Modeling the Sequential Response Time with Item Position and Total Time Limits

Abstract: We presented a general framework for modeling sequential response time (RT) in this research. This framework extends the traditional treatment of RT by considering the truncation of the response distribution and item position effects. These two approaches get a more accurate representation of RT generation. Both the simulation research and the analysis of real data are carried out using a completely Bayesian methodology based on the Markov Chain Monte Carlo (MCMC) technique.

Keywords: Response time, Total Time limits, Adaptive Truncation, Item Position Effect

Introduction

Similar to ability, the construct of speed has a long history in both the psychology of individual differences and educational measurement (e.g., Gulliksen, 1950; Kelley, 1927; Thorndike et al., 1926). Researchers can now automate the recording of students' reaction times (RTs) on items via the use of computer-based assessments. We can use RTs to determine the amount of labor required to finish an item as well as the speed at which tasks are completed (Partchev et al., 2013; van der Linden, 2009). Meanwhile, the information included in RTs could be used to improve normal testing procedures such as item calibration, adaptive item selection, and latent ability estimation, as well as to investigate and analyze factors affecting test performance (Fox, Entink, & van der Linden, 2007).

RTs play a variety of roles in various psychometric models. One strategy is to use the information contained in RTs as a proxy for the latent speed parameter or as an explanatory predictor in item response theory (IRT) and generalized linear mixed models (GLMMs) in order to scale latent ability (Roskam, 1997; Verhelst, Verstraalen, & Jansen, 1997; Thissen, 1983; and; Maris & Van Der Maas, 2012; Goldhanmmer et al., 2014; Van Rijn & Ail, 2017). A second method is to represent the RTs independently of the RAs as a distinct latent variable (Scheiblechner, 1979; Maris, 1993; van der Linden, 2006). The third technique is to use hierarchical or mixed regression to model both RTs and RAs concurrently (Van Breukelen, 2005; van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009; Loeys et al., 2011). In these measurement models, person-specific speed and item-specific parameters are assumed to exist and are fixed to account for the variations in RTs. Meanwhile, various statistical distributions for characterizing the RTs have been examined. Given the fact that RTs have an intrinsic lower bound at zero, the candidate distributions should have positive domains. For instance, the lognormal distribution is a frequently used option

(Thissen, 1983; van der Linden, Scrams, & Schnipke, 1999). In contrast to the exponential model, the mode of lognormal is larger than zero, which is more reasonable given that students need some time to read the item. In contrast to the gamma distribution (Maris, 1993) and the Weibull distribution (Scheiblechner, 1979; Rouder et al., 2003) models, the lognormal model includes separate mean and variance parameters, which enhances its interpretability. Klein Entink et al. (2009) also used a class of Box-Cox transformations to approximate data produced by Weibull, gamma, and exponential models in their RT modeling.

However, the aforementioned RT studies make an implicit assumption that the item parameters remain constant during the assessment. Meanwhile, limited study has been conducted on the influence of item position on RT. In the measurement of response accuracy (RA), it has been shown that changing item order across various test formats has unanticipated consequences on item characteristics (Debber & Janssen, 2013). The item position effect refers to the fact that the difficulty of an item in an accomplishment test is dependent on its position (Weirich, Hecht, & Böhme, 2009). Two different types of item location impacts on item difficulty have been identified: a practice effect, which makes objects easier in a later position, and a tiredness effect, which makes items more difficult in a later position (Kingston & Dorans, 1984). The time-intensity parameter for an item may also be position-dependent. For instance, students might naturally spend more time on the initial items. In an extreme instance, students will be unable to reach the item at the end of the assessment due to a lack of time. Thus, ignoring the item position effect may result in estimate bias for the item time-intensity parameter.

Additionally, the natural upper bound of RT must be regulated (van der Linden, 2005). Without taking possible time restrictions into consideration, conventional modeling of RT assumes that the RTs are conditionally independent given the item and persona parameters. The majority

of past research on time limits has concentrated on item-level treatment. Ranger and Kuhn (2012), for example, introduced the proportional hazard and accelerated failure time models, in which a binary indication indicates if a response occurred prior to the item time limit. According to Goldhammer (2015), specifying the individual speed-ability trade-off involves the development of an appropriate item-level time-limit condition. Additionally, De Boeck, Chen, and Davison (2007) recommended investigating multiple temporal constraints based on reported RT distributions. However, in reality, the total time limit is also common. The traditional assumption of RT modeling may not be reasonable, since students' remaining time on each question is dependent on the amount of RTs used on prior items.

Van der Linden (2005) recommended that the selection method should take the remaining testing time , the time-intensive nature of available items, and the measurement goal into account. Based on this reason, we proposed a general framework for modeling sequential RTs. In comparison to previous research, our primary contributions were (1) expanding reaction time modeling with adaptive truncation to account for remaining time restraints, and (2) including item location effects, which are mostly explored in the literature on RAs, into RT modeling. The paper is organized as follows. We begin by outlining the model structure and discussing how time constraints and item position effects may be included into conventional RT modeling. Then, Bayesian estimation is examined, as well as the model identification problem. Additionally, we build a simulation study to evaluate and compare the proposed models' performance. Finally, the proposed model is applied to a real data example.

Model

In this section, we will introduce the RT modeling framework with item position effect and total time limits. To begin, we will discuss the conventional lognormal model for RT (van der Linden, 2006). We will illustrate how to relax the lognormal model's item parameter stationarity and local independence assumptions. We begin by introducing the adaptive truncation method of the lognormal distribution based on the sequence of RT, which closely resembles the remaining time limitations (named as lognormal-T). Then, we'll discuss how to include item position effects (named as lognormal-P).

Lognormal Model for Response Time

One of the most commonly used framework of RT modeling is the lognormal model proposed by van der Linden (2006, 2007), which share the similar framework of item response theory in RA modeling. The lognormal distribution is used to account for the positively skewed characteristic of RT distributions:

$$f(T_{ij};\tau_i,\alpha_j,\beta_j) = \frac{\alpha_j}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_j\left(lnT_{ij}-(\beta_j-\tau_i)^2\right)\right]\right\} \quad (1)$$

Here, T_{ij} is the observed RT of a student i $(i = 1,\dots,I)$ on item j $(j = 1,\dots,J)$. Under the lognormal distribution, the mean of log RT depends on constant latent speed (τ_i) as the person parameter and time-intensity (β_j) as the item parameter. Increasing the time-intensity leads to a positive shift of the location of the time distribution on the item, while increasing the speed parameter leads to a negative shift. The distribution of the response time has a parameterization close to that of an IRT model for positive continuous RTs (Samejima, 1973). Meanwhile, item discrimination parameter (α_j) is included as the reciprocal of the standard deviation of the normal distribution (van der Linden, 2009).

Lognormal Model with Total Time Limits

In assessment, the students' ability to deal with time pressure or their flexibility to operate at different speed-ability compromises is desired. As a result, overall time limitations are often specified. RT is not defined throughout the whole positive real domain, and its natural upper limits may be position-dependent. The adaptive truncation is introduced into the lognormal distribution. A truncated distribution is a conditional distribution that comes from domain restriction in statistics. With truncation, the sample distribution must be normalized to ensure that it is within the permitted range:

$$f\left(t_{ij};\tau_i,\alpha_j,\beta_j,t_{ij}<L_i\right)=\frac{f\left(t_{ij};\tau_i,\alpha_j,\beta_j\right)I(t_{ij}<L_i)}{F(L_i)}\propto f\left(t_{ij};\tau_i,\alpha_j,\beta_j\right)I\left(t_{ij}<L_i\right) \quad (2)$$

Here, L_i is the time limitation of the ith item. $I(.)$ is the indicator function to check whether the generated RT is under the time limits. $F(.)$ is the cumulative distribution function of lognormal distribution.

In a more general example, only total time limit (L) is constrained. Thus, students must determine the optimal distribution of RT over all items to avoid non-reached missing or quick guessing. Students will not have sufficient time if too much time is used on previous items. Meanwhile, student must consider how much time to provide for subsequent items and whether to proceed. As a result, the RT distributions are not independent. The domain of distribution for the first item has the greatest range, with the upper bound representing the overall time limitations. The domain becomes narrower when the second item is added:

$$f\left(t_{ij};\tau_i,\alpha_j,\beta_j,t_{ij}\le L-\sum_{k=1}^{j-1}t_{ik}\right)\propto f\left(t_{ij};\tau_i,\alpha_j,\beta_j\right)I\left(t_{ij}\le L-\sum_{k=1}^{j-1}t_{ik}\right),j=2,\dots,J \quad (3)$$

The adaptive upper bound ($L - \sum_{k=1}^{j-1} t_{ik}$) is manifest. By including truncation into the lognormal model, we may get a more accurate description of real-world RTs production without introducing any new person or item factors. Large RTs are less likely to be seen for items at later positions.

Lognormal Model with Item Position Effects

Extending the lognormal model by truncating time limitations allows for a more accurate representation of the student's actual sequential RT generating process throughout the evaluation. However, temporal position effects may also contribute to systemic bias in sequential RTs. It has been shown repeatedly that item parameters may vary based on their position within a test form. The item position effect is often detected in two steps: (1) estimating item characteristics (e.g., item difficulty) for each test form, and (2) modeling the variations in item parameters across test forms as a function of item position (Meyers, et al., 2009). Debeer and Janssen (2013) provided a concise overview of the basic frameworks for modeling item position effects using item response theory (IRT). The item position effects and other item features are disentangled in designs with varying test formats using this approach.

In this work, we use the lognormal distribution to quantify the influence of item position on item parameters in RTs (e.g., item time-intensity). In a general framework, we could incorporate latent position factors for each position to represent the item position impacts across items.

$$f(t_{ijk};\tau_i,\alpha_j,\beta_{jk}) = \frac{\alpha_j}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_j\left(\ln t_{ij} - \left(\beta_{jk} - \tau_i\right)^2\right)\right]\right\},\ \ \beta_{jk} = \beta_j^* + \lambda_k \qquad (4)$$

, where k represents position of the jth item for ith student. We decompose the item time intensity β_{jk} into two components: β_j^* and λ_k. β_j^* represents the pure item time intensity and λ_k represents the effect of presenting the item in position k. The specification of formula (4) imposes

no structure on the impacts of various places, which may cause identification problems for some datasets. When each student responds to an identical set of items in an identical sequence, position effects are indistinguishable from the item time-intensity because each item occurs only at a specific position for all students. Thus, a relatively larger sample of students and randomly shuffled item position is needed to distinguish these two effects. Otherwise, additional information or restriction should be included.

Instead, we may use an explanatory style to constrain the amount of position impacts as a function of item position (De Boeck & Wilson, 2004). For instance, if we assume a linear position impact, we may express the model as:

$$\beta_{jk} = \beta_j^* + \lambda_j(k-1) \tag{5}$$

Here item time intensity β_{jk} is decomposed into two components: β_j^* and $\lambda_j(k-1)$. λ_j is the linear weight of the position at item level. γ_j represents the item time intensity when the item is administered in the first position (i.e., $k = 1$). We can further $\gamma_j = \gamma$ for all items to represents the general linear weight of position across different items.

Similarly, we may deconstruct item discrimination into pure item time discrimination and the effect of item position on discrimination. The item position impact may be more readily captured via process data (i.e., RTs) than through RAs, since process data contain more detailed information (along with noise). For example, differences in RTs caused by item position could be detected since RT is a continuous variable. For simplicity, we only investigate the item position effects across items on difficulty in this study.

Alternatively, the individual differences in the effects of item position on difficulty can be taken into account (Debber & Janssen, 2013). Following the idea of Rijmen and De Boeck (2002), person dimension could be incorporated:

$$\tau_i = \lambda_i(k-1) - \tau_i^* \quad (6)$$

Here, latent speed is decomposed into two components: τ_i^* and $\lambda_i(k-1)$. λ_i is a normally distributed linear weight of position effect at person level, which represents to what extent a student's speed change through the test. τ_i^* represents the represents the speed for the first position (i.e., $k = 1$). Fox and Marianti (2016) define the framework of Formula (6) as the differential personal speed using latent curve analysis. Quadratic position effects or higher-order effects might potentially be introduced to depict more complicated differential speed trajectories. We maintain the assumption of latent speed stationarity in this study and assume that the impact of item position is mostly on item parameters.

Due to the sequential nature of RTs, total time limitations and item position have a relationship but are different effects. For instance, the item position impact may be explained in part by the degree of time pressure applied relative to the remaining time. Thus, in the most generic framework (lognormal-PT), the predicted item position impacts are those that extend beyond remaining time constraints.

Estimation

The Markov Chain Monte Carlo (MCMC) technique can be used to estimate parameters in lognormal-PT in a Bayesian approach. In Bayesian estimation, the joint posterior distribution is created by the prior distribution of parameters and the probability of observed data. When all conditional posterior distributions are provided, the Gibbs sampler can be used to simulate draws, producing a series of random variables that converges to the joint posterior distribution of all free parameters in the "target distribution." The Gibbs sampler (Gelfand & Smith, 1990) is utilized in

this work in conjunction with the R2jags program (Version 0.6.1; Su & Yajima, 2015). (Version 4.0.2; R Core Team, 2016). The posterior median could be used as a point estimate.

When using an MCMC algorithm, lognormal models are usually identified by specifying the prior distribution of latent speed parameters as a normal distribution with the mean as zero and variance as a unit. In this way, the mean of time-intensity parameters of the items equates to the mean log RTs. Alternatively, the variance of the speed scale could be identified by restricting the product of time discrimination to one. These identification restrictions are discussed in Entink, Fox, et al. (2009) and Fox (2010).

Assuming local independence, $\log(T_{ij})$ is conditionally and independently distributed as $\log(T_{ijk}) \sim^{iid} N(\mu_{ijk}, a_j^{-2})$, where μ_{ijk} is the mean of log RTs and can be decomposed into the latent item, person, and position parameter. The prior of item parameters are assume to follow a weekly informative distribution. We assume item time intensity $\lambda_j \sim^{iid} N(\mu_\lambda, \sigma_\lambda^2)$ and $\mu_\lambda \sim N(\overline{\log(T)}, 10)$, where $\overline{\log(T)}$ is the average of observed log RTs. σ_λ^2 follow a vague uniform Gamma distribution. For linear weight of the position, we assume it follows the vague student's T distribution: $\lambda_i \sim^{iid} T_7(0,1)$ for formula (5) and $\lambda_i \sim^{iid} T_7(0,1)$ for formula (6).

Simulation Study

In this simulation study, we compare the performance of four models: the log-normal model (lognormal), the log-normal model with truncation only (lognormal-T), the log-normal model with item position effects only (lognormal-P), and the lognormal model with both item position effects and truncation (lognormal-TP). For data generation, we generate the data based on the specification of formula (5) when all item shares the same linear weight. Two factors are discussed in this simulation study: (1) the number of items (i.e., 10, 20, and 30), and (2) the number

of students (i.e., 200, 1000, and 2000). The linear weight for each of the 9 cases is independently and randomly sampled from the uniform distribution from -0.1 to 0. Thus, we assume there is a practice effect of item position. Students' latent speed is sampled from the uniform distribution from -1 to 1. Item time-intensity is sampled from the uniform distribution from 3 to 4. This setting corresponds to the finding from empirical data (van der Linden, Scrams & Schnipke, 1999). Meanwhile, we fixed the discrimination $\alpha_j = \alpha = 1.875$ to be the same across all items, which was chosen following the empirical results of van der Linden (2006). With the same length of test, the order of the item is shuffled randomly for each student independently. If the remaining time is less than 0.01 seconds, the nonreacted items will have missing RT records. The total time limit equals 60 times the number of items. With these simulated items and students' parameters, we generate the RTs using the lognormal-PT model.

Table 1 summarized the performance of parameter recovery. For estimation bias, we calculate the root mean square errors (RMSE) for item time intensity, person speed, and item position effect. As expected, the lognormal-PT model has the best estimation performance for item and person parameters, while the lognormal model performs worst. The estimation performance of lognormal-PT tends to be better when more students or items are included. Compared with other models, the advantage of lognormal-PT is more obvious when the number of students and items is large. In most cases, adding adaptive truncation significantly improves the performance of parameter recovery, compared with the conventional lognormal model. In this simulation, we assume the item time-intensity change linearly and match the specification of lognormal-P. We expect to see the performance of lognormal-P becomes worse if the item position effect does not change in a linear format.

Table 1. Summary of Root Mean Square Error

# Item	# Students	Model	Root Mean Square Error		
			Item Time Intensity	Latent Speed	Item Position Effect
10	200	Lognormal	.293	.174	NA
		Lognormal-T	.207	.035	NA
		Lognormal-P	.303	.091	.047
		Lognormal-TP	.204	.032	.007
	1000	Lognormal	.276	.196	NA
		Lognormal-T	.185	.071	NA
		Lognormal-P	.280	.063	.042
		Lognormal-TP	.183	.054	.004
	2000	Lognormal	.223	.382	NA
		Lognormal-T	.204	.297	NA
		Lognormal-P	.231	.035	.025
		Lognormal-TP	.185	.024	.002
20	200	Lognormal	.141	.327	NA
		Lognormal-T	.125	.263	NA
		Lognormal-P	.142	.069	.007
		Lognormal-TP	.123	.046	.001
	1000	Lognormal	.128	.766	NA
		Lognormal-T	.124	.072	NA
		Lognormal-P	.125	.043	.001
		Lognormal-TP	.123	.036	.001
	2000	Lognormal	.167	.183	NA
		Lognormal-T	.126	.094	NA
		Lognormal-P	.177	.062	.013
		Lognormal-TP	.121	.018	0
30	200	Lognormal	.106	1.454	NA
		Lognormal-T	.091	.638	NA
		Lognormal-P	.093	.066	0
		Lognormal-TP	.090	.057	0
	1000	Lognormal	. 112	1.095	NA
		Lognormal-T	. 103	.130	NA
		Lognormal-P	. 106	.045	0
		Lognormal-TP	.102	.049	0
	2000	Lognormal	. 121	.351	NA
		Lognormal-T	. 097	.308	NA
		Lognormal-P	. 111	.029	.004
		Lognormal-TP	.097	.013	.001

Real Data Analysis

We used one specific sample of the United States drawn from the computer-based assessment (CBA) Program for International Student Assessment (PISA) 2015 computer-based items. Without loss of generality, we focus on the students who took a survey about scientific and reading literacy (i.e., test forms from 31 to 66), which made up approximately 96% of the participants. Each student received a form that consisted of four 30-minute clusters assembled from two domains: science and reading. Rotation design is also applied to create each cluster. Thus, different items are sampled for different students under different item position orders. To ensure the true response sequence match the designed response sequence of PISA, we remove the records if the student revisits any item. After listwise delectation of missing or invalid records of RTs, there are 904 students and 184 unique items from the science domain. The degree of convergence of a random Markov Chain can be estimated using the Gelman-Rubin convergence statistic (i.e., $\hat{R}$) and all estimators should have the statistics smaller than 1.1.

Figure 1 includes four visualizations to summarize the main features of RTs. The first plot indicates the distribution of RT across all students and items. The density line in the first plot is the lognormal distribution with the empirical mean and standard deviation of observed log RTs. In general, lognormal distribution matches the histogram well. The second plot is the distribution of total response time across students. The vertical line (intercept = 7,200 seconds) is the total time limit. It may be because PISA is a low-stake assessment, most students did not make full use of the testing time. For the third plot, we calculate the median response time for each item position across items and students. Based on the evidence of the linear regression line, we expect to see a smaller median response time for later items. The last plot represents the relationship between the item position and median remaining RTs across items and students. As expected, the remaining RTs decrease smoothly as the position increase.

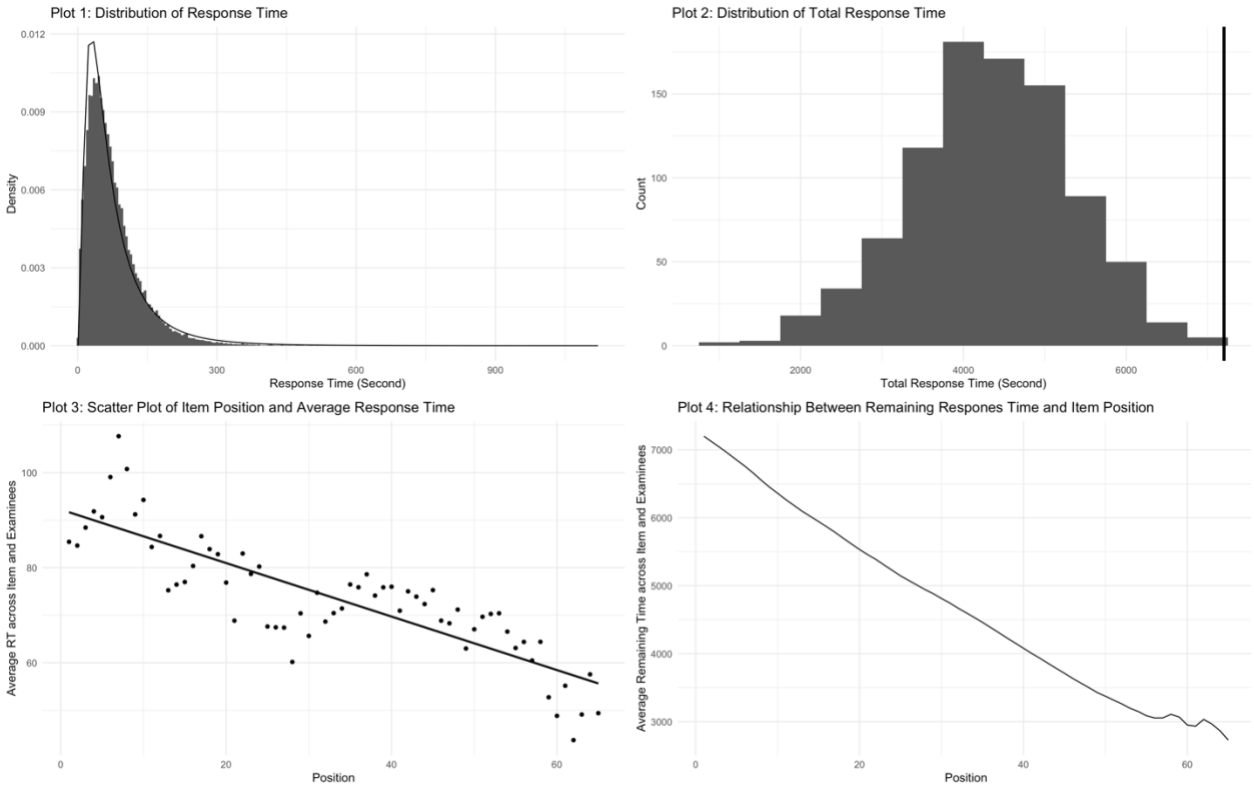

Figure 1. Response Time Analysis through Visualization

Four models formats (i.e., lognormal, lognormal-P, lognormal-T, and lognormal-PT) are applied in this study. Meanwhile, we also compare the specification of the item position effect in formula (5) and formula (6). Formula (5) assumes that the changes in time-intensity follow a linear trajectory, while formula (6) assumes the change in time-intensity depends both on the student. Alternatively, formula (6) could also be used for describing the differential personal speed. Figure 2 indicates the latent person speed (Plot 1) and item time-intensity (Plot 2) trajectory for 100 random sampled students and items. As mentioned, latent speed trajectory is generated from the lognormal-PT model with item position effect specified as formula (6), while the latent time-intensity effect is generated from the lognormal-PT model with item position effect specified as formula (5). 76.65% of a person's linear weight of the position and 100% of the item's linear weight

of position is negative. This result matches our finding in Plot 3 of Figure 1.

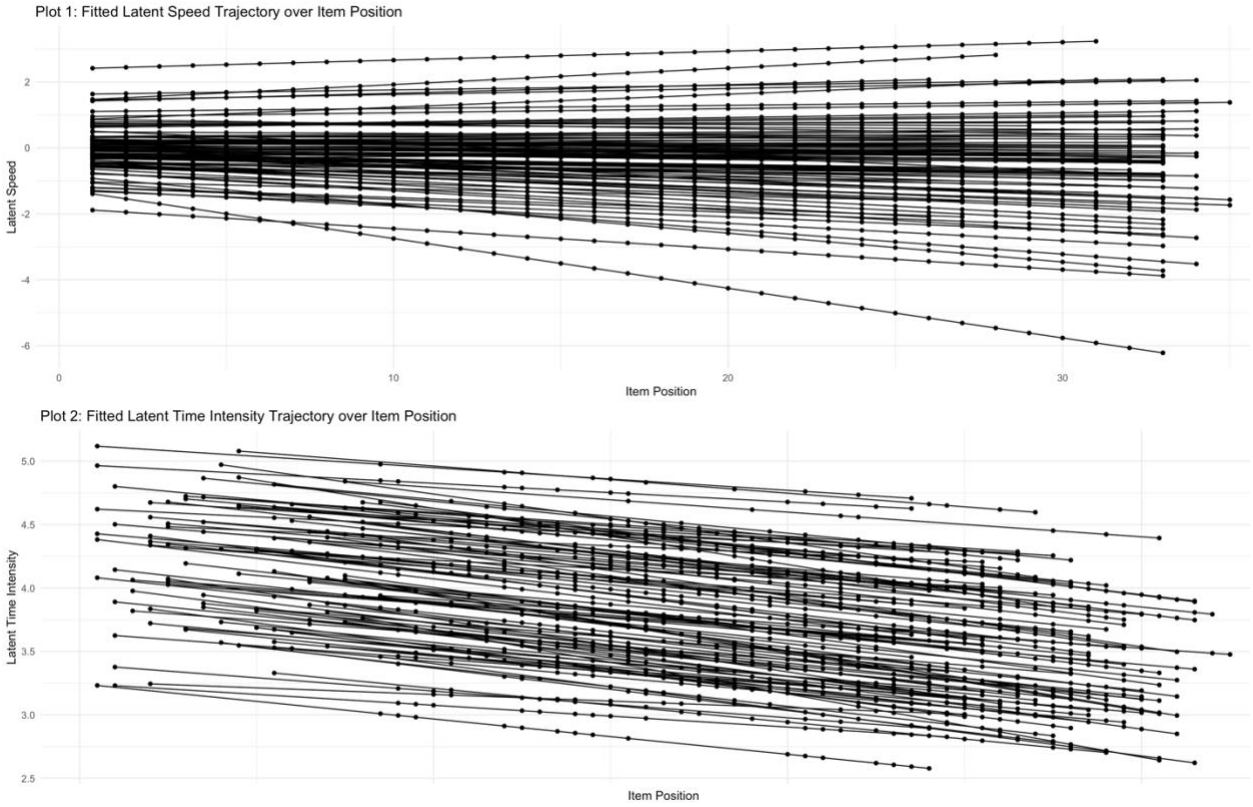

Figure 2. Fitted Latent Speed and Time Intensity Trajectory over Item Position (N=100)

The x-axis in Figure 3 indicates the fitted latent speed for the lognormal-PT model, while the y-axis represents the fitted latent speed for the lognormal, lognormal-P, and lognormal-T models distinguished by respective point shapes. The first plot focuses on the estimate of latent speed, whereas the second plot focuses on the estimation of item time-intensity. The estimates from models with and without truncation almost exactly overlap in both plots, regardless of whether the item position effect is included. This pattern may well be explained in part by the fact that the majority of students do not reach time limit (see Plot 2 in Figure 1). In particular, under low-stack settings, students are often not under time constraints, and hence the effect of overall time constraints is negligible. In the case of latent speed, the estimate bias is clear, but the direction of the bias is equally likely to be positive or negative. However, neglecting the item position impact

tends to overstate the item time intensity, implying that the linear weight of the position effect should be positive and that students spend fewer RTs in later positions. This conclusion confirms that PISA is a low-stack assessment for the sampled students. In summary, the item position effect is unignorable for this real data since it led to significant effects in the estimation of both item and person parameters. While adding truncation does not improve the model fit significantly for the low-stake assessment.

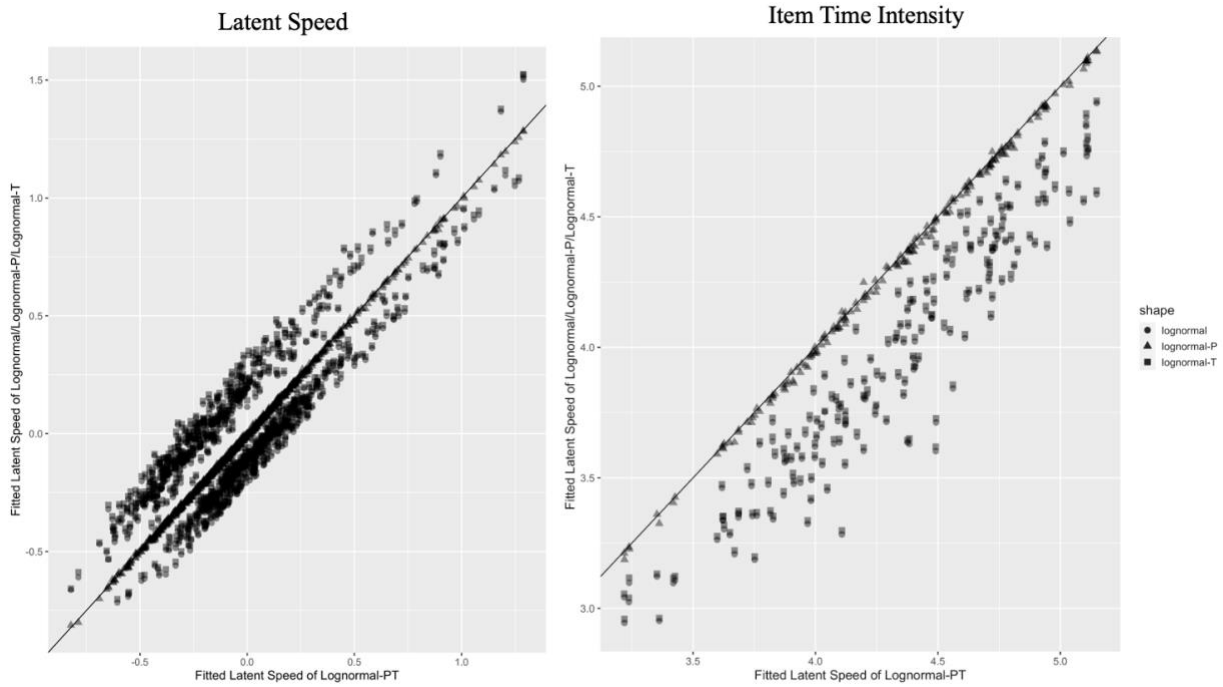

Figure 3. Scatter Plot of Fitted Latent Speed of lognormal-PT and other models

Discussion

We presented a general framework for modeling sequential RTs with total time restraints in this paper. According to evidence from both simulation studies and real-world data analysis, incorporating truncation and item position effects has the potential to enhance model fit and violate the stationarity of latent parameters and local independence assumptions associated with standard RT treatment. Students may take multiple test forms and the position of each item is not similar

depending on the rotation design or computer adaptive testing technology used for assessment. Under this circumstance, incorporating item position effects and truncation can eliminate the lognormal model's systematic noise. In contrast to item position parameters, adaptive truncation makes no assumptions about the total item time-possible intensity's variation between positions. However, when students are not under time constraints owing to low motivation or high time constraints, the effect of adaptive truncation on both item and person parameter estimates may be minimal.

It is critical to minimize context effects in linking and equating research and practice in order to acquire reliable item parameters that perform consistently across administrations (Yen, 1980). The assumption of item parameter invariance is critical for connecting and equating methods that use similar items (Kolen & Brennan, 2004; Meyers, Miller, & Way, 2009). To address possible bias in RAs, item position impacts on item difficulty are examined and discussed using IRT (Debeer & Janssen, 2013) or GLMM (Weirich, Hecht, & Böhme, 2009). If a trade-off exists between speed and accuracy, neglecting time intensity and its variation with item position would be skewed. This paper establishes a general framework for modeling the effects of item position on item parameters in RT, which may be utilized to enhance linking and equating. Meanwhile, additional elements may need to be explored in future research to adequately understand the sequential process of RTs. For instance, let us suppose that each item is only visited once throughout the evaluation. However, students may choose to solve the easier problems first and then go on to the more difficult ones after reading the item. How much time RTs spend on revisiting items and if this increases RAs remain unknown. The time limit, the sequence of visits, and the RTs all provide information on the various testing procedures.

While we describe RT modeling in terms of the lognormal model, the suggested handling of truncation and item position effects is applicable to different RT frameworks. For instance, future research may use the lognormal-PT model to assess the speed-ability trade-off. Partchev et al. (2013) found from posterior time-limit analyses of reasoning data that timed tasks always assess a mix of speed and ability. Van der Linden (2009) developed a hierarchical framework for concurrently modeling latent speed and ability. By integrating item position effects and truncation in the modeling of RAs and RTs, we are able to eliminate the possibility of effective ability confusing with the speed choice. Collaborative efforts including a range of skills are required for building tests, optimizing relevant psychometric instruments for data analysis, and evaluating process results for decision making. We hope that this work establishes a viable new route for future research aimed at improving RT measurement.

Reference

Debeer, D., & Janssen, R. (2013). Modeling Item-Position Effects Within an IRT Framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY

Fox, J. P., Entink, R. K., & Linden, W. V. D. (2007). Modeling of Responses and Response Times with the Packagecirt. *Journal of Statistical Software*, *20*(7). https://doi.org/10.18637/jss.v020.i07

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-ba117a

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608–626. doi:10.1037/a0034716

Goldhammer, F. (2015). Measuring, ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement*, 13, 133–164. https://doi.org/10.1080/15366367.2015.1100020.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8, 147

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74(1), 21–48. doi:10.1007/s11336-008-9075-y

Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practice*. New York, NY: Springer.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*, 50, 31 1-327.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, 69, 232-2

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, 3, 359–379.

Maris E (1993). "Adaptive and Multiplicative Models for Gamma Distributed Variables, and Their Application as Psychometric Models for Response Times." *Psychometrika*, 58, 445–469.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4), 615–633. doi:10.1007/s11336-012-9288-y

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education*, 22, 38-60.

Partchev, I., De Boeck, P., & Steyer, R. (2013). How much power and speed is measured in this test? *Assessment*, 20(2), 242–252. doi:10.1177/1073191111411658

Ranger, J., & Kuhn, J. T. (2011). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31–47. https://doi.org/10.1007/s11336-011-9231-7

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, 26, 271-285

Roskam E.E. (1997). "Models for Speed and Time-Limit Tests." In WJ van der Lin- den, RK Hambleton (eds.), "*Handbook of Modern Item Response Theory*," pp. 187–208. Springer, New York.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68, 589–606.

Samejima F (1973). "Homogeneous Case of the Continuous Response Level." *Psychometrika*, 38, 203–219.

Scheiblechner H (1979). “Specific Objective Stochastic Latency Mechanisms.” *Journal of Mathematical Psychology,* 19, 18–38.

Shi JQ, Lee SY (1998). “Bayesian Sampling-based Approach for Factor Analysis Models with Continuous and Polytomous Data.” *British Journal of Mathematical and Statistical Psychology*, 51, 233–252.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002). “Bayesian Measures of Model Complexity and Fit.” *Journal of the Royal Statistical Society B*, 64, 583–639.

Thissen D (1983). “Timed Testing: An Approach Using Item Response Theory.” In DJ Weiss (ed.), “Latent Trait Test Theory and Computerized Adaptive Testing,” pp. 179–203. Academic Press, New York.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. New York, NY: Teachers College Bureau of Publications.

Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford, UK: Oxford University Press.

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, 70(2), 359–376. doi:10.1007/s11336-003-1078-0

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). “Using Response-Time Constraints to Control for Speededness in Computerized Adaptive Testing.” *Applied Psychological Measurement*, 23, 195–210.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 195–210. https://doi.org/ 10.1177/01466219922031329

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

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. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3), 247–272. doi:10.1111/j.1745-3984.2009.00080.x

van Rijn, P. W., & Ali, U. S. (2017). A Generalized Speed–Accuracy Response Model for Dichotomous Items. *Psychometrika*, *83*(1), 109–131. https://doi.org/10.1007/s11336-017-9590-9

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for timelimit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York, NY: Springer.

Verhelst ND, Verstraalen HHFM, Jansen MG (1997). "A Logistic Model for Time Limit Tests." In WJ van der Linden, RK Hambleton (eds.), "Handbook of Modern Item Response Theory," pp. 169–185. Springer, New York.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling Item Position Effects Using Generalized Linear Mixed Models. *Applied Psychological Measurement*, *38*(7), 535–548. https://doi.org/10.1177/0146621614534955

Yen, W. M. (1980, March). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17, 297-311.