Note: This is a conference paper presented at the Texas Universities Educational Statistics and Psychometrics (TUESAP) annual meeting, March 2023, Austin, TX. This work is funded by Grant R305D200038 from Institute of Education Sciences - U.S. Department of Education.

Beta-binomial Model for Count Data: An Application in Estimating Model-based Oral Reading Fluency

Xin Qiao
University of South Florida

Akihito Kamata

Yusuf Kara
Southern Methodist University

Cornelis Potgieter
Texas Christian University

Joseph Nese
University of Oregon

Abstract

In this article, the beta-binomial model for count data is proposed and demonstrated in terms of its application in the context of oral reading fluency (ORF) assessment, where the number of words read correctly (WRC) is of interest. Existing studies adopted the binomial model for count data in similar assessment scenarios. The beta-binomial model, however, takes into account extra variability in count data that has been neglected by the binomial model. Therefore, it accommodates potential overdispersion in count data compared to the binomial model. To estimate model-based ORF scores, WRC and response times were jointly modeled. The full Bayesian Markov Chain Monte Carlo (MCMC) method was adopted for model parameter estimation. A simulation study showed adequate parameter recovery of the beta-binomial model and evaluated the performance of model fit indices in selecting the true data-generating models. Further, an empirical analysis illustrated the application of the proposed model using a dataset from a computerized ORF assessment. The obtained findings were consistent with the simulation study and demonstrated the utility of adopting the beta-binomial model for count-type item responses from assessment data.

Beta-binomial Model for Count Data: An Application in Estimating Model-based Oral Reading Fluency

1. Introduction

Assessments consisting of many simple, similar and often repetitive tasks where aggregation of successes or errors is used for scoring are commonly seen in educational, psychological and neurocognitive contexts. For example, in the testing of psychomotor skills, the same task is attempted for a limited number of times and counts of successes are recorded (Spray, 1990). In addition, in oral reading assessments, counts of errors or successes on several items are used for scoring (Jansen, 1997; Rasch, 1960). In these two types of assessments, the Binomial Trials Model (Rasch, 1972) and the Rasch Poisson Counts Model (RPCM, Rasch, 1960) are typically used to model the count variables, respectively. Both models belong to the Rasch family and have the same algebraic basis (Masters & Wright, 1984). Mathematically, if we consider number of successes u out of n trials with probability of success on each trial as p, the binomial model becomes the Poisson model in the limit when p becomes small and n becomes large with the expected number of successes np being constant (Masters & Wright, 1984). Therefore, the Poisson model is usually used for rare events while the binomial model is used for count data with the upper limit and the probability of success for each trial being large.

There has been extensive development related to the applications of the Poisson model but much less attention has been paid to the binomial model in the literature on psychometric modeling of count data. For example, the RPCM has been extended to the negative binomial regression model to accommodate the overdispersion in response data from a reading accuracy test (Hung, 2012). In addition, Liu, Liu, Shi, and Jiang (2022) has adapted the RPCM into a diagnostic classification model with discrete latent variables for response data from an English recognition assessment. In these two examples among many other applications related to the RPCM, the number of reading errors (i.e., misread words) was modeled even when the success of events was clearly defined, which led to less straightforward interpretations of latent ability parameters.

In this article, we argue that the binomial family models should be considered for modeling count-type item response data when appropriate. There are several benefits of modeling count data with binomial family distributions. First, unlike the Poisson family models, count data with an upper bound can be modeled by adopting a binomial family distribution. This is thought to be more in line with typical psychological assessments, where a natural limit on the tasks exists. For example, in reading tasks consisting of a fixed number of words, it is reasonable to consider each attempt (i.e., reading a word) as a binomial trial, where the maximum number of attempts is bounded by the total number of words. Second, the binomial family models can model the number of successes directly when successes are common, thus, providing more straightforward model parameter interpretations. In fact, large success probabilities are very likely given the simplicity of the repetitive tasks. For example, empirical data have shown that a majority of students can recognize most words correctly in the English recognition assessment (Liu et al., 2022).

To our knowledge, the only recent applications of the binomial model for count-type item response data are in Potgieter, Kamata, and Kara (2017) and Kara, Kamata, Potgieter, and Nese (2020) in the context of oral reading fluency (ORF) assessments. These studies jointly modeled counts of words read correctly (WRC) and response times (RTs) for reading each passage to estimate words read correctly per minute (WCPM) scores as the measure of latent ability. The purpose of the current study is to extend previous work and propose the beta-binomial model to accommodate potential overdispersion in count data that has been neglected by the binomial model. In this article, we refer to the binomial model in Kara et al. (2020) as JRT-Bin and our proposed beta-binomial model as JRT-BetaBin, where JRT stands for joint modeling of responses and response times. Through a simulation study, we aim to evaluate the impact of ignoring overdispersion in the binomial modeling of count data on parameter recovery and the performance of model fit indices in selecting the true data-generating model. Then, we illustrate the application of the proposed model in the context of ORF assessments. It is worth noting that passages in the ORF contexts are analogous to the concepts of items in traditional assessments. Thus, we use the terms items and passages interchangeably in this article.

The rest of the article is organized as follows. First, we present the formulation, assumptions, and parameter estimation of the proposed model. Second, we describe the simulation study design and present our findings. Third, we conduct an empirical data analysis to demonstrate the use of the proposed model and compare its performance with the JRT-Bin. Lastly, conclusions and future directions are discussed.

2. The Proposed Model

Similar to van der Linden (2007), the proposed model JRT-BetaBin is a joint model consisting of the accuracy component and the speed component. Specifically, the accuracy and speed components are measurement models for count data and response time data, respectively.

Beta-binomial Model for Count Data

We adopt the beta-binomial distribution as the link function for count data. The probability mass function for the beta-binomial distribution with number of trials $n \in \mathbb{N}_0$ and parameters $\phi > 0$, $\psi > 0$ is given by:

$$P(U = u|n, \phi, \psi) = \binom{n}{u} \frac{B(u + \phi, n - u + \psi)}{B(\phi, \psi)}, \tag{1}$$

where $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the beta function and $u \in 0, ..., n$. In the context of ORF assessment, we denote U_{ij} as a random variable describing the count of WRC for person j in passage i that has a total of n_i words. We further parameterize

$$\phi_{ij} = \frac{p_{ij}(n_i - \nu_i)}{\nu_i - 1} \tag{2}$$

and

$$\psi_{ij} = \frac{(1 - p_{ij})(n_i - \nu_i)}{\nu_i - 1}, \tag{3}$$

where $p_{ij} = \Phi(a_i(\theta_j - b_i))$ is the success probability following the same parameterization as the two-parameter normal-ogive IRT model with passage discrimination parameter $a_i > 0$, passage difficulty parameter $b_i \in \mathbb{R}$, and reading accuracy ability $\theta_j \in \mathbb{R}$; $\nu_i > 1$

denotes the dispersion parameter for passage i. In summary, the beta-binomial model is expressed as:

$$u_{ij}|\theta_j \sim \text{BetaBin}\Big(n_i, \frac{p_{ij}(n_i - \nu_i)}{\nu_i - 1}, \frac{(1 - p_{ij})(n_i - \nu_i)}{\nu_i - 1}\Big). \tag{4}$$

Given this model formulation, we have $n_i p_{ij}$ as the conditional mean and $n_i p_{ij}(1 - p_{ij})\nu_i$ as the conditional variance of the counts of WRC given a specific latent reading accuracy level and a specific passage. It can be seen that the conditional mean based on the beta-binomial model is the same as that based on the binomial model, while the conditional variance is ν_i times of that indicated by the binomial model. Therefore, ν indicates overdispersion and takes into account extra variability in the count data that is ignored by the binomial model.

Lognormal Model for Response Time Data

We adopt the lognormal model as in (van der Linden, 2006), where passage RTs (in seconds) are assumed to have a lognormal distribution:

$$f(t_{ij}|\tau_j) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp(-\frac{1}{2}[\alpha_i(\text{log}t_{ij} - (\beta_i - \tau_j))]^2), \tag{5}$$

in which t_{ij} denotes RT for passage i and person j; $\tau_j \in \mathbb{R}$ is the latent speed ability for person j; $\beta_i \in \mathbb{R}$ is the time intensity for item i, indicating the labor required for reading that passage; and $\alpha_i > 0$ represents the time discrimination parameter that reflects the variance of $\text{log}t_{ij}$.

Higher-order Model

The reading accuracy θ and latent speed τ are further assumed to follow a bivariate normal distribution in the population:

$$\begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim MVN \begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix} \begin{pmatrix} \sigma_\theta^2 & \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix} \tag{6}$$

The diagonal of the matrix indicates the variances of the latent accuracy and speed abilities. The off-diagonal of the matrix indicates the covariance between the latent variables. Passage parameters are modeled as fixed effects. It is possible to treat passage parameters as random effects as in the hierarchical model (van der Linden, 2007). However, it is challenging to obtain unbiased parameter estimates with the extreme short test length (i.e., less than 10 passages) in ORF assessments. Further, treating passage parameters as fixed effects would not affect the accuracy and efficiency of parameter estimates based on previous studies (e.g., Molenaar, Tuerlinckx, & van der Maas, 2015).

There are several assumptions made on the proposed model. First, conditional independence (CI) among observed variables is assumed. Similar to what is stated in van der Linden and Glas (2010), three CI assumptions exist in the proposed model: 1) independence between counts given θ; 2) independence between RTs given τ; and 3) independence between counts and RTs given θ and τ. Further, θ and τ are assumed to be constant across passages. Lastly, similar to the binomial model, the beta-binomial count model assumes that each attempt (e.g., reading each word in a passage) is independent of one another and only the number of successes is recorded regardless of the order of successes and failures (Masters & Wright, 1984).

Model Parameter Estimation

We used the full Bayesian approach with the Markov chain Monte Carlo (MCMC) method to estimate model parameters. Specifically, we used the R2jags package to interface with JAGS (Plummer, 2015), which implements the slice-with-Gibbs sampler algorithm (Gelfand & Smith, 1990). For all data-fitting models, four Markov chains using 40,000 iterations and 20,000 burn-in were run with random starting points. The thinning was set as 4 to reduce autocorrelations among draws. Further, the potential scale reduction factor (PSRF, Brooks & Gelman, 1998) smaller than 1.1 was used to indicate an adequate model convergence. In addition, the effective sample size (ESS) larger than 400 was considered as satisfactory precision of the Bayesian estimation (Zitzmann & Hecht, 2019).

Assuming conditional independence, U_{ij} and $\log T_{ij}$ are conditionally and independently distributed as $U_{ij}|\theta_j, a_i, b_i, \nu_i \sim \text{BetaBin}\Big(n_i, \frac{p_{ij}(n_i-\nu_i)}{\nu_i-1}, \frac{(1-p_{ij})(n_i-\nu_i)}{\nu_i-1}\Big)$, where

$p_{ij} = \Phi(a_i(\theta_j - b_i))$ and $\log T_{ij}|\tau_j, \beta_i, \alpha_i \sim N(\beta_i - \tau_j, \frac{1}{\alpha_i^2})$. We used noninformative normal priors for all passage parameters as follows: $a_i \sim N(0, 100)T(0, +\infty)$, $b_i \sim N(0, 100)$, $\alpha_i \sim N(0, 100)T(0, +\infty)$, $\beta_i \sim N(0, 100)$, $\nu_i \sim N(0, 100)T(1, +\infty)$, where $T(\cdot)$ indicates truncation distribution with the two values in the parenthesis as lower and upper bounds. A Gamma distribution was used for the conditional precision of the speed parameter as $\sigma_\tau^{-2}|\theta_j \sim Gamma(0.01, 0.01)$. Finally, a normal prior with zero mean and an arbitrarily large variance was assigned to the covariance between the speed and accuracy parameters as $\sigma_{\theta\tau} \sim N(0, 100)$.

3. Simulation Study

The purpose of the simulation study was two-fold: 1) to evaluate the parameter recovery of JRT-BetaBin and the impact of ignoring overdispersion on parameter recovery of JRT-Bin; 2) to evaluate the performance of model fit indices in terms of correctly selecting the true data-generating model.

Fixed Factors

In data generation, fixed factors included: population distribution of latent variables and population values of passage parameters. Specifically, the latent variables (i.e., θ and τ) were drawn from

$$\begin{pmatrix} \theta \\ \tau \end{pmatrix} \sim MVN \begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0.16 \\ 0.16 & 0.16 \end{pmatrix}.$$

That is, we fixed $\rho_{\theta\tau} = 0.4$, $\sigma_\tau^2 = 0.16$. We also fixed $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta^2 = 1$ for model identification purposes. We selected population values of passage parameters from a calibrated pool as part of a previously conducted study in Potgieter et al. (2017). The population values are presented in Table 1.

Manipulated Factors

We manipulated three factors in the simulation study: sample size (N), test length (i.e., number of passages read in the ORF assessment context) (I), and overdispersion levels in count data. For sample size, we set factor levels as 100, 300, and 500 to represent small, medium, and large sample sizes. We expect sample size would affect the parameter recovery of passage parameters and true WCPM scores. For test length, we set factor levels as 3 and 6 passages to represent short and long test lengths based on real scenarios in ORF assessments. We expect test length would affect the parameter recovery of reading accuracy, latent speed, and true WCPM scores. For overdispersion levels, we manipulated three global dispersion levels (i.e., $\nu = 1$, $\nu = 3$, and $\nu = 5$) and one item-specific dispersion level (i.e., $\nu_1 = \nu_4 = 1$, $\nu_2 = \nu_5 = 3$, $\nu_3 = \nu_6 = 5$). Specifically, for passages with $\nu_i = 1$, the true data-generating model was JRT-bin. For passages with $\nu_i > 1$, the true data-generating model was JRT-BetaBin. The three factors were fully crossed, yielding $3 \times 2 \times 4 = 24$ conditions.

Table 1

True Passage Parameters

Passage (i)	a_i	b_i	α_i	β_i	N_i
1	0.500	-2.771	6.176	3.270	47
2	0.444	-2.923	5.747	3.357	49
3	0.557	-2.760	4.841	3.320	50
4	0.505	-2.856	5.064	3.359	50
5	0.447	-2.949	4.408	3.315	50
6	0.476	-2.877	3.936	3.341	49

Note. First three passages were used under simulation condition $I = 3$.

Data-fitting Models

We simulated thirty datasets (i.e., 30 replications) for each simulated condition. Further, we fit two data-fitting models to each dataset: JRT-BetaBin and JRT-Bin models. As mentioned earlier, the two models differed in their link functions for count data: JRT-BetaBin included the beta-binomial model, while JRT-Bin included the binomial model for count data, respectively. The RT model was the same lognormal model in both JRT-BetaBin and JRT-Bin.

Outcome Measures

We investigated the performance of both relative and absolute model fit measures. For relative model fit, we investigated the performance of the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) by recording the frequency of replications where DIC correctly identified the true data-generating model. In addition, we conducted posterior predictive model checking (PPMC, Gelman, Meng, & Stern, 1996; Meng, 1994; Rubin, 1984) to examine the absolute model fit and item (i.e., passage) fit for both data-fitting models. We examined the absolute fit for the count data model and RT model separately. Specifically, the discrepancy measure employed for the model fit of the beta-binomial count data model is calculated as

$$D(u_{ij};\theta_j,a_i,b_i,\nu_i)=\sum_{i=1}^{I}\sum_{j=1}^{N}\frac{\left(u_{ij}-n_ip_{ij}\right)^2}{n_ip_{ij}(1-p_{ij})\nu_i}. \tag{7}$$

The discrepancy measure employed for the fit of the binomial count data model is calculated as

$$D(u_{ij};\theta_j,a_i,b_i,\nu_i)=\sum_{i=1}^{I}\sum_{j=1}^{N}\frac{\left(u_{ij}-n_ip_{ij}\right)^2}{n_ip_{ij}(1-p_{ij})}. \tag{8}$$

The discrepancy measure employed for the response time model is based on the statistic proposed by Marianti, Fox, Avetisyan, Veldkamp, and Tijmstra (2014):

$$D(t_{ij};\tau_j,\alpha_i,\beta_i)=\sum_{i=1}^{I}\sum_{j=1}^{N}\Big(\alpha_i(\log t_{ij}-\beta_i+\tau_j)\Big)^2. \tag{9}$$

All three discrepancy measures are standardized residual-based measures. To assess item fit, we only took summations across persons in the above equations as discrepancy measures at the item level. A posterior predictive p (PPP, Gelman et al., 1996; Meng, 1994) value was used to summarize the information in PPMC. A PPP value near 0.5 indicates adequate model or item fit. A PPP value < 0.05 or > 0.95 indicates a model or item misfit, which is analogous to conducting a two-tailed hypothesis testing with a significance level of 0.1. The proportion of extreme PPP values among replications where the data-fitting model is the same as the data-generating model is then the empirical Type I error rate.

We considered absolute bias, empirical standard error (SE), and root mean squared error (RMSE) as outcome measures to evaluate model parameter recovery and the recovery of model-based WCPM scores, which is a function of model parameters. A detailed derivation of the model-based WCPM scores can be found in Kara et al. (2020). Specifically, the absolute bias, SE, and RMSE are calculated as:

$$\text{Absolute Bias}(y) = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{y} - y_{\text{true}}\right|, \tag{10}$$

$$\text{SE}(y) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{y} - \frac{\sum_{r=1}^{R}\hat{y}}{R})^2}, \tag{11}$$

$$\text{RMSE}(y) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{y} - y_{\text{true}})^2}, \tag{12}$$

where y is the parameter to evaluate, $\hat{y}$ is the parameter estimate, y_{true} is the true parameter value, and R is total number of replications. The bias and SE quantify the systematic and random error in the parameter estimate, respectively. The RMSE quantifies the total error in the parameter estimate.

Simulation Study Results

Convergence and Computation Time. Convergence was achieved for all model parameters in both data-fitting models with PSRF < 1.1 in all replications. In addition, we found ESS > 400 for most model parameters under all conditions with exceptions for some passage discrimination a and passage difficulty b parameters in both data-fitting models. Given that small ESS is related to high autocorrelation, this indicates that MCMC draws were less efficient for these parameters in both models. In particular, the autocorrelation issue was more severe for a and b parameters in the JRT-Bin given smaller ESS values in most cases.

Computation time for JRT-BetaBin was about 15 minutes when $N = 100$ and $I = 3$ and it was 80 minutes when $N = 500$ and $I = 3$. In addition, the computation time was proportional to dataset sizes. The computation time for JRT-BetaBin was about 3 times the computation time for JRT-Bin in all situations[1].

Model Fit. In terms of relative model fit, DIC performed perfectly in correctly identifying the true data-generating model in all replications under all conditions. That is, when $\nu = 1$, DIC values for JRT-Bin were smaller than those of the JRT-BetaBin. Otherwise, when $\nu = 3$ or $\nu = 5$ or item-specific dispersion was generated, DIC values for JRT-BetaBin were smaller than those of the JRT-Bin.

We examined both absolute model fit and item fit by summarizing PPP values based on discrepancy measures for count data and RT data, respectively. Given that data-generating models were the same as the data-fitting models for RT data, PPP values were all close to 0.5 which indicated adequate model and item fit.

We found that count data model discrepancy measures detected model misfit adequately. When the true data-generating model was JRT-Bin (i.e., $\nu = 1$), the empirical Type I error for the binomial model was 0 in all situations except being 0.033 when $N = 500$. For the beta-binomial model, proportions of extreme PPP values ranged from 0.2 to 0.4 for small to large sample sizes when $I = 6$ while no larger than 0.033 when $I = 3$. This suggests that the beta-binomial model can fit binomial data well especially when test length is short. When the true data-generating model was JRT-BetaBin (i.e., $\nu = 3$ or $\nu = 5$) or item-specific dispersion was simulated (i.e., $\nu =$ various), the proportion of extreme PPP values was 1 for JRT-Bin and 0 for JRT-BetaBin under all conditions. This indicates that the discrepancy measures can detect model misfit perfectly when the binomial model was fit to beta-binomial data.

Similar patterns were found for item fit for both models under $\nu = 1$, $\nu = 3$ and $\nu = 5$. When $\nu =$ various, however, the proportion of extreme PPP values was 0 for the beta-binomial model, while high empirical Type I error rates ranging from 0.4 to 1 for small to large sample sizes were found for the binomial model. This indicates that, when some of the passage counts had overdispersions, the passage-level discrepancy measure may mistakenly suggest that the binomial model misfit passages with no overdispersion.

[1] Computation time was obtained from analyses run on a high-performance cluster featuring the CentOS 7 operating system.

Impact of Dispersion on Parameter Recovery. We report parameter recovery of person and passage parameters for both data-fitting models. For person parameters, we focus on discussing reading accuracy θ and speed τ. Given that sample sizes did not affect much of the recovery of θ and τ, we present the absolute bias, SE, and RMSE of θ and τ under $N = 300$ in combination with test lengths and dispersion conditions in Figures 1 to 3. We found longer test lengths led to smaller estimation errors in θ and τ, which is consistent with previous studies (e.g., Qiao & Jiao, 2021). As expected, dispersion levels in the count data did not affect the recovery of τ. However, when global overdispersion existed (i.e., $\nu > 1$) or item-specific overdispersion existed (i.e., $\nu =$ various), JRT-BetaBin tended to yield smaller absolute bias and smaller SE in θ than JRT-Bin. As a result, JRT-Bin yielded a larger total error in θ given larger RMSE in general. When the true data-generating model was JRT-Bin (i.e., $\nu = 1$), recovery of θ was similar for both data-fitting models.

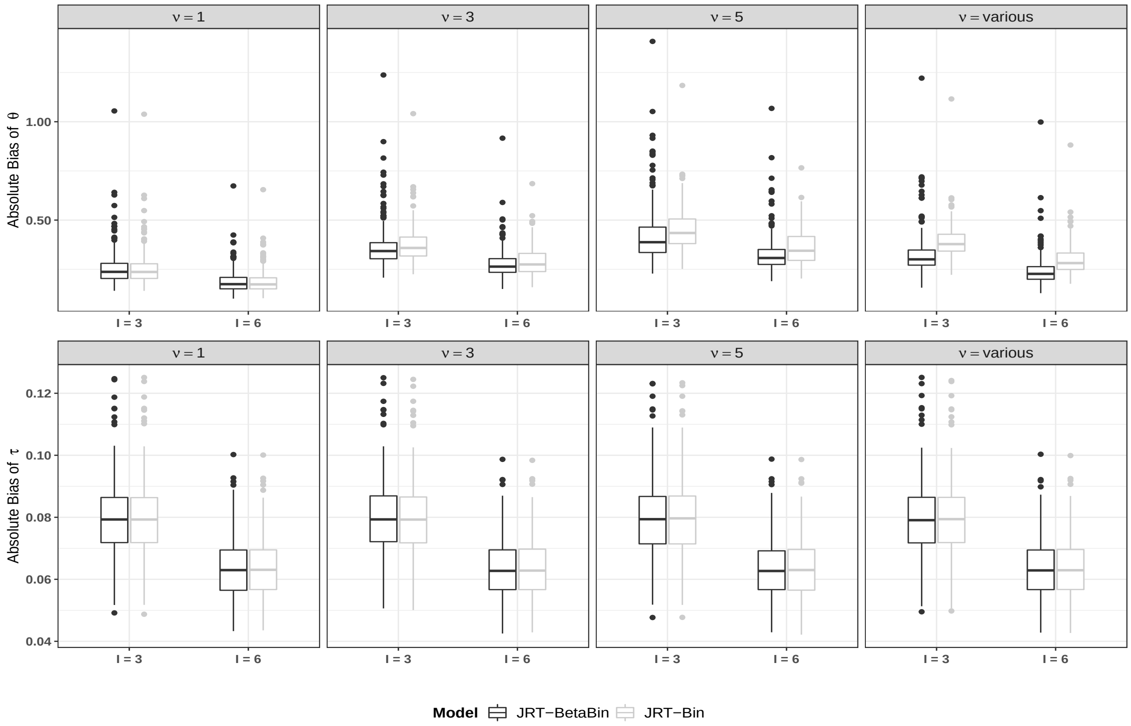

Figure 1. Absolute Bias of θ and τ when $N = 300$.

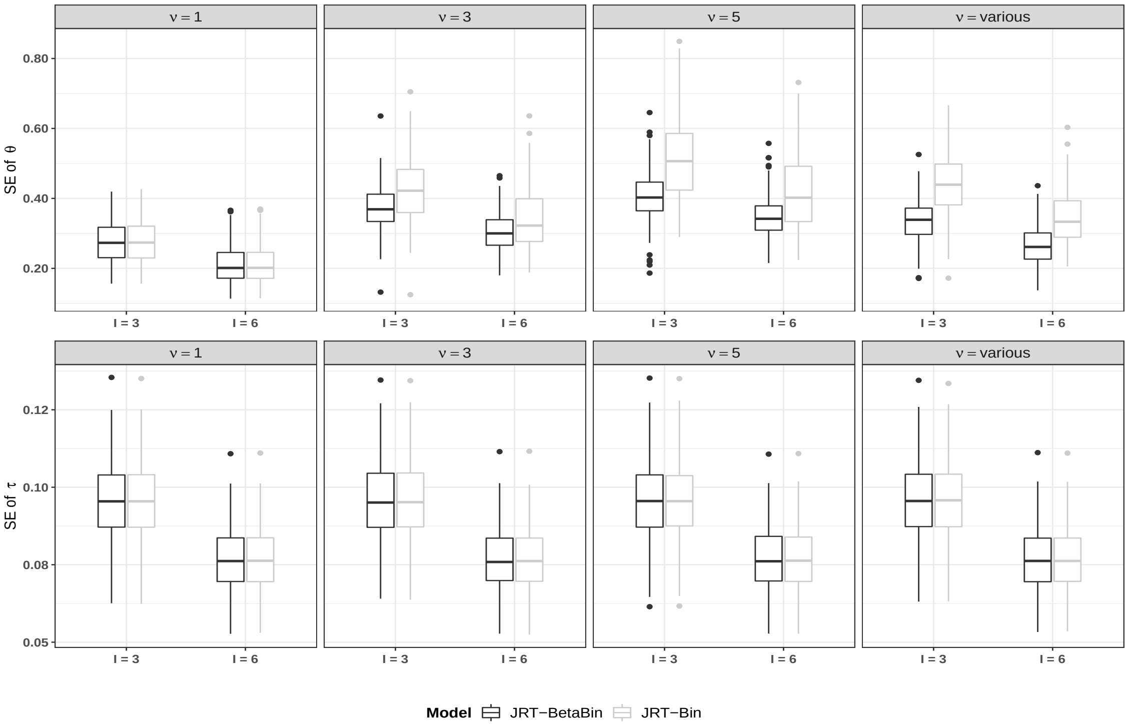

Figure 2. Standard Error of θ and τ when $N = 300$.

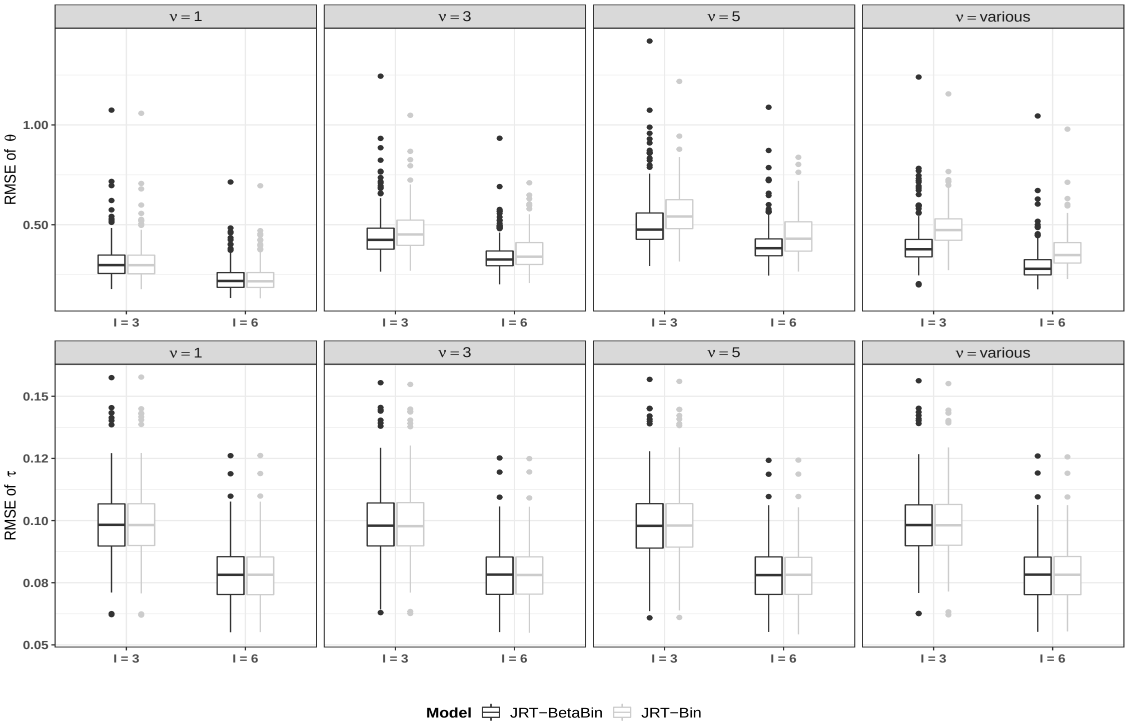

Figure 3. Root Mean Square Error of θ and τ when $N = 300$.

Passage parameters include passage discrimination a, passage difficulty b, time discrimination α, and time intensity β in both models and dispersion parameter ν in JRT-BetaBin. Given that test lengths did not affect much of the recovery of passage parameters, we present the absolute bias, SE, and RMSE of passage parameters under $I = 6$ in combination with sample sizes and dispersion conditions in Figures 4 to 6. We found that parameter recovery for α and β were similarly adequate for both data-fitting models in all conditions. Thus, we focus on discussing the impact of overdispersion levels and sample sizes on the recovery of a, b, and ν. These parameters in both data-fitting models recovered well when JRT-Bin was the true data-generating model (i.e., $\nu = 1$). However, parameters a and b in JRT-Bin had larger absolute bias and RMSE on average as the overdispersion level in count data increased and smaller SE on average as sample size increased. While recovery of a and b in JRT-BetaBin was not affected by overdispersion levels, ν had larger estimation errors as the magnitude of ν increased. This is reasonable because the magnitude of true parameters is usually proportional to estimation errors. As expected, recovery of passage parameters improved as sample size increased on average based on bias, SE, and RMSE.

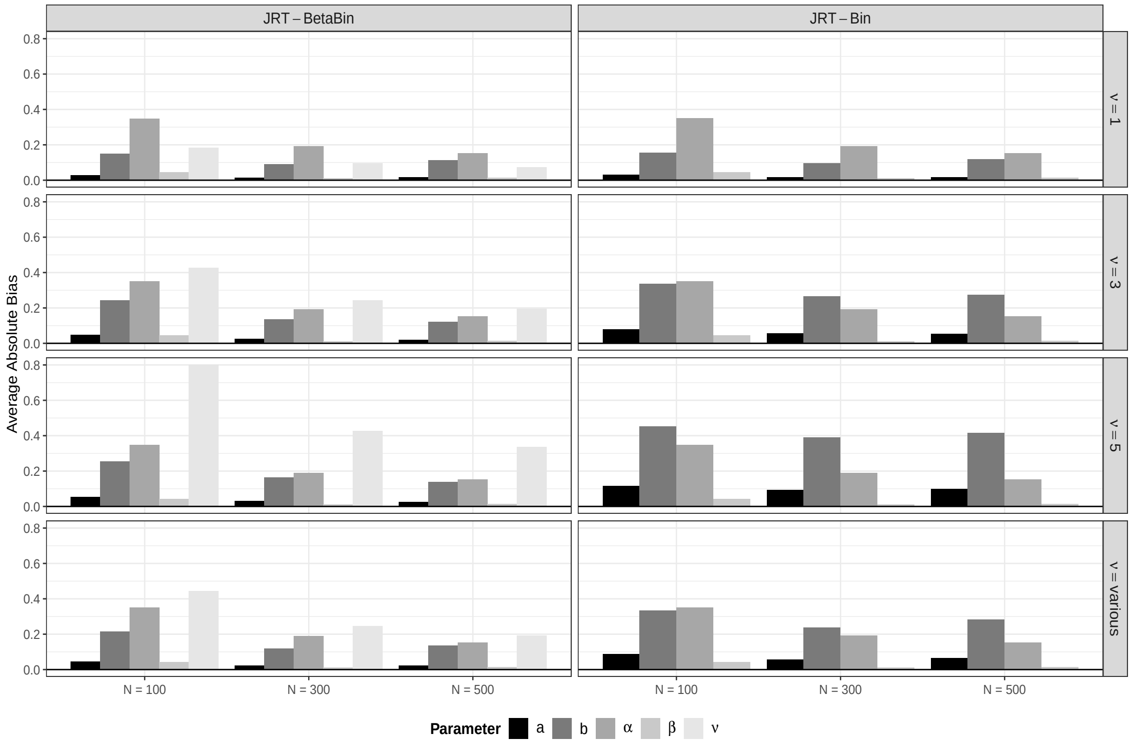

Figure 4. Average Absolute Bias of Passage Parameters when $I = 6$.

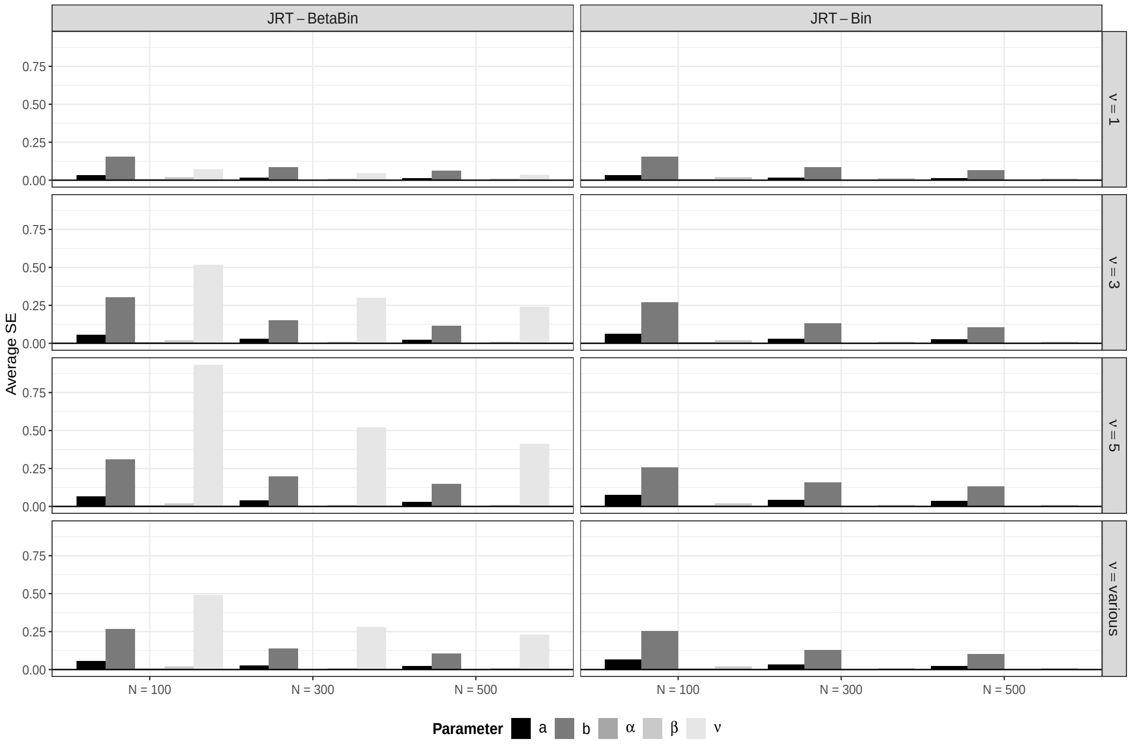

Figure 5. Average Standard Error of Passage Parameters when $I = 6$.

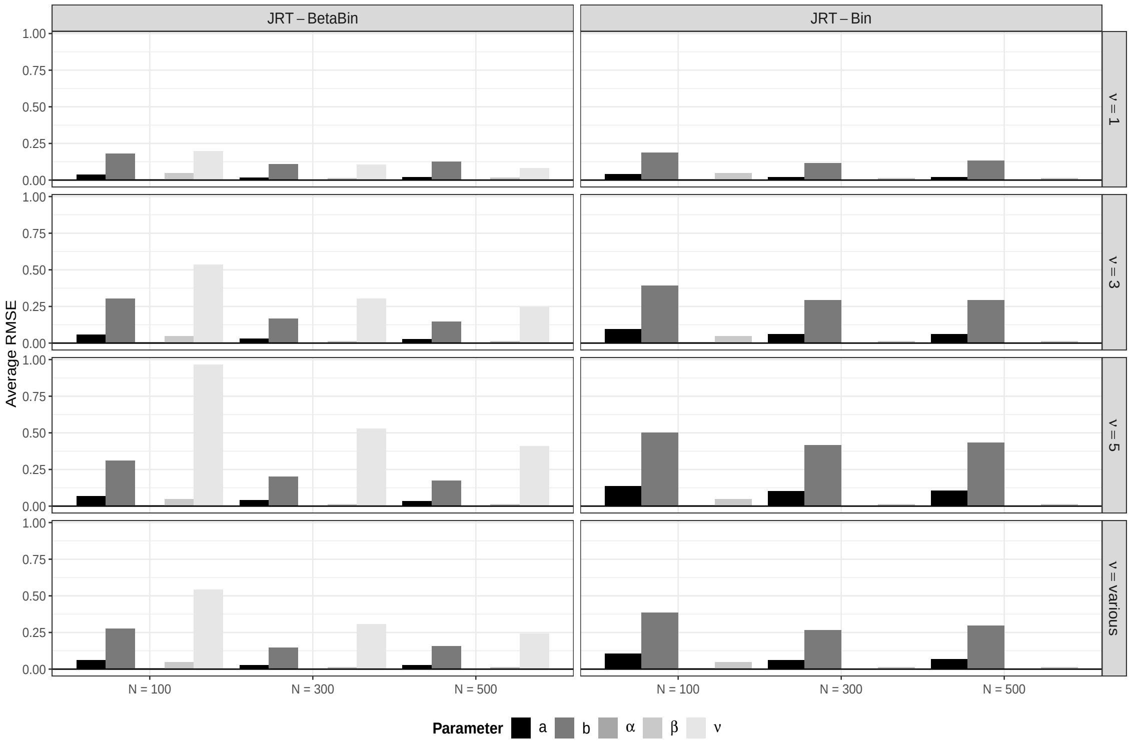

Figure 6. Average Root Mean Square Error of Passage Parameters when $I = 6$.

Impact of Dispersion on WCPM Estimates. We examined the recovery of model-based WCPM scores in terms of absolute bias, SE, and RMSE. We observed similar patterns of WCPM score recovery under different sample sizes. Therefore, we present only conditions of $N = 300$ in combinations with test lengths and overdispersion levels in Figures 7 to 9. In general, longer tests (i.e., more passages) resulted in better recovery of WCPM scores for both data-fitting models. In addition, when the true data-generating model was JRT-Bin (i.e., $\nu = 1$), we observed comparable recovery levels of WCPM scores for JRT-Bin and JRT-BetaBin. When $\nu > 1$, JRT-BetaBin yielded smaller absolute bias and RMSE values for WCPM score estimates on average especially for larger true WCPM scores. More specifically, when true WCPM scores were small (< 50), JRT-BetaBin exhibited either similar or smaller median absolute bias and RMSE values but with larger variability comparing to those obtained from JRT-Bin. As true WCPM scores became larger, absolute bias and RMSE values obtained from JRT-BetaBin became smaller than those from JRT-Bin in terms of both medians and ranges. JRT-Bin, on the other hand, tended to have larger SEs in WCPM score estimates, especially when overdispersion levels increased in the count data.

4. Empirical Data Analysis

Dataset and Analysis

In the empirical data analysis, we aim to evaluate the model fit and consistency of parameter estimates from the JRT-BetaBin and JRT-Bin using the same dataset as in Kara et al. (2020). The dataset consists of 4 passages with 47, 47, 80, and 86 words read by $N = 58$ students. Counts of WRC and response time (in seconds) per passage were recorded for each person. We further took the natural logarithm of the time variables for the analysis. Descriptive statistics of the count variables and time variables are shown in Table 2.

Two models, JRT-BetaBin and JRT-Bin, were fit to the dataset. The prior specifications and Bayesian MCMC estimation setup for both models remained the same as those in the simulation study. Below, we report model convergence,

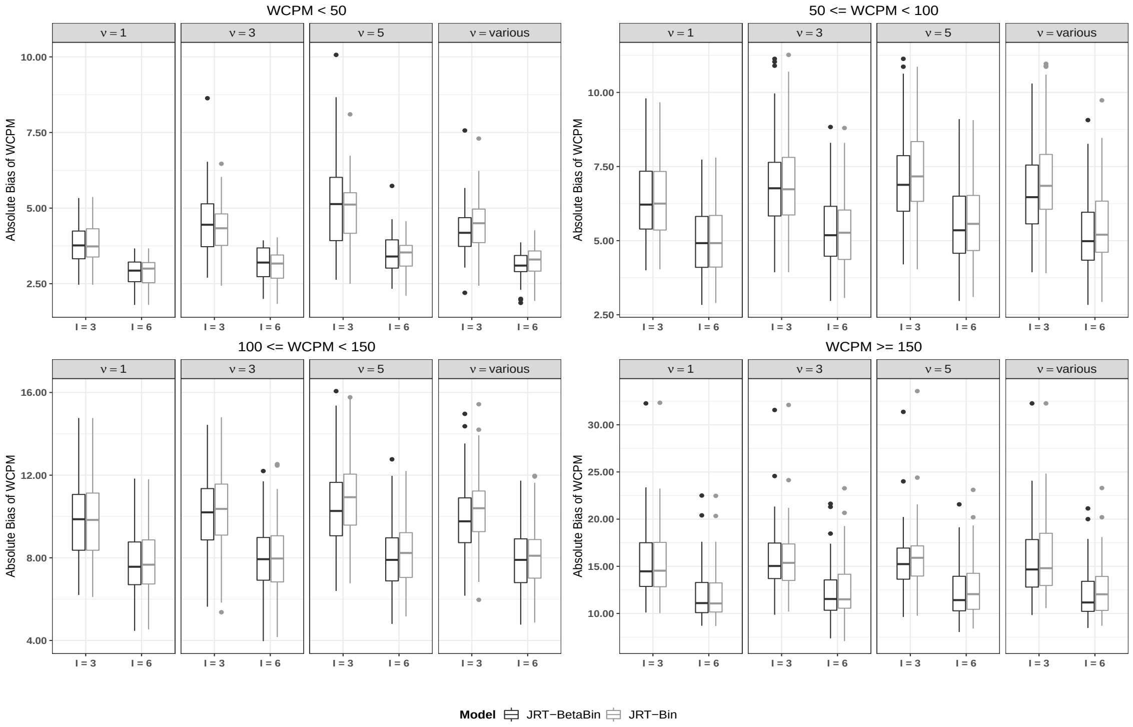

Figure 7. Absolute Bias of Model-based WCPM Estimates when N = 300.

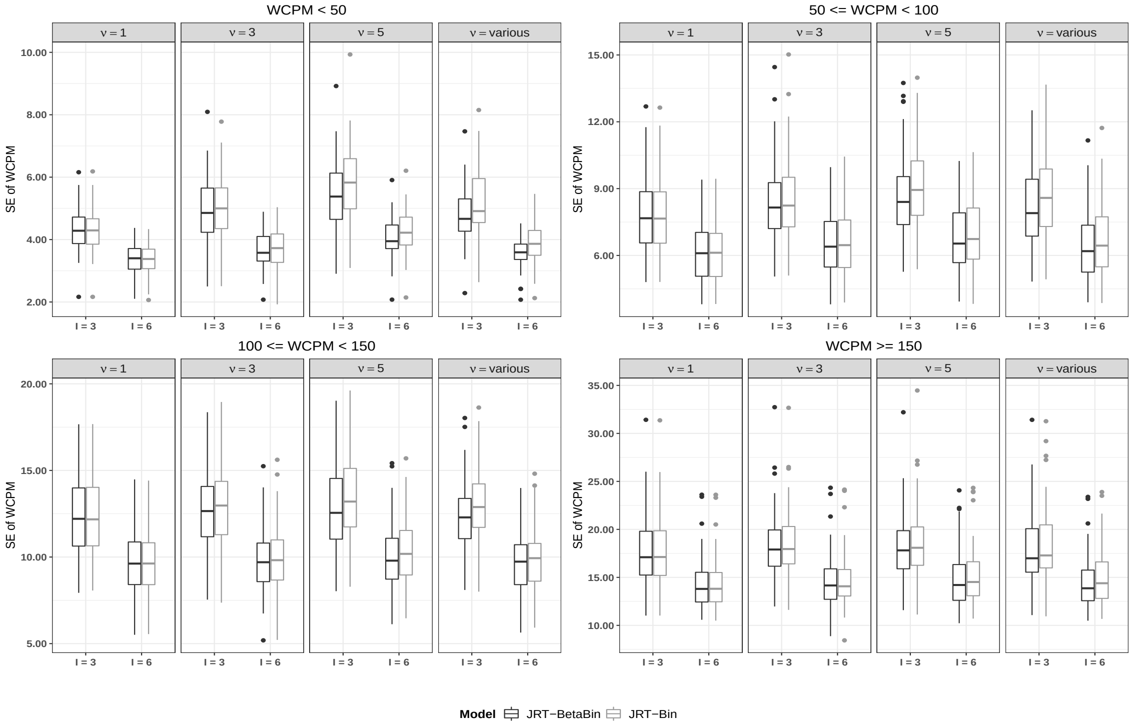

Figure 8. Standard Error of Model-based WCPM Estimates when N = 300.

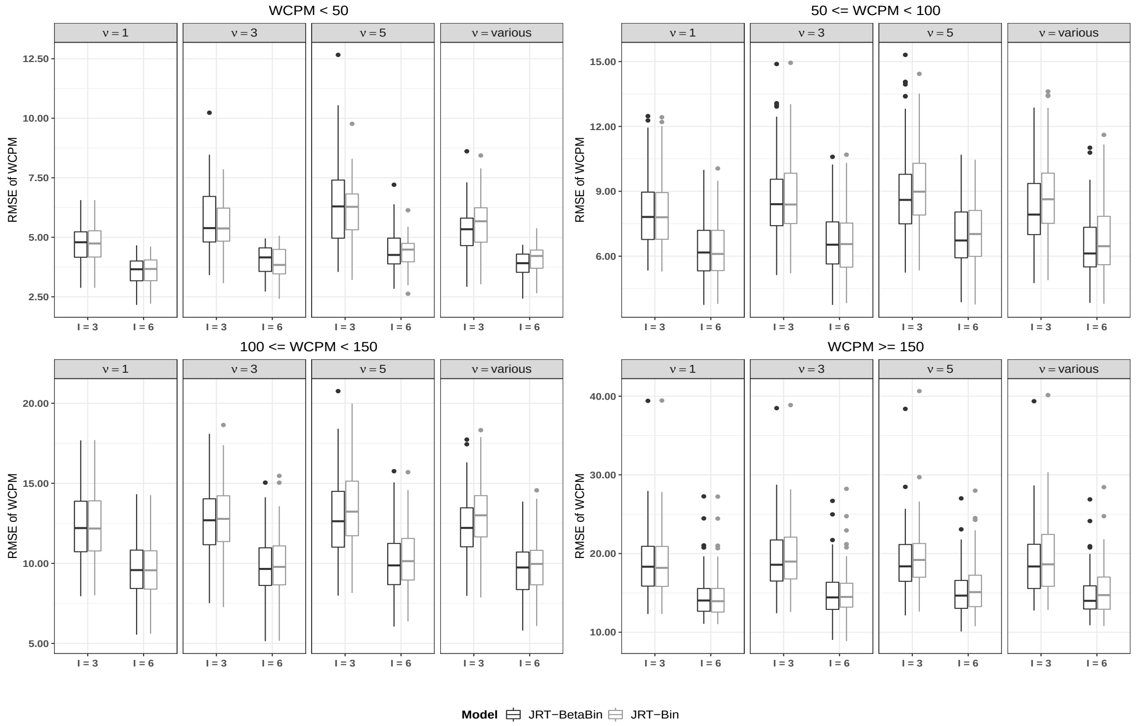

Figure 9. Root Mean Square Error of Model-based WCPM Estimates when $N = 300$.

computation time, model fit, and consistency of parameter estimates from two models.

Table 2

Descriptive Statistics of the Empirical Dataset

Variable	Mean	SD	Min	Max	% Missing
U1	43.552	3.803	31	47	0
U2	42.966	3.464	31	47	0
U3	71.569	8.249	43	80	0
U4	79.500	7.294	54	86	0
T1	3.227	0.336	2.670	3.935	0
T2	3.172	0.296	2.570	3.999	0
T3	3.833	0.348	3.250	4.501	0
T4	3.678	0.318	3.166	4.501	0

Note. U = Count; T = log-time; SD = standard deviation.

Empirical Data Analysis Results

Convergence and Computation Time. Both JRT-BetaBin and JRT-Bin converged well with PSRF < 1.05 for all estimated parameters. Consistent with the simulation study results, most parameter estimates had adequate estimation precision with ESS > 400 except for several passage discrimination a and passage difficulty b parameters. Specifically, $\hat{b}_1$ had ESS $= 300$ in JRT-BetaBin and $\hat{a}_1$, $\hat{a}_2$, $\hat{b}_1$, and $\hat{b}_2$ had ESS as 280, 330, 310, and 310 in JRT-Bin. Computation time for JRT-BetaBin and JRT-Bin was 11.117 and 3.030 minutes, respectively.

Model Fit. Both absolute and relative model fit indices are presented in Table 3. As per the RT model, PPP values were close to 0.5, indicating adequate model fit. The PPP value for the beta-binomial model was close to 0.5, while the PPP value for the binomial model was 0. Passage-level PPP values showed similar patterns, as shown in Table 4. In addition, both deviance and DIC values for the JRT-BetaBin were smaller than those of the JRT-Bin, which indicated that JRT-BetaBin had better fit to the empirical data. Based on the simulation study results, the model fit evidence in the empirical analysis suggested either item-specific ($\nu =$ various) or global overdispersion ($\nu > 1$) in count data.

Table 3

Model Fit Results in Empirical Data Analysis

Model	PPP of Count Model	PPP of RT Model	Deviance	DIC
JRT-BetaBin	0.397	0.475	**910.065**	**1126.724**
JRT-Bin	0	0.474	970.535	1134.985

Note. Smaller values of deviance and DIC are bold face; RT = response time; PPP = posterior predictive *p*-value; DIC = deviance information criterion.

Parameter Estimates. The overdispersion parameter estimates from the JRT-BetaBin model for the four passages were $\hat{\nu}_1 = 1.993$, $\hat{\nu}_2 = 2.012$, $\hat{\nu}_3 = 5.055$, and $\hat{\nu}_4 = 3.775$, respectively. These values can be interpreted as the number of times the conditional variances of the passage-level count data suggested by the JRT-BetaBin

Table 4

Passage Fit Results in Empirical Data Analysis

Model	Passage	PPP of Count Model	PPP of RT Model
JRT-BetaBin	1	0.537	0.491
	2	0.434	0.496
	3	0.439	0.479
	4	0.354	0.476
JRT-Bin	1	0.006	0.485
	2	0.002	0.495
	3	0.000	0.472
	4	0.000	0.482

Note. RT = response time; PPP = posterior predictive p-value.

were those suggested by the JRT-Bin. As a result, JRT-BetaBin suggested overdispersion of the count data for all passages compared to JRT-Bin given all overdispersion parameter estimates larger than 1.

Consistency of passage discrimination $\hat{a}$, difficulty $\hat{b}$, time discrimination $\hat{\alpha}$, and time intensity parameters $\hat{\beta}$ is presented in Figure 10. We observed discrepancies in $\hat{a}$ and $\hat{b}$ for all passages. In addition, we found almost identical $\hat{\alpha}$ and $\hat{\beta}$ obtained from the two data-fitting models. These findings are consistent with the simulation study results where we found JRT-Bin had worse parameter recovery for a and b under overdispersion while comparable parameter recovery for α and β compared to JRT-BetaBin under all simulated conditions.

Lastly, the correlations between $\hat{\theta}$, $\hat{\tau}$ and WCPM score estimates between JRT-BetaBin and JRT-Bin were 0.968, 1, and 0.997, respectively. This finding is close to the $\nu =$ various condition in the simulation study where estimation errors of θ and WCPM scores were slightly larger for JRT-Bin than JRT-BetaBin. Given that recovery of τ was identical between the two models in all conditions, a perfect correlation in the empirical analysis is expected.

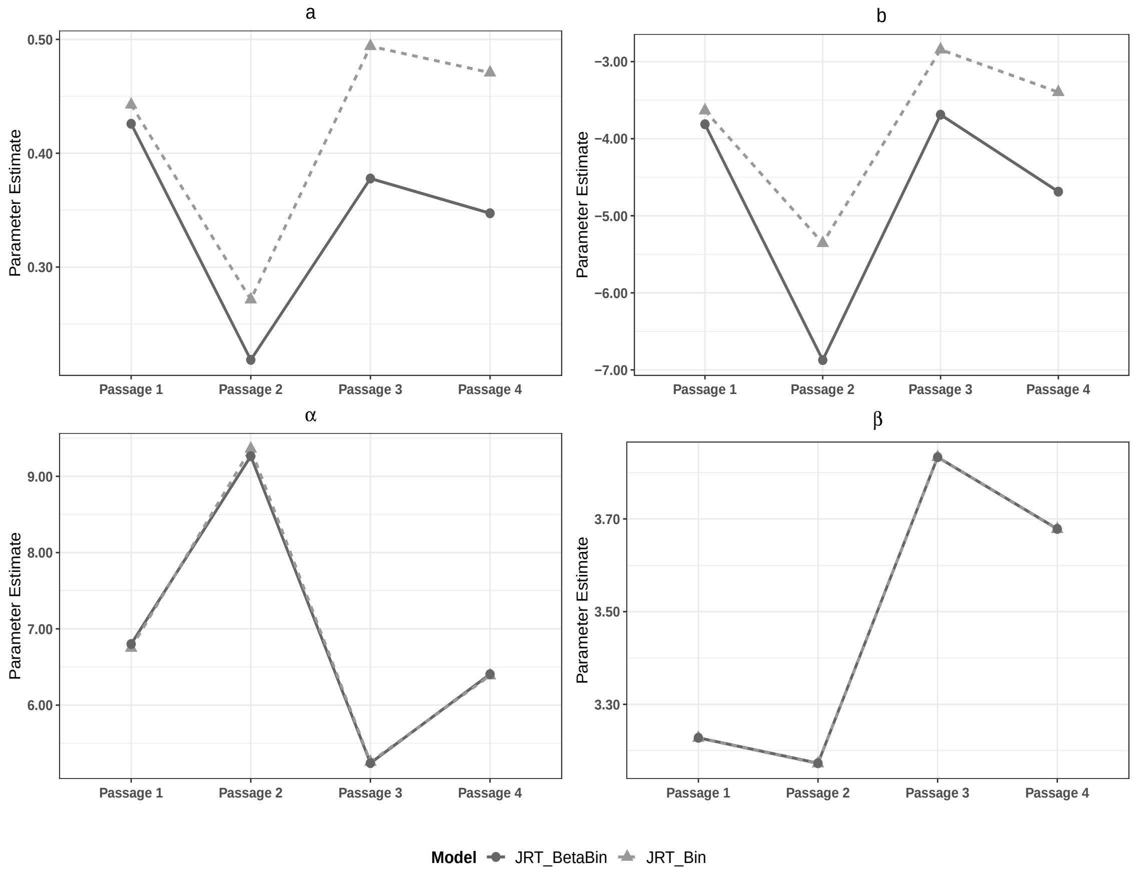

Figure 10. Consistency of Passage Parameter Estimates in Empirical Data Analysis.

5. Discussion

Binomial family models can be good candidates for modeling count-type item response data with an upper bound where each attempt is assumed to be independent with large success probabilities. Observed count data that have such characteristics are commonly seen in educational and psychological assessments such as ORF and word recognition assessments. In this article, we proposed the beta-binomial model that accommodates potential overdispersion in count data, which is ignored in the binomial model. We illustrated the application of the proposed model in the context of ORF assessments. We presented the beta-binomial model in the form of a joint model of count-type item response data and response times. However, it is important to emphasize that the beta-binomial model can be used as a standalone model for count data. Moreover, the proposed model can be used to analyze count-type item response data from various assessment scenarios rather than limiting to ORF assessments as long as the model assumptions can be met.

The simulation study showed adequate parameter recovery of JRT-BetaBin when the count data was generated from the binomial model or the beta-binomial model. Further, we provided evidence that the relative model fit index DIC performed perfectly in selecting true data-generating models under various simulation conditions. In addition, we employed the PPMC method to evaluate the absolute model and item fit. Through a sum of squared residuals perspective, count data discrepancy measures can adequately detect model misfits in terms of ignoring overdispersion. However, high Type I error rates were observed for the binomial model when some passages had overdispersed count data.

The empirical data analysis showed the utility of JRT-BetaBin in accommodating overdispersion in count-type item response data in comparison to JRT-Bin. Specifically, we observed better model fit and item fit for the JRT-BetaBin and overdispersion in the count data with $\hat{\nu} > 1$ as suggested by JRT-BetaBin. The passage parameter estimates $\hat{a}$ and $\hat{b}$ were discrepant between JRT-Bin and JRT-BetaBin given the potential overdispersion in the count data, while person parameter estimates $\hat{\theta}$, $\hat{\tau}$ and WCPM scores were consistent between JRT-BetaBin and JRT-Bin in general.

Based on the promising application of the beta-binomial model for modeling count data as shown in the current study, some extensions can be made in the future. First, the assumption that constant latent reading accuracy and reading speed across passages need to be evaluated. It has been shown that latent speed can be variable across items (Fox & Marianti, 2016). Similarly, it is reasonable to relax the constant reading speed assumption in the beta-binomial model especially when item characteristics vary. Second, we investigated discrepancy measures based on the sum of squared residuals in the PPMC procedure. It is worth noting that PPMC examines specific aspects of model-data fit depending on discrepancy measures being used. Therefore, more studies on Bayesian model fit evaluation on binomial family models are needed in the future. Lastly, we demonstrated the application of the beta-binomial model in the context of

ORF assessments. Future studies that extend its application to various educational or psychological assessments are needed to further explore the utilities of the beta-binomial model in modeling count-type item response data.

References

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434–455.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate behavioral research*, *51*(4), 540–553.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Hung, L.-F. (2012). A negative binomial regression model for accuracy tests. *Applied Psychological Measurement*, *36*(2), 88–103.

Jansen, M. G. (1997). Rasch's model for reading speed with manifest explanatory variables. *Psychometrika*, *62*(3), 393–409.

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. (2020). Estimating model-based oral reading fluency: A bayesian approach. *Educational and Psychological Measurement*, *80*(5), 847–869.

Liu, R., Liu, H., Shi, D., & Jiang, Z. (2022). Poisson diagnostic classification models: A framework and an exploratory example. *Educational and Psychological Measurement*, *82*(3), 506–516.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*(4), 529–544.

Meng, X.-L. (1994). Posterior predictive p-values. *The annals of statistics*, *22*(3), 1142–1160.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219.

Plummer, M. (2015). *Jags version 4.0. 0 user manual. lyon, france.*

Potgieter, C. J., Kamata, A., & Kara, Y. (2017). An em algorithm for estimating an oral reading speed and accuracy model. *arXiv preprint arXiv:1705.10446*.

Qiao, X., & Jiao, H. (2021). Explanatory cognitive diagnostic modeling incorporating response times. *Journal of Educational Measurement*, *58*(4), 564–585.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Rasch, G. (1972). Objektivitet i samfundsvidenskaberne et metodeproblem. *Nationaløkonomisk tidsskrift*, *110*, 161–196.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583–639.

Spray, J. A. (1990). One-parameter item response theory models for psychomotor tests involving repeated, independent attempts. *Research Quarterly for Exercise and Sport*, *61*(2), 162–168.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139.

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661.