

Article

A Comparative Study of Item Response Theory Models for Mixed Discrete-Continuous Responses

Cengiz Zopluoglu ^{1,*}  and J. R. Lockwood ² ¹ College of Education, University of Oregon, Eugene, OR 97403, USA² Duolingo, Inc., Pittsburgh, PA 15206, USA; jr@duolingo.com

* Correspondence: cengiz@uoregon.edu

Abstract: Language proficiency assessments are pivotal in educational and professional decision-making. With the integration of AI-driven technologies, these assessments can more frequently use item types, such as dictation tasks, producing response features with a mixture of discrete and continuous distributions. This study evaluates novel measurement models tailored to these unique response features. Specifically, we evaluated the performance of the zero-and-one-inflated extensions of the Beta, Simplex, and Samejima's Continuous item response models and incorporated collateral information into the estimation using latent regression. Our findings highlight that while all models provided highly correlated results regarding item and person parameters, the Beta item response model showcased superior out-of-sample predictive accuracy. However, a significant challenge was the absence of established benchmarks for evaluating model and item fit for these novel item response models. There is a need for further research to establish benchmarks for evaluating the fit of these innovative models to ensure their reliability and validity in real-world applications.

Keywords: item response theory; bounded continuous data; continuous response model; dictation task; language assessment; natural language processing; zero-and-one inflated data



Citation: Zopluoglu, Cengiz, and J. R. Lockwood. 2024. A Comparative Study of Item Response Theory Models for Mixed Discrete-Continuous Responses. *Journal of Intelligence* 12: 26. <https://doi.org/10.3390/jintelligence12030026>

Received: 30 September 2023

Revised: 2 February 2024

Accepted: 19 February 2024

Published: 25 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In educational and psychological assessments, continuous response features frequently arise either as an inherent characteristic of an assessment, such as in reading fluency measures (e.g., DIBELS, University of Oregon 2018–2020) or continuous rating scales (Bejar 1977; Brumfitt and Sheeran 1999) or as a consequence of summed dichotomous and polytomous item scores, such as in C-tests (Raatz and Klein-Braley 1981) or Maze assessments (Guthrie et al. 1974). In addition, the rapid evolution of machine learning and AI-driven technologies (Bommasani et al. 2021) offers new opportunities for more frequent and innovative use of constructed-response items from which continuous features can be derived. These features can be used as indicators of person-level proficiency traits and can yield inferences about these traits when combined with appropriate theoretical and measurement models.

In practice, it is common for “continuous” features to have observed distributions that are best described by a mixture of discrete and continuous distributions. For instance, when a continuous feature has a bounded distribution (e.g., between 0 and 1), it may have a nontrivial probability of being 0 and/or 1 when computed from a sample of responses from some target population. An example that we consider in this study arises from the use of edit distance (Levenshtein 1965) to evaluate the accuracy of a response to a dictation task, in which a test taker is asked to listen to a target sentence in English and then type the sentence in English (Buck 2001; Cardwell et al. 2023). If the typed response matches the target sentence, one could assign a grade of 1. Lower grades can be assigned as the typed response deviates more from the target sentence with respect to some edit distance. A grade of 0 can be assigned to responses for which the edit distance exceeds the length of the

target sentence. Depending on the complexity of the target sentence and the distribution of English proficiency in the target population of test takers, the distribution of this grade may be best described by a mixture of discrete probability masses at 0 and 1 and a continuous distribution on the open unit interval (0, 1).

Building on the diverse applications of continuous response features in educational and psychological assessments, the psychometric literature has witnessed a significant evolution of Item Response Theory (IRT) models proposed for continuous response features. The pioneering works by Samejima (1973), who presented a Continuous Response Model as a special case of the Graded Response Model, and Müller's extension (Müller 1987) of the Rating Scale Model, along with a simple linear response model by Mellenbergh (1994), laid the groundwork for subsequent models that catered to more complex data structures. These studies emphasized the need for models that could handle ceiling and floor effects inherent in continuous data. The exploration of bounded continuous response features, such as those ranging from 0 to 1, gained momentum with the contributions of Noel and Dauvier (2007), who presented a Beta Item Response Model. This model addresses the complexities and nuances of response distributions within bounded continuous data. This model incorporates an interpolation process as the response mechanism, leading to a beta distribution for responses. Therefore, unlike early IRT models, it effectively captures the asymmetric nature of such data, thereby enriching the theoretical and practical applications of IRT in bounded continuous data scenarios. Recent contributions have expanded the scope of continuous response modeling in IRT. Flores et al. (2020) introduced a model leveraging the simplex distribution to model response times within a bounded framework. This model significantly advances handling response features that exhibit continuous and discrete characteristics, particularly in time-limited assessment scenarios.

While Item Response Theory (IRT) models for bounded continuous features are well-studied, Molenaar et al. (2022) highlighted a significant limitation in these models, mainly when dealing with data concentrated at the boundaries. For instance, significant clustering at these limits may occur in assessments where responses are naturally bounded between 0 and 1. This clustering poses challenges for traditional IRT models, which typically do not account for such high concentrations at the boundary values. Molenaar et al. (2022) demonstrated that even a small proportion of responses at these boundaries could significantly impact parameter estimates, leading to potential biases in the interpretation of data. They proposed a set of new IRT models that better accommodate these boundary concentrations, providing more accurate and reliable analysis in such scenarios. These models are suitable for features with distributions having a mixture of discrete point masses and continuous values on a bounded interval. As these and other novel IRT models emerge, empirical evaluations of their performance with real data from diverse contexts are essential to the research and practitioner communities to improve measurement science.

In this study, we evaluate the performance of three novel IRT models, zero-and-one-inflated extensions of the Beta, Simplex, and Samejima's Continuous IRT models proposed by Molenaar et al. (2022) by (a) using a dataset characterized by its extreme sparsity, (b) supplementing these models with a latent regression component, and (c) conducting model comparisons using cross-validation and posterior predictive checks. The dataset used in this study comes from a high-stakes English language proficiency assessment. It exhibits a substantial proportion of "1" grades and a smaller occurrence of "0" grades, providing a challenging yet ideal scenario for examining models designed to handle a mixture of discrete and continuous data. To address the sparse nature of our dataset, we incorporate latent regression, using standardized writing and speaking scores as auxiliary variables within the IRT framework to improve the accuracy and precision of parameter estimation. This integration is directly in line with established research suggesting that including auxiliary information leads to more accurate estimates. Moreover, we employ cross-validation and posterior predictive checks as a means of model evaluation.

2. Materials and Methods

2.1. Data

We analyzed responses from a sample of the Duolingo English Test test takers, a computerized-adaptive assessment of English language proficiency designed to support high-stakes decisions in English-medium settings (Cardwell et al. 2023). The sample consists of 295,157 test sessions from 222,568 unique test takers (some repeat the assessment). All sessions in the sample occurred before September 2022. Test takers in the sample represent more than 200 unique countries (India = 22%, China = 18%, all others < 5%) and 146 native languages (Mandarin = 19%, Spanish = 9%, English = 9%, Telugu = 8%, all others < 5%). Approximately 47% of test takers are female, 39% percent report the intention to apply to undergraduate programs, and another 39% percent report the intention to apply to graduate programs.

As previously described, we focus on the test takers' responses to dictation items. This task aims to evaluate the test taker's capacity to identify individual words and retain them in memory for a sufficient amount of time to reproduce them with precision. Each test taker in the dataset has responses to between 4 and 7 dictation items, with the majority (91.6%) responding to 6 items. The first dictation item is randomly assigned for each test taker, whereas the remaining items are assigned adaptively based on the test taker's performance on previous items. As a consequence of adaptive assignment, no two test takers in the sample have responses to the same set of dictation items, making the person-item linkage structure highly sparse. The dataset consists of 1,789,297 responses to 2738 dictation items.

The response of a given test taker to a given item is graded on the interval $[0, 1]$ using a function of the character-based edit distance between the target sentence and what is typed by the test taker. The grades are defined so that a value of 1 indicates an essentially perfect rendering of the target sentence, whereas lower grades correspond to increasing discrepancies. A minimum grade of 0 occurs when the edit distance equals or exceeds the length of the target sentence. Approximately 47% of the item responses receive a grade of 1. Grades of 0 occur but are rare (0.03%). The remaining 53% of grades have a mean of 0.87 and a standard deviation of 0.125. The dataset also includes each test taker's score on the speaking and writing portions of the assessment, which we use in some models as predictors of dictation proficiency via latent regression. Details on the tasks contributing to these speaking and writing scores and other technical details about the assessment can be found in Cardwell et al. (2023).

2.2. Zero-and-One-Inflated Item Response Models for Bounded Continuous Data

Molenaar et al. (2022) discuss some novel IRT models designed for bounded continuous data, addressing the limitations of traditional models when applied to data constrained within the closed interval $[0, 1]$. These novel models are particularly adept at handling situations where responses are clustered at the boundaries of the scale, a common occurrence in educational and psychological assessments. The Beta IRT model assumes a beta distribution for the response propensity, characterized by its flexibility in modeling various shapes of response distributions. It is suitable for data with natural boundaries, like percentages or proportions, common in psychological scales. Based on the S_B distribution, the Continuous Response Model is a special case within Samejima's Graded Response Model framework. The Simplex IRT model utilizes the simplex distribution, which, while less common, offers an alternative modeling approach to bounded continuous data. This model is beneficial in contexts such as response time analysis, where the data is naturally bounded within a specific range. In this section, we provide a general introduction to the overall model structure for all these models. All three models operate under the same structure, and the only difference is the model-specific density function utilized when modeling the continuous part of the distribution.

Let X_{pi} denote the continuous bounded item scores such that $X_{pi} \in [0, 1]$ for the p th person, $p = \{1, \dots, P\}$, on the i th item, $i = \{1, \dots, I\}$. We can define a discrete variable Z_{pi} , representing three possible conditions as the following,

$$Z_{pi} = \begin{cases} 0, & \text{if } X_{pi} = 0 \\ 1, & \text{if } 0 < X_{pi} < 1 \\ 2, & \text{if } X_{pi} = 1 \end{cases}$$

A logistic Graded Response Model (Samejima 1969) can be written for modeling Z_{pi} such that,

$$P(Z_{pi} = 0 | \theta_p, \alpha_i, \gamma_{0i}) = \frac{1}{1 + e^{\alpha_i \theta_p - \gamma_{0i}}}$$

$$P(Z_{pi} = 1 | \theta_p, \alpha_i, \gamma_{0i}, \gamma_{1i}) = \frac{1}{1 + e^{\alpha_i \theta_p - \gamma_{1i}}} - \frac{1}{1 + e^{\alpha_i \theta_p - \gamma_{0i}}}$$

$$P(Z_{pi} = 2 | \theta_p, \alpha_i, \gamma_{1i}) = \frac{e^{\alpha_i \theta_p - \gamma_{1i}}}{1 + e^{\alpha_i \theta_p - \gamma_{1i}}}$$

where $\theta_p \in R$ is a latent person parameter, $\alpha_i \in R^+$ is an item discrimination parameter, and $\gamma_{0i} \in R$ and $\gamma_{1i} \in R$ are category threshold parameters satisfying $\gamma_{0i} < \gamma_{1i}$.

Then, the joint conditional density for the model, which is denoted by $k(\cdot)$, can be written as the following:

$$k(X_{pi} | \theta_p, \alpha_i, \gamma_{0i}) = P(Z_{pi} = 0 | \theta_p, \alpha_i, \gamma_{0i})$$

$$k(X_{pi} | \theta_p, \alpha_i, \gamma_{0i}, \gamma_{1i}, \beta_i, \delta_i) = P(Z_{pi} = 1 | \theta_p, \alpha_i, \gamma_{0i}, \gamma_{1i}) \times f(X_{pi} | \theta_p, \alpha_i, \beta_i, \delta_i)$$

$$k(X_{pi} | \theta_p, \alpha_i, \gamma_{1i}) = P(Z_{pi} = 2 | \theta_p, \alpha_i, \gamma_{1i})$$

where $f(\theta_p, \alpha_i, \beta_i, \delta_i)$ corresponds to the model-specific density function with support on the open interval $(0, 1)$, $\beta_i \in R$ is an item location parameter, and $\delta_i \in R^+$ is an item dispersion parameter. So, in total, each model estimates five parameters per item. Note that the probability distribution of X_{pi} is a mixture of a discrete distribution on $\{0, 1\}$ and a continuous distribution on the open interval $(0, 1)$. All three models are structurally the same except for the fact that, the probability density function, $f(\theta_p, \alpha_i, \beta_i, \delta_i)$, is replaced with the corresponding model-specific function in the above equations. The specific density functions for the Beta, Simplex, and Samejima’s Continuous IRT models can be found in Appendix A.

Figure 1 also visually compares the model-generated response distributions for a population with latent proficiency distributed as a standard normal distribution. This figure includes the three models (Beta, Simplex, and SB IRT models) for hypothetical items with identical item parameters ($\gamma_0, \gamma_1, \alpha$, and β), but they differ in their dispersion parameters (δ) since dispersion parameters have different scales across the models. Despite the distinct mathematical formulations behind the density functions, the figure reveals significant similarity in the generated response distributions. This resemblance underscores the models’ robustness and adaptability to different types of distributions, particularly those that tend to cluster at the scale boundaries. For a comprehensive understanding of these models and their specific density functions, readers are encouraged to consult Molenaar et al. (2022) paper, where technical discussions are thoroughly presented.

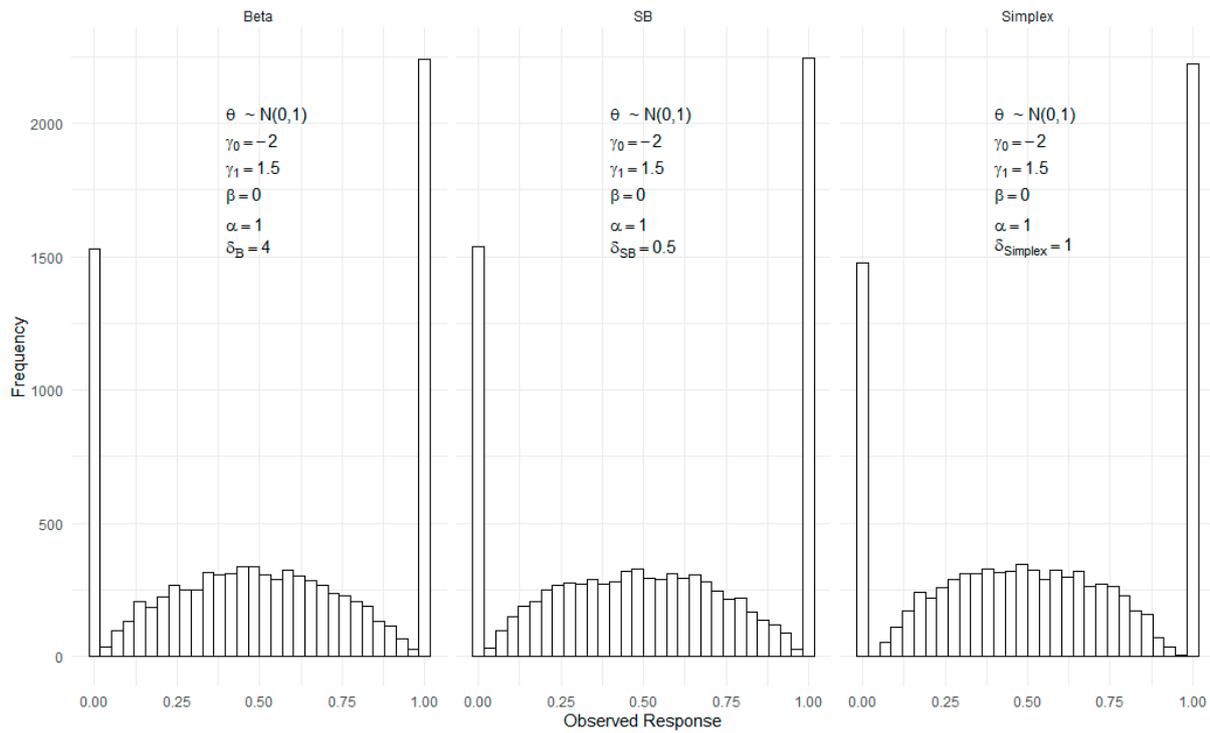


Figure 1. Comparison of model-generated response distributions for the Beta, SB, and Simplex IRT models. Latent proficiency is assumed to follow a standard normal distribution. All item parameters except the dispersion parameter were the same across models.

2.3. Incorporating Collateral Information

Incorporating supplementary data about persons and/or items into IRT models offers numerous benefits. Prior studies have reported that including auxiliary information within the IRT framework can enhance both convergence and the precision of parameter estimation (Adams et al. 1997; de la Torre 2003; Hall 2007; Joo et al. 2022; Mislevy and Sheehan 1989; Tao 2009). Therefore, we also consider extending the models proposed by Molenaar et al. (2022) by incorporating information from two auxiliary variables related to person proficiency: writing and speaking scores. The models above can be extended by proposing a linear regression model of θ_p on the auxiliary variables,

$$\theta_p = \zeta_1 W_p + \zeta_2 S_p + \epsilon_p$$

where W_p and S_p are the observed writing and speaking scores for the p th examinee, ζ_1 and ζ_2 are the associated regression coefficients and ϵ_p is the error term. Both writing and speaking scores were standardized, so they have a mean of zero and unit variance before model fitting.

2.4. Model Fitting in Stan

Prior Specifications. We fit each model using the Stan software (Stan Development Team 2022c). The parameters of each model were estimated by implementing the No-U-Turn Sampler (NUTS) extension of the Hamiltonian Monte Carlo (HMC) algorithm, as implemented in the rstan and cmdstanr packages (Stan Development Team 2022a, 2022b) in R (R Core Team 2022). HMC is more effective for examining the posterior parameter space than conventional Markov Chain Monte Carlo (MCMC) algorithms, especially when dealing with intricate, high-dimensional models. Although traditional MCMC sampling algorithms can investigate the full target distribution for complex, high-dimensional models given sufficient time, approximating the posterior parameter distribution usually takes longer. HMC leverages auxiliary momentum variables, allowing each random draw to

cover more ground in the parameter space, resulting in a quicker exploration of the entire target distribution. McElreath (2017) offers an accessible introduction to HMC with helpful visual aids, and more in-depth technical introductions can be found in works by Hoffman and Gelman (2011) and Betancourt (2018).

Similar to Molenaar et al. (2022), the following priors for the item location, item discrimination, and category thresholds were adopted

$$\beta \sim N(0, 10)$$

$$\log(\alpha) \sim N(0, 1)$$

$$\gamma_0 \sim N(0, 10)$$

$$\gamma_1 \sim N(0, 10), \gamma_0 < \gamma_1$$

For the dispersion parameters,

$$\delta \sim N(0, 10)$$

was used for the Beta IRT model while

$$\log(\delta) \sim N(0, 1)$$

was used for the SB-IRT and Simplex IRT models. The prior for the error term in the regression model was specified as

$$\epsilon \sim N(0, \sigma^2)$$

$$\sigma^2 = 1 - (\zeta_1^2 + \zeta_2^2 + 2\zeta_1\zeta_2r), \sigma > 0$$

where r is the observed correlation between writing and speaking scores. As mentioned, the observed writing and speaking scores were standardized with a mean of zero and unit variance before model fitting. Therefore, this specification implies a standard normal distribution as a prior for the marginal distribution of the latent person parameters. We also fit the models without the latent regression approach for comparison purposes. When no latent regression exists, the latent person parameters are directly modeled with a standard normal distribution.

Parameter Estimation. Parameter estimation was conducted using an out-of-sample prediction approach through cross-validation (Stenhaug and Domingue 2022). The complete dataset was divided randomly into six folds, ensuring each fold included at least one response from every participant with at least six responses (98.9% of respondents). The responses from participants with fewer than six responses were also randomly assigned to one of the six folds. Each model was fitted six times, with one fold excluded in each iteration. At the end, each model was also fitted to the entire dataset. This procedure allowed for comparing models based on their out-of-sample predictive performance. To achieve this, a model was fitted using a combination of five folds, and the estimated parameters were then used to predict observations in the excluded sixth fold. Four chains with random starting values were used when fitting the models, and each chain had 1000 iterations. The first 250 iterations in each chain were used as a warm-up, and the remaining 750 were used for inference. The convergence diagnostic measured by \hat{R} , modified by Brooks and Gelman (1998), was used to assess the convergence of every parameter.

2.5. Disclosure of the Use of AI or AI-Assisted Technologies

In the preparation and revision of this manuscript, the first author, Cengiz Zopluoglu, employed AI-assisted technologies to enhance the readability and linguistic quality of the text. Specifically, he utilized Grammarly for real-time grammar, spelling, punctuation, and clarity enhancements across the entire manuscript. Additionally, he consulted ChatGPT for

specific feedback on cohesion, syntax, vocabulary, and grammar improvements, applying its suggestions to refine his writing. This approach was taken to ensure the manuscript meets the high standards of linguistic quality expected by the academic community. The final version of the entire manuscript after revisions was read and approved by both authors before finalizing the submission.

3. Results

3.1. Model Comparison and Prediction Error

Our approach to comparing models relies on assessing the effectiveness of person and item parameter estimates in predicting out-of-sample observations. We randomly divided the entire dataset into six folds, $X^{(r)}$ for $r = 1, 2, \dots, 6$. Following this, we fitted the models to each training set, excluding the r th fold $X^{(-r)}$. The derived person and item parameter estimates were then used to predict the out-of-sample observations in the excluded r th fold, using the relevant model equations. Subsequently, we computed the sum of squared errors for each r th fold,

$$SSE\left(X_{pi}^{(r)}, \hat{X}_{pi}^{(r)} \mid \theta_p^{(-r)}, \alpha_i^{(-r)}, \beta_i^{(-r)}, \delta_i^{(-r)}, \gamma_{0i}^{(-r)}, \gamma_{1i}^{(-r)}\right) = \sum_p \sum_i \left(X_{pi}^{(r)} - \hat{X}_{pi}^{(r)}\right)^2$$

We performed these calculations at each sampling iteration to estimate the posterior distribution of SSE, utilizing the parameter estimates obtained. To facilitate a more intuitive understanding of these results, we established a baseline SSE. We achieved this by predicting the value of each observation using the corresponding sample average.

Figure 2 illustrates the out-of-sample prediction error for each fold for the Beta, SB, and Simplex IRT models with the collateral information included via latent regression and without collateral information during the estimation process. The outcomes were highly consistent across all six folds, with two key patterns emerging. Primarily, models that integrated collateral information from external variables via latent regression exhibited a significantly smaller prediction error, as measured by the sum of squared error, than those that did not use this collateral information. When not considering the collateral information, the Beta, SB, and Simplex IRT models displayed a proportional reduction in prediction error by 14.4%, 11.4%, and 6.8%, respectively, compared to the baseline SSE (represented by the horizontal lines in the figure). When the collateral information was included, the proportional error reduction improved to 18.6% for Beta IRT, 15.3% for SB-IRT, and 10.3% for Simplex IRT. Consequently, including collateral information contributed to an additional reduction in the prediction error by approximately 4% across models. The second crucial pattern that emerged was the superior predictive performance of the Beta IRT model across all folds regarding unseen responses. This superiority indicates that the Beta IRT model could be a strong candidate for future utilization in processing data from a similar assessment.

3.2. Model Fit

We evaluated certain aspects of the model fit by using the posterior predictive model-checking approach (PPMC; Rubin 1984). Essentially, PPMC contrasts actual data with model-predicted or generated data, utilizing various metrics that pinpoint areas where the model might not align well. If there is a noticeable divergence between the real data and the model's predictions, it suggests that the model is not adequately capturing certain data facets. For this approach, we first generated data following the models by using the draw of person and item parameters from their respective posterior distributions provided by Stan. Visual representations are often the most user-friendly means to conduct these posterior predictive assessments. So, we created several visualizations to check the alignment between real data and model predictions. Given the superior predictive fit exhibited by models incorporating latent regression, we will present the model fit only for those models with latent regression.

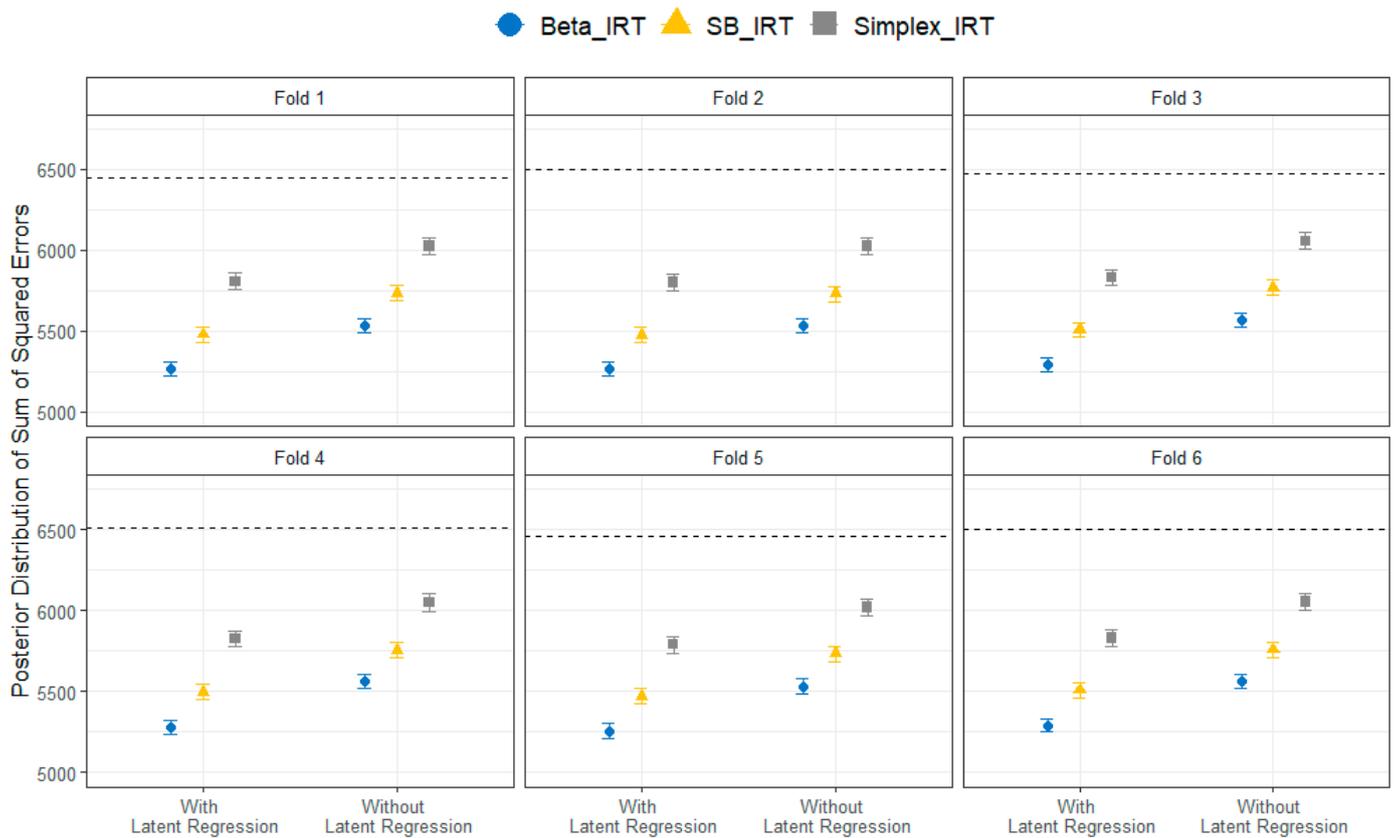


Figure 2. Comparison of the sum of the squared error of predictions across six folds for the Beta, SB, and Simplex IRT models with and without latent regression. The horizontal line for each fold represents the baseline prediction error when an average response is used. A smaller sum of squared error indicates better performance.

Table 1 and Figure 3 compare each model’s observed and posterior predictive score distributions, specifically for test takers who received six items. Given the continuous nature of scores, which can range from 0 to 6, both the table and figure illustrate that all three models—Beta IRT, SB IRT, and Simplex IRT—exhibit very similar performance characteristics. They commendably predict the observed score distribution, albeit with a noted reduction in skewness and kurtosis. However, the data seem to be heavier tailed than the model’s predictions, indicating possible limitations in the parametric models employed. Such models are, in essence, an approximation of any real data-generating model (DGM), and discrepancies tend to become more pronounced in higher moments, which amplify differences in the underlying distributions.

Table 1. Descriptive statistics for the sum scores from observed data and the average of posterior distribution of sum scores.

	Mean	SD	Skewness	Kurtosis
Beta IRT	5.56	0.36	−1.96	7.76
SB IRT	5.56	0.36	−1.87	6.72
Simplex IRT	5.56	0.37	−1.93	6.90
Observed Data	5.56	0.38	−2.90	22.63

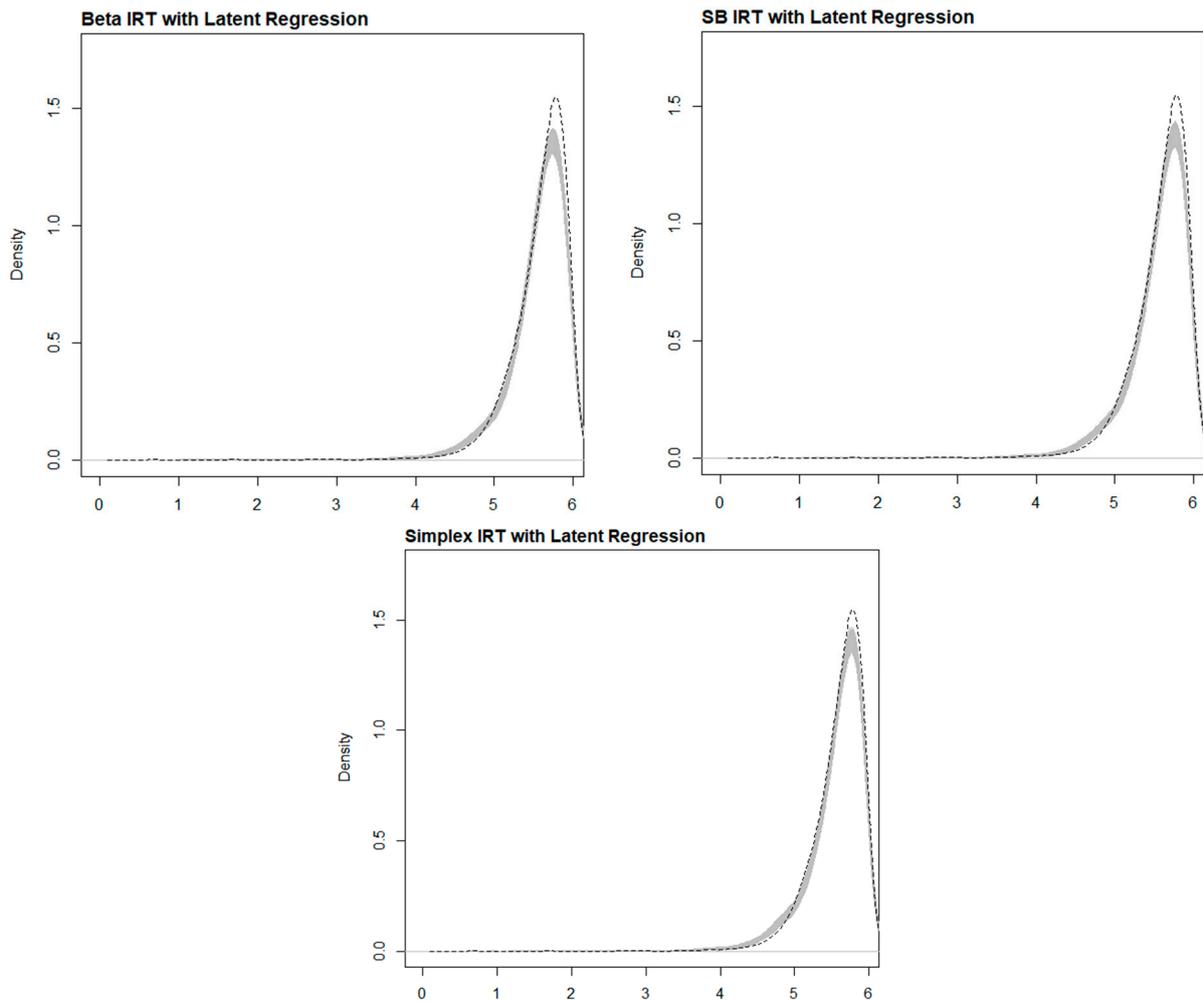


Figure 3. Density plots of observed sum score distribution (dashed line) and distributions of sum scores from 3000 posterior samples (gray area) for each model.

Furthermore, even if the parametric model accurately represented the majority of the data, real data are often subject to contamination from outliers due to idiosyncratic events. For instance, a test taker might become distracted and make an atypical error that would not replicate under different measurement conditions for the same person or item combination. This is particularly pertinent given that the test takers were required to type their responses for the items in this dataset; hence, even if they have completely understood the stimulus, there is still a potential for error during the input phase. Such outliers have a pronounced impact on higher moments, as they are probably more sensitive to this kind of data contamination.

Figure 4 illustrates the comparison between the average scores of actual observed data and those generated by the posterior predictive distributions of the models. This calculation is performed for each item, and Figure 4 summarizes across items. All three models—Beta, SB, and Simplex IRT—closely matched the observed average scores. The mean observed response for all items stood at 0.9307. In contrast, the averages for the model-predicted responses for the Beta, SB, and Simplex IRT models were 0.9315, 0.9318, and 0.9320, respectively.

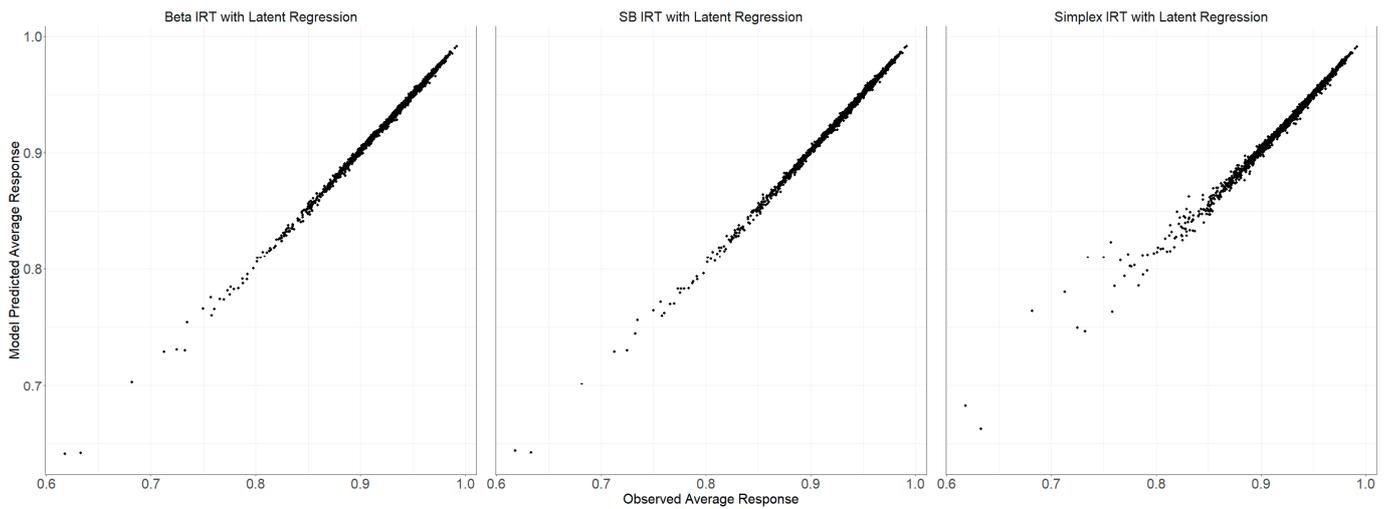


Figure 4. Comparison of average item scores from observed data and posterior predictive distributions of model-generated data.

Figure 5 similarly presents a comparison across items between the standard deviations of observed item scores and those generated by the posterior predictive distributions of the models. The observed responses had an average standard deviation of 0.1021 across all items. On the other hand, the model-predicted responses slightly higher average standard deviations: 0.1053 (Beta), 0.1067 (SB), and 0.1110 (Simplex IRT). This data also highlights the Beta IRT model’s superior prediction accuracy for unseen data discussed earlier compared to the other two models. While all models were relatively consistent in reproducing average item scores, the Beta-IRT produced predictions with the smallest variance.

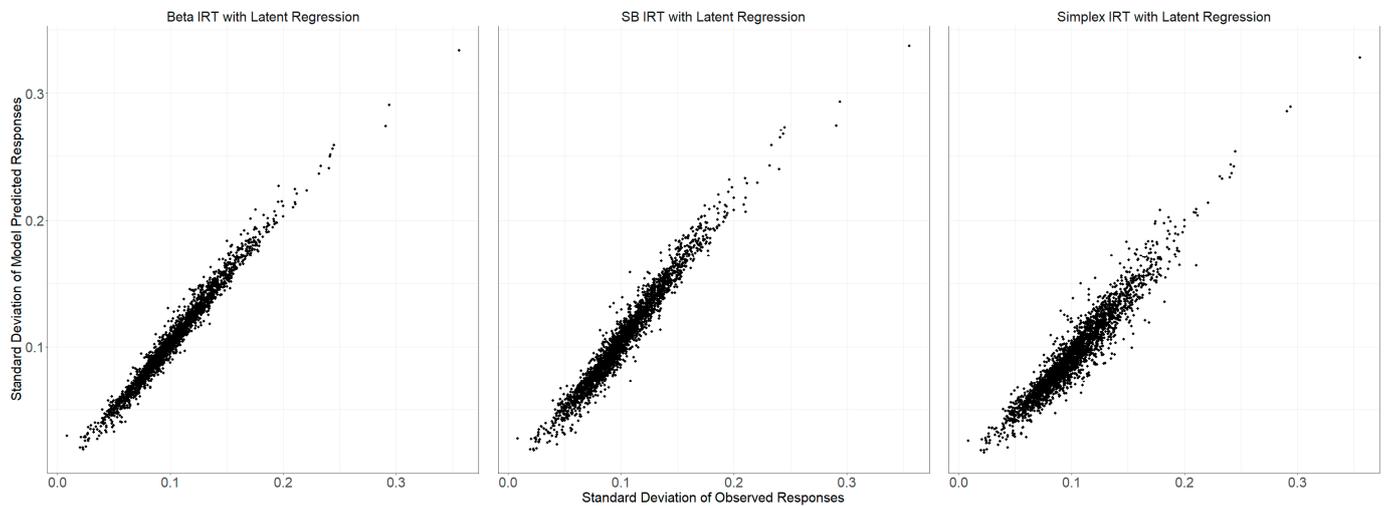


Figure 5. Comparison of standard deviations of item scores from observed data and posterior predictive distributions of model-generated data.

3.3. Parameter Estimates

Given the superior predictive fit exhibited by models incorporating latent regression, we will summarize the findings regarding parameter estimates specifically for these models. Each item in these models has five parameters, resulting in many item parameters per model. We evaluated the convergence of these item parameters, utilizing R-hat values. The majority of item parameters demonstrated high-quality convergence. Specifically, the R-hat values for 97.5%, 100%, and 99.9% of all item parameters for the Beta, SB, and Simplex IRT models were below the threshold of 1.05, indicating good convergence. We also examined

the convergence for all person parameter estimates for each model and observed a similar quality. The R-hat values for 99.4%, 100%, and 99.9% of all person parameters for the Beta, SB, and Simplex IRT models were below the threshold of 1.05.

Across the models, the item and person parameters exhibited strong similarities. The descriptive statistics for these model parameter estimates are presented in Tables 2 and 3. Figures 6 and 7 illustrate the relationships and correlations between item and person parameter estimates among the Beta, SB, and Simplex IRT models. Except for the α parameter, the parameter estimates from different models aligned very closely, with correlation values ranging from 0.980 to 0.999. However, the correlations were relatively lower for the α parameter, falling within a range of 0.91 to 0.94. The supplemental writing and speaking scores were significant predictors of the latent person parameters, and the estimated parameters were similar across models.

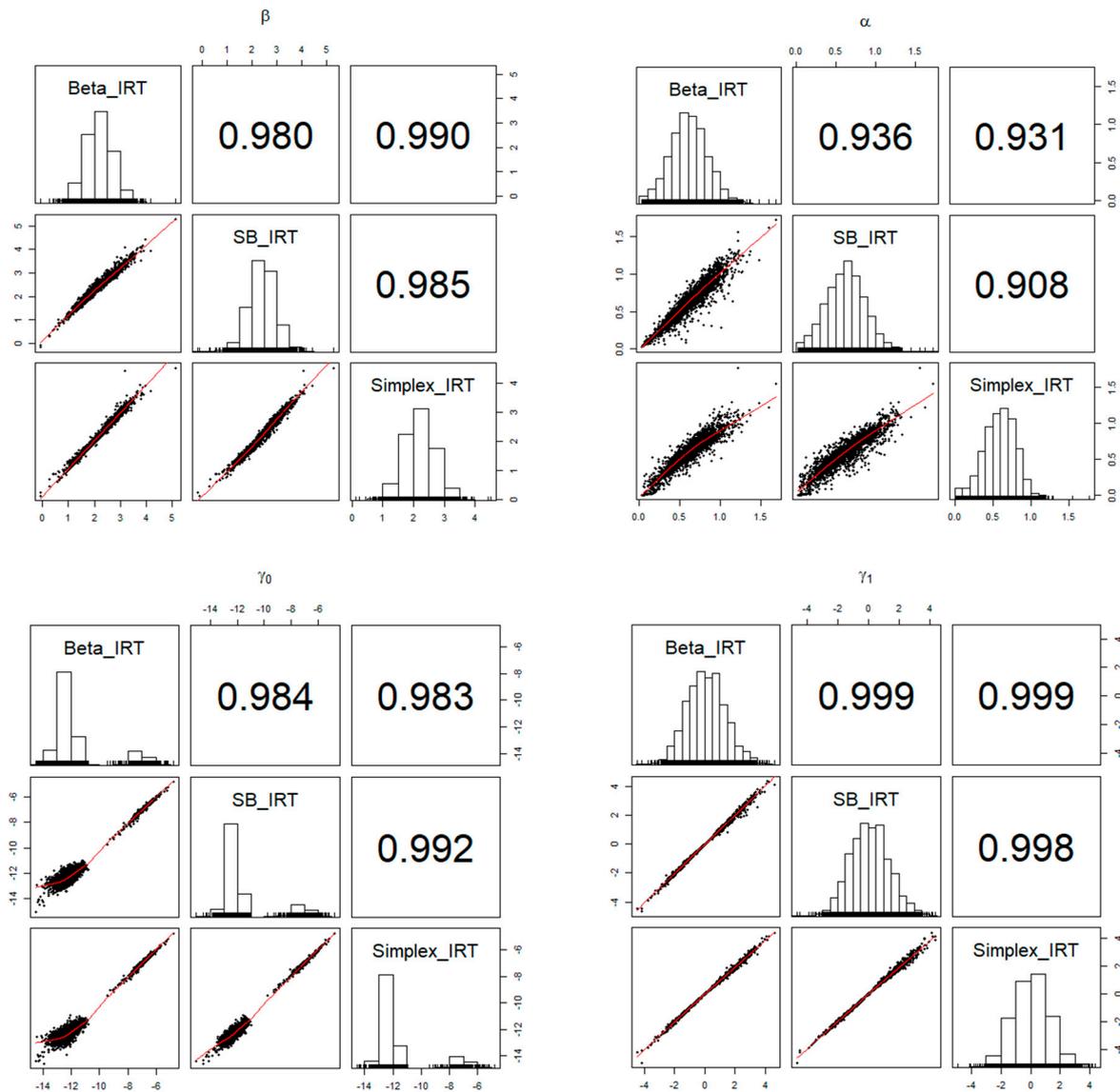


Figure 6. The relationships among the item parameter estimates obtained from Beta, SB, and Simplex IRT models.

Table 2. Descriptive statistics for item and person parameters estimated from the Beta, SB, and Simplex IRT models with latent regression.

Parameters	Beta IRT Model with Latent Regression				SB IRT Model with Latent Regression				Simplex IRT Model with Latent Regression			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
θ	0.00	0.86	-11.86	3.05	0.00	0.85	-6.99	3.17	0.00	0.85	-10.17	3.20
β	2.18	0.53	-0.09	5.15	2.41	0.54	-0.16	5.27	2.20	0.51	0.12	4.52
α	0.61	0.23	0.02	1.69	0.62	0.25	0.03	1.72	0.59	0.22	0.00	1.77
γ_0	-11.62	1.97	-14.51	-4.77	-11.65	1.95	-15.05	-4.81	-11.64	1.96	-14.56	-4.77
γ_1	0.10	1.27	-4.47	4.62	0.09	1.28	-4.64	4.35	0.08	1.25	-4.93	4.38
δ^*	3.53	0.97	-0.05	13.21	0.66	0.28	0.04	3.30	8.41	5.16	0.26	36.45

* The disturbance parameters (δ) are not comparable across the model.

Table 3. Estimated coefficients from the latent regression.

	Writing (ζ_1)		Speaking (ζ_2)	
	Posterior Mean	95% Credible Interval	Posterior Mean	95% Credible Interval
Beta IRT	0.351	(0.346, 0.354)	0.345	(0.340, 0.349)
SB IRT	0.347	(0.342, 0.351)	0.352	(0.347, 0.356)
Simplex IRT	0.330	(0.325, 0.334)	0.346	(0.342, 0.351)

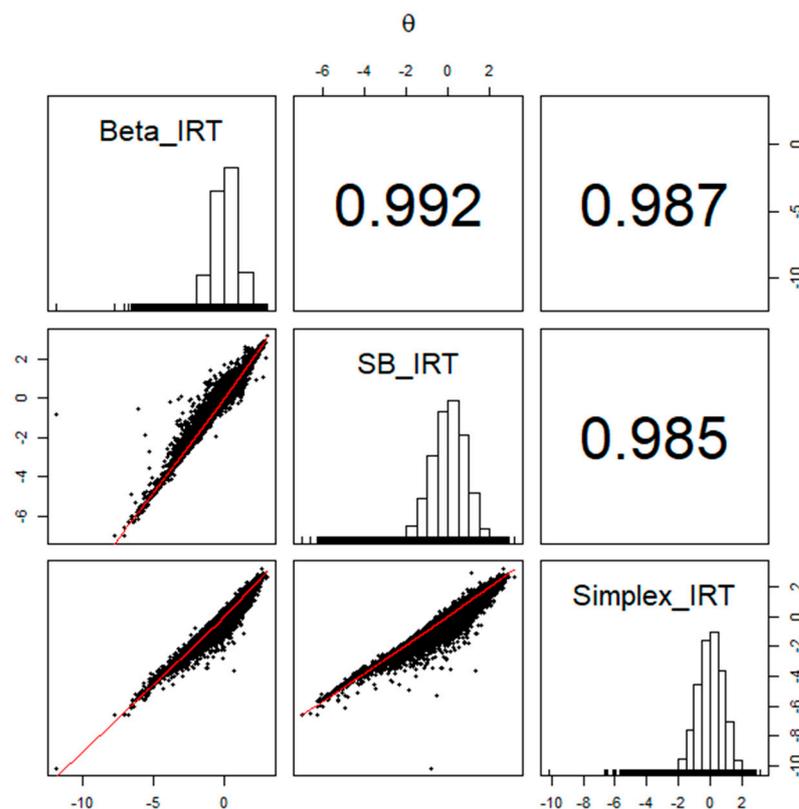


Figure 7. The relationships among the person parameter estimates obtained from Beta, SB, and Simplex IRT models.

4. Discussion

The present study evaluated the performance of novel IRT models, specifically zero-and-one-inflated extensions of the Beta IRT, Simplex IRT, and Samejima’s Continuous IRT models, modeling grades of dictation tasks in a high-stakes English language proficiency

assessment. We adopted a “predictive fit” approach, discussed and advocated by [Stenhaug and Domingue \(2022\)](#), to compare the models through cross-validation. Each independent fold of the whole response dataset included a randomly selected subset of responses from all test takers. The goal of the comparison was to measure how well a model predicts a person’s missing response in a certain fold using the item and person parameters estimated from the remaining folds.

Our findings underscore the potential of these models, especially when they incorporate collateral information, to provide accurate estimates and predictions. One of the most salient takeaways from our research is the superior predictive performance of the Beta IRT model regarding out-of-sample responses. This finding aligns with [Molenaar et al. \(2022\)](#), who found the Beta IRT model best fitting for 11 out of 22 scales they included in their study. We can argue that, among the models evaluated, the Beta IRT model holds the most promise for future applications in a similar assessment context, particularly when the task is administered adaptively. However, it is crucial to note that while the Beta IRT model demonstrated superior predictive accuracy for unseen data, the differences in model fit and parameter estimates among the models were not substantial. It is also important to consider the concept of fitting propensity when evaluating these models. While our study highlights the effectiveness of these models, especially the Beta IRT model, in high-stakes English assessments, a detailed comparison of their relative parsimony and flexibility remains an area for future exploration. Understanding whether the Beta IRT’s success is due to its greater flexibility than the other models could be crucial for its application. Future studies could focus on a fitting propensity analysis, like the work by [Bonifay and Cai \(2017\)](#) and [Ergin \(2020\)](#), to comprehensively evaluate these models.

Our research has also highlighted specific gaps in the current understanding of novel IRT models studied in this paper. A notable challenge was the scarcity of established benchmarks or guidelines for evaluating model fit and item fit specifically for these advanced IRT models. This issue is particularly pressing given the fundamental importance of accurate model fit assessment in ensuring the validity and reliability of a model’s predictions. The absence of well-established evaluative criteria or methods leaves researchers and practitioners uncertain about the appropriateness of these newer models. In addressing this gap, our study primarily relied on heuristic visual checks based on model-generated data from respective posterior distributions of the item and person parameters. While established discrepancy measures for assessing model fit are relatively better researched and documented for traditional dichotomous and polytomous IRT models (e.g., [Sinharay et al. 2006](#)), such measures are not yet fully developed for these recent IRT models, particularly those for continuous responses or those introduced in recent research, such as by [Molenaar et al. \(2022\)](#). There is a need for dedicated research to develop and validate a comprehensive set of discrepancy measures for assessing model fit for these novel IRT models. Such research would enhance the credibility and utility of these models and equip researchers and practitioners with the essential tools they need to make informed decisions about model selection and interpretation. Considering the generally modest differences among the models in various performance metrics reported in our study, the selection in real-world settings could be based on alternative criteria such as interpretability, computational stability, and simplicity.

As high-stakes computerized adaptive assessments for language proficiency continue to evolve, the integration of recent NLP technologies is anticipated to introduce novel item types. These advancements necessitate analytical models capable of accurately processing and interpreting data from such items. The models evaluated in our study, particularly their ability to handle bounded continuous outcomes, are well-suited for this emerging landscape. Our research indicates that these IRT models could be instrumental in managing the data complexities presented by NLP-driven assessment items, thereby supporting more nuanced and effective language proficiency evaluations.

Author Contributions: Conceptualization, C.Z. and J.R.L.; methodology, C.Z. and J.R.L.; software, C.Z.; formal analysis, C.Z.; data curation, J.R.L.; writing—original draft preparation, C.Z.; writing—review

and editing, C.Z. and J.R.L.; visualization, C.Z.; funding acquisition, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Duolingo, Inc. through the Competitive Research Grant Program to support topics of interest related to Duolingo’s English Test’s ongoing research agenda.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: You can access our GitHub repository at https://github.com/czopluglu/Duolingo_paper. It contains code samples that showcase the analyses performed on a simulated dataset that mimics the structure of real data analyzed in the study.

Acknowledgments: We would like to thank the Research Advanced Computing Services that maintain Talapas, The University of Oregon’s performance computing cluster, for providing resources to run analysis that contributed to the research results reported within this article. URL: <https://racs.uoregon.edu/talapas>. We also acknowledge the use of Grammarly for grammar and clarity enhancements and ChatGPT for feedback on cohesion, syntax, and vocabulary during the preparation of this manuscript. These AI tools were instrumental in refining the manuscript to meet academic standards, replacing traditional editorial services with advanced, accessible technology.

Conflicts of Interest: Co-author Lockwood is employed by Duolingo. He extracted data from Duolingo data systems to support the study and contributed to technical analyses and writing. These activities were conducted in a manner consistent with professional standards for methodological research.

Appendix A. Model-Specific Probability Density Functions

SB_IRT Model

$$f(X_{pi}|\theta_p, \alpha_p, \beta_p, \delta_p) = \frac{1}{\sqrt{2\pi\delta_i}} \frac{1}{X_{pi}(1-X_{pi})} \exp\left(-\frac{\left(\log\left(\frac{X_{pi}}{1-X_{pi}}\right) - \alpha_i\theta_p - \beta_i\right)^2}{2\delta_i}\right)$$

Beta IRT Model

$$f(X_{pi}|\theta_p, \alpha_p, \beta_p, \delta_p) = \frac{\Gamma(a_{pi} + b_{pi})}{\Gamma(a_{pi})\Gamma(b_{pi})} X_{pi}^{a_{pi}-1} (1 - X_{pi})^{b_{pi}-1}$$

where $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(d) = \int_0^\infty t^{d-1}e^{-t}dt$, and

$$a_{pi} = e^{\frac{\alpha_i\theta_p + \beta_i + \delta_i}{2}}$$

$$b_{pi} = e^{\frac{-(\alpha_i\theta_p + \beta_i) + \delta_i}{2}}$$

Simplex IRT Model

$$f(X_{pi}|\theta_p, \alpha_i, \beta_i, \delta_i) = \frac{1}{\sqrt{2\pi\delta_i} [X_{pi}(1-X_{pi})]^3} \exp\left(-\frac{(X_{pi} - \mu_{pi})^2}{2\delta_i X_{pi}(1-X_{pi})\mu_{pi}^2(1-\mu_{pi})^2}\right)$$

$$\mu_{pi} = \frac{1}{1 + e^{-(\alpha_i\theta_p + \beta_i)}}$$

References

- Adams, Raymond J., Mark Wilson, and Margaret Wu. 1997. Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics* 22: 47–76. [CrossRef]
- Bejar, Isaac I. 1977. An Application of the Continuous Response Level Model to Personality Measurement. *Applied Psychological Measurement* 1: 509–21. [CrossRef]
- Betancourt, Michael. 2018. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv* arXiv:1701.02434. Available online: <http://arxiv.org/abs/1701.02434> (accessed on 25 December 2020).
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv* arXiv:2108.07258. Available online: <http://arxiv.org/abs/2108.07258> (accessed on 18 August 2021).
- Bonifay, Wes, and Li Cai. 2017. On the complexity of item response theory models. *Multivariate Behavioral Research* 52: 465–84. [CrossRef] [PubMed]
- Brooks, Steve P., and Andrew Gelman. 1998. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7: 434–55. [CrossRef]
- Brumfitt, Shelagh M., and Paschal Sheeran. 1999. The development and validation of the Visual Analogue Self-Esteem Scale (VASES). *British Journal of Clinical Psychology* 38: 387–400. [CrossRef] [PubMed]
- Buck, Gary. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.
- Cardwell, Ramsey, Ben Naismith, Geoffrey T. LaFlair, and Steven Nydick. 2023. *Duolingo English Test: Technical Manual*. Pittsburgh: Duolingo. Available online: https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf (accessed on 30 March 2023).
- de la Torre, Jimmy. 2003. Improving the Accuracy of Item Response Theory Parameter Estimates through Simultaneous Estimation and Incorporation of Ancillary Variables. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL, USA. Available online: <https://www.proquest.com/docview/288199771/abstract/5B88C5006775440APQ/1> (accessed on 29 April 2023).
- Ergin, Ezgi Ayturk. 2020. Fitting Propensities of Item Response Theory Models. Ph.D. dissertation, Fordham University, New York, NY, USA. Available online: <https://www.proquest.com/docview/2416910996> (accessed on 17 January 2024).
- Flores, Sandra, Jorge Luis Bazán, and Heleno Bolfarine. 2020. A Hierarchical Joint Model for Bounded Response Time and Response Accuracy. In *Quantitative Psychology*. Edited by Marie Wiberg, Dylan Molenaar, Jorge González, Ulf Böckenholt and Jee-Seon Kim. Cham: Springer International Publishing, pp. 95–109. [CrossRef]
- Guthrie, John T., Mary Seifert, Nancy A. Burnham, and Ronald. I. Caplan. 1974. The Maze Technique to Assess, Monitor Reading Comprehension. *The Reading Teacher* 28: 161–68.
- Hall, Erika. 2007. Using Collateral Item and Examinee Information to Improve IRT Item Parameter Estimation. Ph.D. dissertation, The University of Iowa, Iowa City, IA, USA. Available online: <https://www.proquest.com/docview/304856966/abstract/CA86530008C542E7PQ/1> (accessed on 29 April 2023).
- Hoffman, Matthew D., and Andrew Gelman. 2011. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv* arXiv:1111.4246. Available online: <http://arxiv.org/abs/1111.4246> (accessed on 25 December 2020).
- Joo, Seang-Hwane, Philseo Lee, and Stephen Stark. 2022. The Explanatory Generalized Graded Unfolding Model: Incorporating Collateral Information to Improve the Latent Trait Estimation Accuracy. *Applied Psychological Measurement* 46: 3–18. [CrossRef] [PubMed]
- Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163: 845–48, English translation in *Soviet Physics Doklady* 10: 707–10, 1966.
- McElreath, Richard. 2017. Markov Chains: Why Walk When You Can Flow? *Elements of Evolutionary Anthropology*, November 28. Available online: <https://eleanth.org/blog/2017/11/28/build-a-better-markov-chain/> (accessed on 25 December 2020).
- Mellenbergh, Gideon J. 1994. A Unidimensional Latent Trait Model for Continuous Item Responses. *Multivariate Behavioral Research* 29: 223–36. [CrossRef] [PubMed]
- Mislevy, Robert J., and Kathleen M. Sheehan. 1989. The role of collateral information about examinees in item parameter estimation. *Psychometrika* 54: 661–79. [CrossRef]
- Molenaar, Dylan, Mariana Cúri, and Jorge L. Bazán. 2022. Zero and One Inflated Item Response Theory Models for Bounded Continuous Data. *Journal of Educational and Behavioral Statistics* 47: 693–735. [CrossRef]
- Müller, Hans. 1987. A rasch model for continuous ratings. *Psychometrika* 52: 165–81. [CrossRef]
- Noel, Yvonnick, and Bruno Dauvier. 2007. A Beta Item Response Model for Continuous Bounded Responses. *Applied Psychological Measurement* 31: 47–73. [CrossRef]
- R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. [Computer Software]. Vienna: R Foundation for Statistical Computing. Available online: <https://www.R-project.org/> (accessed on 11 May 2023).
- Raatz, Ulrich, and Christine Klein-Braley. 1981. The C-Test—A Modification of the Cloze Procedure. In *Practice and Problems in Language Testing*. Edited by Terry Culhane, Christine Klein-Braley and Douglas Keith Stevenson. Essex: University of Essex Occasional Papers, pp. 113–48.
- Rubin, Donald B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 12: 1151–72. [CrossRef]
- Samejima, Fumiko. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 34: 1–97. [CrossRef]

- Samejima, Fumiko. 1973. Homogeneous case of the continuous response model. *Psychometrika* 38: 203–19. [CrossRef]
- Sinharay, Sandip, Matthew S. Johnson, and Hal S. Stern. 2006. Posterior predictive assessment of item response theory models. *Applied Psychological Measurement* 30: 298–321. [CrossRef]
- Stan Development Team. 2022a. CmdStan User’s Guide (2.32) [Computer Software]. Available online: https://mc-stan.org/docs/2_32/cmdstan-guide-2_32.pdf (accessed on 11 May 2023).
- Stan Development Team. 2022b. RStan: The R interface to Stan [Computer Software]. Available online: <https://mc-stan.org/rstan/index.html> (accessed on 11 May 2023).
- Stan Development Team. 2022c. Stan User’s Guide (2.29) [Computer Software]. Available online: <https://mc-stan.org> (accessed on 14 July 2021).
- Stenhaug, Benjamin A., and Benjamin W. Domingue. 2022. Predictive fit metrics for item response models. *Applied Psychological Measurement* 46: 136–55. [CrossRef] [PubMed]
- Tao, Shuqin. 2009. Using Collateral Information in the Estimation of Sub-Scores—A Fully Bayesian Approach. Ph.D. dissertation, University of Iowa, Iowa City, IA, USA. [CrossRef]
- University of Oregon. 2018–2020. *8th Edition of Dynamic Indicators of Basic Early Literacy Skills*. [Technical Manual]. Eugene: University of Oregon. Available online: https://dibels.uoregon.edu/sites/default/files/DIBELS8-TechnicalManual_04152020.pdf (accessed on 27 April 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.