Explanatory Item Response Models for Continuous Data: A Tutorial in R

Joshua B. Gilbert [1]

[1]Harvard Graduate School of Education

October 15, 2025

Abstract

The explanatory item response model (EIRM) is a common tool in psychometrics to model person and item characteristics as functions of covariates. Existing tutorials demonstrate how to model dichotomous or polytomous item responses. In this tutorial, we show how to fit the extended two-parameter logistic (E2PL) item response model for continuous item responses using the `brms` package in R. Using a worked example with visual analog scale data, we demonstrate data exploration, model building, and interpretation strategies. By following this tutorial, researchers will be able to fit and interpret the EIRM for continuous item response data.

Keywords: item response theory, psychometrics, Bayesian multilevel models, explanatory item response model, R

Corresponding author: joshua_gilbert@g.harvard.edu

The author wishes to thank Ben Domingue and Bill Young for helpful comments on drafts of this paper.

1 Introduction

Item response theory (IRT) models are ubiquitous in psychometrics for developing and validating measures. Traditional IRT models are applied to categorical data, such as correct or incorrect answers on educational tests or Likert scale ratings on psychological surveys (Baker, 2001; Embretson & Reise, 2000). With the advent of response formats such as visual analog scales, common in ecological momentary assessments and digital measurement contexts (Heller et al., 2016; Newhouse & Njiru, 2009; Shiffman et al., 2008), IRT models that can accommodate continuous responses are of increasing interest to researchers (Li & Shin, 2025).

In this tutorial, we consider the extended two-parameter logistic (E2PL) IRT model for continuous responses, recently proposed by Li and Shin (2025). We focus on extending the explanatory item response model (EIRM)—an approach in which person and item parameters are modeled as functions of person and item covariates (De Boeck et al., 2016; Wilson & De Boeck, 2004; Wilson et al., 2008)—from categorical to continuous responses using the E2PL. The EIRM is commonly applied in psychometric research to address areas such as differential item functioning, group differences, causal inference, response time analysis, network psychometrics, and careless responding, among others (Briggs, 2008; Gilbert, Domingue, & Kim, 2025; Gilbert, Himmelsbach, Soland, et al., 2025; Gilbert, Young, et al., 2025; Randall et al., 2011; Ulitzsch et al., 2022), but it has not yet been extended to continuous item responses. Accordingly, existing tutorials for fitting the EIRM in R with the `lme4` and `brms` software packages have only considered dichotomous and polytomous categorical data (Bürkner, 2017; De Boeck et al., 2011; Gilbert, 2024), thus limiting the applicability of the EIRM to diverse data-analytic contexts.

The purpose of this tutorial is to provide an accessible overview of the E2PL model and demonstrate how to fit and interpret the EIRM for continuous responses using the Bayesian multilevel modeling software `brms` in R (Bürkner, 2017, 2021). Using a running example of visual analog scale (VAS) data from a recent empirical study (Yu et al., 2025), we emphasize exploratory data analysis, model building, and interpretation strategies.

The tutorial is organized as follows. We begin with a review of categorical and continuous IRT models, the EIRM framework, and past tutorials and software for fitting the EIRM in R. We then describe the empirical data, modeling strategy, and R syntax. We proceed with the illustrative results with an emphasis on the substantive interpretation and conclude with a discussion of potential extensions of the framework.

1.1 IRT Models for Categorical and Continuous Data

We define y_{ij} as the response of person j to item i. y_{ij} may be dichotomous, polytomous, or continuous. A standard approach to dichotomous responses (e.g., correct or incorrect answers on an educational achievement test coded as 1 or 0) is the two-parameter logistic (2PL) IRT model:

$$\text{logit}(\Pr(y_{ij} = 1)) = a_i(\theta_j - b_i). \tag{1}$$

Here, the log-odds that $y_{ij} = 1$ is a function of latent person trait θ_j, item discrimination a_i, and item location (or difficulty) b_i. a_i provides the difference in log-odds of a positive response per unit difference in θ_j, while b_i provides the point on the θ_j continuum at which $\Pr(y_{ij} = 1) = 0.5$.

The 2PL can be extended to ordered polytomous responses such as Likert scale items with the graded response model (GRM). For items with K ordered categories, the GRM models the log-odds of being greater than or equal to category k as

$$\text{logit}(\Pr(y_{ij} \geq k)) = a_i(\theta_j - b_{ik}), \tag{2}$$

in which the b_{ik} represent $K-1$ threshold parameters at which point the respondent has even odds of responding in category k or higher versus less than k for item i. When $K=2$, the GRM reduces to the 2PL.[1]

Various IRT models have been proposed for continuous responses (Molenaar et al., 2022; Müller, 1987; Noel & Dauvier, 2007; Samejima, 1973; Veldkamp & Sluijter, 2019). In this tutorial, we focus on the "extended 2PL" (E2PL) model proposed by Li and Shin (2025), which uses a Beta regression framework to accommodate the bounded nature of many continuous item response formats such as VAS ratings. The E2PL is specified as follows, adapting the notation of Li and Shin (2025):

$$y_{ij} = \text{logit}^{-1}(a_i(\theta_j - b_i)) + \varepsilon_{ij} \tag{3}$$

$$= \mu_{ij} + \varepsilon_{ij} \tag{4}$$

$$\sigma^2_{\varepsilon_{ij}} = \frac{\mu_{ij}(1-\mu_{ij})}{\nu_i + 1} \tag{5}$$

$$\mu_{ij} + \varepsilon_{ij} \sim \text{Beta}(\nu_i \mu_{ij}, \nu_i(1-\mu_{ij})). \tag{6}$$

Note that the E2PL models the item response itself (y_{ij}) rather than the log-odds of a positive response in the 2PL ($\text{logit}(y_{ij}=1)$) or the log-odds of a cumulative response in the GRM ($\text{logit}(y_{ij} \geq k)$). Because the Beta distribution is not defined for values of exactly 0 or 1, y_{ij} must be rescaled to the (0,1) interval for analysis (we return to this issue in our empirical application).[2]

The use of the logistic function for the mean (μ_{ij}) plus the additive error term (ε_{ij}) conveniently means that the interpretations of the a_i and b_i item parameters are analogous across the E2PL and the standard 2PL models. Specifically, a_i represents the difference on

[1] Many other polytomous IRT models exist, such as the rating scale and partial credit models. For the present discussion, we consider only the GRM because it is most analogous to the standard 2PL and we are primarily interested in continuous data in this study. See Nalbandyan et al. (2024) and Domingue, Kanopka, et al. (2025) for more detail on differences between polytomous IRT models.

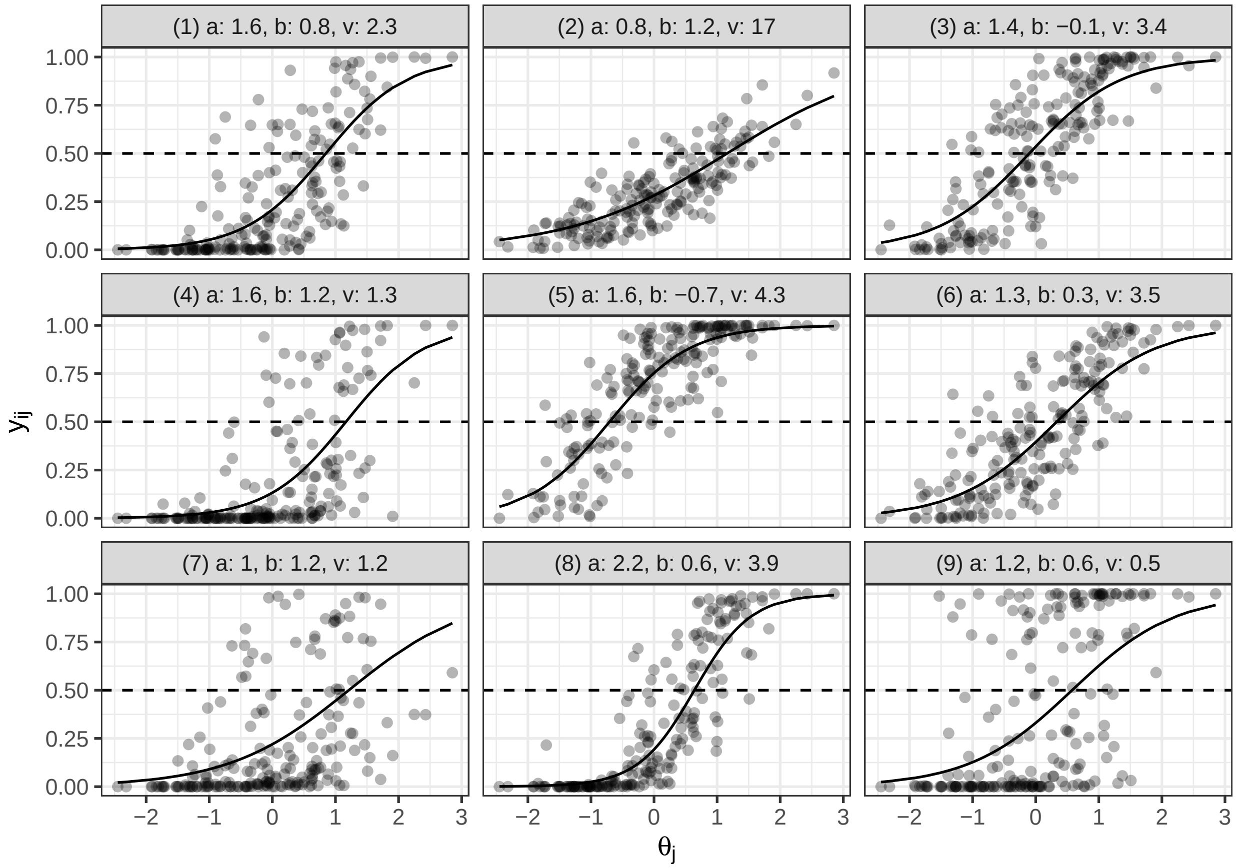
[2] The E2PL can also model polytomous responses. For example, a 0-4 Likert scale item could be rescaled to .1, .3, .5, .7, .9 and then analyzed with the E2PL. See Li and Shin (2025) for a discussion.

the logit scale of the mean response per one-unit difference in θ_j, and b_i is the point on θ_j at which $\mu_{ij} = 0.5$. Unlike the 2PL, the E2PL includes an error term ε_{ij} to allow each y_{ij} to deviate from its mean μ_{ij}. The variance of ε_{ij} is governed by the item-specific precision parameter ν_i. As $\nu_i \to \infty$, the error variance approaches zero, so the responses fall on μ_{ij}. In contrast, as $\nu_i \to 0$, the responses fall near the extreme values, approaching the standard 2PL for dichotomous responses.

To illustrate the interpretation of the item parameters under the E2PL, consider Figure 1, which shows simulated item response data drawn from an E2PL data-generating process. For the a_i parameter, contrast items 2 and 8, where $a_8 > a_2$ means that the slope of the logistic function for item 8 is much steeper than that of item 2. For b_i, contrast items 4 and 5, where the mean response for item 4 is 0.5 when $\theta_j = 1.2$, compared to $\theta_j = -.7$ for item 5, thus shifting the item response function to the left. For ν_i, contrast items 2 and 9, where $\nu_2 > \nu_9$ yields points much more tightly clustered around the line for item 2, compared to points that fall much more towards the extremes for item 9.

The E2PL is closely related to the linear confirmatory factor analysis (CFA) model in that both the item responses and latent trait are continuous. Indeed, both IRT and CFA can be considered special cases of a generalized latent variable modeling framework. In IRT, the latent trait is continuous, the indicators are categorical, and the link function is logistic. In CFA, the indicators are continuous and the link function is linear (Gilbert, 2025; Skrondal & Rabe-Hesketh, 2004). As such, the E2PL is somewhat of a hybrid model that combines characteristics of both IRT and CFA. The key differences between the E2PL and the linear CFA are the logistic functional form, non-normal error distribution, and bounded response in the E2PL compared to the linear functional form, normal error distribution, and unbounded response in the linear CFA. Given that common continuous item response formats such as VAS are bounded from above and below, the E2PL is likely to be more theoretically appropriate in many applications, including the empirical VAS data used as an illustration in this tutorial. In contrast, when the indicators are themselves composite variables—e.g.,

Figure 1: Simulated Item Response Data from the E2PL Model



The y-axis shows the item response y_{ij} and the x-axis shows the latent trait θ_j for 200 subjects responding to 9 items. The item parameters are drawn from $a_i \sim \text{lognormal}(.5, .25), b_i \sim N(0, 1), \nu_i \sim \text{lognormal}(1, 2)$.

math, reading, and science test scores in a model for academic achievement—the assumptions underlying linear CFA may be more realistic. We return to this issue in our Discussion.

1.2 The Explanatory Item Response Model (EIRM)

The IRT models described above are purely descriptive in that they solely estimate the person and item parameters, either as fixed effects or as sources of variation. Often, however, researchers are interested in the extent to which the person and item parameters are themselves functions of observable characteristics. For example, in intervention contexts, θ_j may increase or decrease as a result of treatment (Gilbert, Himmelsbach, Soland, et al., 2025). In assessment design, a_i may depend on whether an item is positively or negatively worded (Gilbert, Zhang, et al., 2025; Min et al., 2018). Interactions between person and item predictors can capture differential item functioning (DIF), whereby groups show differences in item performance conditional on θ_j (De Boeck et al., 2011; Randall et al., 2011).

We can extend Equation 4 into an EIRM by specifying each parameter as a function of covariates. For illustration, consider person covariate X_j (e.g., gender, age, treatment status, etc.) and item covariate X_i (e.g., item modality, item wording, etc.), predicting θ_j and b_i, respectively:

$$
\begin{aligned}
y_{ij} &= \text{logit}^{-1}(a_i(\theta_j + b_i)) + \varepsilon_{ij} && (7)\\
&= \mu_{ij} + \varepsilon_{ij} && (8)\\
\sigma^2_{e_{ij}} &= \frac{\mu_{ij}(1-\mu_{ij})}{\nu_i + 1} && (9)\\
\theta_j &= \beta_1 X_j + \theta^*_j && (10)\\
b_i &= \beta_0 + \beta_2 X_i + b^*_i && (11)\\
\log(a_i) &= \gamma_0 + a^*_i && (12)\\
\log(\nu_i) &= \delta_0 + \nu^*_i && (13)\\
\theta^*_j &\sim N(0, 1) && (14)\\
\begin{bmatrix} b^*_i \\ a^*_i \\ \nu^*_i \end{bmatrix} &\sim N\left(0, \begin{bmatrix} \sigma^2_b & & \\ \sigma_{ab} & \sigma^2_a & \\ \sigma_{b\nu} & \sigma_{a\nu} & \sigma^2_\nu \end{bmatrix}\right) && (15)
\end{aligned}
$$

We provide more detail on the interpretation of each term of Equation 7 in Table 1. Note that in the EIRM formulation, we add rather than subtract b_i to the model because, as a regression model, the terms enter into the model additively. The stars indicate residual item parameter random effects after accounting for the effects of any covariates. For clarity of exposition, Equation 7 includes the person and item covariates only in the equations for θ_j

and b_i, respectively, but an advantage of the EIRM framework is that *any* item parameter can be specified as a function of covariates. For example, item discrimination and precision may also vary as functions of covariates, as we will demonstrate in our empirical analysis.

Here, θ_j^* is constrained to a standard normal distribution for model identification and a_i is modeled on the log scale to constrain the item discriminations to be positive (Bürkner, 2021). β_1 reveals whether θ_j varies as a function of person covariate X_j while β_2 reveals whether b_i varies as a function of item covariate X_i. Equation 7 is therefore a Generalized non-linear mixed model (GNLMM) that combines a non-linear link function with fixed and random effects (Gilbert, Domingue, & Kim, 2025; Molenberghs & Verbeke, 2004).

Table 1: Interpretation of Terms in Equation 7

Term	Interpretation
β_0	μ_{ij} value when $X_j, X_i = 0$
β_1	Effect of a 1-unit difference in X_j on the θ_j scale
β_2	Effect of a 1-unit difference in X_i on the b_i scale
γ_0	log discrimination of the average item
δ_0	log precision of the average item
σ_b^2	residual variance of item location
σ_a^2	residual variance of item discrimination
σ_ν^2	residual variance of item precision
σ_{ab}	residual covariance between location and discrimination
$\sigma_{b\nu}$	residual covariance between location and precision
$\sigma_{a\nu}$	residual covariance between discrimination and precision

1.3 Past EIRM Software and R Tutorials

Several tutorials exist for fitting the EIRM in R. Early examples use the multilevel modeling package `lme4` (Bates et al., 2015) to fit 1PL or Rasch models with multilevel structures (De Boeck et al., 2011; Doran et al., 2007; Gilbert, 2024). One limitation of `lme4` is that it cannot estimate item discriminations from the data, limiting its utility in many settings, though `lme4` can accommodate item discriminations when they are known in advance (Rockwood & Jeon, 2019). While not designed for polytomous data, `lme4` can fit rating scale and partial credit models when the data is reshaped to represent pairwise contrasts between categories (Bulut et al., 2021; Gilbert, Hieronymus, et al., 2024). Extensions to `lme4` include `PLmixed` and `galamm` (Rockwood & Jeon, 2019; Sørensen, 2024), which allow for varying item discriminations. However, these two packages only allow for fixed effects for item discriminations, limiting their applicability when we wish to specify the item discrimination itself as a function of covariates (Cho et al., 2014; Gilbert, Zhang, et al., 2025).

The Bayesian multilevel modeling package `brms` (Bürkner, 2017) provides the most flexible approach and is the subject of a highly detailed IRT tutorial (Bürkner, 2021). `brms` allows for a wide range of dichotomous and polytomous models, such as 1PL, 2PL, 3PL, and the GRM, and item parameters can be specified as either fixed or random effects. To our knowledge, no tutorials exist for fitting the explanatory E2PL model using `brms` or other packages, though `IRTest` can fit the E2PL without covariates (Li, 2024). The explanatory E2PL is easily fit in the `brms` framework, whereas `lme4`, `PLmixed`, and `galamm` can only accommodate continuous responses in a CFA or Generalizability Theory framework.

2 Methods

2.1 Data Source

In this study, we use a running example based on visual analog scale (VAS) data from Yu et al. (2025) to motivate and organize our analysis. The authors examine VAS ratings of object size. Participants completed a digital task in which they were presented with circles of varying sizes and estimated the circle's size in millimeters with a VAS. The data also include an analogous task in which participants reproduce the circle's size by interacting with the assessment interface, but for clarity of exposition, we limit our analysis to the VAS data and the first and second attempts to each item (we show the cleaning code in Appendix A). Thus, we analyze 1,489 item responses from 167 participants. Table 2 provides a codebook for the subset of the data we explore in this study, and Table 3 shows the first ten rows of the data.

Table 2: Codebook for the Yu et al. (2025) Data

Variable Name	Variable Description
`id`	Participant Identifier
`item`	Item Identifier
`resp`	Raw Response [0-200]
`resp01`	Rescaled Response (0-1)
`cov_female`	1 = Participant is Female
`item_cov_second`	1 = Item Response is Second Attempt

Table 3: First 10 Rows of the Yu et al. (2025) Data

id	item	resp	resp01	cov_female	item_cov_second
1	S1	58	0.29	0	0
1	S1	60	0.30	0	1
1	S2	63	0.31	0	0
1	S2	68	0.34	0	1
1	S3	97	0.48	0	0
1	S3	74	0.37	0	1
1	S4	86	0.43	0	0
1	S4	92	0.46	0	1
1	S5	100	0.50	0	0
1	S5	107	0.53	0	1

The raw VAS scores range from 0 to 200mm. Because the Beta distribution cannot accommodate values of exactly 0 or 1, we rescale the raw item responses to the (0, 1) interval by adding .5 and dividing by 201 (Li & Shin, 2025). We examine participant gender and item attempt to illustrate the interpretations of both person and item covariates in the model. To guide our analysis, we consider the following research questions:

1. To what extent do item location, discrimination, and precision vary as a function of person gender and item attempt?

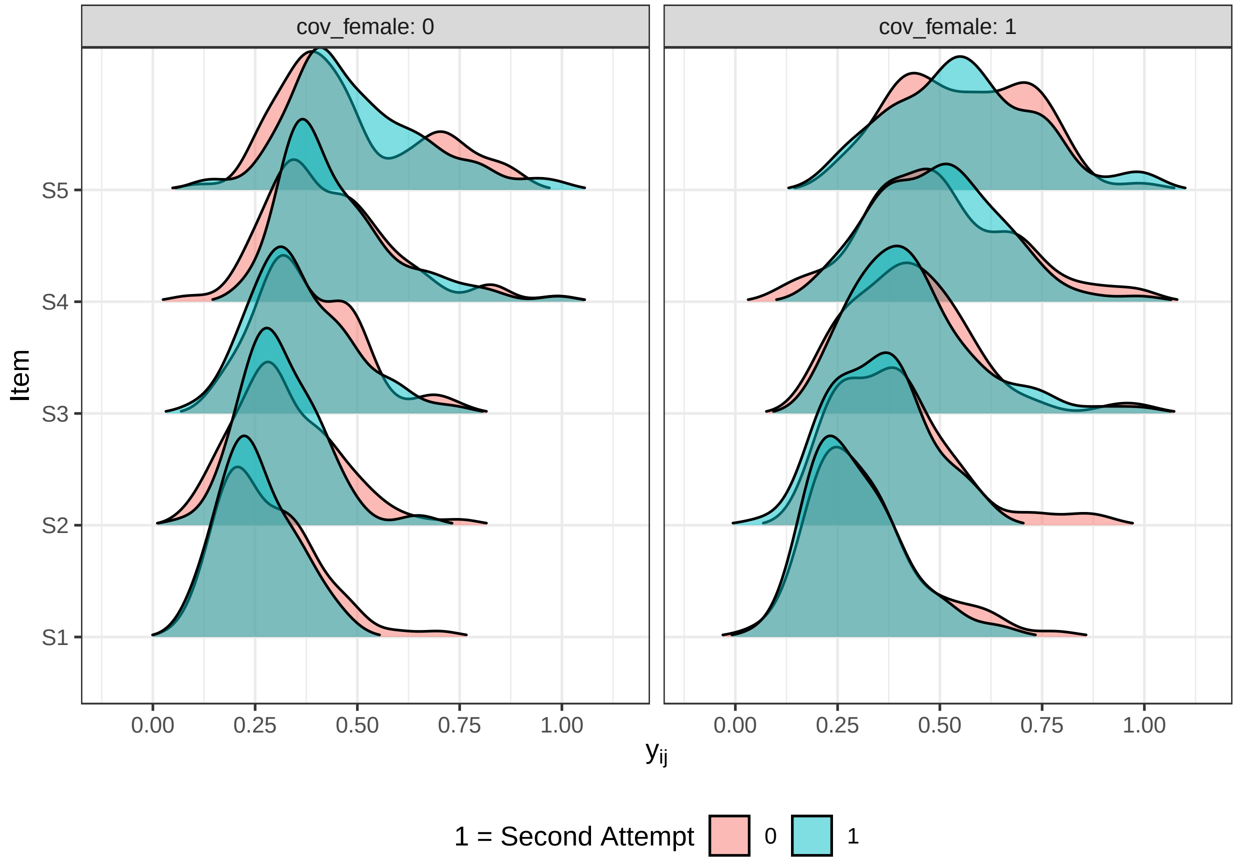
2. To what extent do gender and item attempt interact in their prediction of item parameters?

2.2 Exploratory Data Analysis

We begin by loading the data and performing some exploratory data analysis. The R code below loads the relevant libraries, sets the aesthetic themes for the figures, and loads the empirical data. We then generate the density plot shown in Figure 2, which shows the distribution of the rescaled VAS response by item, person gender, and item attempt. We see slightly right-skewed distributions with higher item numbers (corresponding to larger circles in the stimulus) showing greater mean responses. Distributions for first and second attempts are mostly overlapping.

```
# load libraries
library(tidyverse)
library(brms)
library(IRTest)
library(ggridges)

# set ggplot theme
theme_set(theme_bw())
theme_update(legend.position = "bottom")

# load the empirical data
yu_long <- read_csv("data/clean/yu_brm.csv") |>
  # turn covariates into factors
  mutate(cov_female = factor(cov_female),
         item_cov_second = factor(item_cov_second))

# create a density plot
ggplot(yu_long, aes(x = resp01, y = item, fill = item_cov_second)) +
  geom_density_ridges(rel_min_height = .01, alpha = .5) +
  facet_wrap(~cov_female, labeller = label_both) +
  labs(x = expression(y[ij]),
       y = "Item",
       fill = "1 = Second Attempt")
```

Figure 2: Density of VAS Ratings by Item, Attempt, and Participant Gender



The y-axis shows the item and density and the x-axis shows the item response values. The color fills indicate whether the item response is from the first or second attempt and the plot is faceted by participant gender (1 = female).

2.3 Standard E2PL Model with `IRTest`

Before fitting the EIRM, we first estimate a standard E2PL model using the `IRTest` package (Li, 2024). The code below reshapes the data from long to wide and fits the standard E2PL model, treating the first and second attempt items as independent items. This modeling decision likely violates the local independence assumption of the IRT model; we proceed with the example as an illustration to gain insight into whether item parameters may differ upon the second attempt, a result we formally test in the EIRMs that follow.

```
# pivot to wide for IRTest
yu_wide <- yu_long |>
```

```r
  pivot_wider(names_from = item_unique,
              values_from = resp01,
              names_prefix = "item_",
              id_cols = id)

# fit the E2PL model
e2pl <- IRTest_Cont(yu_wide |>
                      select(starts_with("item_")) |>
                      as.matrix())

# get the coefficients
e2pl |>
  coef() |>
  as_tibble(rownames = "item")
```

Table 4 shows the estimated E2PL item parameters. The substantive interpretations of parameter estimates for the first attempt for item 1 (S1-1) are as follows. $\widehat{a_i} = .50$ indicates that a one SD difference in θ_j predicts a .50 difference in y_{ij}, on average, on the logit scale. $\widehat{b_i} = 1.78$ indicates that when $\theta_j = 1.78$, participants respond at 50% of the maximum VAS rating (i.e., 100mm), on average. $\widehat{\nu_i} = 42.75$ indicates that the points are very tightly clustered around mean line. In the context of these data where participants rate circle lengths, higher values of θ_j represent the tendency of participants to rate circles as larger.

To aid interpretability of the E2PL item parameters, Figure 3 shows the empirical item characteristic curves (ICCs), using the expected a posteriori (EAP) $\widehat{\theta_j}$ scores on the x-axis and fitted logistic curves superimposed, generated with the code below. We see that the logistic functions fit the observed data well, all discriminations are positive, and precision is high, with most points clustered tightly around the line. As would be expected from this very simple perceptual task that only varied the size of the circle presented as the stimulus, the a_i and ν_i parameter estimates are similar across items whereas the b_i parameter estimates vary more extensively due to the varying sizes of the circles in the assessment.

```r
# get the EAP scores
yu_wide <- yu_wide |>
  mutate(eap = factor_score(e2pl)$theta)
```

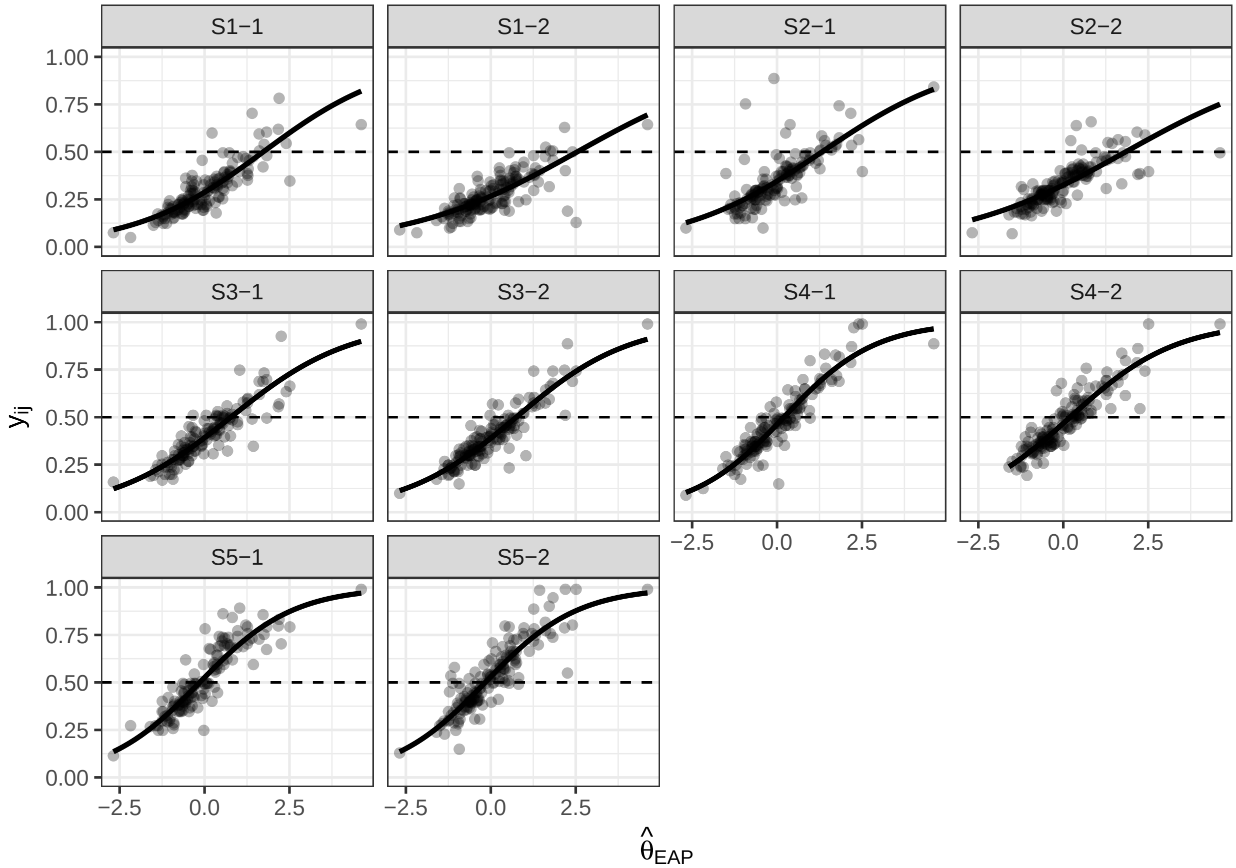
Table 4: Item Parameter Estimates from the E2PL Model

Item	a_i	b_i	ν_i
S1-1	0.50	1.78	42.75
S1-2	0.37	2.74	42.72
S2-1	0.45	1.37	23.16
S2-2	0.39	1.88	47.49
S3-1	0.57	0.75	42.30
S3-2	0.59	0.75	47.83
S4-1	0.76	0.17	30.21
S4-2	0.64	0.18	38.74
S5-1	0.69	-0.16	30.26
S5-2	0.75	-0.19	22.63

The table shows parameter estimates from a standard E2PL model applied to the Yu et al. (2025) data. Each row is a separate item, where the first number is the item number and the number after the dash is the attempt number.

```r
# graph the empirical ICCs
yu_wide |>
  pivot_longer(starts_with("item_"),
               names_to = "item",
               values_to = "resp") |>
  mutate(item = str_remove_all(item, "item_|VAS-")) |>
  ggplot(aes(x = eap, y = resp)) +
  facet_wrap(~item) +
  geom_point(alpha = .3) +
  geom_hline(yintercept = .5, linetype = "dashed") +
  geom_smooth(se = FALSE,
              method = "glm",
              method.args = list(family = "binomial"),
              color = "black") +
  labs(y = expression(y[ij]),
       x = expression(hat(theta)[EAP])) +
  ylim(0, 1)
```

Figure 3: Empirical Item Characteristic Curves



The y-axis shows the rescaled item response and the x-axis shows the $\widehat{\theta_j}$ EAP scores from the E2PL model. The fitted curves are logistic functions. Each panel is a separate item, where the first number is the item number and the number after the dash is the attempt number.

2.4 EIRMs with `brms`

We continue our exploration by estimating three EIRMs. The most complex model (Model 3) is as follows:

$$y_{ij} = \text{logit}^{-1}(a_i(\theta_j + b_i)) + \varepsilon_{ij} \tag{16}$$

$$= \mu_{ij} + \varepsilon_{ij} \tag{17}$$

$$\sigma^2_{e_{ij}} = \frac{\mu_{ij}(1 - \mu_{ij})}{\nu_{ij} + 1} \tag{18}$$

$$\theta_j + b_i = \beta_0 + \beta_1 \text{female}_j + \beta_2 \text{attempt}_i + \beta_3 \text{female}_j \times \text{attempt}_i + \theta^*_j + b^*_i \tag{19}$$

$$\log(a_{ij}) = \gamma_0 + \gamma_1 \text{female}_j + \gamma_2 \text{attempt}_i + \gamma_3 \text{female}_j \times \text{attempt}_i + a_i^* \tag{20}$$

$$\log(\nu_{ij}) = \delta_0 + \delta_1 \text{female}_j + \delta_2 \text{attempt}_i + \delta_3 \text{female}_j \times \text{attempt}_i + \nu_i^* \tag{21}$$

$$\theta_j^* \sim N(0,1) \tag{22}$$

$$\begin{bmatrix} b_i^* \\ a_i^* \\ \nu_i^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_b^2 & & \\ 0 & \sigma_a^2 & \\ 0 & 0 & \sigma_\nu^2 \end{bmatrix}\right). \tag{23}$$

Note that a_{ij} and ν_{ij} now have j subscripts because we are including person predictors in the relevant equations. Furthermore, unlike the descriptive E2PL above where we treat the repeated attempts to the stimuli as independent items, here, we treat each circle stimulus as a single item, yielding 5 unique items total (thus relaxing the local independence assumption above). Compared to Model 3, Model 2 constrains $\beta_3, \gamma_3, \delta_3 = 0$ so that we examine main effects only, and Model 1 further constrains $\beta_1, \beta_2, \gamma_1, \gamma_2, \delta_1, \delta_2 = 0$ so that we only estimate the intercepts and variances of the item parameters as a baseline model for further comparison.

We begin by setting the priors for the models, shown in the code below. `eta` refers to the sum of the θ_j and b_i parameters and `phi` is a built-in `brms` distributional parameter that is equivalent to ν_i. Note that the `normal` priors on the SDs represent half-normal distributions, as SDs must be positive. These priors are moderately informative, following other examples of Bayesian EIRMs using `brms` (Bürkner, 2021; Gilbert, Zhang, et al., 2025).[3]

```r
# set priors
prior <-
  # sd of item easiness
  prior(normal(0,1), class = "sd", group = "item", nlpar = "eta") +
  # sd of person ability
  prior(constant(1), class = "sd", group = "id", nlpar = "eta") +
  # sd of item discrimination
  prior(normal(0, .5), class = "sd", group = "item", nlpar = "logalpha") +
  # coefs on eta
  prior(normal(0, .5), class = "b", nlpar = "eta") +
```

[3]The prior on the coefficients for v_i must be excluded from Model 1. See our replication materials for the full code to fit all models.

```r
  # coefs on disc
  prior(normal(0, .5), class = "b", nlpar = "logalpha") +
  # coefs on phi
  prior(normal(0, .5), class = "b", dpar = "phi") +
  # eta intercept
  prior(normal(0, 1), class = "b", coef = "Intercept", nlpar = "eta") +
  # disc intercept
  prior(normal(0, .5), class = "b", coef = "Intercept", nlpar = "logalpha") +
  # phi intercept
  prior(normal(0, 1), class = "Intercept", dpar = "phi") +
  # sd of phi
  prior(normal(0, 1), class = "sd", group = "item", dpar = "phi")
```

The code below first declares the formulas for Models 1, 2, and 3, then shows how to fit a single model using the `brm` function. We specify `family = Beta()` for the Beta error distribution, declare `nl = TRUE` for a non-linear model, and use 2,000 iterations across 4 chains, with 1,000 for burn in, yielding 2,000 posterior draws for each model.

```r
# baseline model
mod1 <- bf(
  # model for response
  resp01 ~ exp(logalpha)*eta,
  # model for linear predictor eta
  eta ~ 1 + (1|id) + (1|item),
  # model for a
  logalpha ~ 1 + (1|item),
  # model for nu
  phi ~ 1 + (1|item),
  # declare non-linear model
  nl = TRUE
)

# main effects
mod2 <- bf(
  resp01 ~ exp(logalpha)*eta,
  eta ~ 1 + cov_female + item_cov_second + (1|id) + (1|item),
  logalpha ~ 1 + cov_female + item_cov_second + (1|item),
  phi ~ 1 + cov_female + item_cov_second + (1|item),
  nl = TRUE
)

# interactions
mod3 <- bf(
```

```
  resp01 ~ exp(logalpha)*eta,
  eta ~ 1 + cov_female*item_cov_second + (1|id) + (1|item),
  logalpha ~ 1 + cov_female*item_cov_second + (1|item),
  phi ~ 1 + cov_female*item_cov_second + (1|item),
  nl = TRUE
)

# fit the models
brm(
    formula = mod, # model
    data = yu_long, # data
    family = Beta(), # distribution family
    prior = prior, # prior
    backend = "cmdstanr", # backend fitting software
    chains = 4, # N chains
    iter = 2000, # N iterations
    cores = 4, # N CPUs
    threads = threading(4), # N threads per CPU
    refresh = 50 # report updates every 50 iterations
)
```

3 Results

3.1 EIRM Results

Table 5 shows the results of the three EIRMs fit to the data. Model 1 includes only the intercepts and random terms. Model 2 adds main effects for female and item attempt and we find coefficients that differ significantly from 0 for female for β_1 and δ_1 and for item attempt for δ_2. Model 3 shows interaction effects near 0. We therefore proceed with Model 2 to illustrate interpretation of the results.

The parameter estimates from Model 2 are interpreted as follows, beginning with the intercept terms and then considering only the main effects that significantly differ from 0. $\widehat{\beta}_0 = -.87$ means that for the average first-attempt item and non-female person the mean response is -.87 on the scale of the linear predictor. This corresponds to a rating of approximately $\text{logit}^{-1}(e^{-.55} \times -.87) = .38$ on the transformed 0-1 response scale. $\widehat{\gamma}_0 = -.54$

Table 5: Results of Explanatory Item Response Models

Label	Parameter	Model 1	Model 2	Model 3
Intercept	β_0	-0.72 (0.45)	-0.87 (0.48)	-0.91 (0.46)
Female	β_1		0.36 (0.14)	0.39 (0.16)
Attempt	β_2		-0.03 (0.04)	0 (0.05)
Interaction	β_3			-0.06 (0.07)
Intercept	δ_0	3.43 (0.22)	3.44 (0.24)	3.47 (0.22)
Female	δ_1		-0.25 (0.08)	-0.29 (0.11)
Attempt	δ_2		0.22 (0.08)	0.18 (0.11)
Interaction	δ_3			0.07 (0.16)
Intercept	γ_0	-0.53 (0.16)	-0.55 (0.18)	-0.55 (0.17)
Female	γ_1		-0.03 (0.04)	-0.02 (0.05)
Attempt	γ_2		0.02 (0.03)	0.03 (0.04)
Interaction	γ_3			-0.01 (0.06)
SD(a)	σ_a	0.34 (0.14)	0.36 (0.14)	0.35 (0.14)
SD(b)	σ_b	1.07 (0.35)	1.13 (0.36)	1.13 (0.35)
SD(θ)	σ_θ	1	1	1
SD(ν)	σ_ν	0.31 (0.22)	0.31 (0.24)	0.3 (0.22)

The table shows the results of the fitted EIRMs. Standard errors are in parentheses. Parameters correspond to Equation 17. σ_θ is constrained to 1 for model identification.

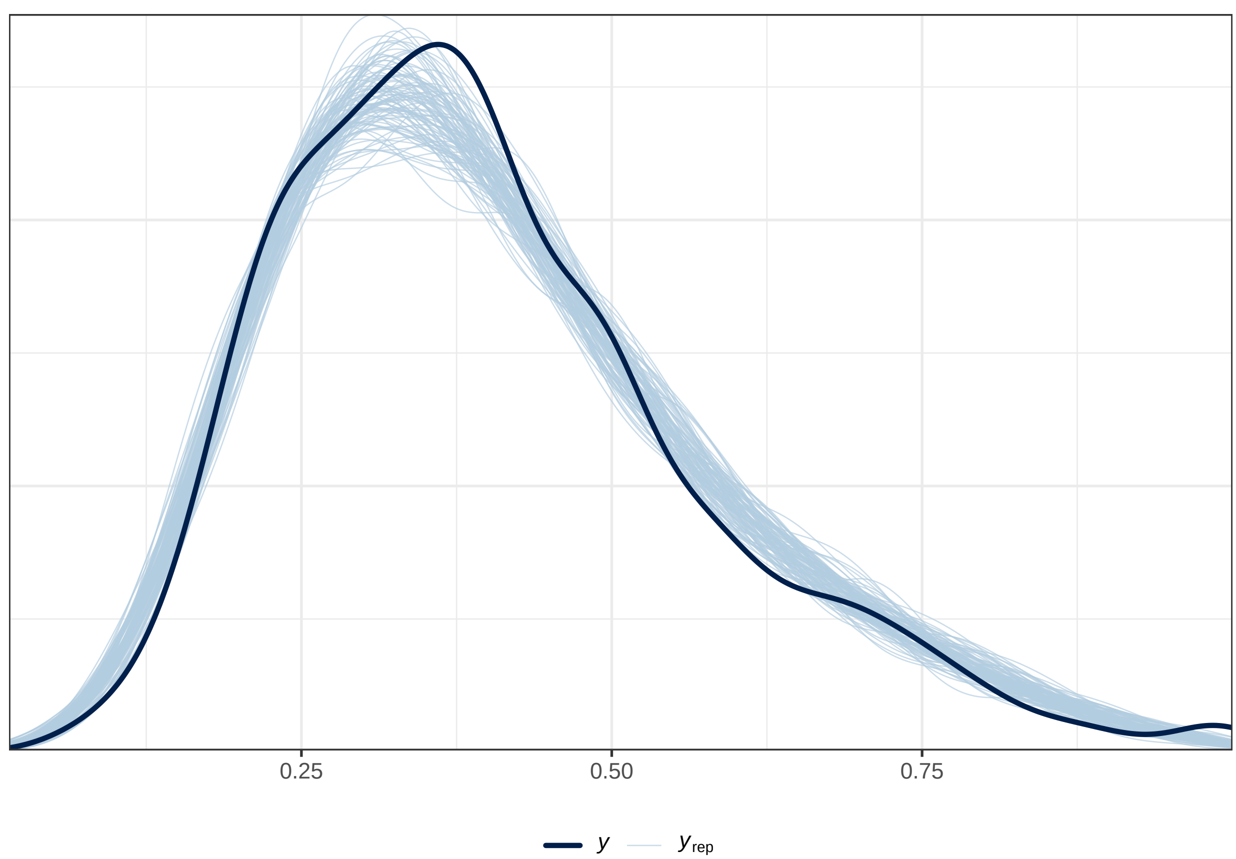
means that a 1SD difference in θ_j predicts a $e^{-.55} = .58$ difference (on the logit scale) for first attempts of non-female participants for the average item. $\widehat{\delta_0} = 3.44$ means that the precision for first attempts of non-female participants to the average item is $e^{3.53} = 31.2$. $\widehat{\beta_1} = .36$ means that female participants rate the circles as larger, on average. $\widehat{\delta_1} = -.25$ means that female responses were $e^{-.25} = 78\%$ as precise as male respondents (see, e.g., Jansen and Heil, 2009; Neubauer et al., 2010). In other words, responses from female respondents show greater residual variation. $\widehat{\delta_2} = .22$ means that second attempts were $e^{.22} = 25\%$ more precise than first attempts. Such an effect could represent, for example, increased familiarity with the task.

3.2 Model Diagnostics

Having identified our final model, we examine several model diagnostics. We first examine model convergence. All $\widehat{R}$ statistics for Model 2 are less than or equal to 1.01, well below the general guideline of 1.05 for acceptable convergence. Next, we conduct posterior predictive checks, which compare simulated data drawn from the model's posterior predictive distribution to the observed data (McElreath, 2020). The code below implements the check and Figure 4 shows the results, and we find that the posterior distribution approximates the observed distribution well. We include analogous plots by gender and item attempt in Appendix B, and find similar results. These two standard diagnostic checks provide confidence that the model results are trustworthy and stable.

```
pp_check(fit2, ndraws = 100)
```

Figure 4: Posterior Predictive Checks for Model 2



The y-axis shows the relative probability density and the x-axis shows the item response. The dark blue line represents the observed data and the light blue lines represent 100 draws from the posterior distribution.

3.3 Extensions

The model-building strategy outlined above provides a concise overview of using the EIRM to address specific research questions. Here, we provide a few extensions to highlight the flexibility of the EIRM framework for continuous responses. We show the `brms` code to fit these extensions below and discuss the relevant interpretation and example research questions they could address, but do not show the results applied to the empirical data.

3.3.1 Correlated Random Effects

For simplicity and given the small number of items, we modeled the (residual) item random effects (b_i^*, a_i^*, ν_i^*) as mutually independent. We can easily allow the item random effects to be correlated to assess whether, for example, more precise items are more discriminating $(\sigma_{b\nu} > 0)$, or easier items are less discriminating $(\sigma_{ba} < 0)$. To allow for correlated random effects in `brms`, we simply replace `(1|item)` with `(1|i|item)` in each equation with item random effects (Bürkner, 2021), as shown in the code below.

```
eta ~ 1 + (1|id) + (1|i|item),
logalpha ~ 1 + (1|i|item),
phi ~ 1 + (1|i|item)
```

3.3.2 Differential Item Functioning

Measurement invariance—when participants with equal standing on the latent trait have identical response distributions—is an important property of psychometric measures (Meredith, 1993; Schmitt & Kuljanin, 2008). For example, ensuring that item parameters are equivalent across groups may be necessary for accurate inference and validity (Olivera-Aguilar & Rikoon, 2023; Soland, 2021). While CFA-based approaches tend to provide global estimates of invariance (Thissen, 2025), an advantage of IRT-based approaches such as the EIRM is that measurement invariance can be tested with respect to single items or item clusters (De Boeck et al., 2011; Gilbert et al., 2023).

In IRT, differences in item parameters by some person covariate are known as differential item functioning, or DIF. In the E2PL model, DIF occurs when $\mathbb{E}(y_{ij}|\theta_j) \neq \mathbb{E}(y_{ij}|\theta_j, X_j)$, where X_j is a person covariate. The code below shows how to test for uniform DIF—a shift in item location b_i based on a person covariate—between the female indicator and item 5. Assuming the other items do not show DIF, the main effect of `cov_female` captures the difference in θ_j between females and males and the `cov_female:item5` interaction provides the increment to μ_{ij} for females on item 5. For example, a positive value of this coefficient would suggest females tend to give higher responses to item 5, conditional on θ_j. Note that because we are testing DIF for a single item, we include `item` as a fixed effect in the model rather than as a random effect as in prior models.

```
eta ~ 1 + item + cov_female + cov_female:item5 + (1|id)
```

We can easily extend the code above to explore potential non-uniform DIF (Montoya & Jeon, 2020) by replicating the equation above for the $\log(a_i)$ parameter. The code below allows both the item intercept and slope to differ between females and males for item 5, assuming again the other four items are invariant. Similarly, this approach could be applied to the ν_i parameter to test if precision differs between groups. Note that, technically, group differences in ν_i do not satisfy the standard definition of DIF given above because $\mathbb{E}(y_{ij}) = \mu_{ij}$ does not depend on ν_i, as shown in Equation 4. However, if we expand our definition of DIF to refer to the *distribution* of y_{ij} rather than the expectation of y_{ij}, such group differences in ν_i are similar in spirit to standard DIF frameworks (Van Der Linden, 2019).

```
eta ~ 1 + item + cov_female + cov_female:item5 + (1|id),
logalpha ~ 1 + item + cov_female + cov_female:item5
```

3.3.3 Random Slopes Models

The DIF tests above focus on single item indicators. An alternative approach to DIF uses a random slopes framework where item-specific effects are conceptualized as deviations from the mean difference in μ_{ij} (Adams et al., 1997). For example, in causal inference contexts, treatments may have differential effects across items above and beyond an overall average effect (Ahmed et al., 2024; Gilbert, Himmelsbach, Miratrix, et al., 2025; Gilbert, Himmelsbach, Soland, et al., 2025; Gilbert, Kim, & Miratrix, 2024; Gilbert, Miratrix, et al., 2025; Gilbert et al., 2023; Halpin & Gilbert, 2024; Student, 2025; Student et al., 2025). Using gender again as an example, the code below shows how to add a random slope for `cov_female` across items. In this model, the main effect provides the difference in μ_{ij} between females and males for the average item, and these effects are assumed to be normally distributed around the mean effect.

```
eta ~ 1 + cov_female + (cov_female|item)
```

3.3.4 Generalized Additive Models

The person and item covariates examined in the current application are binary, but the EIRM can also accommodate continuous covariates, such as person age or response time. Including continuous covariates in the EIRM introduces a linearity assumption in the relationship between the covariate and the person or item parameters. Individual participant meta-analyses of item response data suggest that such assumptions are often unrealistic, in, for example, the response time case (Domingue et al., 2022; Gilbert, Young, et al., 2025). To relax the linearity assumption, recent developments show how to fit flexible non-linear relationships between covariates and person and item parameters by introducing a generalized additive model (GAM) within the EIRM framework (Cho et al., 2024; Gilbert, Young, et al., 2025). GAMs allow for flexible relationships between covariates and outcomes at all levels of the model (Pedersen et al., 2019). In `brms`, we specify a non-linear relationship between covariates and outcomes by wrapping the continuous covariate in the `s()` helper function. For example, the code below allows for a non-linear effect of age (`cov_age`) on $\log(a_i)$.

```
logalpha ~ 1 + s(cov_age) + (1|item)
```

4 Discussion

As continuous response formats such as visual analog scales (VAS) become more common in digital assessments, there is a need for appropriate IRT models that accommodate the bounded nature of the continuous responses. To address this need, Li and Shin (2025) propose the extended two-parameter logistic (E2PL) IRT model, which uses Beta regression to flexibly accommodate the bounded continuous response and provides a convenient interpretation analogous to the standard 2PL for binary responses.

Beyond descriptive IRT models to calibrate person and item parameters, the EIRM specifies person and item parameters as functions of covariates to answer a wide range of research questions in psychometrics and related disciplines. This tutorial outlines a full data analysis pipeline using VAS data from Yu et al. (2025), using the Bayesian multilevel modeling software `brms` (Bürkner, 2021) to fit a series of explanatory E2PL models with person and item covariates and their interactions. Illustrative results show that female participants had both higher mean ratings of circle size with greater residual variability, and that second attempt items show less residual variability. We do not have evidence that the person gender and item attempt effects interact in predicting person or item parameters.

While flexible, the approach outlined in this study has several limitations. First, as a Bayesian model, the Markov Chain Monte Carlo (MCMC) estimation is computationally intensive, particularly when datasets are large. For example, even with only about 1,500 item responses, the models explored here take about 90 seconds on the author's personal computer; computational demand would be much greater with larger datasets.[4] Second, model results can be difficult to interpret, especially when the same covariates predict both μ_{ij} and a_i, because the covariate predicts both the linear predictor $\theta_j + b_i$ *and* the extent to which the linear predictor translates to differences in the observed outcome. That is, simultaneous predictors of $\theta_j + b_i$ and a_i capture a type of moderated mediation analysis,

[4] All analyses in this study were conducted on a 2021 MacBook Pro with an 8-core 3.2 GHz M1 Pro CPU with 16 GB RAM.

where predictors can interact with themselves (Gilbert, Domingue, & Kim, 2025; Montoya & Jeon, 2020). Last, the models explored in this study are highly parameterized, relying on a Beta distribution for the errors and a logistic functional form for the mean. While Figure 3 suggests that the functional form assumptions appear reasonable and Figure 4 suggests that the fitted model generally captures the distribution of observed data, verifying that the model assumptions is an essential step with all latent variable models. For example, if the functional form appears linear and the error distribution is more symmetrical, a linear CFA model may be more appropriate than the E2PL.

We discussed the code and interpretation of several direct extensions to our modeling framework in the Results section; we highlight some more substantive potential extensions here. For example, the models examined in this study are unidimensional; multidimensional extensions of the EIRM have been applied to dichotomous responses using `lme4` (De Boeck & Wilson, 2014; De Boeck et al., 2011) and analogous extensions would apply to the E2PL. Similarly, further development of connections between explanatory E2PL and structural equation modeling (SEM) (Kline, 2023) could be a fruitful avenue of research, combining the type of analysis demonstrated in this study with, for example, multiple outcomes, mediation, and latent predictors. Last, extension of the EIRM framework to linear CFA models could provide a valuable approach when item responses are functionally unbounded and error distributions are normal, such as when the indicators are themselves composite variables.

In sum, this tutorial extends the EIRM framework to bounded continuous response formats by demonstrating how to fit and interpret the E2PL model using Bayesian multilevel modeling in `brms`. By providing practical code examples, illustration with empirical VAS data, and discussion of extensions such as DIF testing and generalized additive models, we aim to make the proposed methodology accessible to researchers facing the growing prevalence of continuous item response data in digital assessment contexts. As continuous response formats become more common, the ability to model item and person characteristics as functions of observed covariates will further enhance the scope and rigor of psychometric analyses.

5 Declarations

Funding: The authors report no funding.

Conflicts of Interest: The authors report no conflicts of interest.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of Data and Materials: The original dataset from Yu et al. (2025) is available at the following URL: https://osf.io/f97rz/. The data are also available in the Item Response Warehouse under the name `yu2025` (Domingue, Braginsky, et al., 2025).

Code Availability: Our code, analysis output, and supplemental materials are available at the following URL: https://dataverse.harvard.edu/previewurl.xhtml?token=dda6d25c-48d7-4ee7-9dc9-e55d4

Author Contributions: Conceptualization: JG; Methodology: JG; Software: JG; Formal Analysis: JG; Writing—original draft preparation: JG; Writing—review and editing: JG.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Benjamin, W. (2024). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2024.2361337

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC. https://eric.ed.gov/?id=ED458219

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118. https://doi.org/10.1080/08957340801926086

Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych*, *3*(3), 308–321. https://doi.org/10.3390/psych3030023

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Cho, S.-J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, *79*(1), 84–104. https://doi.org/10.1007/s11336-013-9360-2

Cho, S.-J., Goodwin, A., Naveiras, M., & De Boeck, P. (2024). Modeling nonlinear effects of person-by-item covariates in explanatory item response models: Exploratory plots and modeling using smooth functions. *Journal of Educational Measurement*, *61*(4), 595–623. https://doi.org/10.1111/jedm.12410

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12

De Boeck, P., Cho, S.-J., & Wilson, M. (2016). Explanatory item response models. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 247–266). Wiley.

De Boeck, P., & Wilson, M. (2014). Multidimensional explanatory item response modeling. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 252–271). Routledge.

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L. A., Gilbert, J., Kanopka, K., Kapoor, R., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2025). An introduction to the Item Response Warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*, *57*. https://doi.org/10.3758/s13428-025-02796-y

Domingue, B. W., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradović, J., Piech, C., Porter, T., Consortium, P. i., Soland, J., Weeks, J., Wise, S. L., & Yeatman, J. (2022). Speed–accuracy trade-off? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *Journal of Educational and Behavioral Statistics*, *47*(5), 576–602. https://doi.org/10.3102/10769986221099906

Domingue, B. W., Kanopka, K., Ulitzsch, E., & Zhang, L. (2025). Implied probabilities of polytomous response functions for model-based prediction and comparison. *Behaviormetrika*, *52*, 683–705. https://doi.org/10.1007/s41237-025-00262-9

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2). https://doi.org/10.18637/jss.v020.i02

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Gilbert, J. B. (2024). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. *Behavior Research Methods*, *56*(5), 5055–5067. https://doi.org/10.3758/s13428-023-02245-8

Gilbert, J. B. (2025). How measurement affects causal inference: Attenuation bias is (usually) more important than outcome scoring weights. *Methodology*, *21*(2), 91–122. https://doi.org/10.5964/meth.15773

Gilbert, J. B., Domingue, B. W., & Kim, J. S. (2025). Estimating causal effects on psychological networks using item response theory. *Psychological Methods*. https://doi.org/10.1037/met0000764

Gilbert, J. B., Hieronymus, F., Eriksson, E., & Domingue, B. W. (2024). Item-level heterogeneous treatment effects of selective serotonin reuptake inhibitors (SSRIs) on depression: Implications for inference, generalizability, and identification. *Epidemiologic Methods*, *13*(S2). https://doi.org/10.1515/em-2024-0006

Gilbert, J. B., Himmelsbach, Z., Miratrix, L. W., Ho, A. D., & Domingue, B. W. (2025). Item-level heterogeneity in value added models: Implications for reliability, cross-study comparability, and effect sizes [edworkingpapers.com]. https://doi.org/10.26300/EZ4Q-FS31

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*, *44*(4), 1417–1449. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2024). Leveraging item parameter drift to assess transfer effects in vocabulary learning. *Applied Measurement in Education*, *37*(3), 240–257. https://doi.org/10.1080/08957347.2024.2386934

Gilbert, J. B., Miratrix, L. W., Joshi, M., & Domingue, B. W. (2025). Disentangling person-dependent and item-dependent causal effects: Applications of item response theory to the estimation of treatment effect heterogeneity. *Journal of Educational and Behavioral Statistics*, *50*(1), 72–101. https://doi.org/10.3102/10769986241240085

Gilbert, J. B., Young, W. S., Himmelsbach, Z., Ulitzsch, E., & Domingue, B. W. (2025). Conditional dependencies between response time and item discrimination: An item-level meta-analysis. https://doi.org/10.31234/osf.io/rp34w_v1

Gilbert, J. B., Zhang, L., Ulitzsch, E., & Domingue, B. W. (2025). Polytomous explanatory item response models for item discrimination: Assessing negative-framing effects in social-emotional learning surveys. *Behavior Research Methods*, *57*(4), 1–21. https://doi.org/10.3758/s13428-025-02625-2

Halpin, P., & Gilbert, J. (2024). Testing whether reported treatment effects are unduly dependent on the specific outcome measure used. https://doi.org/10.48550/ARXIV.2409.03502

Heller, G. Z., Manuguerra, M., & Chow, R. (2016). How to analyze the Visual Analogue Scale: Myths, truths and clinical relevance. *Scandinavian Journal of Pain*, *13*(1), 67–75. https://doi.org/10.1016/j.sjpain.2016.06.012

Jansen, P., & Heil, M. (2009). Gender differences in mental rotation across adulthood. *Experimental Aging Research*, *36*(1), 94–104. https://doi.org/10.1080/03610730903422762

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Publications.

Li, S. (2024). IRTest: An R package for item response theory with estimation of latent distribution. *The R Journal*, *16*(4), 23–41.

Li, S., & Shin, H. J. (2025). An extended two-parameter logistic item response model to handle continuous responses and sparse polytomous responses. *Psychometrika*, 1–26. https://doi.org/10.1017/psy.2025.10044

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315372495

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Min, H., Zickar, M., & Yankov, G. (2018). Understanding item parameters in personality scales: An explanatory item response modeling approach. *Personality and Individual Differences*, *128*, 1–6. https://doi.org/10.1016/j.paid.2018.02.012

Molenaar, D., Cúri, M., & Bazán, J. L. (2022). Zero and one inflated item response theory models for bounded continuous data. *Journal of Educational and Behavioral Statistics*, *47*(6), 693–735. https://doi.org/10.3102/10769986221108455

Molenberghs, G., & Verbeke, G. (2004). An introduction to (generalized (non)linear mixed models. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (pp. 111–153). Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_4

Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, *44*(2), 118–136. https://doi.org/10.1177/0146621619835496

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*(2), 165–181. https://doi.org/10.1007/BF02294232

Nalbandyan, R., Gilbert, J. B., Franco, V. R., & Domingue, B. W. (2024). Signposts on the path from nominal to ordinal scales. https://doi.org/10.31234/osf.io/zbv8f

Neubauer, A. C., Bergner, S., & Schatz, M. (2010). Two- vs. three-dimensional presentation of mental rotation tasks: Sex differences and effects of training on performance and brain activation. *Intelligence*, *38*(5), 529–539. https://doi.org/10.1016/j.intell.2010.06.001

Newhouse, C. P., & Njiru, J. N. (2009). Using digital technologies and contemporary psychometrics in the assessment of performance on complex practical tasks. *Technology, Pedagogy and Education*, *18*(2), 221–234. https://doi.org/10.1080/14759390902992626

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47–73. https://doi.org/10.1177/0146621605287691

Olivera-Aguilar, M., & Rikoon, S. H. (2023). Intervention effect or measurement artifact? Using invariance models to reveal response-shift bias in experimental studies. *Journal of Research on Educational Effectiveness*, 1–29. https://doi.org/https://doi.org/10.1080/19345747.2023.2284768

Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, *7*, e6876. https://doi.org/10.7717/peerj.6876

Randall, J., Cheong, Y. F., & Engelhard, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129–147. https://doi.org/10.1177/0013164410391577

Rockwood, N. J., & Jeon, M. (2019). Estimating complex measurement and growth models using the R package PLmixed. *Multivariate Behavioral Research*, *54*(2), 288–306. https://doi.org/10.1080/00273171.2018.1516541

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203–219. https://doi.org/10.1007/BF02291114

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. STATA Press.

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, *45*(5), 346–360. https://doi.org/10.1177/01466216211013102

Sørensen, Ø. (2024). Multilevel semiparametric latent variable modeling in R with "galamm". *Multivariate Behavioral Research*, *59*(5), 1098–1105. https://doi.org/10.1080/00273171.2024.2385336

Student, S. R. (2025). Using moderated nonlinear factor analysis to separate, and estimate, treatment effects and DIF. https://doi.org/10.31234/osf.io/bkafj_v2

Student, S. R., Gilbert, J., Eze, J. U., Young, W., & Domingue, B. (2025). Instrumental variables regression with latent variables: Accounting for treatment-based differential item functioning as item-level heterogeneity or item parameter moderation. https://doi.org/10.31234/osf.io/sudgt_v1

Thissen, D. (2025). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research*, *60*(2), 211–235. https://doi.org/10.1080/00273171.2024.2396148

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

Van Der Linden, W. J. (2019). Lord's equity theorem revisited. *Journal of Educational and Behavioral Statistics*, *44*(4), 415–430. https://doi.org/10.3102/1076998619837627

Veldkamp, B. P., & Sluijter, C. (Eds.). (2019). *Theoretical and practical advances in computer-based educational measurement*. Springer International Publishing. https://doi.org/10.1007/978-3-030-18480-3

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). Springer.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Hogrefe & Huber Publishers.

Yu, K., Lin, T.-Y., Zaman, J., Tuerlinckx, F., & Vanhasbroeck, N. (2025). Consistency of perceptual response variability in size estimation and reproduction tasks. *Behavior Research Methods*, *57*(5), 127. https://doi.org/10.3758/s13428-025-02650-1

Appendices

A Cleaning Code for the Empirical Data

The code below cleans the raw data from Yu et al. (2025) (`yu_data.Rds`) and exports the subset of data used in this study (`yu_brm.csv`). The data are also available on the Item Response Warehouse (IRW) under the name `yu2025` (Domingue, Braginsky, et al., 2025).

```
# load libraries
library(tidyverse)
library(data.table)

# set seed
set.seed(2025)

# clear memory
rm(list = ls())

yu_long <- readRDS("data/raw/yu_data.Rds") |>
  # rename with IRW conventions
  rename(id = participant,
         item = stim,
         resp = esti,
         item_cov_truth = size,
         cov_gender = gender,
         cov_age = age,
         item_cov_phase = phase) |>
  select(-c(file_name)) |>
  # get trial number
  drop_na(resp) |>
  group_by(id, item) |>
  mutate(item_cov_rep = frank(trial, ties.method = "dense")) |>
  ungroup() |>
  mutate(item_unique = glue("{item}-{item_cov_phase}-{item_cov_rep}")) |>
  # transform resp to 0/1
  mutate(resp_prop = 100*resp/max(resp),
         resp01 = (resp_prop) / max(resp_prop + 1)) |>
  arrange(id, item, item_cov_phase, item_cov_rep)

# clean up gender
gender <- yu_long |>
```

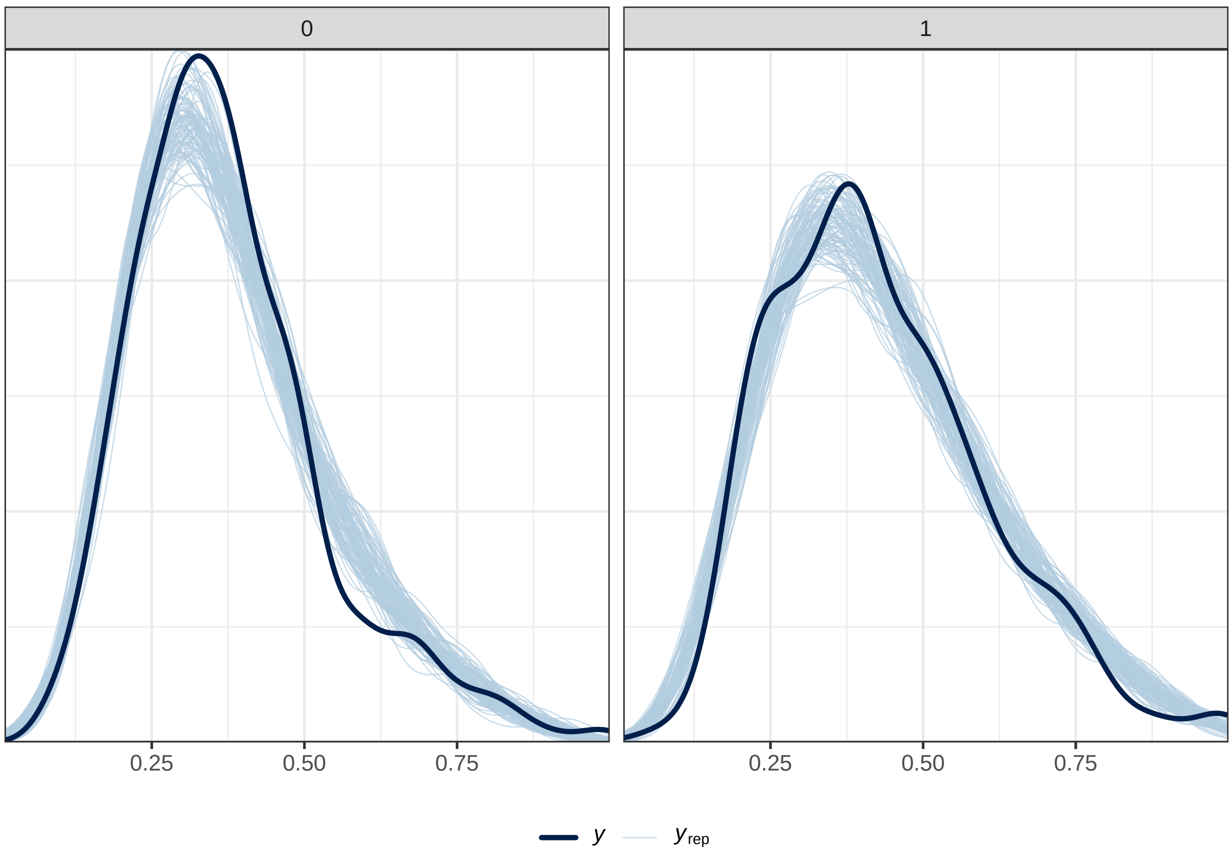
```r
  distinct(cov_gender) |>
  mutate(cov_female = c(0, 1, 1, rep(0, 6)))

# export full data
yu_long <- yu_long |>
  left_join(gender, by = "cov_gender") |>
  select(-cov_gender)

write_csv(yu_long, "data/clean/yu2025.csv")

# take only first trial for this study
yu_long_brm <- yu_long |>
  filter(item_cov_phase == "VAS",
         item_cov_rep < 3) |>
  mutate(item_cov_second = if_else(item_cov_rep == 2, 1, 0))

write_csv(yu_long_brm, "data/clean/yu_brm.csv")
```

B Additional Model Diagnostics

Figures B1 and B2 show posterior predictive checks from Model 2, stratified by gender and item attempt, respectively. The code to generate the first plot is below.

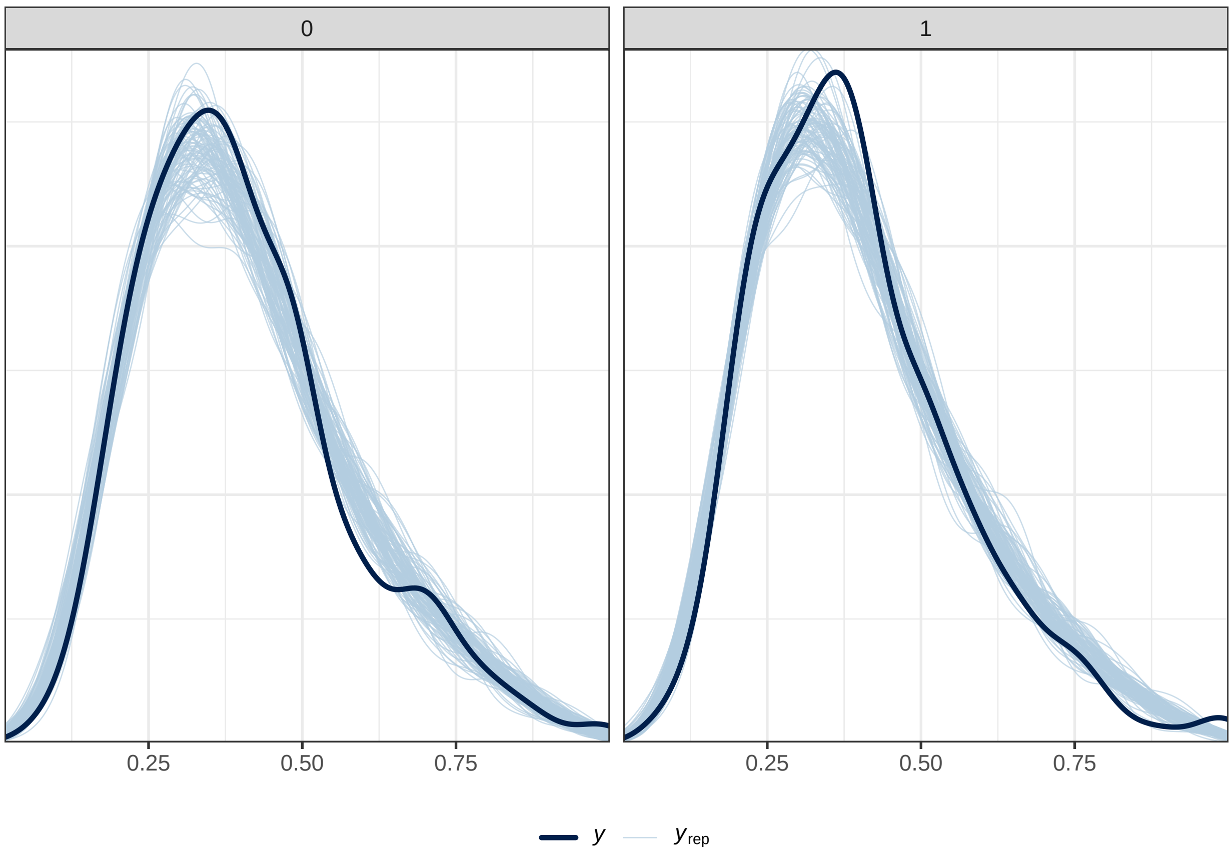
```r
pp_check(fit2,
         group = "cov_female",
         type = "dens_overlay_grouped",
         ndraws = 100)
```

Figure B1: Posterior Predictive Checks by Gender



The y-axis shows the relative probability density and the x-axis shows the item response. The dark blue line represents the observed data and the light blue lines represent 100 draws from the posterior distribution. 0 = male, 1 = female.

Figure B2: Posterior Predictive Checks by Item Attempt



The y-axis shows the relative probability density and the x-axis shows the item response. The dark blue line represents the observed data and the light blue lines represent 100 draws from the posterior distribution. 0 = first attempt, 1 = second attempt.