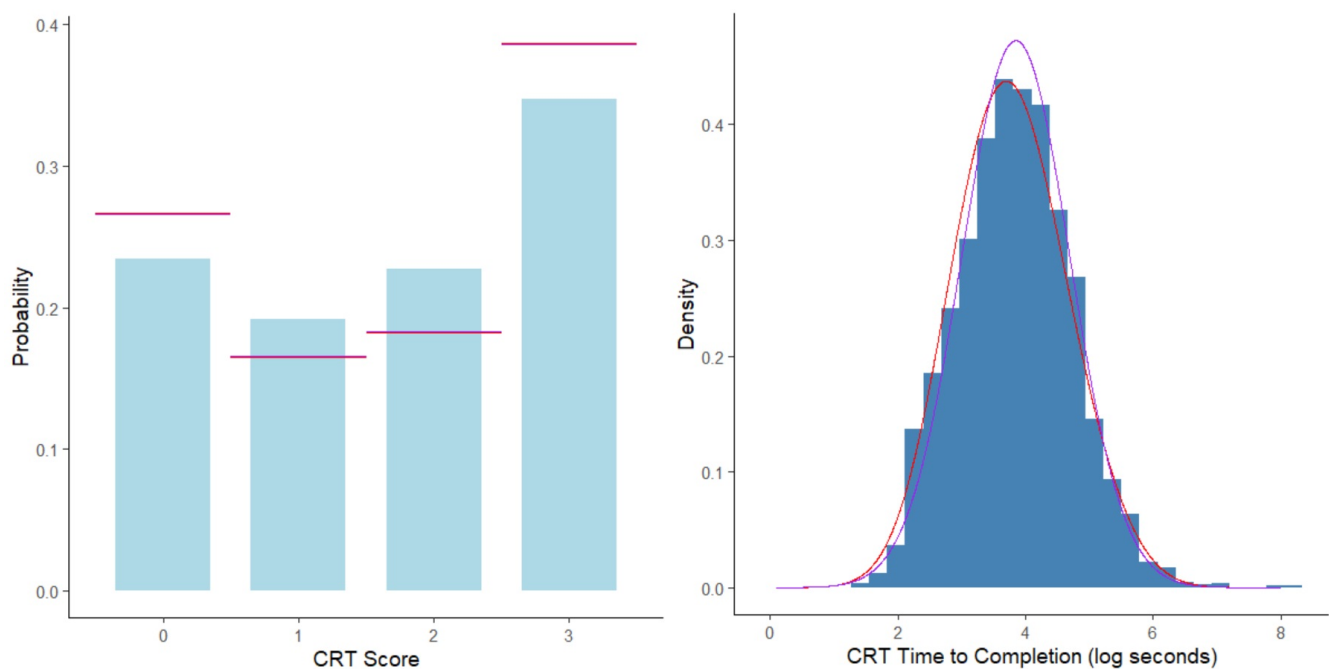
## Citation

Berkowitz, M., & Altman, R. M. (2022). A bivariate longitudinal cluster model with application to the Cognitive Reflection Test. *The Quantitative Methods for Psychology*, 18(1), 21–38. doi:10.20982/tqmp.18.1.p021

Copyright © 2022, Berkowitz and Altman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 01/09/2021 ~ Accepted: 21/12/2021

Figure 7 follows.



**Figure 7** ■ Empirical and estimated distributions of CRT score (left) and time to completion (right) at  $n_{PrevS} = 1$  (1-cluster model estimate in purple; 4-cluster model estimate in red). The estimated CRT score distributions based on the 1- and 4-cluster models are close to identical, i.e., the purple and red lines overlap.