RICE UNIVERSITY

STAT 525

Bayesian Statistics

Shashank Sonkar

Project Report

1 Introduction

Let us say we are given some questions, and students' responses to these questions (responses are either marked correct or incorrect). Given this information, one of the most important tasks in the field of educational data mining is to predict each question's (also called an item) difficulty and each student's proficiency. Rasch laid the foundation of item response theory (IRT) in around 1960s [9, 10]. Using Markov Chain Monte Carlo (MCMC) methods, one can estimate the item difficulty and student proficiency parameters.

However, one of the main drawbacks of the IRT models is the lack of interpretability. To alleviate this drawback, the paper under review for this project [5] provides an elegant multilevel extension to traditional IRT models. This paper extends the standard IRT model further by grouping students into schools, and proposing a multilevel regression model to study the properties of groups (e.g. schools, in this case) as well. It adopts a fully Bayesian approach, unlike its predecessors, that provides it benefits like uncertainty quantification, and ease of incorporating information about schools through priors.

2 Project Outline and Contributions

We start by understanding the conceptual development of IRT models in section 3. One key idea, that we understood by taking this route, is the decision to either use the normal ogive or the logit function to model the success probability of a student's response to an item.

Rasch model, 1-PL, 2-PL, and 3-PL models use the logit function, while the paper under review for this project [5] uses the normal ogive function.

We list down the full conditionals for the normal ogive model, derived in [5], in section 5. To explore the multilevel IRT models through the lens of the logit function, and not just normal ogive function [5], we use Stan which provides us with the flexibility to model a logit multilevel IRT model without deriving the full conditionals.

We experiment on a small dataset [11] for which the Stan implementation of 2PL logit multilevel model is open-source [6]. We can easily modify the codebase for normal ogive model to replicate [5], and compare the two models. The primary motivation to compare the two models is to check which one of the two - logit or normal-ogive, provides a better fit to the data.

For implementations in Stan, [6] highlights the importance of non-centered parametrization for hierarchical models (see [8] for in-depth review). In our experiments, we also test the impact of non-centered parametrization on effective sample size in the case of normal ogive 2PL multilevel model.

Our experiments to run the models on the larger PISA 2003 dataset were unsuccessful, mainly because of the huge running time needed for successful convergence. Efforts to install and run the package (/source code) of 'mlirt' in R provided by the author of [5] in 2007 [4] did not pay off. However, we could extract the code required to pre-process and load the 2003 PISA dataset from the 'mlirt' package source code.

3 Conceptual Development of x-PL IRT Models

This section is based on the readings of chapter 1, 2, 5, and 6 of [3].

3.1 1-PL IRT Model

In an IRT model, one estimates the properties of the item which are responsible for the differing responses to that item. In education, these items are generally questions. In the 1-PL model, the most characteristic of the item is measured, which is the difficulty of the item. The probability that a student will answer j^{th} item correctly is modeled by:

$$P(y_j = 1|\theta, \delta_j) = \frac{e^{\theta - \delta_j}}{1 + e^{\theta - \delta_j}}, \tag{1}$$

where y_j models the correctness of the response of the student to item j (1 if correct, 0 otherwise), θ is the student's ability parameter, and δ_j is the item' difficulty parameter.

3.2 2-PL IRT Model

One can observe from (1) that when a student's ability parameter, θ, is equal to δ_j, the probability of the student answering the question j correctly is 0.5. As θ moves away from δ_j, the change in success probability ($dP/d\theta$) is not a function of the item (sigmoid function only gets shifted by δ_j). However, in educational settings, some items may require a large change in θ to achieve some fixed change in success probability from 0.5. To elucidate, δ_j is fixed at 1 in figure 1. The blue line indicates the change in success probability across item 1 and item 2, when the ability parameter of the student changes from 1 to 2. Thus, item 1 is said to be better at 'discriminating' students (more sensitive to θ), since a large change in θ is needed for the same increase in success probability from 0.5.

The probability that a student will answer j^{th} item correctly in the 2PL model is modeled by:

$$P(y_j = 1|\theta, \delta_j, \alpha_j) = \frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}, \tag{2}$$

where α_j is the item's discrimination parameter.

Note on use of sigmoid and its ramifications for multilevel IRT models:

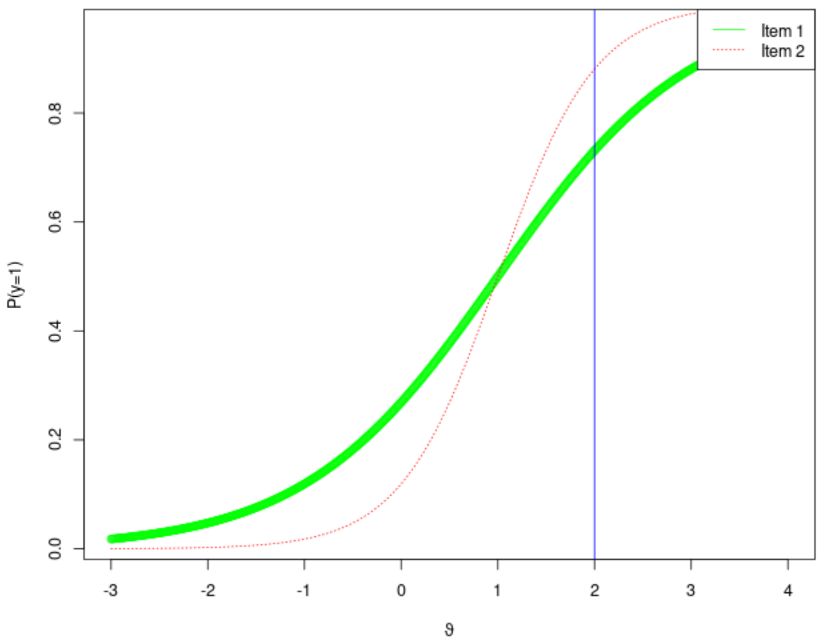

Figure 1: 2PL model

Fitting the probability of the correct response using the sigmoid function is not an arbitrary choice. [3] provides data studies to back this decision. Shape of the sigmoid curve also has resemblance to the cumulative distribution function (CDF) of the standard normal distribution. Thus, one can also use the standard normal CDF to model this probability. In that case, the probability that a student will answer j^{th} item correctly is modeled by:

$$P(y_j = 1|\theta, \delta_j, \sigma_i) = \Phi(\frac{\theta - \delta_j}{\sigma_i}), \tag{3}$$

where Φ is the CDF of the standard normal distribution, δ_i is the item's difficulty parameter, and σ_i is the item's discrimination parameter. **These models are termed as normal ogive models.**

Note that the paper under review, [5] provides multilevel extension of the normal ogive models only.

In this project, we also modify the code 2PL Stan model for the logit model to transform it into the normal ogive 2PL Stan model, and compare the models by their fit on the observed data. Our findings suggest that normal ogive 2PL model provides better fit for the dataset in study.

3.3 3-PL IRT Model

In educational settings, multiple-choice questions (MCQs) are the easiest to grade, since open-form questions require human intervention to grade the response. However, designing the choices for MCQs is challenging. A question may be hard, but looking at the choices, it may be easy to guess the correct response. 'By chance alone', it may be possible to correctly answer the item. To model this probability of correctly answering the item, purely by chance, 3-PL model is used.

The probability that a student will answer j^{th} item correctly in the 3-PL model is modeled by:

$$P(y_j = 1|\theta, \delta_j, \alpha_j) = \chi_j + (1 - \chi_j)\frac{e^{\alpha_j(\theta-\delta_j)}}{1 + e^{\alpha_j(\theta-\delta_j)}}, \tag{4}$$

where χ_j models the probability of correctly answering the item by pure guessing.

4 Multilevel IRT Model Specifications

4.1 Motivation for multilevel IRT model

Just like deep learning models, the key drawback with the IRT models is their lack of interpretability. Assigning meaning to θ and δ parameters is upto to the modeler. In fact, the origin of 2PL and 3PL models was to fit the observed data more accurately, which the simple Rasch model did not provide [3]. The trade-off is that the IRT models fail to answer simple, yet necessary questions, for instance, does gender affect the students' ability in particular subjects, how much is the school responsible for poor performance of students, etc. These are quite relevant questions for policy makers, however, IRT models fail to offer satisfactory answers to these questions. On the other hand, multilevel IRT models alleviate this drawback due to their explanatory model design.

4.2 Description of ‘levels’ in multilevel IRT model

Let students be indexed by variable i, which ranges 1 to n_j, where j represents the index of the school from 1 to J. Let items be indexed by variable k, which ranges from 1 to K. Then, the probability that the student i studying in school j will answer the item k correctly is modeled by:

$$P(Y_{ijk} = 1)|\theta_i, a_k, b_k) = \Phi(a_k \theta_{ij} - b_k), \tag{5}$$

where θ_{ij} is the ability parameter of student i in school j, a_k is the item's discriminative power parameter, and b_k is the item's difficulty parameter. Equation (5) models the first level of the multi-level IRT model.

At level 2, let there be Q covariates, denoted by $\boldsymbol{x}$, to model θ_{ij}. Then,

$$\theta_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + ... + \beta_{Qj} x_{Qij} + e_{ij}. \tag{6}$$

x_{qij}, where $q \in [1, Q]$, contains the properties of the student indexed by ij. It can be 0 or 1, denoting if the student is male or female. It can be an integer to model the age of the student. Once the parameters β_{qj} are learned, one can explain the influence of students' properties on their abilities and how they vary across schools. For instance, one can analyze the impact of student's age (modeled by β) on his or her ability (modeled by θ), depending on the school (modeled by the index j in β).

Note that this level captures variance in student abilities within the same school, and provides explanations for students' abilities based on their characteristics. **In our experiments, we restrict our analysis to level 2 multilevel IRT models, and choose gender of the**

student as the covariate.

At level 3, let there be S covariates, denoted by $\boldsymbol{w}$.

$$\beta_{qj} = \gamma_{0q} + \gamma_{1q} w_{1j} + ... + \gamma_{Sq} w_{Sj} + u_{qj}, \tag{7}$$

for all $q \in [1, Q]$, and for all $j \in [1, J]$. w_{sj}, where $s \in [1, S]$, contains the s^{th} property of school j. For instance, it can be an integer to model the social, economic, and cultural status of the school. Then, γ_{sq} will explain the influence of the status of the school in relation to the q^{th} property of the student. For instance, if the q^{th} property denotes if the student is female or not (x_{qij} is 1 if the i^{th} student in j^{th} school is female, otherwise 0), one can analyze the importance of status of the school for better learning of the female students.

Note that this level captures variance in student abilities across different schools. Thus, it can explain the contribution of the school on difference in abilities of students.

Also, $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, while $\boldsymbol{u_j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{T})$.

Distinction from the paper notation and possible error: Notation from the paper [4] is followed, however, the paper reviewed is [5] from the same author. [5] was published in 2001, while [4] was published in 2007. This was done because [5] made an error (as far as I understood), and treated w to be indexed by q as well. Equation 5 from [5], which I feel has an indexing error, is written below.

$$\beta_{qj} = \gamma_{0q} + \gamma_{1q} w_{11j} + ... + \gamma_{Sq} w_{Sqj} + u_{qj}, \text{ for } q \in [0, .., Q].$$

4.3 Common Priors

Generally, items are not considered independent. They test a concept or are often tagged under an unifying skill. The below prior helps to capture the inter dependency between them.

$$(\log a_k, b_k) \sim \mathcal{N}(\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})$$

Hyperpriors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by:

$$\boldsymbol{\Sigma_I} \sim Inv - Wishart_v(V^{-1}),$$
$$\boldsymbol{\mu_I}|\boldsymbol{\Sigma_I} \sim \mathcal{N}(\mu_0, \Sigma_I/\kappa),$$

where v and V are the degrees of freedom and scale matrix of the inverse Wishart distribution, μ_0 is the prior mean, and κ is the number of observations.
Equations and notations for these priors are again taken from [4].

4.4 Identifiability Issues

For the feasibility of Gibbs sampling algorithm for normal ogive models (even without the multilevel extension), [1] proved that a variable transformation is needed. One needs to introduce a new variable, z_{ijk}, which is sampled using the following equation:

$$p(z_{ijk}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\xi_k}) = \mathcal{N}(a_k\theta_{ij}, 1),$$

where $\xi_k = (a_k, b_k)^t$.

5 MCMC algorithm for estimating Multilevel IRT

To implement the Gibbs sampling algorithm the full conditionals for multilevel IRT model parameters needed are given by (note that the equations are taken directly from [4]):

1. The derivation for the following equations are provided in [7].

$$\begin{aligned} p(\xi_k|\boldsymbol{\theta}, \boldsymbol{z_k}, \boldsymbol{\mu_I}, \boldsymbol{\Sigma_I}) &= p(\boldsymbol{z_k}|\boldsymbol{\xi_k}, \boldsymbol{\theta})p(\xi_k|\boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})/p(\boldsymbol{z_k}|\boldsymbol{\theta}, \boldsymbol{\mu_I}, \boldsymbol{\Sigma_I})) \\ &= \phi(\boldsymbol{\xi_k}|\hat{\boldsymbol{\xi_k}}, \Omega_I), \end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_{\boldsymbol{I}} &= \Omega_I(\boldsymbol{H}^{\boldsymbol{t}}\boldsymbol{z}_{\boldsymbol{k}} + \Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{I}}), \\
\Omega_I^{-1} &= \boldsymbol{H}^{\boldsymbol{t}}\boldsymbol{H} + \Sigma^{-1}, \\
\boldsymbol{H} &= [\boldsymbol{\theta}, \mathbf{1}],
\end{aligned}$$

and ϕ is the normal distribution.

2. The derivation for the following equations are provided in [1]. Note that this follows similar steps from the paper [7].

$$\begin{aligned}
p(\theta_{ij}|\boldsymbol{z'_{ij}}, \boldsymbol{\xi}, \boldsymbol{\beta_j}, \sigma^2) &= p(\boldsymbol{z'_{ij}}|\theta_{ij}, \boldsymbol{\xi})p(\theta_{ij}|\boldsymbol{\beta_j}, \sigma^2)/p(\boldsymbol{z'_{ij}}|\boldsymbol{\xi}, \boldsymbol{\beta_j}, \sigma^2) \\
&= \phi(\theta_{ij}|\mu_\theta, \Sigma_\theta),
\end{aligned}$$

where

$$\begin{aligned}
\mu_\theta &= \Sigma_\theta(\boldsymbol{a}^{\boldsymbol{t}}\boldsymbol{z'_{ij}} + \boldsymbol{x_{ij}}\boldsymbol{\beta_j}/\sigma^2), \\
\Sigma^{-1} &= \boldsymbol{a}^{\boldsymbol{t}}\boldsymbol{a} + \sigma^{-2},
\end{aligned}$$

and $\boldsymbol{z'_{ij}} = \boldsymbol{z_{ij}} + \boldsymbol{b}$.

3. Paper under review [5] provided the derivations for the regression coefficients which are listed below:

$$\begin{aligned}
p(\boldsymbol{\beta_j}|\boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{T}) &= p(\boldsymbol{\theta_j}|\boldsymbol{\beta_j}, \sigma^2)p(\boldsymbol{\beta_j}|\boldsymbol{\gamma}, \boldsymbol{T})/p(\boldsymbol{\theta}|\sigma^2, \boldsymbol{\gamma}, \boldsymbol{T}) \\
&= \phi(\boldsymbol{\beta_j}|\mu_\beta, \Sigma_\beta),
\end{aligned}$$

where

$$\begin{aligned}\mu_\beta &= \Sigma_\beta(\boldsymbol{x_j^t\theta_j}/\sigma^2 + \boldsymbol{T^{-1}w_j\gamma}),\\ \Sigma_\beta &= \boldsymbol{x_j^t x_j}/\sigma^2 + \boldsymbol{T^{-1}}.\end{aligned}$$

For regression coefficients at level 2, following equations are used:

$$\begin{aligned}p(\boldsymbol{\gamma}|\boldsymbol{\beta, T}) &= p(\boldsymbol{\beta}|\boldsymbol{\gamma, T})p(\boldsymbol{\gamma})/p(\boldsymbol{\beta}|\boldsymbol{T})\\ &= \phi(\gamma|\mu_\gamma, \Sigma_\gamma)\end{aligned}$$

where

$$\begin{aligned}\mu_\gamma &= \sum_j \boldsymbol{w_j^t T^{-1} w_j}(\sum_{\boldsymbol{j}} \boldsymbol{w_j^t T^{-1}\beta_j})\\ \Sigma_\gamma^{-1} &= \sum_j \boldsymbol{w_j^t T^{-1} w_j}\end{aligned}$$

To circumvent these tedious calculations for the multilevel logit IRT models, we use open-source logit IRT models Stan implementations in our experiments.

6 Experiment

6.1 Motivation

The purpose of this experiment is firstly, compare the model fit of normal ogive Stan 2PL model against the logit Stan 2PL model on observed data, secondly, compare the model fit of normal ogive multilevel Stan 2PL model against the logit multilevel Stan 2PL model on observed data, and lastly, measure the effect on non-centered parametrization on effective sample size in the case of normal ogive Stan 2PL model [8].

The paper under review [5] only derives the full conditionals for normal ogive model, and not for logit models. Using Stan, we can analyze if logit models provides a better fit of the data as compared to the normal ogive models.

To compare the model fits on observed data, we adopt standard techniques (for which open-source implementations are already available [6]) which try to replicate the data by sampling from posterior predictive distribution and subsequently, compare the replicated data against the original data test statistics using the following strategies.

1. Use a violin plot to graph the distribution of test scores.

2. For each replicated dataset, [6] uses χ^2_{NC} test statistic [2] to summarize the shift of the replicated test score distribution from the observed test score distribution, where χ^2_{NC} is defined as:
$$\chi^2_{NC} = \sum_{s=0}^{S} \frac{[NC_s - E(NC_s)]^2}{E(NC_s)},$$
   where S is the number of items, NC_s is the number of correct responses to s^{th} item generated by a replicated dataset.

   An additional posterior predictive p-value (PPP) is be defined to quantify the amount of replicated datasets that produced χ^2_{NC} values greater than the χ^2_{NC} of the observed dataset. PPP-values is given by:
$$PPP = p(\chi^2_{NC^{rep}} > \chi^2_{NC^{obs}})$$

6.2 Dataset

We use a small scale spelling dataset [11], which contain the test results of 658 students (284 male, 374 female) on a spelling task. Task involved spelling four words - girder, succumb, infidelity, and panoramic. We only model level 2 IRT model and check if the student ability varies with the gender of the student.

6.3 Results

6.3.1 Normal Ogive 2PL vs Logit 2PL model

Comparing figure 2 with 4, one can observe that posterior predictive medians of normal ogive 2PL model align better with observed data as compared to logit 2PL model, especially for item 3 and 4.

Also, the chi square discrepancy test produces heavier tails with largely incorrect PPP values for logit 2PL model, as compared to normal ogive 2PL model (figure 3 vs 5).

Thus, we can conclude from these tests that for this dataset in particular, using the normal ogive model provides a better fit to the data as compared to the logit model.

6.3.2 Normal Ogive Multilevel 2PL vs Logit 2PL Multilevel model

There is no clear winner - both models perform similarly, and seem to provide an equivalent good(/bad) fit to the observed data (figures 6, 8, 7, and 9).

6.3.3 Normal Ogive Multilevel 2PL: Centered vs Non-centered Parametrization

As stated earlier [6] suggests using non-centered parametrization for hierarchical models for better effective sample size. However, for normal ogive 2PL hierarchical models, we found that the mean effective size for all parameters was more for centered parametrization as compared to non-centered parametrization (465 vs 447 with standard deviation of 148 vs 184 respectively).

We found similar statistics for logit 2PL hierarchical models as well, however, the stan documentations [6] writes otherwise. There is a possibility of a bug in my normal ogive 2PL hierarchical Stan model implementation, but the implementation for logit 2PL hierarchical models used for this experiment is taken directly from [6].

6.4 Code Details

The code for 2PL logit, 2PL multilevel Logit model, plotting violin plots, and chi square discrepancy test results is provided in the case study [6] available in mc-stan documentation. The code has been minimally modified to model and sample from normal ogive function. File twopl_edstan_mlirt_l2_ogive_nc.stan contains the necessary changes.

6.5 Future Work

The decision to go with the Spelling dataset, instead of PISA 2003 dataset, which was originally the plan, was because of ease of use of the dataset. 'mlirt' package [4], being an old package (2007), had difficulty installing with new R versions and other dependent packages versions. However, the code to load PISA 2003 dataset has been extracted out. It took a lot of time to run a simple 2PL Stan model for the dataset (code included in zip). Thus, in future work, we can compare the normal ogive vs logit three-level IRT model fit fit to 2003 PISA dataset.

References

[1] James H Albert. Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of educational statistics*, 17(3):251–269, 1992.

[2] Anton A Béguin and Ceec AW Glas. Mcmc estimation and some model-fit analysis of multidimensional irt models. *Psychometrika*, 66(4):541–561, 2001.

[3] Rafael Jaime De Ayala. *The theory and practice of item response theory.* Guilford Publications, 2013.

[4] Jean-Paul Fox et al. Multilevel irt modeling in practice with the package mlirt. *Journal of Statistical Software*, 20(5):1–16, 2007.

[5] Jean-Paul Fox and Cees AW Glas. Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, 66(2):271–288, 2001.

[6] Daniel C Furr, S Lee, J Lee, and Sophia Rabe-Hesketh. Two-parameter logistic item response model. *Stan Case Studies*, 2016.

[7] Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

[8] Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

[9] Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press, 1960.

[10] Georg Rasch. An item analysis which takes individual differences into account. *British journal of mathematical and statistical psychology*, 19(1):49–57, 1966.

[11] David Thissen, Lynne Steinberg, and Howard Wainer. Detection of differential item functioning using the parameters of item response models. *Lawrence Erlbaum Associates, Inc*, 1993.

Appendix

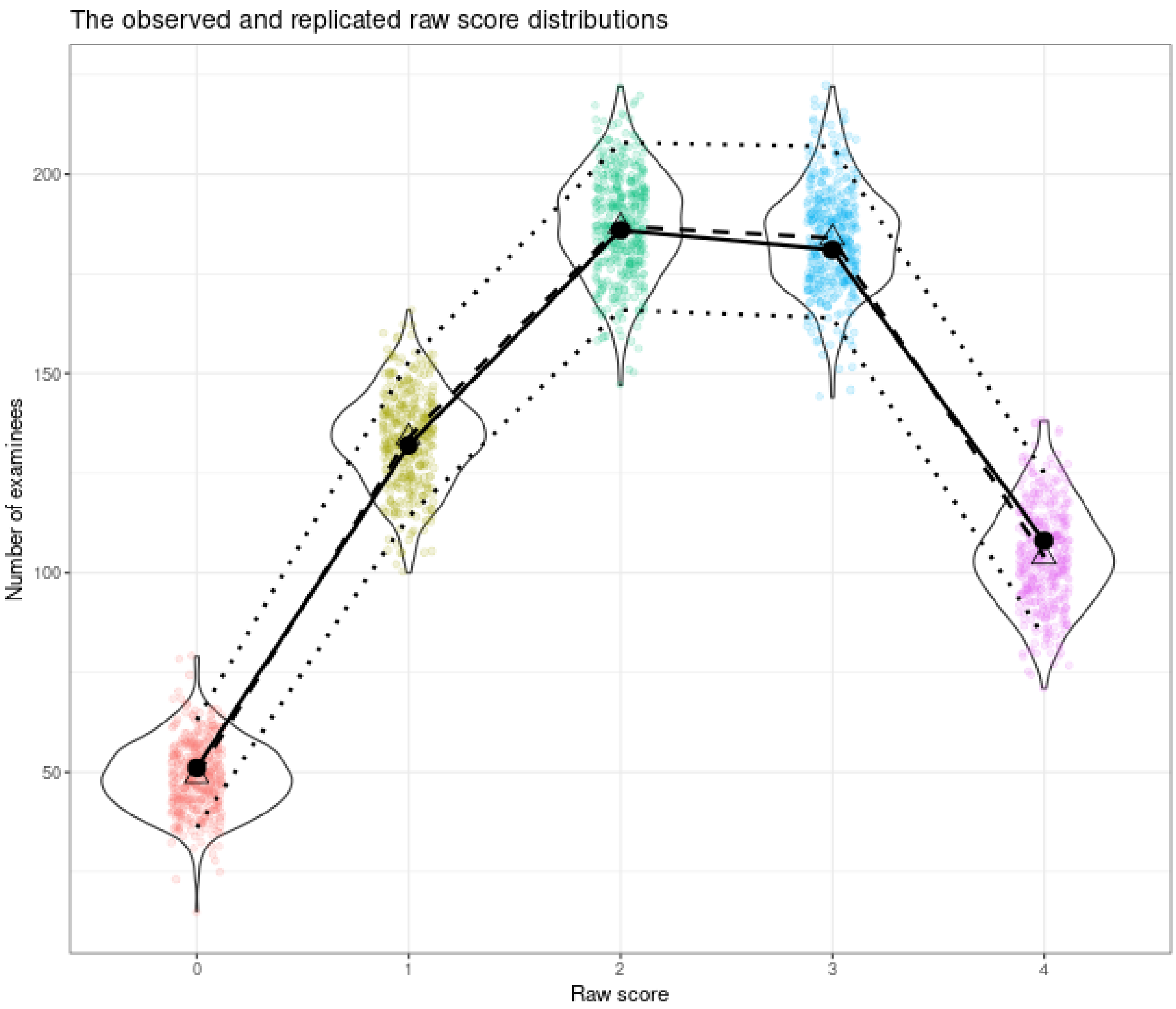

Figure 2: Raw Score Distribution for 2PL Logit model. Hollow triangles are posterior predictive medians, while black dots are the number of correct responses to that item.

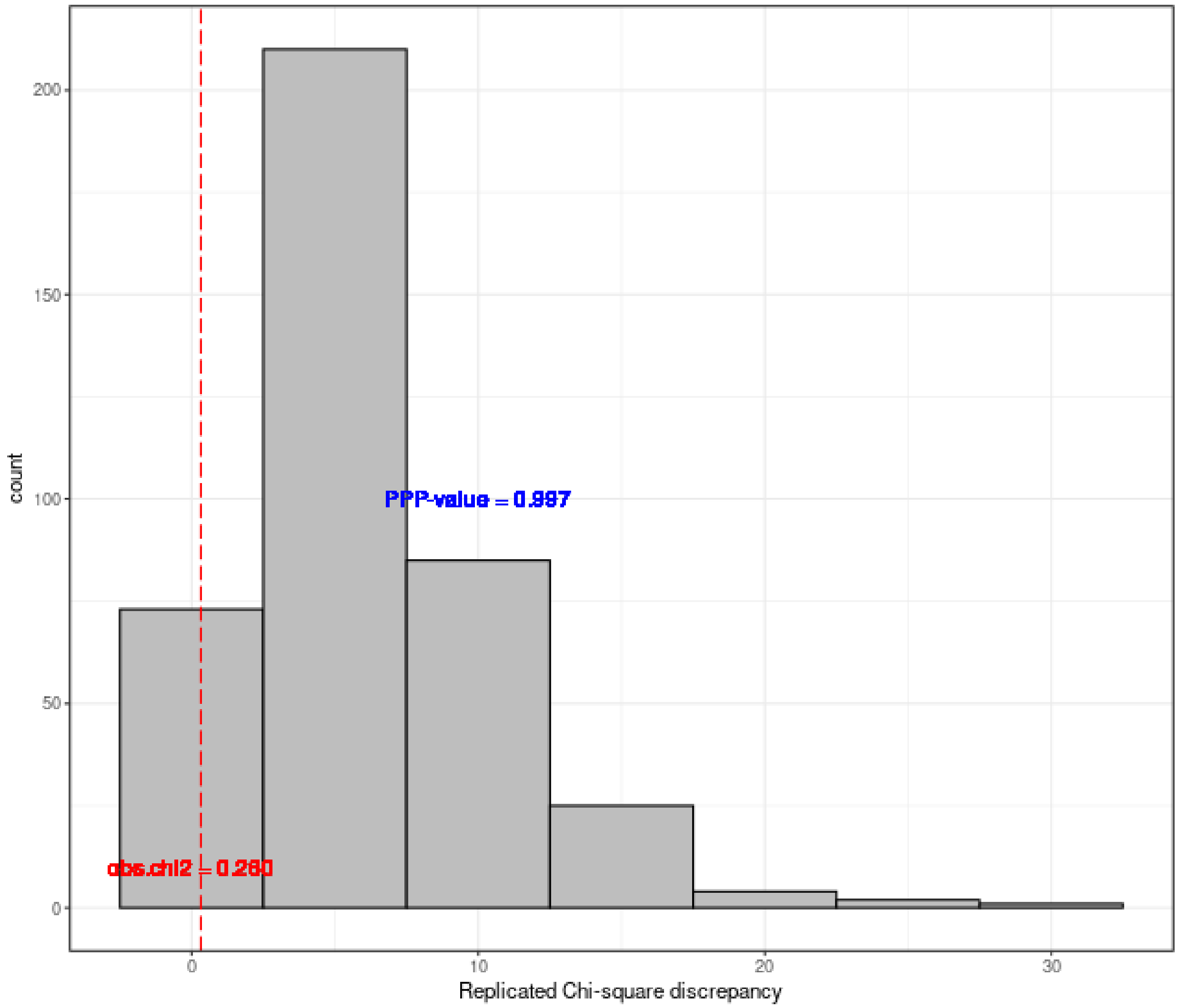

Figure 3: Chi Square Discrepancy for 2PL Logit model

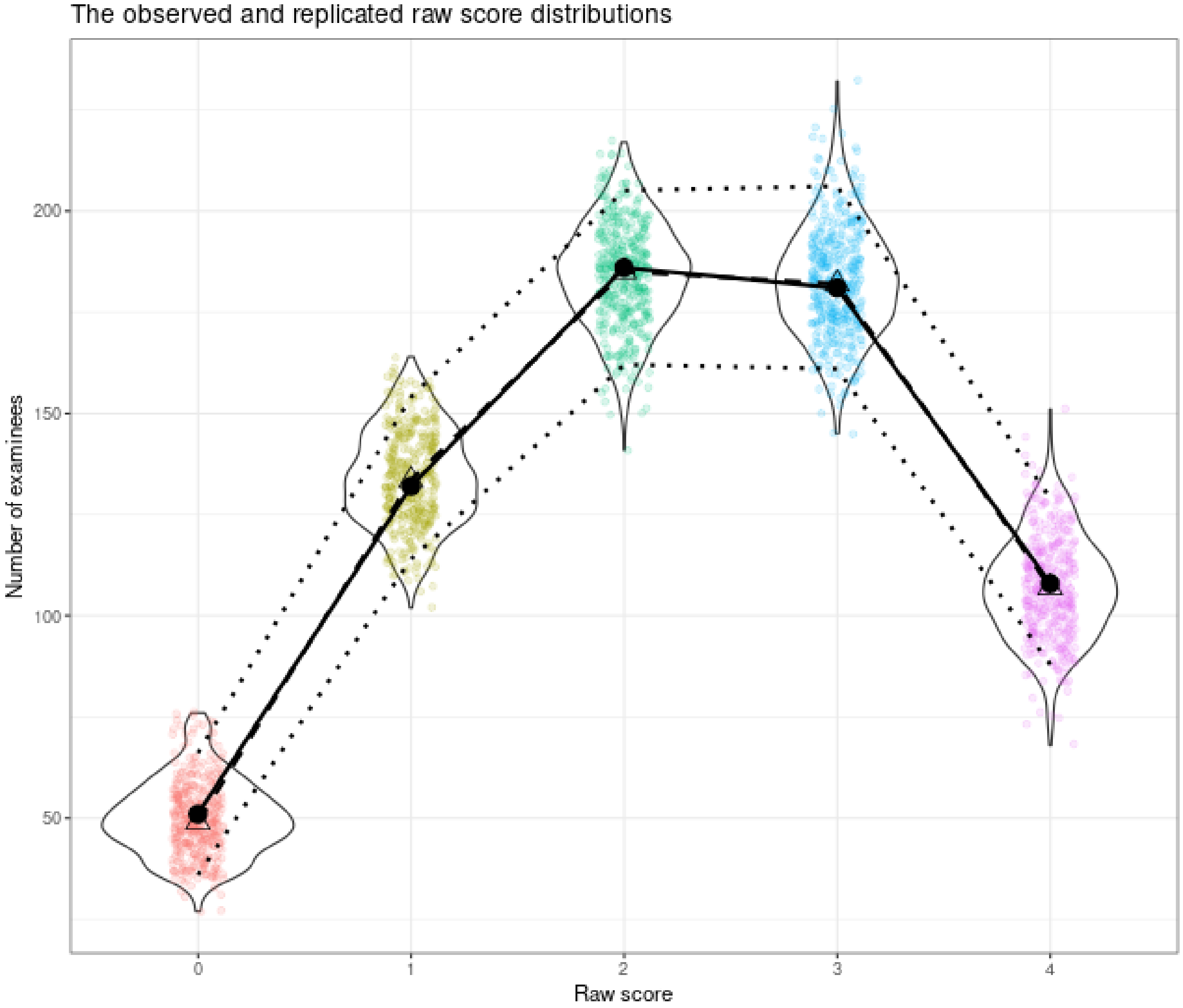

Figure 4: Raw Score Distribution for 2PL Normal Ogive model. Hollow triangles are posterior predictive medians, while black dots are the number of correct responses to that item.

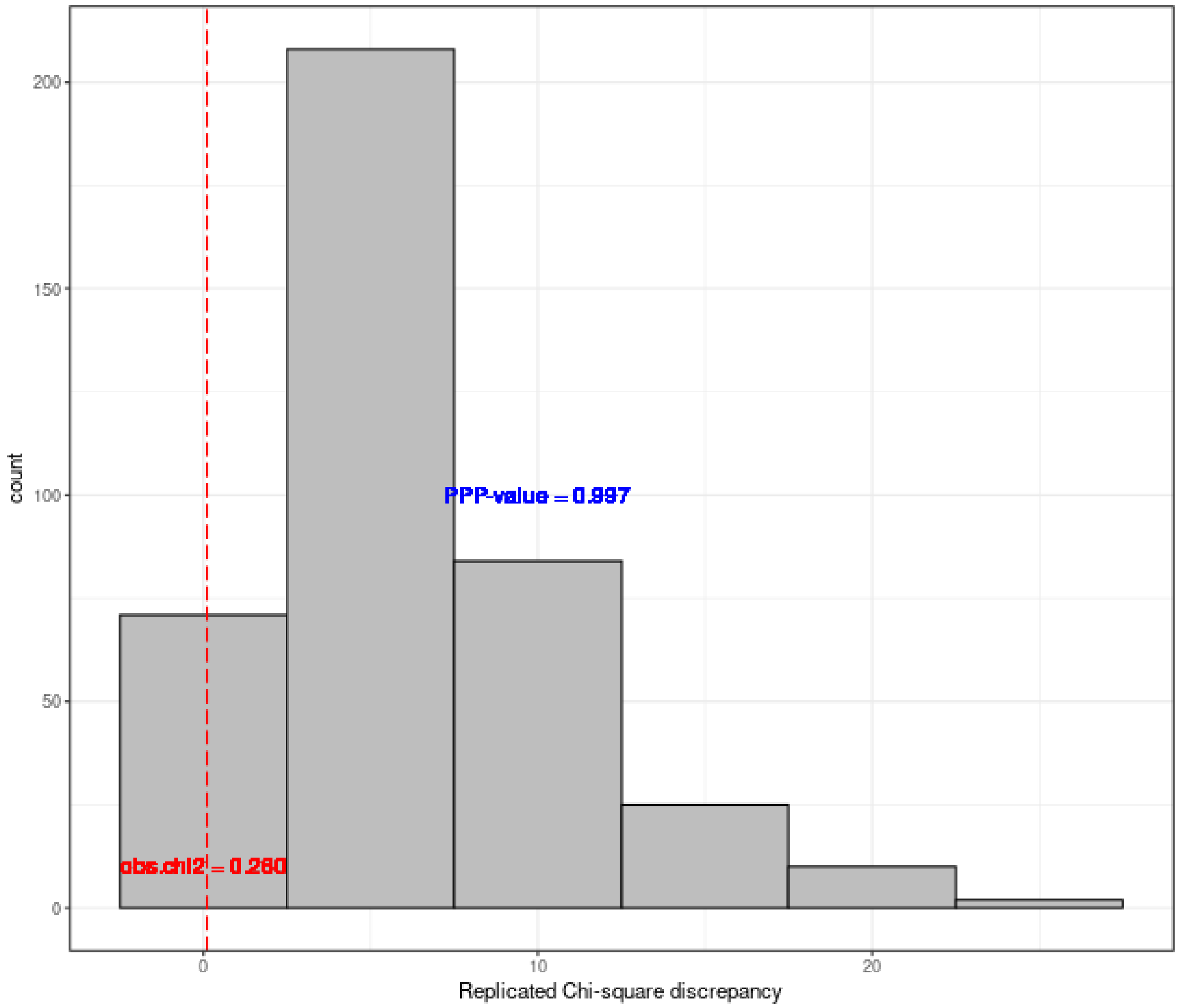

Figure 5: Chi Square Discrepancy for 2PL Normal Ogive model

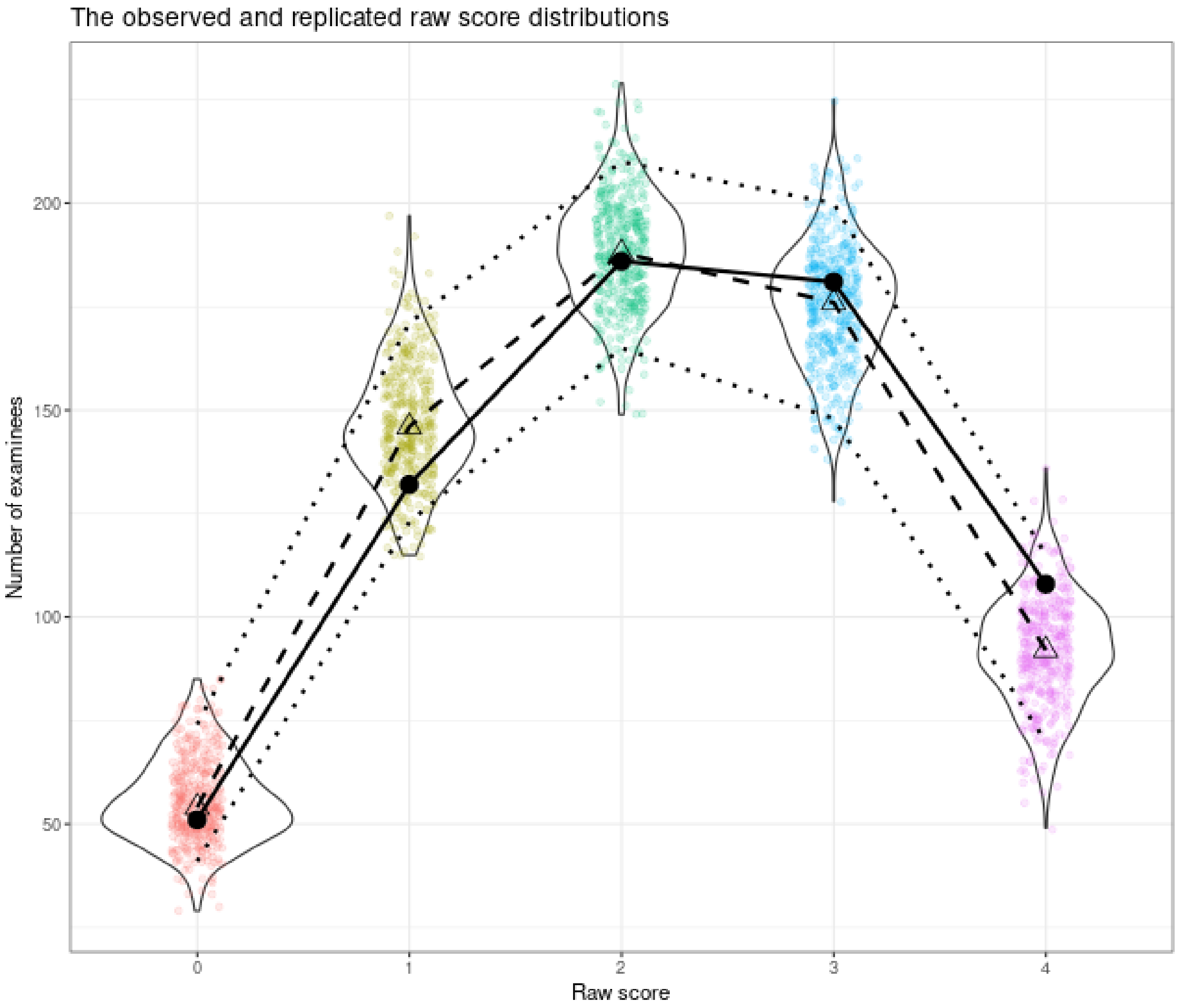

Figure 6: Raw Score Distribution for 2PL Multilevel Logit model. Hollow triangles are posterior predictive medians, while black dots are the number of correct responses to that item.

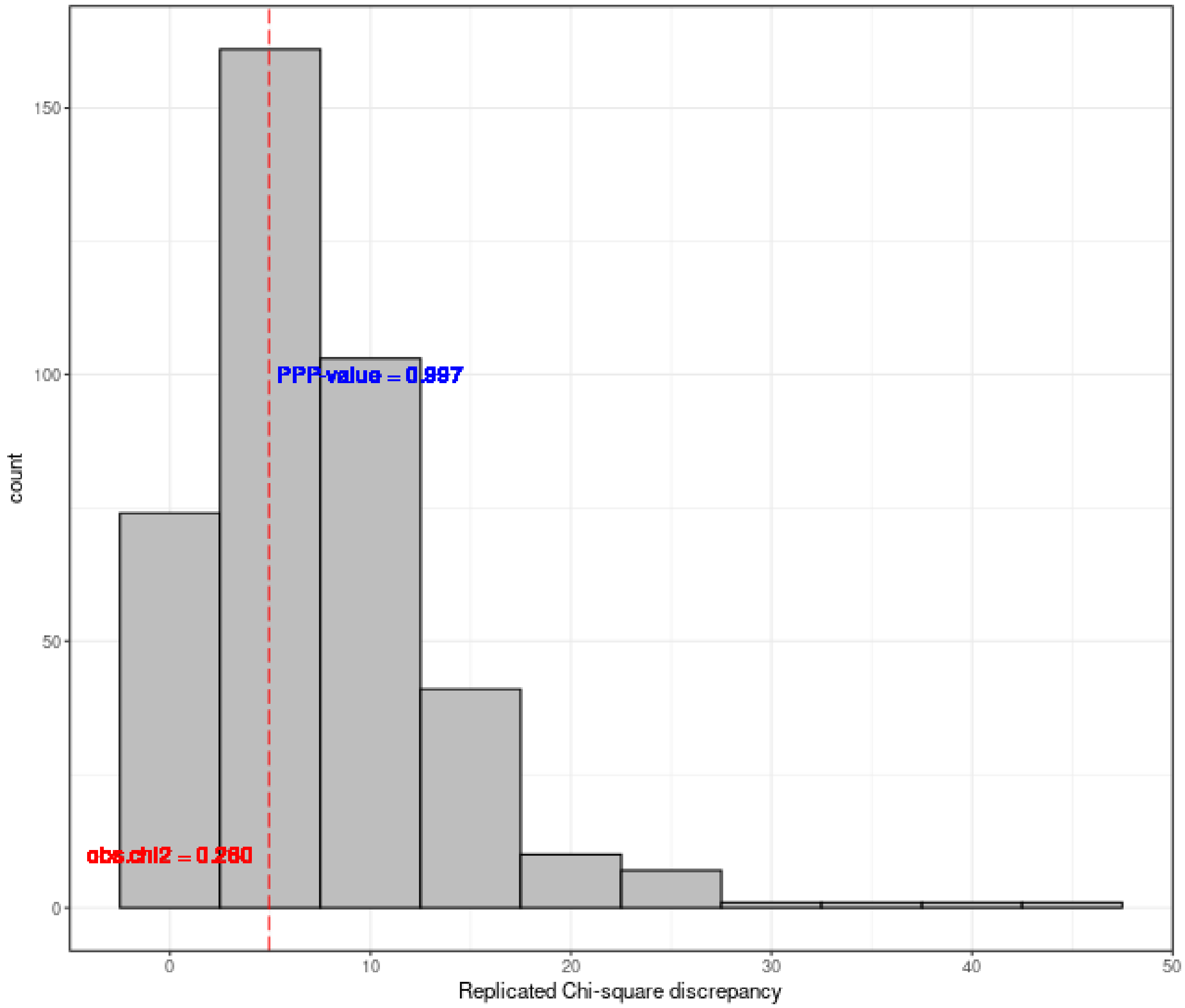

Figure 7: Chi Square Discrepancy for 2PL Multilevel Logit model

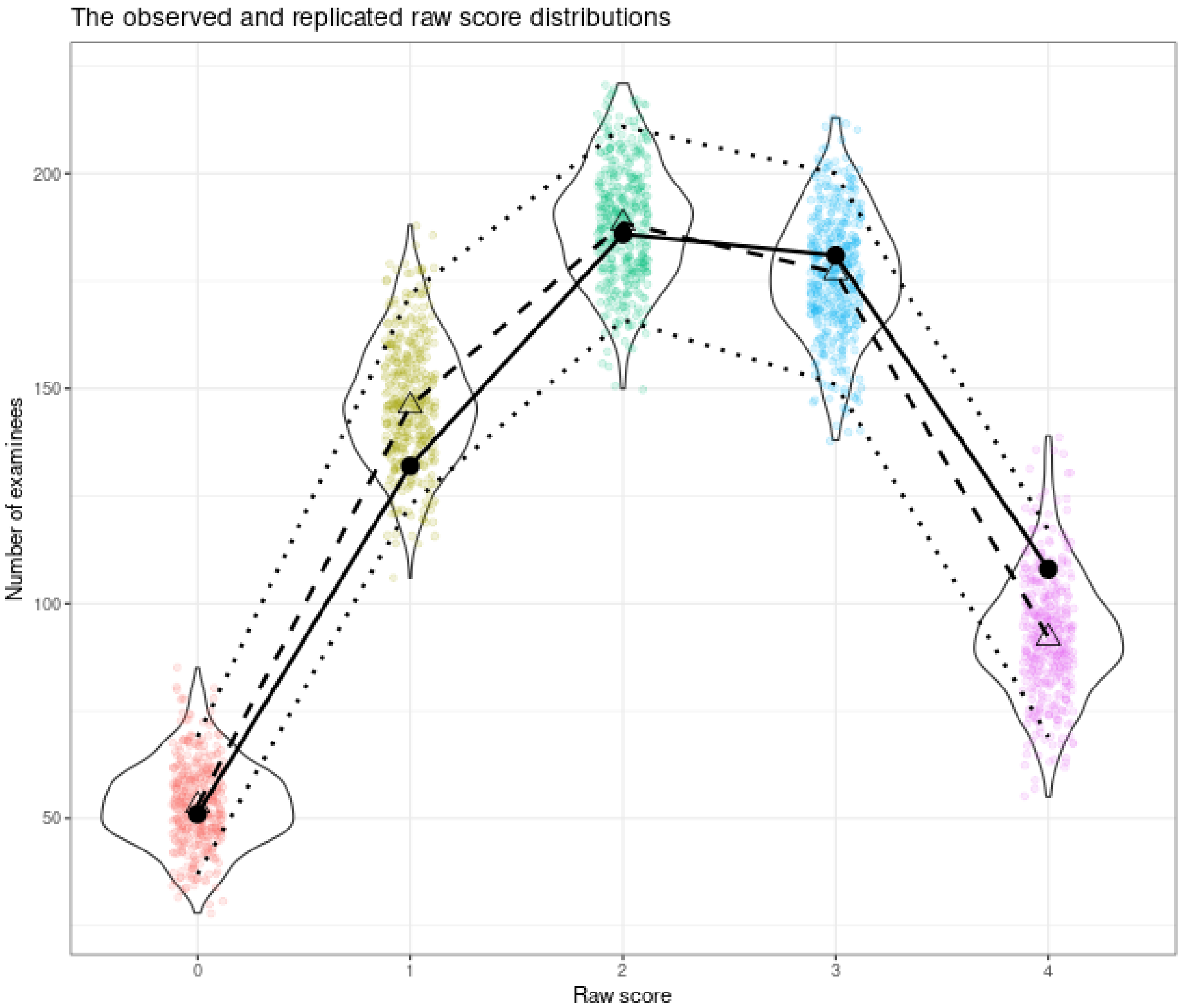

Figure 8: Raw Score Distribution for 2PL Multilevel Normal Ogive model. Hollow triangles are posterior predictive medians, while black dots are the number of correct responses to that item.

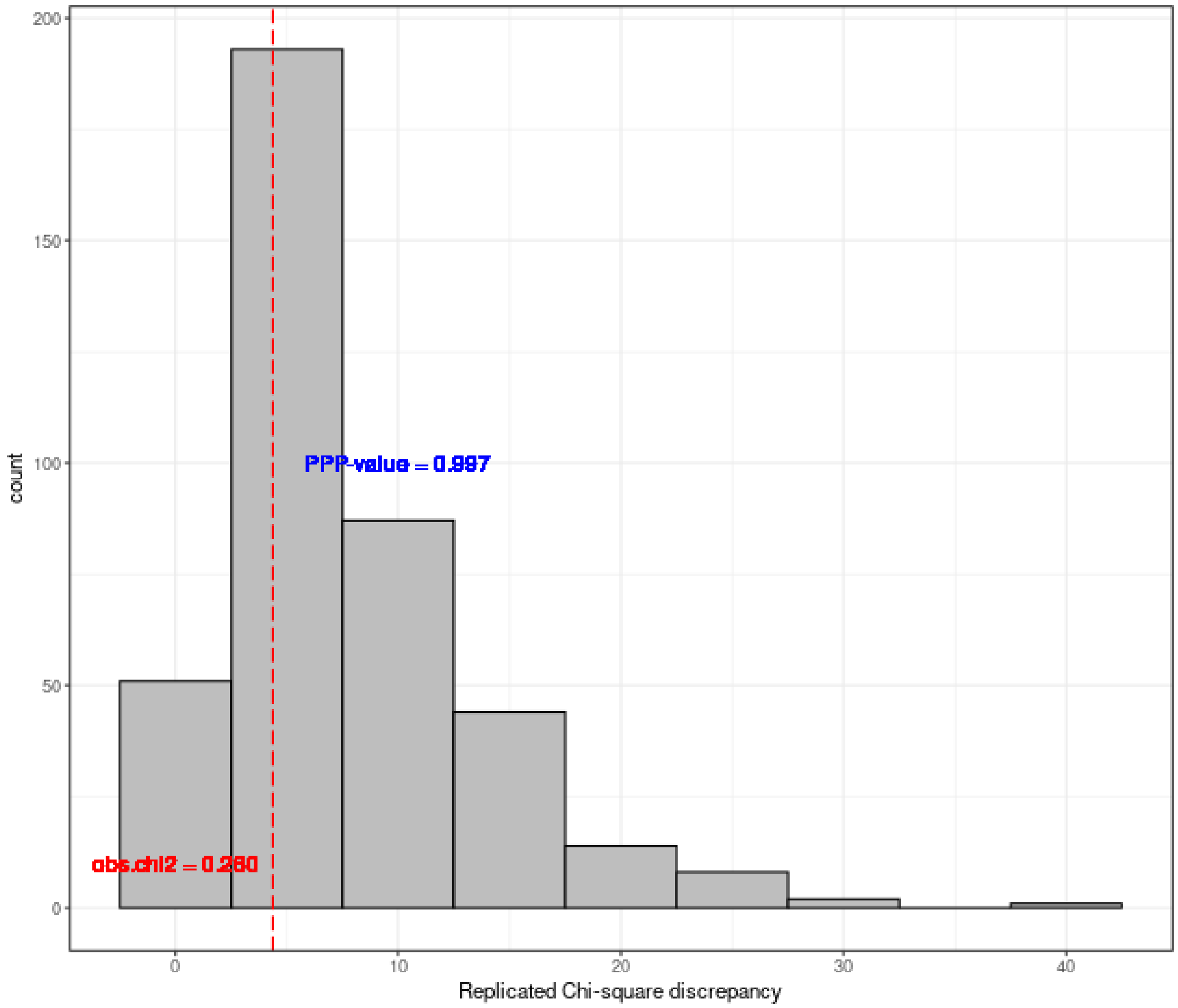

Figure 9: Chi Square Discrepancy for 2PL Multilevel Normal Ogive model