博士論文

New item response models for multi-alternative forced choice questionnaire data with response time

(回答時間データを利用した多肢強制選択型質問紙データの項目反応モデルの開発)

分寺杏介

(Bunji, Kyosuke)

Contents

List of Figures

List of Tables

Abstract

The Likert scale is the most frequently used scale format in personality measurement. However, the Likert scale is subject to particular response biases. One promising solution to response biases is to use a forced-choice format. Additionally, there is a growing interest in using response time (RT) information. However, many models that use the RT do not consider underlying cognitive processes. The main objective of this series of studies was to investigate the combination of item response theory and cognitive process models, particularly for forced-choice paradigms.

In Study 1, a combination of the Thurstonian IRT model and the D-diffusion IRT model was proposed. The Thurstonian IRT model has been successfully used to analyze forced-choice personality measurements by means of Thurstone's Law of Comparative Judgment. In contrast, the D-diffusion IRT model is a novel cognitive-process model that can naturally use RT information, and thereby estimate concurrently item and respondent parameters. Consequently, two-alternative forced-choice personality measurement data with RT information can be analyzed using a combination of the models, namely the Thurstonian D-diffusion IRT model. In addition, the success of this model would support the applicability of the diffusion model to personality measurement, even though the original diffusion model did not focus on this type of assessment.

Study 2 investigated a combination of the IRT model and linear ballistic accumulator (LBA) model. The LBA model is considered to be simpler than the diffusion model, yet the parameters in the LBA model can be interpreted as the corresponding parameters in the diffusion model. Therefore, parameter decomposition of the LBA model was proposed similarly to the D-diffusion IRT model. The proposed model was named the (unidimensional binary) D-LBA IRT model. Several simulation results revealed that the proposed D-LBA IRT model can estimate parameters more quickly and efficiently than the D-diffusion IRT model. Moreover, simulation and experimental results showed that the D-LBA model is expected to be more robust than the D-diffusion IRT model when the true model is unknown.

Study 3 examined an extension of the D-LBA IRT model, namely, the multiple-alternative multidimensional D-LBA IRT model. To date, there exist no models that can be used to

multiple-alternative forced-choice personality measurement with RT information. A real data example showed that the proposed model can generate slightly different trait scores compared to other models that do not use RT information. Besides, the proposed model appears more robust than the Thurstonian D-diffusion IRT model, even for two-alternative forced-choice data.

The three proposed models are designed for different situations; yet, the success of these models suggests the possibility of using cognitive process models in the field of psychometrics.

Chapter 1

Introduction

1.1 Likert Scale and Forced-Choice

The Likert scale (Likert, 1932), in which respondents express the level of agreement towards an item by choosing one of the ordinal choice categories, is perhaps one of the most widely used instruments to measure respondents' personality, attitude, and other individual psychological differences. It is not very evident why it became very popular in psychological and educational assessments, but one of the reasons may be attributed to a simple fact that, as Likert (1932, p. 42) mentioned, it is less laborious to construct the scales and to analyze the obtained data. Conceptually, the Likert scale is able to assess the respondents' attitude or other psychological properties.

However, it is also known that the data obtained from a Likert scale tend to be affected by response biases, which is "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991, p.17). In fact, many previous studies have shown that measurement with the Likert scale can be distorted when the respondents have some particular response biases (e.g., Chen, Lee, & Stevenson, 1995; Ferrando & Lorenzo-Seva, 2010; Kam & Meyer, 2015; Masuda & Sakagami, 2017; Phelps, Schmitz, & Boatright, 1986; Rammstedt & Farmer, 2013; Viswesvaran & Ones, 1999). Examples include the uniform response biases, which consist of several categorical tendencies to distort a

response in particular directions regardless of the content (He, Bartram, Inceoglu, & van de Vijver, 2014), and the socially desirable response bias, which is a tendency to give answers that make the respondent look good (Paulhus, 1991). These response biases are considered as one of the main sources of measurement error and may therefore harm the validity of the measurement (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

One promising solution for dealing with response biases is to move away from the Likert measurement and instead present items in a comparative manner. In the multiple-alternative forced-choice (MAFC) format, several statements are simultaneously presented as a questionnaire item. Respondents are asked to choose one statement that best describes themselves, or to rank the statements. An example of two-alternative forced choice items is the pair of statements "*Get stressed out easily*" and "*Don't talk a lot*." Each statement in an MAFC questionnaire item corresponds to different dimensions of psychological traits to be measured, such as emotional stability and extraversion in the preceding example. The MAFC format should reduce the impact of uniform response biases because it makes it impossible to process uniformly elevated or decreased judgment across all items (A. Brown & Maydeu-Olivares, 2018; Cheung & Rensvold, 2002). In addition, there is evidence that the use of the MAFC format can also reduce the influence of socially desirable responding and improve predictive validity (e.g., Cao & Drasgow, 2019; Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000; Salgado & Táuriz, 2014).

One of the common methods of analyzing the forced-choice format data in the classical test theory is the binary scoring (e.g., Baron, 1996; Ghiselli, 1954; Hirsh & Peterson, 2008; Meade, 2004). In typical binary scoring, each choice option is scored as 1 when the option is chosen, and 0 when not. Of course, each respondent can obtain only one point from each item when they are asked to pick out the most suitable statement. This means when the forced-choice responses are coded as binary scores, the resultant variables become *ipsative*, that is, the sum of all dimension scores on a questionnaire is always a constant (typically the same as the number of items) across all respondents (Hicks, 1970). This leads to a number of psychometric challenges (e.g. Cattell, 1944; Cornwell & Dunlap, 1994; Meade, 2004). For instance, the summed

binary scores are relative in nature and cannot be used for absolute comparison between respondents (Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988). Furthermore, the construct validity of the summed binary scores should be distorted. This is because the average intercorrelation between D latent trait dimensions with equal variances analytically equals $-1/(D-1)$, regardless of the true relationship among them (A. Brown, 2010; Hicks, 1970). Similarly, conventional reliability coefficients such as the coefficient alpha (Cronbach, 1951) should also be distorted. This happens because the key assumptions in their derivation, such as consistent coding, which refers to the assumption that the observed score is always higher when the true score is higher, are not met by the summed binary scores (A. Brown & Maydeu-Olivares, 2011).

A promising solution to the problem of ipsative data is to apply the framework of item response theory (IRT; Lord, 1980; Rasch, 1960) instead of summing the binary-coded scores. For this purpose, A. Brown and Maydeu-Olivares (2011) proposed a novel and general IRT model of forced-choice format observations called the Thurstonian IRT (TIRT) model by extending Thurstone's Law of Comparative Judgment (Thurstone, 1927) to an IRT formulation. By incorporating a Thurstone's concept that is also applicable for forced-choice data (Thurstone, 1931), the TIRT model succeeded in simultaneously estimating both item and respondent parameters and is free from the abovementioned problems of binary-coded ipsative data. In the wake of this development, personality scales of forced-choice format have been actively developed in recent years (e.g. Anguiano-Carrasco, MacCann, Geiger, Seybert, & Roberts, 2015; Joubert, Inceoglu, Bartram, Dowdeswell, & Lin, 2015; Guenole, Brown, & Cooper, 2018; P. Lee, Lee, & Stark, 2018).

To date, several models (e.g., Hontangas et al., 2015; H. Lee & Smith, 2019; Revuelta, 2014; Stark, Chernyshenko, & Drasgow, 2005; W.-C. Wang, Qiu, Chen, Ro, & Jin, 2017) have been proposed that are considered to be readily applicable to data collected with existing multidimensional MAFC format questionnaires. Still, TIRT is the most well-known and most-used model in the field of psychometrics.

1.2 Response Time

In recent years, personality assessments have been widely administered via computers and tablets. Through such computerized tests, researchers can obtain not only the item responses but also the response times (RTs) of respondents. Considering the fact that every testing situation typically involves at least some time pressure, the time required to respond to items can reveal useful information about respondents' traits (Tuerlinckx, Molenaar, & van der Maas, 2016).

Various studies have discussed the utility of incorporating RT information in, e.g., classifying response behaviors, removing aberrant responses, and improving the estimation accuracy (e.g., Bertling & Weeks, 2018; Kong, Wise, & Bhola, 2007; C. Wang, Xu, & Shang, 2018; Wise & Kong, 2005). The general validity of online measurement has been demonstrated for both personality characteristics and RTs (e.g., Condon, 2018; Raz, Bar-Haim, Sadeh, & Dan, 2014).

In personality measurement, the relationship between the RT and the respondent's latent trait is characterized by the *inverted-U* relationship (Akrami, Hedlund, & Ekehammar, 2007; Ferrando & Lorenzo-Seva, 2007a; Kuiper, 1981; Ranger & Ortner, 2011). This indicates that a respondent processes self-relevant information faster if they are more extreme (high or low) from the viewpoint of that respondent's trait level. For the personality measurement in binary and Likert scale formats, several studies have developed the joint models of item response and RT by incorporating the inverted-U relationship (e.g., Ferrando & Lorenzo-Seva, 2007b; Meng, Tao, & Shi, 2014; Ranger, 2013; Ranger & Ortner, 2011).

Tuerlinckx et al. (2016) conceptually classified previously proposed models for dealing with both the item responses and RTs into three types. First, there exist models that regard the RT as collateral information. These models are primarily concerned with the item response model and use the RT to improve the efficiency of parameter estimation or to provide additional information such as the detection of cheating and guessing. For example, Roskam (1987, 1997) added the logarithm of the RT as an independent variable to the logistic item response model. This class of models essentially represents the probability of a correct item response and uses the RT merely as a source of additional information. The second type of model includes normative models, which typically employ some scoring rules. For example, van der Maas and Wagen-

makers (2005) introduced the correct item summed residual time (CISRT) scoring rule, which assigns the sum of all residual times for correct responses as the score of the respondent for the measurement of chess expertise. Although scoring rules are popular in games and sports, they are not commonly adopted in psychometrics because it is difficult to establish reasonable and widely acceptable scoring rules.

The third type of model includes process models, which make specific assumptions about the underlying psychological process that result in the item responses and RTs. Such models enable us to apply accumulated scientific knowledge in psychology for the direct modeling of the relationships between the observed test performance and the inner psychological process that drives the item-answering behavior. By contrast, the first and second types of models mentioned above have no obvious connection to the psychological models for representing the underlying mechanisms of such information processing; hence, it is difficult for them to consider the underlying cognitive process of the item response and RT. Nevertheless, most previously proposed RT-incorporated item response models are of the first or second type (van der Linden, 2016). As Schulte-Mecklenbeck et al. (2017) noted, "Decision research has experienced a shift from simple algebraic theories of choice to an appreciation of mental processes underlying choice." This means there would be a growing interest in considering the third type of item response models.

1.3 Cognitive Process Model

In the field of cognitive and mathematical psychology, which has a rich tradition of RT modeling (Luce, 1986; Voss, Nagler, & Lerche, 2013), psychological models have been developed to explain the course of information processing that leads to the observed human response and the associated RT. Two of the best-known models are the diffusion model and the linear ballistic accumulation (LBA) model. By introducing several cognitive parameters, as elaborated in the next chapter, the diffusion and LBA models address the tradeoff between the response accuracy and the speed, thereby facilitating the quantification of individual performance. The

diffusion model, originally formalized by Ratcliff (1978) on the basis of preceding studies such as M. Stone (1960) and Laming (1968), is a representative model that considers the underlying response generation process. By contrast, the LBA model was proposed by S. D. Brown and Heathcote (2008), who aimed to propose a simple yet effective alternative to the well-known models in the field of cognitive psychology. They demonstrated that the LBA model accounts for many important empirical phenomena from choice tasks, such as the speed difference between correct and incorrect responses and the shape of the speed–accuracy tradeoff function.

By explicitly taking into account the cognitive process underlying the observed responses, researchers can obtain psychologically meaningful parameter estimates such as the speed of information uptake and the amount of information used to make a decision. This helps us to quantify and empirically test psychological theories in an applied context. In fact, process models have been successfully applied to understand and explain the processes underlying a variety of topics in human sciences, such as memory, familiarity effects, perceptual judgments, and decision making (see van Ravenzwaaij & Oberauer, 2009). For example, Ratcliff, Thapar, and Mckoon (2007) applied a process model in order to examine why elder people take more time than the younger ones in lexical decision tasks. Their finding based on the obtained parameter estimates is that this is not due to the decreased rate of information processing, but because elder respondents set their response boundaries more conservatively. Palada et al. (2016) also applied process models to response data with complex stimuli. In their study, respondents tended to perform the task faster when the stimulus is more complex. Based on the parameter estimates of process models, they found out that this is caused not because the respondents' capacity increases with the complexity of stimuli (which is a phenomenon called "super-capacity"), but because respondents set lower threshold for response when the task is complex. As Donkin, Brown, Heathcote, and Wagenmakers (2011, p. 61) state, "this kind of conclusion would have been very difficult to draw without a cognitive model of choice RT."

The main objective of this thesis is to develop new cognitive process IRT models that can be applied to forced-choice personality measurement data with RT information. In addition, simulation and empirical studies are conducted to characterize the performance of the proposed

models and to compare the models with selected extant models. The model development process consider the possibility of combining IRT and two cognitive process models: the diffusion and LBA models.

The remainder of this thesis is organized as follows. In Chapter 2, we briefly review selected extant models relevant to the studies in this thesis. The series of studies is fundamentally structured as the combination of extant frameworks.

Chapter 3 extends the D-diffusion IRT model, which is a cognitive-based IRT model previously proposed by Tuerlinckx and De Boeck (2005), for two-alternative forced-choice questionnaire items. Similar to the TIRT model, the proposed model considers the drift rate as the difference of latent utilities of both choice alternatives. The proposed extension is thus named the Thurstonian D-diffusion IRT (TDIRT) model.

Chapter 4 considers a combination model of the IRT and LBA models, namely the (unidimensional binary) D-LBA IRT model. The LBA model is one of the most complete cognitive models that can be applied to tasks involving more than two choice alternatives. To provide a new cognitive-based IRT model for MAFC questionnaire data with RT information in Chapter 5, this chapter first investigates the combination of IRT and LBA models and proposes the D-LBA IRT model as a preparatory study. The performance of the D-LBA IRT model is compared with that of the D-diffusion IRT model.

In Chapter 5, a new cognitive-based model is proposed that is applicable to multidimensional MAFC personality measurement data with RT information, namely the multidimensional MAFC D-LBA IRT model. Performance of the proposed model is compared with that of the TIRT, multidimensional nominal response model (Revuelta, 2014), and TDIRT models. To the best of my knowledge, there exist no models that address MAFC questionnaire data with the RT information, even considering psychometric or algebraic models. The proposed multidimensional MAFC D-LBA IRT model is therefore a novel model in the field of psychometrics and personality measurement.

Finally, a summary of the series of studies and future directions is discussed in Chapter 6.

Chapter 2

Existing Models

2.1 Diffusion Model

The diffusion model provides us with detailed information about the cognitive process underlying the respondent's answers to items based on the observed data. Figure 2.1 shows a schematic of the diffusion model. When an item is presented to the respondent, the cognitive information required to answer the item (or conviction towards the item response) accumulates over time from the starting point (z). Once it reaches the upper (α) or lower (0) boundary, the respondent answers the item. The upper boundary corresponds to the correct response, and the lower boundary corresponds to the incorrect response. The increase in the amount of information in unit time follows a normal distribution with a mean v and within-trial variance s^2. Here, the mean parameter v, which indicates the average rate of information accumulation, is usually called *drift rate*.

The diffusion model has four main parameters. α represents the distance between the upper and lower boundaries. A larger value of α means that a longer time is required to answer the item, which suggests that the respondent's choice is more deliberate. z denotes the starting point. If the subject has a positive response bias to the item from the beginning, z becomes larger; thus, the starting point approaches the upper bound (α), which leads to a higher probability of giving the correct answer. When there is no response bias, i.e., when the respondent does not

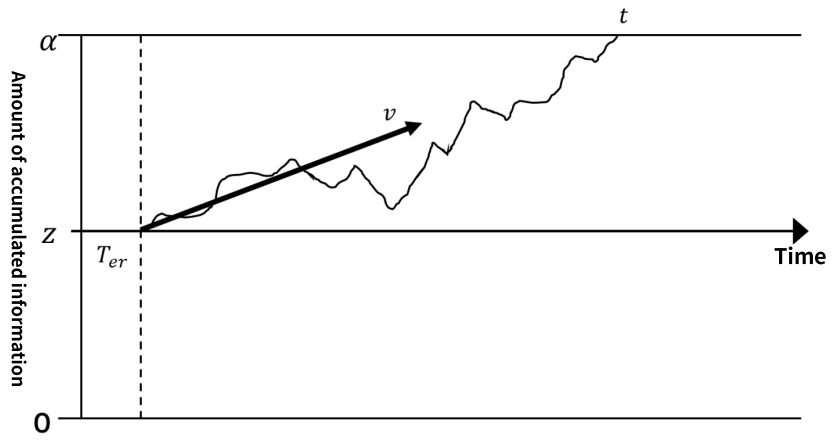

Figure 2.1: Schematic of the diffusion model. The accumulator starts from the starting point (z) and moves upwards and downwards. The direction is randomly drawn from $N(v, s^2)$ at each unit of time. The accumulation continues until it reaches upper or lower bound (α or 0), which each corresponds to different response options.

favor or avoid one of the choice alternatives at the beginning (Leite & Ratcliff, 2011), z equals $\alpha/2$. v represents the average slope of the information accumulation process. The process approaches the upper (resp. lower) bound when v is positive (resp. negative). τ represents the nonresponse time, i.e., the duration of nondecisional processes, which may comprise basic encoding processes.

Let x be a binary observed variable that takes the value 1 when the correct answer is observed (or the accumulation reaches the upper bound) and 0 when the wrong answer is observed (or the accumulation reaches the lower bound). To derive the formulation of the diffusion model, consider a discrete random walk process at first. In the process, the accumulator goes up one unit with a probability p and goes down one unit with a probability $q = 1 - p$. Let us consider the probability that the accumulator reaches the lower boundary when it starts from z, which we denote by $P(x = 0 \mid z)$. From the above assumption, $P(x = 0 \mid z)$ must satisfy the recursive equation $P(x = 0 \mid z) = pP(x = 0 \mid z+1) + qP(x = 0 \mid z-1)$. By solving this equation, $P(x = 0 \mid z)$ becomes

$$P(x = 0 \mid z) = \begin{cases} \frac{\left(\frac{q}{p}\right)^{\alpha} - \left(\frac{q}{p}\right)^{z}}{\left(\frac{q}{p}\right)^{\alpha} - 1} & \text{when } q \neq p \\ 1 - \frac{z}{\alpha} & \text{when } q = p. \end{cases} \tag{2.1}$$

In a similar manner, let us consider $P(x = 0, n \mid z)$, which is the probability that the accumulator reaches the lower boundary at the n-th step when it starts from z. $P(x = 0, n \mid z)$ must satisfy the recursive equation $P(x = 0, n+1 \mid z) = pP(x = 0, n \mid z+1) + qP(x = 0, n \mid z-1)$. As a result, $P(x = 0, n \mid z)$ becomes

$$P(x = 0, n \mid z) = \frac{2^{n+1}}{\alpha} p^{\frac{n-z}{2}} q^{\frac{n-z}{2}} \sum_{k<\alpha/2} \cos^{n-1} \frac{\pi k}{\alpha} \sin \frac{\pi k}{\alpha} \sin \frac{\pi z k}{\alpha}. \tag{2.2}$$

The diffusion process is expressed as the continuous version of this random walk process. The limiting form of the joint distribution of the item response (x) and RT (t) in the diffusion model is represented as

$$\begin{aligned} f(x,t) = {} & \frac{\pi s^2}{\alpha^2} \exp\left(\frac{(\alpha x - z)v}{s^2} - \frac{v^2}{2s^2}(t-\tau) \right) \\ & \times \sum_{m=1}^{\infty} m \sin\left(\frac{\pi m(\alpha x - 2zx + z)}{\alpha} \right) \exp\left(-\frac{1}{2} \frac{\pi^2 s^2 m^2}{\alpha^2} (t-\tau) \right). \end{aligned} \tag{2.3}$$

For more details, refer to Ratcliff (1978) and Feller (1968, chap. 14).

2.2 Diffusion IRT Model

Tuerlinckx and De Boeck (2005) extended the diffusion model to the IRT framework. Equation 2.1 indicates the probability that the accumulator reaches the lower bound in the random walk process. Again, by considering the continuous version of the random walk process, Equation 2.1 becomes

$$P(x = 0 \mid z) = \frac{\exp(-2\alpha v) - \exp(-2zv)}{\exp(-2\alpha v) - 1}. \tag{2.4}$$

In typical IRT applications (e.g., educational testing and personality questionnaires), it would be reasonable to assume that there is no response bias at the starting point. Under the assumption that z is set to $\alpha/2$ (no response bias), the probability that respondent i answers the

j-th item correctly is given as

$$P(x_{ij} = 1 \mid z_{ij} = \alpha_{ij}/2) = 1 - P(x_{ij} = 0 \mid z_{ij} = \alpha_{ij}/2) = \frac{\exp(-2z_{ij}v_{ij}) - 1}{\exp(-2\alpha_{ij}v_{ij}) - 1} = \frac{\exp(\alpha_{ij}v_{ij})}{1 + \exp(\alpha_{ij}v_{ij})}. \quad (2.5)$$

This equation corresponds to the standard two-parameter logistic IRT model when α_{ij} corresponds to the discriminability and v_{ij} corresponds to the difference between the respondent's trait and the item difficulty parameters.

Note that in this thesis, following Tuerlinckx and De Boeck (2005) and van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011), we use the subscripts i and j for the parameters of the IRT-based models but suppress them for the original diffusion and LBA model parameters. This is because, whereas the separation between item and respondent parameters is the key characteristic of IRT models, the original diffusion and LBA models essentially do not distinguish them. In the diffusion IRT model, this separation can be represented as

$$\alpha_{ij} = w(\gamma_i,\, \xi_j), \qquad\qquad v_{ij} = u(\theta_i,\, b_j), \quad (2.6)$$

where γ_i and ξ_j represent the respondent factor (e.g., deliberateness and potential speed to answer) and the item factor (e.g., item complexity and item length) involved in the boundary parameter, respectively. θ_i and b_j can be interpreted in almost the same manner as those in the original IRT model. That is, θ_i represents the respondent's latent trait to be measured, and b_j represents the item threshold (difficulty or severity level).

IRT models are typically used for two different types of psychological measurement: personality measurement such as the Big Five scale and ability measurement such as college entrance examinations. Corresponding to these two scenarios, two types of diffusion IRT models with different functional forms have been developed.

In the first type—namely, the D-diffusion IRT model, v is expressed as the *difference* between the item threshold and the respondent's latent trait. In this model, the specific form of

Equation 2.6 is given as

$$\begin{cases} \alpha_{ij} = \dfrac{\gamma_i}{\xi_j} & with \;\; \gamma_i, \xi_j \in \mathbb{R}_{>0} \\ v_{ij} = \theta_i - b_j & with \;\; \theta_i, b_j \in \mathbb{R}, \end{cases} \tag{2.7}$$

where α is the ratio of the respondent properties (e.g., attentiveness and deliberateness) and the item properties (e.g., complexity and length). As a result, Equation 2.5 is rewritten as

$$P(x_{ij} = 1 \mid \theta_i, \gamma_i) = \frac{\exp\left(\dfrac{\gamma_i}{\xi_j}(\theta_i - b_j)\right)}{1 + \exp\left(\dfrac{\gamma_i}{\xi_j}(\theta_i - b_j)\right)}, \tag{2.8}$$

which is the form of the two-parameter logistic IRT model except that the discrimination parameter γ_i/ξ_j depends on both the respondent and item. τ_j retains the same meaning as in the original diffusion model, and is basically considered as an item parameter in this thesis. In this way, the D-diffusion IRT model generates important predictions for personality measurement. In the original diffusion model, the expected RT is the largest when $v = 0$. This means that the expected RT is large when the respondent's latent trait is close to the item threshold. This is consistent with Ferrando and Lorenzo-Seva's (2007a) model, which extends Thissen's (1983) model on the basis of the distance–difficulty hypothesis (Ferrando & Lorenzo-Seva, 2007a) of personality measurement. Because the D-diffusion IRT model generates many important predictions for personality measurement as described above, it is typically used for personality tests.

The second type of diffusion IRT model is the Q-diffusion IRT model, which expresses v as the *quotient* of the item threshold and respondent's ability. In this model, the specific form of

Equation 2.6 is given as

$$\begin{cases} \alpha_{ij} = \dfrac{\gamma_i}{\xi_j} & with \;\; \gamma_i, \xi_j \in \mathbb{R}_{>0} \\ v_{ij} = \dfrac{\theta_i}{b_j} & with \;\; b_j \in \mathbb{R}_{>0}, \;\; \theta_i \in \mathbb{R}_{\geq 0}. \end{cases} \tag{2.9}$$

This model corresponds to typical applications in ability measurement for which it has a number of attractive properties. For example, the expected RT is the largest when $\theta_i = 0$ because θ_i and b_j are restricted to be nonnegative and positive, respectively. This corresponds to the assumption that a more competent respondent tends to answer faster than a less competent one. In addition, when a respondent has the lowest competence and answers a two-alternative forced choice item at random, the probability of obtaining correct response should equal 50%. In Equation 2.9, the lower bound of v_{ij} is 0 (when $\theta_i = 0$), and in this case, the probability of reaching upper bound (choose correct response) is exactly 50%. In this manner, the Q-diffusion model takes into account the effect of guessing. Refer to van der Maas et al. (2011) and Tuerlinckx et al. (2016) for further details regarding the difference in D- and Q- diffusion models.

2.3 Linear Ballistic Accumulation Model

S. D. Brown and Heathcote (2008) proposed a simple cognitive model called the LBA model, which is schematically illustrated in Figure 2.2. In the LBA model, information regarding a certain choice alternative of an item is linearly accumulated with time, whereas the amount of accumulation is normally distributed in the diffusion model. Once the item is presented to the respondent, the evidence toward each choice alternative of the item accumulates independently (the accumulation of one choice alternative is irrelevant to that of any other choice) and linearly (the amount of accumulation does not change with time). This means that the model assumes no within-trial variance ($s^2 = 0$). Instead, the LBA model introduces the between-trial variance η^2, which indicates that the amount of information accumulation over time varies between each trial, even if a respondent answers the same item repeatedly. When the information for any

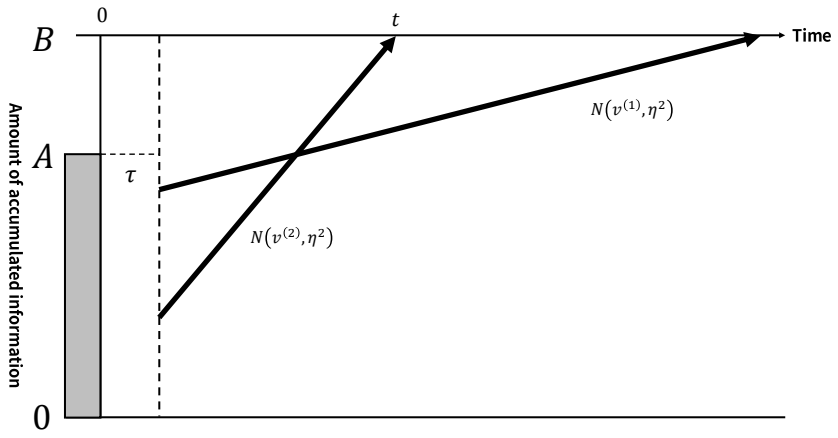

Figure 2.2: Schematic of the LBA model. Contrary to diffusion model, each response option has corresponding different accumulator. The starting points are i.i.d random variables with $U(0,A)$. The slopes are randomly drawn from $N(v^{(k)},\eta^2)$ and do not fluctuate within a trial. t is the time when any accumulator reaches the upper bound (B). The resulting RT is the summation of t and τ.

one choice reaches the boundary (B), a corresponding response is provided by the respondent. The starting point of evidence accumulation for each choice alternative is a random realization between 0 to A, and the amount of evidence accumulated in unit time is a realization from a normal distribution with a mean v and between-trial variance η^2. All choices have A and B in common, whereas v differs among the choices.

In terms of the relationship with the diffusion model, v, τ, and $B-\frac{A}{2}$ in the LBA model can be interpreted in a similar manner as v, τ, and α in the diffusion model, respectively. However, the LBA model differs from the diffusion model in two major aspects. First, each choice has its own v. On the contrary, in the diffusion model, the information for each choice accumulates dependently; for example, approaching one choice (e.g., the upper bound) indicates drifting away from the other choice (i.e., the lower bound), which is shown in Figure 2.1. Hence, the diffusion model has only one drift rate parameter (v). In addition, because of the problem of scaling, the standard deviation of the amount of information accumulation is typically fixed. Second, the parameter A indicates the upper bound of the starting point, whereas z in the diffusion model corresponds to the starting point of information processing.

To derive the LBA model, let c be the random value derived from the uniform distribution

$U(B-A,B)$. This c represents the distance from the start point, which is randomly derived from $U(0,A)$, to the threshold B. Moreover, let $\Delta^{(k)}$ be the random value derived from $N(v^{(k)},\eta^2)$. Then, the cumulative distribution function for the k-th choice alternative ($k = 1,\cdots,K$) and RT is given as

$$\begin{aligned} F(k,t) &= \text{prob}\left(\frac{c}{\Delta^{(k)}} < t\right) \\ &= \text{prob}(c < \Delta^{(k)}t) \\ &= \int_{-\infty}^{\infty} U(m \mid B-A,B)\phi\left(\frac{m-tv^{(k)}}{t\eta}\right)dm \\ &= \int_{B-A}^{B} \frac{m-B+A}{A}\phi\left(\frac{m-tv^{(k)}}{t\eta}\right)dm + 1 - \Phi\left(\frac{B-tv^{(k)}}{t\eta}\right), \end{aligned} \tag{2.10}$$

where $\phi(.)$ and $\Phi(.)$ represent the density and cumulative distribution functions of the standard normal distribution, respectively. By transforming Equation 2.10, the joint cumulative distribution function of the item response for the k-th response category and RT (T) is given as (for the detailed derivation, see Appendix A in S. D. Brown & Heathcote, 2008)

$$\begin{aligned} LBA_{CDF}(k,t \mid B,A,v^{(k)},\eta) = 1 &+ \frac{B-A-tv^{(k)}}{A}\Phi\left(\frac{B-A-tv^{(k)}}{t\eta}\right) - \frac{B-tv^{(k)}}{A}\Phi\left(\frac{B-tv^{(k)}}{t\eta}\right) \\ &+ \frac{t\eta}{A}\phi\left(\frac{B-A-tv^{(k)}}{t\eta}\right) - \frac{t\eta}{A}\phi\left(\frac{B-tv^{(k)}}{t\eta}\right). \end{aligned} \tag{2.11}$$

The corresponding joint probability density function is derived by the differentiation of Equation 2.11 with respect to t:

$$LBA_{PDF}(k,t \mid B,A,v^{(k)},\eta) = \frac{1}{A}\left[-v^{(k)}\Phi\left(\frac{B-A-tv^{(k)}}{t\eta}\right) + \eta\phi\left(\frac{B-A-tv^{(k)}}{t\eta}\right) + v^{(k)}\Phi\left(\frac{B-tv^{(k)}}{t\eta}\right) - \eta\phi\left(\frac{B-tv^{(k)}}{t\eta}\right)\right]. \tag{2.12}$$

These are the functions concerning the time when a certain accumulator of the choice alternative k reaches the boundary, regardless of any other alternative. However, in typical appli-

cations of the LBA model, researchers can only observe the time for a single (chosen) choice alternative to reach the threshold; the times for the other choice alternatives are not known. Therefore, the *defective* (meaning not summing to 1) distribution, which is the distribution of a certain accumulator reaching the threshold before any other ones at time t, is needed. From Equations 2.12 and 2.11, the defective density function of choice alternative k with RT t can be obtained as

$$\text{PDF}(k,t) = LBA_{PDF}(k, t-\tau \mid B, A, v^{(k)}, \eta) \prod_{l \neq k} \Big[1 - LBA_{CDF}(k, t-\tau \mid B, A, v^{(l)}, \eta)\Big]. \tag{2.13}$$

The corresponding cumulative distribution function can be obtained by the numerical integration of Equation 2.13. The probability of the respondent's choice k can be derived by the integration of Equation 2.13 over all t ($0 \leq t \leq \infty$) or by evaluating the cumulative distribution function defined from the density of Equation 2.13 at $t \to \infty$.

S. D. Brown and Heathcote (2008) have pointed out several differences between the LBA model and other decision-making models. First, the EZ-diffusion model (Wagenmakers, van der Maas, & Grasman, 2007) is simpler than the LBA model, but Wagenmakers et al. note that the EZ-diffusion model was developed for expressing data in a simple form rather than fully modeling the cognitive process. Therefore, it may not reflect all of the important features of the RT. For example, the EZ-diffusion model assumes that the RT distributions of the correct and incorrect responses are identical, which can be an unrealistically strong assumption in practice. On the other hand, the LBA model is considered as the simplest yet complete decision-making model of the RT because the model successfully accounts for important empirical phenomena of the choice RT, such as the speed–accuracy tradeoff and the relative speed of correct vs. incorrect responses (S. D. Brown & Heathcote, 2008). Similarly, other choice RT models such as the latency model (Grice, 1968) and the LATER model (Reddi & Carpenter, 2000) can be seen as simplifications of the diffusion model that assume no trial-to-trial variability among evidence accumulation processes. In these models, however, the model-predicted RT distribution of the incorrect choice is negatively skewed, and this would never occur in the observed data. This

inadequate negative skew is caused by projecting the tail of a normal distribution that generates negative slopes (S. D. Brown & Heathcote, 2008). On the other hand, the LBA model represents accumulation processes for every choice; hence, the RT distributions of all choices can be fitted well. Finally, the LBA model has simple analytic solutions for choices among any number of different alternatives.

2.4 Thurstonian IRT Model

As previously stated, the Thurstonian IRT model (A. Brown & Maydeu-Olivares, 2011) extends Thurstone's Law of Comparative Judgment (Thurstone, 1927) to the IRT framework. Let j_k be the k-th ($k = 1, \cdots, K$) statement presented in the j-th item, and Let $x_{ij}^{(l\_m)}$ ($1 \leq l, m \leq K \wedge\ l < m$) be the binary coded response of respondent i to item j, that takes 1 when the statement l is chosen over the statement m and 0 otherwise. For the TIRT model, the observed item responses to MAFC questionnaire items were first re-coded into binary forced-choice format. For instance, when respondent i chose the second statement for a four-alternative forced-choice ($K = 4$) item j, this observed item response produces three binary re-coded item responses: $x_{ij}^{(1\_2)} = 0$, $x_{ij}^{(2\_3)} = 1$, and $x_{ij}^{(2\_4)} = 1$. In this thesis, the rest of possible binary-coded responses, which are $y_{ij}^{(1\_3)}$, $y_{ij}^{(1\_4)}$, and $y_{ij}^{(3\_4)}$ in this example, are treated as missing data. This is because this thesis only focuses on the situation that respondents were asked to choose only the best statement.

In addition, let $u_{ij}^{(l)}$ and $u_{ij}^{(m)}$ be the latent utilities of respondent i for j_l and j_m, respectively. In the Thurstonian IRT model, the observed response is modeled to be determined by the difference between $u_{ij}^{(l)}$ and $u_{ij}^{(m)}$ of that respondent; that is,

$$x_{ij}^{(l\_m)} = \begin{cases} 1 & \text{if } u_{ij}^{(l)} \geq u_{ij}^{(m)} \\ 0 & \text{if } u_{ij}^{(l)} < u_{ij}^{(m)}. \end{cases} \tag{2.14}$$

This means that the observed item response depends on the sign of the difference between the latent utilities: $u_{ij}^{(l)} - u_{ij}^{(m)}$. In the MAFC measurement of personality, each statement is designed

to measure different trait dimensions; for instance, the first statement measures emotional stability, and the second statement measures extraversion. Let $d_j^{(l)}$ and $d_j^{(m)}$ denote the latent trait dimensions that are measured by j_l and j_m, respectively, where D is the total number of latent traits to be measured in the questionnaire, $1 \leq d_j^{(k)} \leq D$ $(k = 1, \cdots, K)$, and $d_j^{(l)} \neq d_j^{(m)}$.

Here, $u_{ij}^{(l)}$ and $u_{ij}^{(m)}$ can be written as the Thurstonian factor model:

$$\begin{aligned} u_{ij}^{(l)} &= \tilde{\mu}_j^{(l)} + \tilde{\beta}_j^{(l)} \theta_{id_j^{(l)}} + \epsilon_{ij}^{(l)}, \\ u_{ij}^{(m)} &= \tilde{\mu}_j^{(m)} + \tilde{\beta}_j^{(m)} \theta_{id_j^{(m)}} + \epsilon_{ij}^{(m)}, \end{aligned} \tag{2.15}$$

where $\tilde{\mu}_j^{(k)}$ represents the mean of the latent utility $u_{ij}^{(k)}$, $\tilde{\beta}_j^{(k)}$ represents a factor loading, $\theta_{id_j^{(k)}}$ represents the latent trait of respondent i, and $\epsilon_{ij}^{(k)}$ represents the residual or unique factor. Both $\boldsymbol{\theta}_i$ and $\epsilon_{ij}^{(k)}$ are assumed to be (multivariate) normally distributed.

In order to facilitate understanding of the notation, let us consider the example in Table 3.4 on page 36, in which $D = 5$ latent traits are measured using $J = 25$ items (i.e., pairs of statements). In item 2, the first and second statements measure agreeableness and intellect/imagination, respectively. They correspond to the third and fifth traits (as indicated in the footnote of the table). Therefore, in this case, $d_2^{(1)} = 3$, and $d_2^{(2)} = 5$. As a result, Equation 2.15 for item 2 in Table 3.4 is given as

$$\begin{aligned} u_{i2}^{(1)} &= \tilde{\mu}_2^{(1)} + \tilde{\beta}_2^{(1)} \theta_{i3} + \epsilon_{i2}^{(1)}, \\ u_{i2}^{(2)} &= \tilde{\mu}_2^{(2)} + \tilde{\beta}_2^{(2)} \theta_{i5} + \epsilon_{i2}^{(2)}. \end{aligned} \tag{2.16}$$

A. Brown and Maydeu-Olivares (2011) proposed their Thurstonian IRT model by reparameterizing the above second-order Thurstonian factor model. Specifically, its item response function is given by

$$P(x_{ij}^{(l\_m)} = 1 \mid \boldsymbol{\theta}_i) = \Phi\left[\frac{\left(\tilde{\mu}_j^{(l)} + \tilde{\beta}_j^{(l)} \theta_{id_j^{(l)}}\right) - \left(\tilde{\mu}_j^{(m)} + \tilde{\beta}_j^{(m)} \theta_{id_j^{(m)}}\right)}{\sqrt{\Psi_j^{2(l)} + \Psi_j^{2(m)}}} \right], \tag{2.17}$$

where $\Psi_j^{2(k)}$ represents the variance of the unique factor $\epsilon_{ij}^{(k)}$. By using the reparameterization

$$\begin{aligned}
\mu_j^{(l)} &= \frac{\tilde{\mu}_j^{(l)}}{\sqrt{\Psi_j^{(l)2} + \Psi_j^{(m)2}}}, \quad \mu_j^{(m)} = \frac{\tilde{\mu}_j^{(m)}}{\sqrt{\Psi_j^{(l)2} + \Psi_j^{(m)2}}}, \\
\beta_j^{(l)} &= \frac{\tilde{\beta}_j^{(l)}}{\sqrt{\Psi_j^{(l)2} + \Psi_j^{(m)2}}}, \quad \beta_j^{(m)} = \frac{\tilde{\beta}_j^{(m)}}{\sqrt{\Psi_j^{(l)2} + \Psi_j^{(m)2}}},
\end{aligned} \tag{2.18}$$

Equation (2.17) can be rewritten as

$$P(x_{ij}^{(l\_m)} = 1 \mid \boldsymbol{\theta}_i) = \Phi\left[\left(\mu_j^{(l)} + \beta_j^{(l)}\theta_{id_j^{(l)}}\right) - \left(\mu_j^{(m)} + \beta_j^{(m)}\theta_{id_j^{(m)}}\right)\right], \tag{2.19}$$

which corresponds to the form of the two-dimensional normal ogive IRT model. When an item consists of three or more statements ($K > 2$), i.e., when a respondent needs to include comparisons between more than two statements simultaneously, $\mu_j^{(k)}$ and $\beta_j^{(k)}$ become mathematically dependent parameters. However, when an item consists of only two statements ($K = 2$), which is the case considered in Chapter 3, all $\Psi_j^{(k)}$ have to be fixed (A. Brown & Maydeu-Olivares, 2011). By setting all values to $\sqrt{0.5}$, all denominators in Equation 2.18 become one, hence $\tilde{\mu}_j^{(1)}$, $\tilde{\mu}_j^{(2)}$, $\tilde{\beta}_j^{(1)}$, and $\tilde{\beta}_j^{(2)}$ are equivalent to $\mu_j^{(1)}$, $\mu_j^{(2)}$, $\beta_j^{(1)}$, and $\beta_j^{(2)}$, respectively.

Chapter 3

Study 1: Thurstonian D-diffusion IRT Model

3.1 Objective of the Study

The D-diffusion IRT model, which was introduced in Section 2.2, is a promising cognitive psychometric model (Batchelder, 1998; Ranger, Kuhn, & Szardenings, 2017; Vandekerckhove, 2014) for the joint modeling of item responses and RTs for personality questionnaire items. It combines two key modeling ideas: IRT models that have the virtue of disentangling respondent and item characteristics and cognitive process models that are grounded in cognitive theory to implement validated mechanisms for performing response tasks.

The D-diffusion IRT model implements the inverted-U relationship of personality measurement in a natural and quantitative manner. The model has a good parameter recovery property (Ranger, Kuhn, & Szardenings, 2016) and has begun to be applied to empirical psychological data such as the measurement of extraversion (Molenaar, Tuerlinckx, & van der Maas, 2015).

However, so far, Thurstonian IRT and D-diffusion IRT models have taken two different pathways of model development. In fact, we are not aware of any cognitive psychometric models that can jointly model the item response and RT information of personality measurements collected in the two-alternative forced-choice (2AFC) format. Consequently, the primary objective

of this study is to propose and examine such a model by integrating the Thurstonian IRT and D-diffusion IRT models. The present study focuses on the 2AFC personality measurement as the proposed model is based on the diffusion model. A Bayesian modeling approach (M. D. Lee & Wagenmakers, 2013) is adopted in the proposed method.

The remainder of this chapter consists of four sections. In Section 3.2, the proposed model is introduced, which is named the Thurstonian D-diffusion IRT (TDIRT) model. Section 3.3 describes a simulation study that was conducted to check parameter recovery. Section 3.4 provides the application of the proposed TDIRT model to real psychological data to demonstrate that it successfully combines the item response and RT information. Finally, a general discussion and possible directions for further research are provided in Section 3.5.

3.2 Thurstonian D-diffusion IRT Model

3.2.1 Parameter Settings

Obviously, Equations 2.8 and 2.19 have similar functional forms, as a logistic function can be considered as an approximation to the normal ogive function. In addition, both respondent parameters, θ_i in the diffusion IRT model and $\boldsymbol{\theta}_i$ in the Thurstonian IRT model, have a similar empirical meaning and the same distributional form. Therefore, when combining these two models, $(\theta_i - b_j)$ in Equation 2.8 can simply be replaced with $\left(\mu_j^{(1)} + \beta_j^{(1)}\theta_{id_j^{(1)}}\right) - \left(\mu_j^{(2)} + \beta_j^{(2)}\theta_{id_j^{(2)}}\right)$ in Equation 2.19. However, these two models have different forms for the discrimination parameter. In the diffusion IRT model, the discrimination parameter is the quotient of the respondent parameter γ_i divided by the item parameter ξ_j. Therefore, in the diffusion IRT model, the discriminability depends on both item factors such as the statement length and time limit and respondent factors such as the response caution and temper (van der Maas et al., 2011). On the other hand, the slope parameter in the original Thurstonian IRT model is a statement parameter. Therefore, the discriminability in the Thurstonian IRT model is not affected by respondent factors.

Besides, in this study, the difference between the means of the latent utility $\mu_j^{(1)} - \mu_j^{(2)}$ is

replaced with $\mu_j^{(1-2)}$. This is because when $K = 2$, both $\mu_j^{(1)}$ and $\mu_j^{(2)}$ cannot be estimated at the same time. The TIRT and proposed models therefore estimates a single parameter $\mu_j^{(1-2)}$ instead of $\mu_j^{(1)}$ and $\mu_j^{(2)}$.

On the basis of the above fact, in introducing the proposed model, we start from the Thurstonian IRT model and extend it so that the slope and intercept parameters in Equation 2.19 depend on both the item and respondent factors. The Thurstonian IRT model maintains its identifiability even if $\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}$ in Equation 2.19 is simply multiplied by (γ_i/ξ_j) because $\beta_j^{(k)}$ is a *statement* parameter; that is, each statement independently has its own $\beta_j^{(k)}$. This parameter may be interpreted as the discriminability of the specified statement. On the other hand, ξ_j is an *item* parameter that is specific to the pair of statements. This means that ξ_j represents an effect that is inherent to the paired statements that are simultaneously presented to respondents.

The above arguments lead us to introduce the proposed Thurstonian D-diffusion IRT model, which was developed as an extension of the D-diffusion IRT model. Specifically, in the proposed model, the joint distribution of the item response and RT is given as Equation 2.3. However, in its marginalized IRT-type form of Equation 2.5, the two parameters a_{ij} and v_{ij} were reparametrized in the same manner as in the case of Thurstonian IRT as

$$\begin{aligned} a_{ij} &= \frac{\gamma_i}{\xi_j} && \text{with} \quad \gamma_i,\ \xi_j \in \mathbb{R}_{>0} \\ v_{ij} &= \beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)} && \text{with} \quad \theta_{id_j^{(k)}},\ \beta_j^{(k)},\ \mu_j^{(1-2)} \in \mathbb{R}, \end{aligned} \tag{3.1}$$

instead of Equation 2.8.

Then, the joint likelihood of the item response and RT in the proposed model is given as

$$\begin{aligned} f(x_{ij}, t_{ij}) = \frac{\pi s^2 \xi_j^2}{\gamma_i^2} \exp\left(\frac{\gamma_i\left(\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}\right)(2x_{ij}-1)}{2s^2\xi_j} - \frac{\left(\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}\right)^2 (t_{ij}-\tau_j)}{2s^2} \right) \\ \times \sum_{m=1}^{\infty} m \sin\left(\frac{\pi m}{2}\right) \exp\left(-\frac{1}{2}\frac{\pi^2 s^2 m^2 \xi_j^2}{\gamma_i^2}(t_{ij}-\tau_j) \right). \end{aligned} \tag{3.2}$$

In practice, Equation 3.2 cannot be directly calculated because it involves the sum of an infinite series. Therefore, Navarro and Fuss's (2009) approximation is commonly used for its evaluation. This approach is also employed in `stan`, which is the software used in this study.

From Equation 3.2, the probability of choosing the first statement, which is a counterpart of both Equations 2.8 and 2.19, is then given as

$$P(x_{ij} = 1 \mid \boldsymbol{\theta}_i) = \frac{\exp\left(\frac{\gamma_i}{\xi_j}\left(\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}\right)\right)}{1 + \exp\left(\frac{\gamma_i}{\xi_j}\left(\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}\right)\right)}. \tag{3.3}$$

In this study, it is assumed that the researcher already knows the keyed direction of each statement, because this is natural in personality measurement using existing scales. Therefore, the sign of $\beta_j^{(k)}$ is known. Parameter $\beta_j^{(k)}$ was first estimated under the restriction of positivity. Then, the sign of the obtained negatively keyed item parameter estimates were manually inverted in order to facilitate their interpretations.

A. Brown and Maydeu-Olivares (2011) recommends designing a forced-choice questionnaire by preparing both types of items that are keyed in the same direction (e.g., both statements are positively keyed) and those in the opposite direction (one statement is positively keyed and the other negatively keyed) in order to ensure accurate measurement of traits. We recommend the same in the proposed approach.

When the number of latent traits measured in a 2AFC questionnaire is two ($D = 2$), the proposed model, as well as the Thurstonian IRT model, suffers from the identification problem. This can be readily understood because in this case, all statements have nonzero factor loadings to all traits, which is essentially the case of an exploratory factor analysis. A. Brown and Maydeu-Olivares (2011, 2012) dealt with this problem by fixing the factor loadings of the first two statements to their true values in the simulation study. Obviously, this approach is not realistic in practice because a researcher never knows the true factor loading parameter val-

ues. Therefore, this chapter only consider the case when $D > 2$ and recommend applying the proposed model for measuring more than two traits.

3.2.2 Prior Distribution

The Bayesian model is completed by placing prior distributions over the parameters. The following priors were used throughout this chapter:

$$\begin{aligned} &\xi_j \sim t_{[0,\infty)}(4,0,2.5), && \gamma_i \sim LN(0,1), \\ &\beta_j^{(k)} \sim t_{[0,\infty)}(4,0,2.5), && \boldsymbol{\theta}_i \sim MVN(\mathbf{0},\boldsymbol{\Sigma}), \\ &\mu_j^{(1-2)} \sim N(0,2.5^2), && \tau_j \sim U(0,\min(\mathrm{RT})_j), \\ & && \boldsymbol{\Sigma} \sim LKJCorr(1), \end{aligned} \tag{3.4}$$

where $\min(\mathrm{RT})_j$ represents the minimum observed RT for item j. $LN(\cdot)$, $MVN(\cdot)$, $U(\cdot)$ denote lognormal, multivariate normal, and uniform distributions, respectively. The basic principle in specifying the priors here is to place standardized distributions for the respondent parameters and weakly informative priors that give very low probabilities to very implausible parameter values for item parameters. The priors for the respondent parameters are standardized for identification reasons. Hence, the variance of each latent trait is fixed to 1. This means that the covariance matrix $\boldsymbol{\Sigma}$ reduces to a correlation matrix. Hence, $\boldsymbol{\Sigma}$ is called as the correlation matrix hereafter and we denote its elements by $\rho_{dd'}$. Moreover, γ_i follows the standard lognormal distribution, and the variance of the amount of information accumulation (s^2) is set to be 1. The LKJ correlation distribution (Lewandowski, Kurowicka, & Joe, 2009) for the correlation matrix here corresponds to a uniform prior over the space of possible $D \times D$ correlation matrices (Stan Development Team, 2018). For the priors of the item parameters, the half$-t$ distribution was adopted for ξ_j and $\beta_j^{(k)}$ as weakly informative priors, which are recommended by Gelman (2006) and Stan Development Team (2018).

The parameters of the proposed Thurstonian D-diffusion IRT model can be estimated using the Markov Chain Monte Carlo (MCMC) algorithm. For all of the estimation results presented in this chapter, we used R 3.5.0 and `stan` 2.17.3 on a Windows 10 PC. Three MCMC chains

were run for each data set. The number of MCMC iterations per chain was 10,000, half of which were discarded as warm-up. The data and computer codes (`stan` and R) that was used in this chapter are available from the Open Science Framework website (`https://osf.io/jswqg/`). The stan code for the TIRT and TDIRT models can also be found in Appendix A and B, respectively. The posterior means of the parameters were chosen as their point estimates.

3.3 Simulation Study

We conducted a simulation study to check the parameter recovery properties of the proposed model. Simulation data were generated from the TDIRT model with some narrower distributions described below and estimated the parameter values from the generated data with the conventional priors described in the previous section.

We considered the following five scenarios in this simulation:

- The first scenario manipulated the two crossed factors—the number of trait dimensions $D = (3,4)$ and the number of items that consist of the same pairs of dimensions $J_{pair} = (3,5,7)$—to examine their effects on parameter recovery. Note that the total number of items J is related to these factors by

$$J = J_{pair} \times \frac{D(D-1)}{2}. \tag{3.5}$$

  The trait dimensions were independent of one another in this scenario and the second scenario.

- The second scenario examined the influence of the number of dimensions D when the questionnaire length was fixed. For this purpose, the number of dimensions $D = (3,4)$ were manipulated under conditions where the total number of items $J = (12,24,36)$.

- The third and fourth scenarios examined if the proposed model can properly recover the parameter values even when the dimensions are correlated. The factor examined in the

third scenario was the correlations between the dimensions, which correspond to the off-diagonal elements of $\boldsymbol{\Sigma}$. All of the off-diagonal elements were set to have the same values, which was denoted by ρ_{pair}. Its three manipulated conditions were $\rho_{pair} = (-0.3, 0.3, 0.5)$. In this scenario, we set $D = 3$ and $J_{pair} = 5$.

- The fourth scenario considered more realistic conditions and attempted to emulate the Big-Five factor structure. Specifically, the correlation matrix $\boldsymbol{\Sigma}$ was specified on the basis of A. Brown and Maydeu-Olivares (2011) as

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & & & & \\ -.21 & 1 & & & \\ 0 & .40 & 1 & & \\ -.25 & 0 & 0 & 1 & \\ -.53 & .27 & 0 & .24 & 1 \end{pmatrix}.$$

  this matrix was denoted as $\rho_{pair} = BF$ for simplicity. In this scenario, $D = 5$, and J_{pair} was manipulated for three conditions, which were $J_{pair} = (3, 5, 7)$.

- The fifth scenario examined the influence of the complexity (cognitive workload) of the items. In this scenario, the true distribution of ξ_j was manipulated to be either $U(0.2, 0.3)$, $U(0.4, 0.5)$, or $U(0.6, 0.7)$. Here, we fixed $D = 3$, $J_{pair} = 5$ and assumed no correlations among traits.

This results in a total of 21 conditions. The number of respondents was kept constant at $I = 300$ for all conditions. In each condition, the cycle of random data generation and parameter estimation was replicated 30 times.

Data generation

The distributions used to generate random data from the proposed model were as follows: $\gamma_i \sim LN(0, 1)$, $\beta_j^{(k)} \sim U(0.5, 1.5)$, $\boldsymbol{\theta}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, $\mu_j^{(1-2)} \sim U(-1.5, 1.5)$, and $\tau_j \sim U(0.2, 1)$. With regard to ξ_j, samples were drawn from $U(0.3, 0.7)$ in the first to fourth scenarios. These

distributions are chosen on the basis of the preliminary analysis and A. Brown and Maydeu-Olivares (2011). Note that the off-diagonal elements of the correlation matrix $\mathbf{\Sigma}$ were treated as separate parameters and estimated as such even when the true correlations had the same values. Moreover, for a practical reason, observations with a RT greater than 120 seconds were deleted by listwise method. We considered this manipulation to be acceptable because an observation with such a large RT may be considered as an irregular response.

Results

Tables 3.1 and 3.2 summarize the averages of the root mean square errors (RMSEs) and the averages of the biases for each condition, respectively. The RMSE for the parameter β, for example, is calculated by

$$\text{RMSE}_\beta = \sqrt{\frac{1}{J}\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{2}\left(\hat{\beta}_j^{(k)} - \beta_j^{(k)}\right)^2}, \tag{3.6}$$

where $\beta_j^{(k)}$ is the realized value that is generated from $U(0.5, 1.5)$ in each iteration and $\hat{\beta}_j^{(k)}$ is its expected a posteriori estimate. Similarly, the bias for the parameter θ, for example, is calculated by

$$\text{bias}_\theta = \frac{1}{I}\frac{1}{D}\sum_{i=1}^{I}\sum_{d=1}^{D}\left(\hat{\theta}_{id} - \theta_{id}\right). \tag{3.7}$$

Since γ_i is lognormally distributed, the RMSEs and biases of the log-transformed estimates for γ are presented. The RMSEs for both β and ρ are acceptably small when $D = 3$ and $J_{pair} = 3$, even when J is as small as 9. These results suggest the proposed model is capable of appropriately estimating the trait loadings $\beta_j^{(k)}$ and the correlations between traits $\rho_{dd'}$.

In the second scenario, the RMSEs for θ become larger when the number of traits increases. Therefore, this suggests that a smaller number of traits would provide greater accuracy in the estimation of θ_{id}. Further, since the RMSE decreases as J increases, it would be desirable to have many items in the questionnaire if possible.

Table 3.1: Mean RMSEs of the parameter estimates in the simulation study

					Mean RMSE						
D	J_{pair}	J	ρ_{pair}	$\xi_j \sim$	ξ	β	μ	τ	$\log(\gamma)$	θ	ρ
3	3	9	0	U(0.3, 0.7)	0.034	0.126	0.106	0.001	0.166	0.564	0.059
3	5	15	0	U(0.3, 0.7)	0.035	0.105	0.108	0.001	0.136	0.487	0.052
3	7	21	0	U(0.3, 0.7)	0.043	0.107	0.095	0.002	0.129	0.433	0.037
4	3	18	0	U(0.3, 0.7)	0.042	0.102	0.106	0.001	0.136	0.494	0.046
4	5	30	0	U(0.3, 0.7)	0.056	0.112	0.102	0.001	0.137	0.422	0.040
4	7	42	0	U(0.3, 0.7)	0.065	0.102	0.107	0.002	0.139	0.390	0.038
3	4	12	0	U(0.3, 0.7)	0.026	0.112	0.106	0.001	0.142	0.513	0.052
4	2	12	0	U(0.3, 0.7)	0.044	0.126	0.113	0.001	0.160	0.558	0.054
3	8	24	0	U(0.3, 0.7)	0.053	0.107	0.100	0.001	0.137	0.417	0.034
4	4	24	0	U(0.3, 0.7)	0.054	0.110	0.143	0.010	0.178	0.478	0.043
3	12	36	0	U(0.3, 0.7)	0.072	0.108	0.107	0.001	0.153	0.366	0.036
4	6	36	0	U(0.3, 0.7)	0.075	0.107	0.142	0.011	0.193	0.429	0.036
3	5	15	-0.3	U(0.3, 0.7)	0.041	0.115	0.108	0.001	0.144	0.475	0.054
3	5	15	0.3	U(0.3, 0.7)	0.040	0.113	0.106	0.001	0.143	0.478	0.056
3	5	15	0.5	U(0.3, 0.7)	0.033	0.113	0.098	0.001	0.132	0.474	0.047
5	3	30	*BF*	U(0.3, 0.7)	0.050	0.111	0.113	0.001	0.128	0.437	0.038
5	5	50	*BF*	U(0.3, 0.7)	0.099	0.116	0.108	0.001	0.189	0.378	0.035
5	7	70	*BF*	U(0.3, 0.7)	0.138	0.129	0.105	0.001	0.246	0.343	0.035
3	5	15	0	U(0.2, 0.3)	0.019	0.084	0.096	0.004	0.138	0.355	0.035
3	5	15	0	U(0.4, 0.5)	0.035	0.106	0.106	0.002	0.142	0.473	0.051
3	5	15	0	U(0.6, 0.7)	0.045	0.113	0.127	0.001	0.137	0.546	0.053

Note: *BF* = correlation between traits based on A. Brown and Maydeu-Olivares (2011).

Table 3.2: Mean biases of the parameter estimates in the simulation study

					Mean bias						
D	J_{pair}	J	ρ_{pair}	$\xi_j \sim$	ξ	β	μ	τ	$\log(\gamma)$	θ	ρ
3	3	9	0	U(0.3, 0.7)	0.010	0.017	0.012	0.000	0.035	0.001	0.000
3	5	15	0	U(0.3, 0.7)	0.023	0.011	0.001	0.000	0.051	-0.002	0.003
3	7	21	0	U(0.3, 0.7)	0.035	0.012	0.002	0.000	0.073	0.003	-0.002
4	3	18	0	U(0.3, 0.7)	0.032	0.011	0.005	0.000	0.067	0.000	0.002
4	5	30	0	U(0.3, 0.7)	0.049	0.013	0.004	0.000	0.101	0.001	-0.003
4	7	42	0	U(0.3, 0.7)	0.055	0.011	0.009	0.000	0.106	-0.006	0.000
3	4	12	0	U(0.3, 0.7)	0.014	0.013	-0.005	0.000	0.038	-0.006	-0.006
4	2	12	0	U(0.3, 0.7)	0.029	0.016	0.014	0.000	0.072	0.009	0.001
3	8	24	0	U(0.3, 0.7)	0.044	0.012	0.011	0.000	0.089	0.002	-0.002
4	4	24	0	U(0.3, 0.7)	0.044	0.015	0.007	0.002	0.087	0.004	-0.007
3	12	36	0	U(0.3, 0.7)	0.066	0.012	0.000	0.000	0.126	0.009	0.006
4	6	36	0	U(0.3, 0.7)	0.066	0.015	-0.014	0.003	0.129	0.000	0.001
3	5	15	-0.3	U(0.3, 0.7)	0.027	0.014	-0.020	0.000	0.060	-0.002	-0.008
3	5	15	0.3	U(0.3, 0.7)	0.025	0.013	0.005	0.000	0.056	0.000	0.007
3	5	15	0.5	U(0.3, 0.7)	0.019	0.014	0.005	0.000	0.047	-0.003	-0.008
5	3	30	*BF*	U(0.3, 0.7)	0.044	0.012	0.008	0.000	0.090	0.008	-0.031
5	5	50	*BF*	U(0.3, 0.7)	0.094	0.014	-0.003	0.000	0.177	-0.005	-0.028
5	7	70	*BF*	U(0.3, 0.7)	0.133	0.017	0.010	0.000	0.239	0.002	-0.034
3	5	15	0	U(0.2, 0.3)	0.011	0.007	0.005	0.000	0.047	-0.002	-0.002
3	5	15	0	U(0.4, 0.5)	0.023	0.012	0.005	0.000	0.059	0.000	-0.001
3	5	15	0	U(0.6, 0.7)	0.024	0.013	-0.001	0.000	0.045	0.009	0.002

Note: *BF* = correlation between traits based on A. Brown and Maydeu-Olivares (2011).

The RMSEs for θ might look rather large for all conditions. However, these values are actually comparable in scale with the values reported by C. A. Stone (1992) when the number of items for each factor is larger than 10 (for example, the number of items for each factor is 10 when $D = 3$ and $J_{pair} = 5$).

A similar estimation accuracy is obtained when nonzero correlations exist between traits, which corresponds to the third and fourth scenarios. In addition, the trait correlations are successfully recovered for all conditions.

From the results of the fifth scenario, it can be seen that the estimation accuracy of θ_{id} increases as the true distribution of ξ_j shrinks. In other words, given the same number of items, the estimation accuracy tends to be better as the items require more cognitive workload. This relationship is consistent with the specifications of the diffusion IRT model, in which the item discrimination is equal to the boundary parameter, because a smaller ξ_j leads to higher discrimination.

The means of the biases for ξ and $\log(\gamma)$ are positive for almost all conditions. They become larger according to the increase in the total number of items J. In the same manner, the mean RMSEs for these parameters also become larger in accordance with the increase in J. The reason for these results can be explained from the form of the parameter constraints, as described below. ξ_j and γ_i originally have a mutual sign indeterminacy because they are given in the form of a quotient. To avoid this indeterminacy, a parameter constraint needs to be introduced. In this study, the standard lognormal distribution was chosen for the prior distribution of γ_i. The effect of the priors tends to be reduced, corresponding to the increase in the data size, and this may explain the result that larger estimates for both ξ_j and γ_i tend to be obtained for larger J. However, we note that these biases and RMSEs are still practically sufficiently small to be acceptable. Moreover, they do not affect the relative order of the respondents or items effects, which should be more practically important. Furthermore, the biases for the other parameters are nearly zero and neither systematically positive nor negative.

Sensitivity analysis

In order to check that our prior specification does not actually have a strong influence on parameter estimation, a sensitivity analysis was conducted. In this sensitivity analysis, the second scenario was considered and changed the priors for the item parameters to be more diffuse ones. Specifically, the priors for the parameters ξ_j, $\beta_j^{(k)}$, and $\mu_j^{(1-2)}$ were set as a diffuse $N(0, 100^2)$ prior, and parameter recovery was comparatively checked. The other conditions of the simulation were the same as before. The results suggest that the change in the priors has only a minor influence on the parameter estimates (see Table 3.3). Considering the estimation efficiency, we think that it is practically adequate to employ the prior specification of Equation 3.4, provided that the sample size is not very small.

To summarize the simulation results, the proposed TDIRT model is found to have sufficient parameter recovery properties.

3.4 Real Data Application: Big-Five Data

In this section, we demonstrate the application of the proposed model to real 2AFC personality measurement data, which was collected along with the RT information.

Data

We collected data from a sample of 500 Japanese respondents through CrowdWorks, a major online crowdsourcing service in Japan. All data collection procedures in this study and subsequent studies were reviewed and approved by the Ethical Committee of the University of Tokyo. The sample consists of 232 males and 262 females (the remaining six did not answer the question), and their ages range from their 20s to their 70s. The survey was conducted in an online survey environment that was developed with the jsPsych library (de Leeuw, 2015). Online informed consent was obtained from all participants. After data collection, the observations from one respondent were eliminated owing to system trouble. As a result, the proposed model and Thurstonian IRT model were applied to the data with a sample size of 499. In ad-

Table 3.3: Comparison of the mean RMSE and mean bias between two prior settings.

				Mean RMSE						
D	J_{pair}	J	Prior	ξ	β	μ	τ	$\log(\gamma)$	θ	ρ
3	4	12	weakly informative	0.026	0.112	0.106	0.001	0.142	0.513	0.052
			diffuse	0.043	0.112	0.119	0.001	0.158	0.512	0.060
4	2	12	weakly informative	0.044	0.126	0.113	0.001	0.160	0.558	0.054
			diffuse	0.030	0.118	0.116	0.001	0.148	0.550	0.054
3	8	24	weakly informative	0.053	0.107	0.100	0.001	0.137	0.417	0.034
			diffuse	0.042	0.104	0.105	0.001	0.123	0.411	0.041
4	4	24	weakly informative	0.054	0.110	0.143	0.010	0.178	0.478	0.043
			diffuse	0.049	0.109	0.100	0.001	0.133	0.453	0.045
3	12	36	weakly informative	0.072	0.108	0.107	0.001	0.153	0.366	0.036
			diffuse	0.077	0.117	0.106	0.001	0.163	0.367	0.032
4	6	36	weakly informative	0.075	0.107	0.142	0.011	0.193	0.429	0.036
			diffuse	0.070	0.111	0.098	0.001	0.150	0.395	0.036

				Mean bias						
D	J_{pair}	J	Prior	ξ	β	μ	τ	$\log(\gamma)$	θ	ρ
3	4	12	weakly informative	0.014	0.013	-0.005	0.000	0.038	-0.006	-0.006
			diffuse	0.025	0.018	0.018	0.000	0.060	0.004	0.003
4	2	12	weakly informative	0.029	0.016	0.014	0.000	0.072	0.009	0.001
			diffuse	0.015	0.018	0.009	-0.001	0.046	0.006	0.004
3	8	24	weakly informative	0.044	0.012	0.011	0.000	0.089	0.002	-0.002
			diffuse	0.035	0.027	0.000	0.000	0.074	-0.003	0.011
4	4	24	weakly informative	0.044	0.015	0.007	0.002	0.087	0.004	-0.007
			diffuse	0.040	0.026	-0.001	0.000	0.084	0.001	0.004
3	12	36	weakly informative	0.066	0.012	0.000	0.000	0.126	0.009	0.006
			diffuse	0.072	0.031	0.004	0.000	0.140	0.004	-0.005
4	6	36	weakly informative	0.066	0.015	-0.014	0.003	0.129	0.000	0.001
			diffuse	0.065	0.036	0.007	0.000	0.127	-0.001	0.003

Note: Under the Prior=*weakly informative* condition, the priors are specified as in Equation 3.4. Note that the results under this condition are already listed in Tables 3.1 and 3.2, but are repeated here for ease of comparison. Under the Prior=*diffuse* condition, the priors for $\xi_j, \beta_j^{(k)}$, and $\mu_j^{(1-2)}$ are changed to a diffuse distribution of $N(0, 100^2)$.

dition, approximately 0.1% of responses for which the RTs were shorter than 300 milliseconds were deleted by listwise method. Woodworth and Schlosberg (1954, Chapter 2) indicate that the minimum RT (they use the term *latency*) for simple visual tasks is approximately 180 ms. However, because the cognitive comparison task of the current study requires a higher cognitive load, more time would be required to answer the task used in this study. Therefore, the lower cutoff was set to 300 ms.

The items we used originate from the Japanese version of the Big-Five factor marker questionnaire (Apple & Neff, 2012). This scale is intended to measure the Big-Five traits, which are emotional stability, extraversion, agreeableness, conscientiousness, and intellect/imagination. Each trait is measured by 10 statements, which add up to a total of 50 statements. In order to rearrange them into the 2MFC format, 25 pairs of items were constructed from the 50 statements without duplication. The pairs were carefully designed in order to maintain balance among pairs of traits and to include pairs of statements that are keyed both in the same and opposite directions. Table 3.4 summarizes the resultant (English-translated) items used in this study. Respondents were required to select the statement of the pair that better represents themselves for all 25 items. The order of the items and the order of statements in each item were randomized across respondents.

Parameter estimation

We comparatively estimated the parameters for two models: the TIRT model, which only uses the item responses as observed variables, and the proposed TDIRT model, which uses both the item responses and their RTs as observed variables. For both models, the number of MCMC iterations per chain was $10,000$, half of which were discarded as warm-up samples. The priors used in the TIRT model were a subset of the priors for the proposed model as follows:

$$\begin{aligned} \beta_j^{(k)} &\sim t_{[0,\infty)}(4,0,2.5), \qquad & \boldsymbol{\theta}_i &\sim MVN(\mathbf{0},\boldsymbol{\Sigma}), \\ \mu_j^{(1-2)} &\sim N(0,2.5^2), & \boldsymbol{\Sigma} &\sim LKJCorr(1). \end{aligned} \tag{3.8}$$

Table 3.4: List of items (pairs of statements) used in this study

	Trait 1	Statement 1	Trait 2	Statement 2
1	Emo	Get stressed out easily.*	Ext	Don't talk a lot.*
2	Agr	Feel little concern for others.*	Int	Have a rich vocabulary.
3	Emo	Am relaxed most of the time.	Int	Have a vivid imagination.
4	Emo	Worry about things.*	Con	Leave my belongings around.*
5	Con	Am always prepared.	Int	Have excellent ideas.
6	Emo	Am easily disturbed.*	Ext	Am the life of the party.
7	Agr	Insult people.*	Int	Have difficulty understanding abstract ideas.*
8	Emo	Get upset easily.*	Int	Am not interested in abstract ideas.*
9	Ext	Keep in the background.*	Con	Pay attention to details.
10	Ext	Have little to say.*	Int	Am quick to understand things.
11	Emo	Change my mood a lot.*	Con	Get chores done right away.
12	Agr	Am interested in people.	Con	Like order.
13	Emo	Have frequent mood swings.*	Agr	Take time out for others.
14	Ext	Feel comfortable around people.	Agr	Sympathize with others' feelings.
15	Con	Make a mess of things.*	Int	Use difficult words.
16	Emo	Seldom feel blue.	Ext	Start conversations.
17	Agr	Am not interested in other people's problems.*	Con	Follow a schedule.
18	Ext	Don't like to draw attention to myself.*	Con	Often forget to put things back in their proper place.*
19	Ext	Talk to a lot of different people at parties.	Agr	Have a soft heart.
20	Con	Shirk my duties.*	Int	Do not have a good imagination.*
21	Emo	Get irritated easily.*	Agr	Am not really interested in others.*
22	Agr	Feel others' emotions.	Con	Am exacting in my work.
23	Emo	Often feel blue.*	Int	Spend time reflecting on things.
24	Ext	Am quiet around strangers.*	Agr	Make people feel at ease.
25	Ext	Don't mind being the center of attention.	Int	Am full of ideas.

Notes: Statements with an asterisk (*) are negative statements; Emo = Emotional Stability (trait number: 1); Ext = Extraversion (2); Agr = Agreeableness (3); Con = Conscientiousness (4); Int = Intellect/Imagination (5); Japanese versions of the statements are available at `https://ipip.ori.org/JapaneseBig-FiveFactorMarkers.htm`.

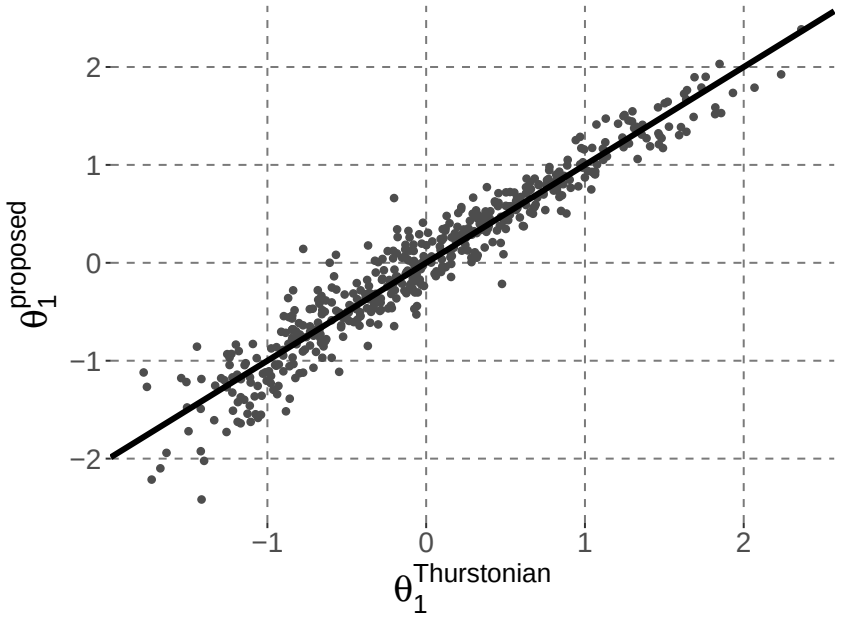

Figure 3.1: Scatter plot between the θ_i estimates from the Thurstonian IRT and proposed models with regard to trait 1 (extraversion). Black line indicates the line of $y = x$.

Results

Table 3.5 summarizes the posterior means of the item parameters estimated by both models. Apparently, the point estimates from the TIRT model are several times larger than those of the proposed model. A major reason for this is that there exist three types of parameters relates to the discriminability ($\beta_j^{(k)}$, ξ_j, and γ_i) in the proposed model. In fact, the item parameters obtained by the proposed model become similar to those obtained by the TIRT model when they are divided by ξ_j. This is because all values of γ_i follow $LN(0, 1)$ a priori, and the expected value of a random variable that follows $LN(0, 1)$ is 1.

The correlations between the estimates from the two models are .995 for $\mu_j^{(1-2)}$ and .902 for $\beta_j^{(k)}$. Therefore, it would be appropriate to consider that the item parameters of the proposed model have similar interpretations as those of the Thurstonian IRT model.

Figure 3.1 shows a scatter plot between the TIRT and TDIRT models in terms of the first trait estimates θ_{i1}, which correspond to the extraversion trait. The first trait was chosen for illustration purposes, and similar tendencies were observed for all five traits.

Table 3.6 summarizes the correlations between the respondent parameter estimates, both

Table 3.5: Posterior means of the item parameter values in the Thurstonian IRT and proposed models

	Thurstonian IRT model			Proposed model			
item	$\mu_j^{(1-2)}$	$\lvert\beta_j^{(1)}\rvert$	$\lvert\beta_j^{(2)}\rvert$	$\mu_j^{(1-2)}$	$\lvert\beta_j^{(1)}\rvert$	$\lvert\beta_j^{(2)}\rvert$	ξ_j
1	-0.728	2.334	1.681	-0.227	0.761	0.511	0.308
2	1.512	1.173	1.199	0.368	0.295	0.335	0.259
3	1.286	1.043	1.270	0.416	0.274	0.399	0.296
4	-1.899	1.142	1.847	-0.555	0.279	0.462	0.307
5	-1.099	1.481	1.900	-0.265	0.387	0.433	0.288
6	-1.978	0.314	1.964	-0.679	0.267	0.577	0.313
7	0.844	0.611	0.271	0.229	0.173	0.094	0.243
8	-1.505	0.994	0.248	-0.462	0.347	0.084	0.249
9	0.403	1.946	0.302	0.100	0.724	0.121	0.308
10	0.494	1.783	0.287	0.133	0.508	0.176	0.277
11	-0.186	1.024	1.689	-0.059	0.332	0.426	0.278
12	0.550	2.983	3.827	0.143	0.490	0.762	0.267
13	-0.316	1.274	0.893	-0.108	0.384	0.251	0.258
14	1.993	1.199	1.140	0.570	0.229	0.272	0.261
15	0.132	0.222	0.106	0.028	0.097	0.043	0.304
16	0.864	1.907	1.996	0.218	0.422	0.488	0.231
17	1.060	0.973	0.561	0.324	0.342	0.141	0.247
18	-0.572	0.980	1.813	-0.168	0.240	0.443	0.254
19	2.161	1.304	0.858	0.694	0.348	0.271	0.244
20	0.102	1.325	1.669	0.026	0.327	0.371	0.247
21	-1.210	1.246	1.353	-0.429	0.464	0.388	0.291
22	-0.263	0.219	0.320	-0.084	0.092	0.121	0.287
23	-0.035	1.100	0.581	-0.054	0.257	0.189	0.244
24	-1.352	1.872	0.241	-0.407	0.591	0.074	0.288
25	0.522	0.812	1.526	0.169	0.193	0.385	0.270

Note: With regard to β, absolute values are displayed. In parameter estimation, β is restricted to be positive. The estimated values for keyed items were manually sign-inverted afterwards.

within and between the models. All of the between-model correlations for the same trait are larger than .95, which can be interpreted that very similar respondent parameter estimates are obtained between the two models. Moreover, as for the correlations between traits within a model, similar correlation patterns are obtained for both models, except that they are slightly smaller for the proposed model.

Figure 3.2 shows the posterior distribution of ξ_j for each item. A clear between-item difference can be seen in this figure. In order to check their validity, we calculated the correlation between the posterior means of ξ_j and the item mean of the readability scores of each statement. The readability score, which quantifies the difficulty in reading the sentence, was computed by jReadability (Hasebe & Lee, 2015; `https://jreadability.net/sys/`). The resulting correlation was .539 (95% CI [.162, .751]). This positive correlation can be evidence that ξ_j reflects empirical item characteristics such as the reading difficulty.

Figure 3.3 shows two scatter plots between the boundary-related parameter estimates (γ_i and ξ_j) and the mean RT. It can be seen that there exist strong associations between the two quantities; the obtained correlation coefficients were .929 and $-.716$, respectively. These strong relationships between the observed RT and the parameter estimates are consistent with the results of Ratcliff, Thompson, and Mckoon (2015) and Tuerlinckx et al. (2016). As shown in Equation 3.1, the boundary of the diffusion process becomes larger according to the increase in γ_i. This means that a respondent with a higher value of γ_i needs more information accumulation to respond to an item. As a result, such a respondent tends to have a longer RT. This relationship is properly reflected in the left panel of Figure 3.3. The opposite is true for ξ_j, which is the denominator of the boundary.

Next, we examined how the estimates from the proposed model reflect the inverted-U relationship. For this purpose, Figure 3.4 shows scatter plots between the respondent's proximity to the comparison threshold (x axis), which is given by $\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}$, and the observed RT ($y$ axis). In the original diffusion IRT model, the expected RT is given by (Tuerlinckx et al., 2016)

$$E(t_{ij}) \approx \tau_j + \frac{1}{2|\theta_i - b_j|}\frac{\gamma_i}{\xi_j}. \tag{3.9}$$

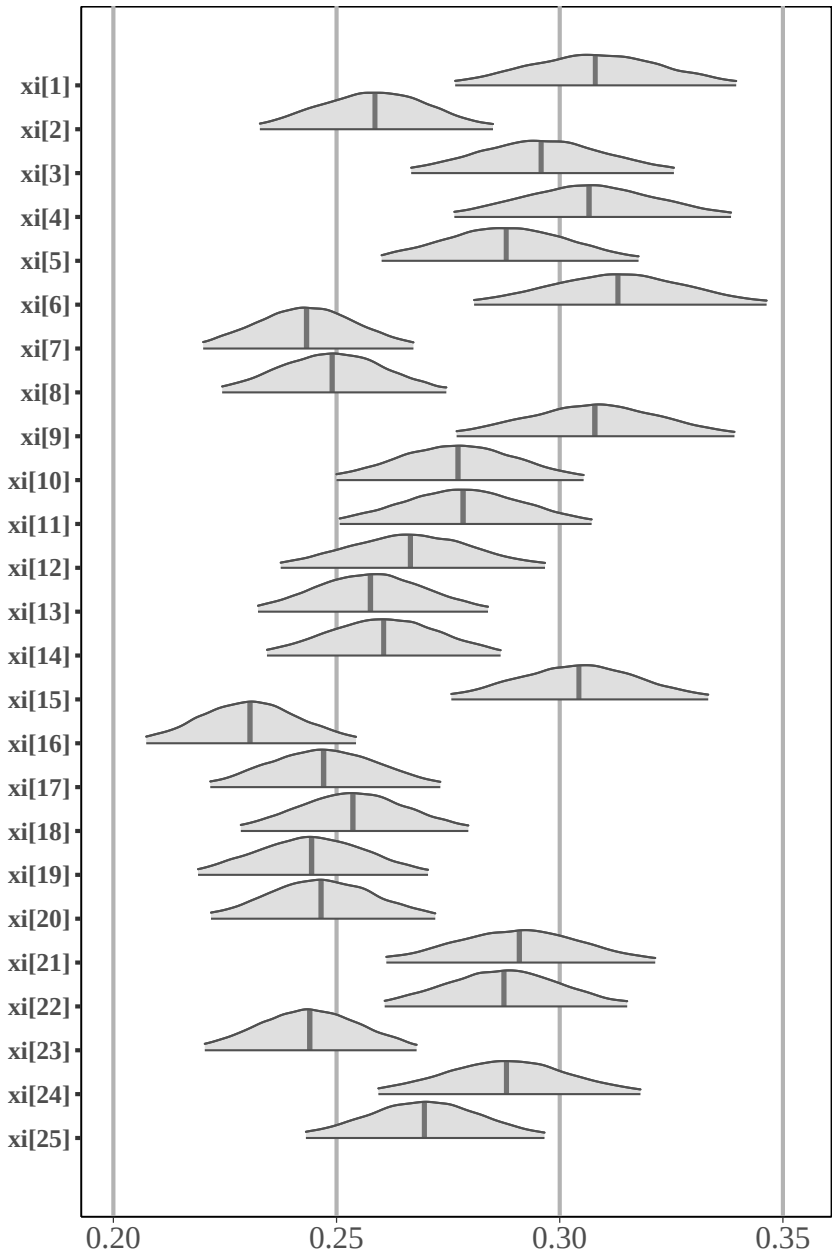

Figure 3.2: Posterior distribution of ξ_j for each item. The gray area indicates the 95% CI. The indices of the items correspond to those in Tables 3.4 and 3.5.

Table 3.6: Estimated correlation matrix in the real data application

	Thurstonian IRT					Proposed model						
	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	Mean	SD
θ_{i1}	-										-0.002	0.815
θ_{i2}	0.632	-									-0.001	0.850
θ_{i3}	0.333	0.624	-								0.002	0.761
θ_{i4}	0.380	0.231	0.396	-							0.002	0.817
θ_{i5}	0.420	0.605	0.269	0.262	-						0.000	0.776
θ_{i1}	0.963					-					-0.002	0.849
θ_{i2}		0.970				0.601	-				-0.001	0.873
θ_{i3}			0.954			0.247	0.532	-			0.001	0.781
θ_{i4}				0.967		0.283	0.143	0.393	-		0.000	0.831
θ_{i5}					0.969	0.366	0.608	0.260	0.198	-	-0.001	0.796

Note: θ_{i1} = emotional stability; θ_{i2} = extraversion; θ_{i3} = agreeableness; θ_{i4} = conscientiousness; θ_{i5} = intellect/imagination

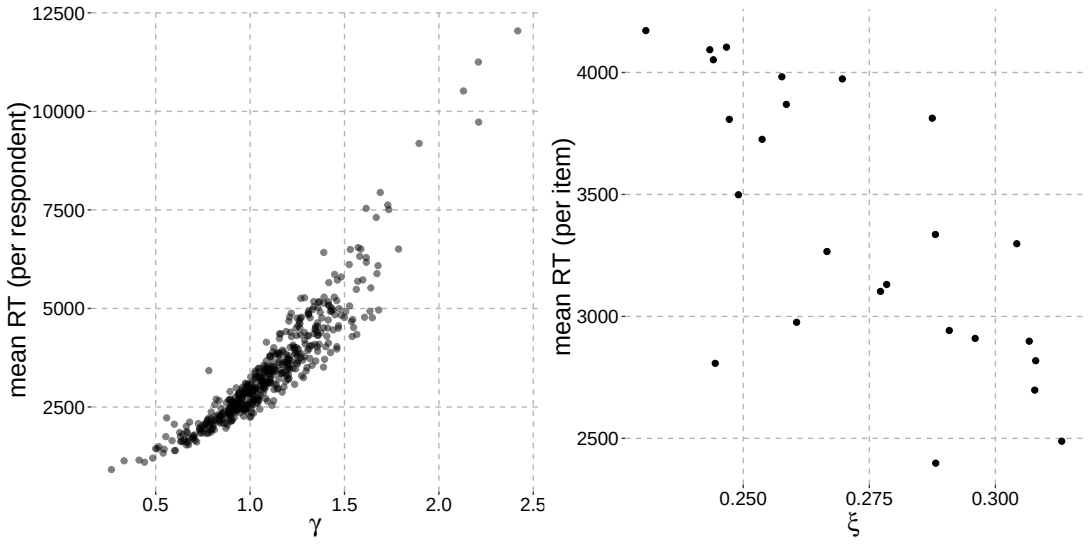

Figure 3.3: Left panel = scatter plot between the estimate γ_i and the mean RT for each respondent. Right panel = scatter plot between the estimate of ξ_j and the mean RT for each item.

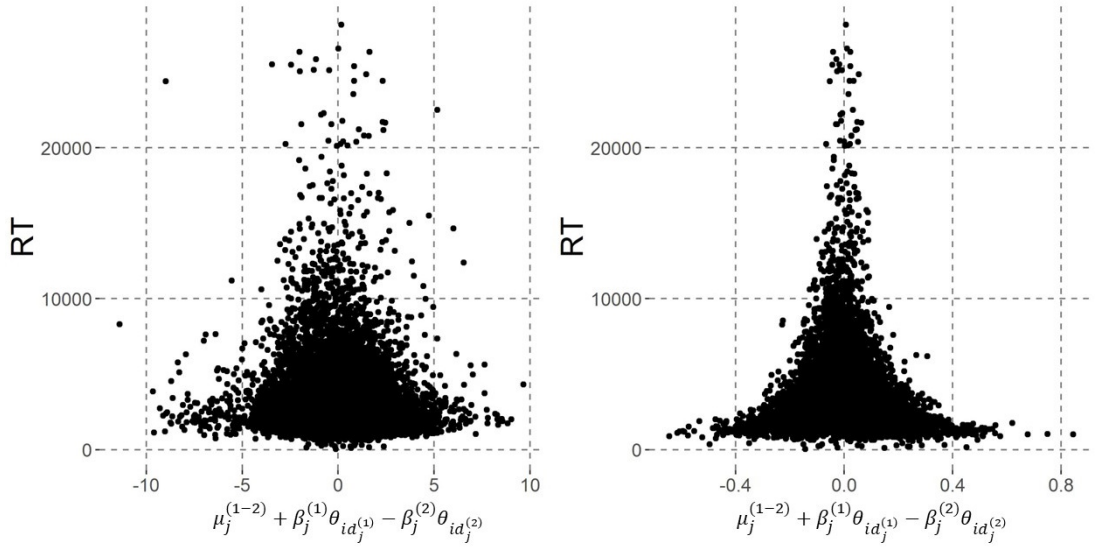

Figure 3.4: Scatter plots between the latent respondent position relative to the item parameters (x axis) and the observed RT (y axis). Left panel = results for the Thurstonian IRT model. Right panel = results for the proposed model (y-axis values are divided by γ_i/ξ_j in order to adjust the scale to the left panel).

From this formula, the expected RT for the proposed model can easily be obtained, i.e.,

$$E(t_{ij}) \approx \tau_j + \frac{1}{2\left|\beta_j^{(1)}\theta_{id_j^{(1)}} - \beta_j^{(2)}\theta_{id_j^{(2)}} - \mu_j^{(1-2)}\right|}\frac{\gamma_i}{\xi_j}. \tag{3.10}$$

It is seen in the left panel of Figure 3.4, which shows the estimates from the Thurstonian IRT model, that the RT tends to be large when the latent positions of the respondent and item parameters are close to each other. This is no surprise given that the inverted-U relation is well-known in personality measurement. However, this tendency is more clearly evident when the parameters are estimated by the proposed model (right panel). This is because the proposed model explicitly accounts for the inverted-U relationship by separating the respondent and item factors that affect the RT, such as the psychological traits to be measured and the item complexity.

Lastly, we examined how the RT information is utilized for estimating $\boldsymbol{\theta}_i$ in the proposed model. In the proposed model, the parameter $\boldsymbol{\theta}_i$ is estimated on the basis of both the item response and RT, whereas in the existing Thurstonian IRT model, the RT information is not used. In order to examine the unique effect of the RT in the proposed model, we first calculated

the *mean signed standardized RT* (MSSRT) for each respondent and each trait by the following procedure:

1. Prepare all 10 item responses and RTs for each statement of a trait.
2. Standardize the RT using the whole-sample mean and standard deviation of the item. This yields the standardized RT.
3. When a statement is positively keyed and not chosen or when it is negatively keyed and chosen, make the sign of its standardized RT negative. This yields the signed standardized RT.
4. Finally, compute the mean of the signed standardized RT for all 10 items. This is the MSSRT.

If the statement that represents a trait is positively keyed and chosen, the signed standardized RT becomes smaller when it is quickly chosen. If it is positively keyed but not chosen, the signed standardized RT becomes smaller when it takes long time to decide not to choose it. The opposite is true when the statement is negatively keyed. That is, the signed standardized RT becomes smaller if a negatively keyed statement is slowly chosen or quickly avoided.

Here, a possible relationship to be found would be the following. For a smaller MSSRT, the estimate obtained by the proposed model tends to be larger compared to that of the Thurstonian IRT model. In other words, a negative correlation would be expected between the MSSRT and the difference of $\boldsymbol{\theta}_i$ estimates between the proposed model and the Thurstonian IRT model. Figure 3.5 shows the scatter plot of the MSSRT and the difference between the two model estimates of $\boldsymbol{\theta}_i$. A negative correlation is evident between the two quantities, and little difference among traits was found. The overall correlation was $-.483$ (95% CI $[-.513, -.452]$). This result suggests that the proposed model successfully incorporates the RT information to supplement the estimation of latent traits.

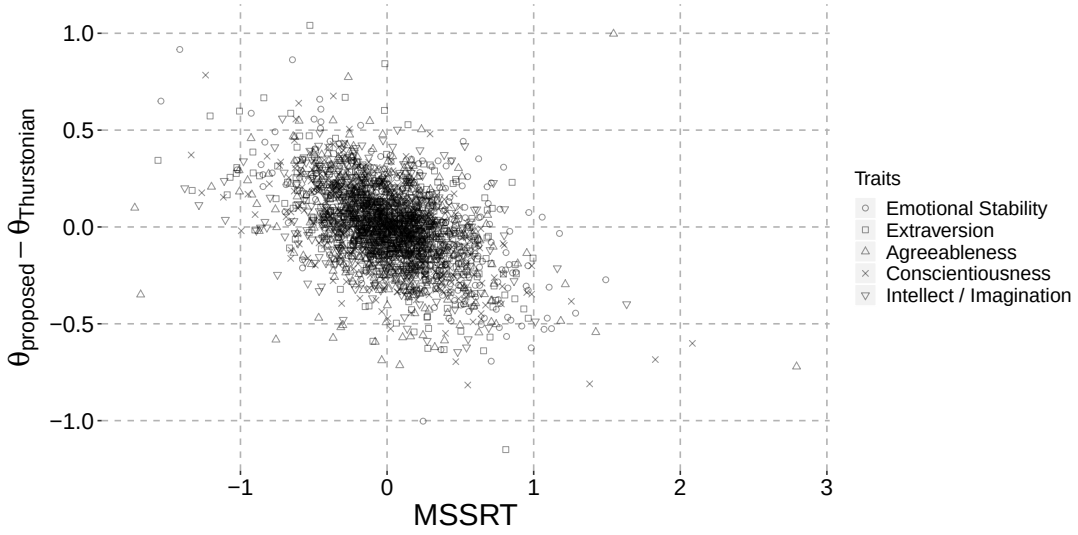

Figure 3.5: Scatter plot of the mean signed standardized RT (x axis) and the difference of trait estimates θ_i between the proposed model and the Thurstonian IRT model (y axis).

3.5 Discussion

The main objective of this study is to propose and examine the Thurstonian D-diffusion IRT model, which is an extension of the Thurstonian IRT model for 2MFC measurement to incorporate the RT information. To achieve this goal, we base our model on the diffusion IRT model, which is a representative cognitive psychometric model, and reparameterized its parameters to match the idea of the Thurstonian IRT model. A simulation study has shown that the parameters of the proposed model can be sufficiently recovered in typical application settings. In the application to Big-Five measurement data, several interesting findings have emerged. The obtained item parameter estimates from the proposed model were similar to those of the Thurstonian IRT model, except for the theoretically expected scale differences. As for the respondent parameters (θ), the estimates of the proposed model were also similar to those of the Thurstonian IRT, but there is a difference, which was explained by the inverted-U relationship. Because this relationship is believed to be generally applicable in personality measurement, we believe that personality estimation based on the proposed Thurstonian D-diffusion IRT model should be meaningful.

When one trait to be measured is clearly dominant over others for a respondent, this respon-

dent is expected to rapidly choose the statements of this trait. In this way, the proposed model should be able to reflect the degree of dominance of one trait over others. Such usage of RT information is not possible for existing Thurstonian IRT models that only use the item response as observations.

According to Bertling and Weeks (2018, p. 328), the motivation to utilize the RT information in conjunction with the item response can be classified into two types. The first is "to obtain more accurate proficiency level estimates," and the second is "to estimate examinee performance on a separate latent trait." Although our situation may be different in that the current study considers personality measurement while Bertling and Weeks (2018) considered ability measurement, we believe that the proposed model is closer to the latter case. That is, the proposed TDIRT model estimates the item and respondent parameters on the basis of psychological theory models of comparative judgment, the diffusion process, and the inverted-U relationship. As a result, the obtained latent traits would be well-isolated from the effects of factors such as the respondents' deliberateness and item length. Moreover, according to Kahneman (2011), comparative judgment, which involves a slower and more deliberative process, better reflects the cognitive component than single evaluations, which often reflect the intensity of an unstable emotional response. Meanwhile, in the diffusion model, the discrimination parameter correspond to the degree of response caution (boundary); a response that has a longer RT better reflects one's latent trait. Such a correspondence between these two theories suggests the relevance of this study's contribution to extend the diffusion model for Thurstonian judgment.

By introducing the approach of cognitive psychometrics, the parameters of the proposed model can quantify different cognitive subskills separately, such as the speed of information uptake and the degree of response caution. This is in contrast to the classical latent variable in psychometrics, which may be an unknown composite of cognitive processes (Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). Although we have not fully investigated the empirical relations between these parameter estimates and other exogenous variables, we consider such a study (e.g. Ratcliff & Rouder, 1998; Voss, Rothermund, & Voss, 2004) would be of great interest.

Another advantage of the application of cognitive psychometric models is that they can incorporate the properties of a rather complex measurement environment such as the time pressure. For example, Milosavljevic, Malmaud, Huth, Koch, and Rangel (2010) revealed that when a respondent is under a high time pressure condition, the mean RT becomes shorter than under the low time pressure condition; correspondingly, the boundary parameter estimates in the diffusion model become smaller. For this type of high time pressure, the responses become faster and, at the same time, more inaccurate. The diffusion-based models can appropriately account for such effects of the time pressure (e.g. Voss et al., 2004). Such an explanatory power would be one of the advantages of the application of cognitive psychometric models.

Finally, we note a few limitations of the current study. First, a forced-choice item typically provides lower information than a Likert-scale item. As noted by A. Brown and Maydeu-Olivares (2013), although a five-point Likert-scale item provides four pieces of information, a 2AFC item provides only one piece of information. Therefore, in practice, care would be needed when designing the forced-choice test, in terms of the number of items and the number of traits, for example, in order to ensure its reliability. Meanwhile, as noted in Chapter 1, one of the major motivations to use the forced-choice format is to reduce the effect of possible biases that originate from the Likert measurement. Recent studies that compared forced-choice Thurstonian IRT measurement and Likert measurement have found evidence that forced-choice measurement has higher validity, even though its obtained information is smaller (e.g. Guenole et al., 2018; P. Lee, Joo, & Lee, 2019). That being said, we have not examined the external validity of the latent traits estimated by the proposed model. Although the obtained parameter estimates are based on psychologically meaningful models, an empirical comparison of the external validity of the trait parameter estimates with existing methods remains an important subject for future study.

In Study 1, we selected the diffusion model as the base model of the current study. A limitation of the diffusion model is that it is only applicable to items that consist of two statements. When considering more than two statements for an item, other models have to be used as the base model. The LBA model would be a strong candidate in such a case because “For multiple

choices between $N > 2$ alternatives, only the LBA has simple-to-use analytic solutions, making it the preferred choice" (S. D. Brown & Heathcote, 2008, p. 173). In fact, the LBA model can be thought to be as popular as the diffusion model in cognitive psychology studies. Therefore, in the following chapters, we examine the combination of the LBA and IRT models in order to develop a new cognitive model for multidimensional MAFC forced-choice personality measurement with the RT.

Chapter 4

Study 2: Unidimensional Binary D-LBA IRT Model

4.1 Objective of the Study

The original diffusion and LBA models do not distinguish respondent and item parameters, which are fundamental characteristics of IRT models. However, considering the fact that RT-incorporating IRT models are meant to be applied to item response data obtained as a result of internal human information processing, it would be reasonable to believe that the combination of traditional IRT models for the item response and psychological models for the RTs would lead to the emergence of a novel, important, and practical class of models. On the basis of this idea, the D-diffusion IRT model (Tuerlinckx & De Boeck, 2005; van der Maas et al., 2011) has been proposed as a novel category of RT-incorporating IRT models and is currently the most representative process model for item-answering behavior for personality assessments. It is an elegant combination of the diffusion and IRT models with both respondent and item parameters. Its respondent parameters provide psychological insights into the underlying cognitive properties of human information processing.

Nevertheless, the diffusion model is a complex nonlinear model. It is expressed as a sum of infinite series, and the parameters vary across trials. Wagenmakers et al. (2007) suggested

that mathematical psychologists use such a complicated model because of the substantial payoff involved; the estimated parameter values of the diffusion model can provide psychological insights that cannot be provided by standard superficial methods of analysis. However, when combined with IRT in the form of the diffusion IRT model, the diffusion model becomes even more complex. The increased complexity of the model might prevent its application to real datasets. This is a critical issue in practice, especially when the data have a large sample size or when the hierarchical structure of the data is to be taken into account.

In addition, in the diffusion IRT model, the discriminability and expected RT have a linear relationship, as elaborated and illustrated in Section 4.3.2. The discriminability is equivalent to the slope of the logistic curve, i.e., how well an item can distinguish high-scoring people from low-scoring ones. However, this linear relationship may be a strong and unrealistic assumption. In the case of a reaction time task in cognitive psychology, which is the original context modeled with the diffusion model, the expected RT is typically shorter than a few seconds; then, the linearity assumption might work out as a primary assumption. However, when the expected RT is longer than a few seconds, which is actually the case in typical personality assessments, the estimates of the discriminability could be considerably inflated owing to this linear relationship. One possible way to deal with this problem is to add another new parameter to moderate the relationship. However, as noted above, the diffusion IRT model is already a complex model. It would be better to not increase the number of parameters further in consideration with the estimation efficiency. Therefore, this study instead focuses on the LBA model, which is a simpler alternative of the diffusion IRT model.

Donkin et al. (2011) investigated whether and to what extent conclusions regarding psychological processes depend on the choice between the diffusion and LBA models. They found a largely straightforward correspondence between the parameters of the two models. In fact, for the diffusion and LBA models, they concluded that "inferences about psychological processes made from real data are unlikely to depend on the model that is used" (p. 61).

On the basis of the abovementioned observations, the central idea of this study is that the combination of the LBA and IRT models would be beneficial as a novel model for modern

test data. In particular, it would provide us with insights into the underlying psychological process, distinguish the item and respondent characteristics, and facilitate faster and more stable estimation than the diffusion IRT models, even when new parameters are included. Thus, the objectives of this study are to propose a new LBA IRT model and to comparatively evaluate it against existing diffusion IRT models using simulated and real data.

The remainder of this chapter is organized as follows. Section 4.2 reviews the relationship between the diffusion and LBA models. Section 4.3 introduces the proposed LBA IRT model. Section 4.4 describes a simulation study conducted to compare the performance of the proposed LBA IRT model with that of existing diffusion IRT models. Section 4.5 provides an empirical illustration of the proposed and existing methods using a real personality dataset. Finally, Section 4.6 concludes the study with a brief discussion of the directions for future research.

4.2 Relation Between the Diffusion and LBA Models

Donkin et al. (2011) conducted a parameter recovery simulation study to examine the relation between the diffusion and LBA models. They considered the relationship between the two models with the following settings. (a) In the LBA model, the sum of v should be one. When the LBA model is compared with the diffusion model, $v^{(2)}$ becomes $1 - v^{(1)}$. Here, $v^{(1)}$ is the mean of the slope for the first response category ($k = 1$; e.g., "*I agree*") and $v^{(2)}$ is that of the second response category ($k = 2$; e.g., "*I disagree*"). (b) In the diffusion model, the distance from the starting point to the boundary is $\frac{\alpha}{2}$ when $z = \frac{\alpha}{2}$. In the LBA model, the starting point is uniformly distributed from 0 to A. Therefore, the expected distance from the starting point to the boundary becomes $B - \frac{A}{2}$. Accordingly, they compared $\frac{\alpha}{2}$ with $B - \frac{A}{2}$ rather than comparing α with B directly. In their simulation study, simulated data were generated and estimated with both models. Their results indicated the existence of a nearly one-to-one relationship with regard to the drift rate or nondecision time parameters, while the boundary parameters did not exhibit simple mapping (even though they have a fairly high correlation).

On the other hand, there are also studies that discuss the differences between the diffusion

and LBA models. For example, Heathcote and Hayes (2012) pointed out that the parameters of the two models would result in equivalent inferences under some conditions and different inferences under other conditions. This is not surprising because the precise functional forms of the diffusion and LBA models are different. Thus, caution may be needed when qualitatively translating a parameter estimate of one model to the other. In general, however, most core parameters of the diffusion and LBA models are comparable and have similar empirical meanings (Heathcote, Brown, & Wagenmakers, 2015).

4.3 Unidimensional Binary D-LBA IRT Model

We apply the LBA framework of modeling the response time data, which has been proved to be useful in the field of cognitive and mathematical psychology, to IRT models which are popular in psychometrics. For this purpose, in this section, we reparameterize the LBA model to yield the proposed LBA IRT model. This reparameterization allows us to combine the strengths of the LBA and IRT models, which are both popular in different fields.

The D-diffusion IRT model was derived from a relationship between the functional forms of the Diffusion and IRT equations, as explained in Section 2.2. However, as shown in Section 2.3, it is impossible to obtain a simple relationship between the functional forms of the LBA and IRT equations from Equation 2.13 just well as the D-diffusion IRT model. Instead, we will propose a D-LBA IRT model as an analogy of the D-diffusion IRT model.

Note that in this chapter, we only focus on 2AFC ($k = 1, 2$) tasks, although the original LBA model applies to MAFC tasks.

4.3.1 Parameter Settings

As stated in Section 2.2, there exist two classes of diffusion IRT models. In this study, we chose to extend the D-diffusion model for the following reasons. First, θ and b of the D-diffusion model can be regarded as nearly the same as those of traditional IRT, whereas those of Q-diffusion are restricted in that they cannot take negative values. Second, v in D-diffusion is

simply the *difference* between θ and b; therefore, it is easier to estimate than in the case of Q-diffusion, where v is a *quotient*. Third, the responses of personality measurement tend to be faster than those of ability measurement. The LBA and diffusion models were typically applied to cognitive tasks, the RT of which is typically less than a few seconds (although there are recent exceptions; see e.g., Palada et al., 2016). In general, ability measurements require a much longer RT than personality measurements, and the model properties under such conditions are less well-known. Therefore, we adopt D-diffusion. Accordingly, the proposed model is called the (unidimensional binary) *D-LBA IRT* model (we may simply call it as D-LBA model in this chapter).

Boundary Donkin et al. (2011) showed that α in the diffusion model can be interpreted as nearly the same as $B - \frac{A}{2}$ of the LBA model. In order to retain the same number of parameters in the D-LBA model as that in the D-diffusion model, a parameter constraint needs to be introduced. For this purpose, A is considered to be fixed in this study; specifically, we set

$$A_{ij} = \frac{B_{ij}}{2}, \tag{4.1}$$

where

$$B_{ij} = \frac{\gamma_i}{\xi_j} \ \ with \ \ \gamma_i, \xi_j \in \mathbb{R}_{>0}. \tag{4.2}$$

Other forms of constraint may also be possible, but we think this constraint is natural because the upper bound of the starting point in this setting is simply half of the distance between 0 and B.

Drift rate the drift rate can be treated in the same manner as in the case of the D-diffusion model, except that it needs to satisfy an identification constraint, which is required for latent variable models. In the original LBA model, a common way of incorporating this constraint is to set the sum of the drift rates among the alternative choices to be one. We follow this approach

by letting v to be the difference between θ and b scaled by a logistic function:

$$\begin{cases} v_{ij}^{(1)} = \left[1 + \exp(-\theta_i + b_j)\right]^{-1} \\ v_{ij}^{(2)} = \left[1 + \exp(\theta_i - b_j)\right]^{-1}, \end{cases} \tag{4.3}$$

Note that in order to study the comparable performance of the proposed D-LBA model with the D-diffusion model, this chapter considers the case of 2AFC data (i.e., $K = 2$), although the D-LBA model can be extended to polytomous choices, as shown in Chapter 5.

Nondecision time The nondecision time is nearly the same as that in the case of D-diffusion. The nondecision time was set as the item parameter τ_j.

Between-trial variability of the slope As with the LBA model, the diffusion model involves the identification issue. To deal with this in the diffusion model, the within-trial variance of the slope, s, has to be fixed at a certain value such as 0.1 or 1. This is due to the indeterminacy of the scale among s, α, and v in the diffusion model. On the other hand, the LBA model has no with-trial variance s; instead, it incorporates the between-trial variance η. By decomposing this variance into the item and respondent parameters, the proposed model can solve the problem of the linear relationship between them, which was, as noted before, one of the major problems of the diffusion IRT model. Specifically, the proposed model decomposes the between-trial variance η into person and item factors:

$$\eta_{ij} = \frac{\sigma_i}{\psi_j}. \tag{4.4}$$

As a result, the total number of parameters in the D-LBA model becomes $4J + 3I$, while the D-diffusion model used in this study has $3J + 2I$ parameters. The joint cumulative distribution

function of the item response and RT is given as

$$
\begin{aligned}
F(1,t) &= LBA_{CDF}\left(1, t-\tau \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{1}{1+\exp(-\theta_i+b_j)}, \frac{\sigma_i}{\psi_j}\right) \\
F(2,t) &= LBA_{CDF}\left(2, t-\tau \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{1}{1+\exp(\theta_i-b_j)}, \frac{\sigma_i}{\psi_j}\right),
\end{aligned}
\tag{4.5}
$$

and the corresponding joint probability density function is given as

$$
\begin{aligned}
f(1,t) &= LBA_{PDF}\left(1, t-\tau \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{1}{1+\exp(-\theta_i+b_j)}, \frac{\sigma_i}{\psi_j}\right) \\
f(2,t) &= LBA_{PDF}\left(2, t-\tau \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{1}{1+\exp(\theta_i-b_j)}, \frac{\sigma_i}{\psi_j}\right).
\end{aligned}
\tag{4.6}
$$

4.3.2 Meanings of the Parameters in the D-diffusion and D-LBA Models

To facilitate the understanding of parameters, in this section, we briefly show the relationships between the parameter values and the observed quantities, which are the RTs and item responses.

Expected RT Figure 4.1 shows the relationships between the expected RTs and $(\theta_i - b_j)$. This $(\theta_i - b_j)$ is equal to the drift rate in the D-diffusion model (Equation 2.7) and is a one-to-one with the drift rate in the D-LBA model (Equation 4.3); thus, we comprehensively call it the drift rate component here. Each of the three lines correspond to different boundary parameter values. We can observe two major characteristics common to both models. First, the expected RT is longer when the absolute difference $\left|\theta_i - b_j\right|$ is close to zero. Second, the expected RT is longer when the boundary (γ_i/ξ_j) is larger. In the D-diffusion model, the expected RT is given

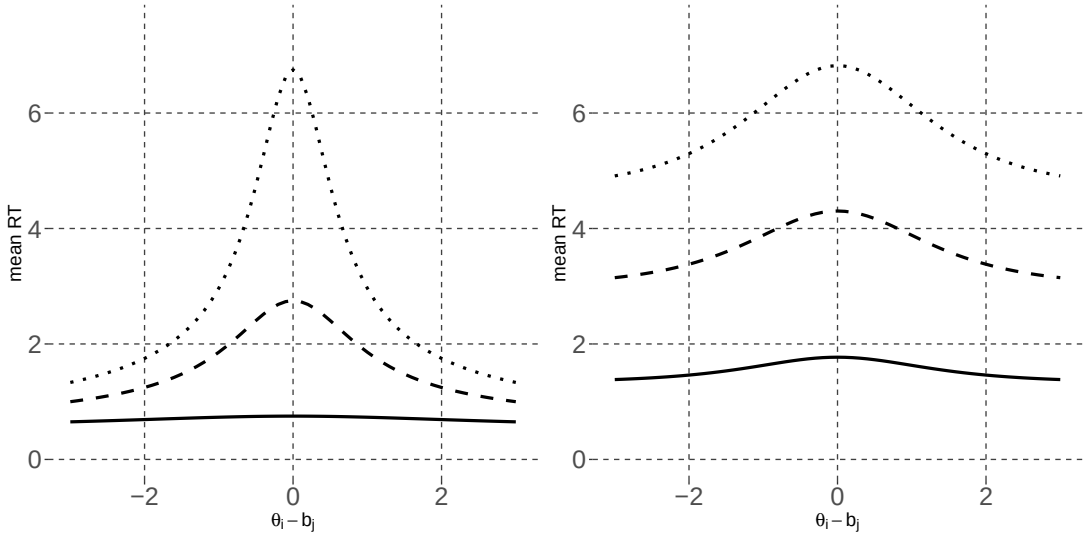

Figure 4.1: Relationship between the drift rate component ($\theta_i - b_j$) and expected RT. Left panel: D-diffusion model; Right panel: D-LBA model; Solid line: $\gamma_i/\xi_j = 1$; Dashed line: $\gamma_i/\xi_j = 3$; Dotted line: $\gamma_i/\xi_j = 5$.

by (Tuerlinckx et al., 2016)

$$E(t_{ij}) \simeq \begin{cases} \dfrac{1}{2\left|\theta_i - b_j\right|}\left(\dfrac{\gamma_i}{\xi_j}\right) & \text{if } \left|\theta_i - b_j\right| \neq 0 \\ \\ \dfrac{1}{4}\left(\dfrac{\gamma_i}{\xi_j}\right)^2 & \text{if } \left|\theta_i - b_j\right| = 0. \end{cases} \tag{4.7}$$

From Equations 4.7, 2.5, and 2.7, we can see that both the expected RT and the discriminability are functions of the boundary parameters. More specifically, the expected RT is approximately quadratic function of the boundary parameter when $\left|\theta_i - b_j\right|$ is zero, and otherwise approximately linear function of it. For example, if the expected RT is five seconds when $\left|\theta_i - b_j\right|$ is equal to zero, the discriminability becomes $\sqrt{20} \simeq 4.47$. Figure 4.2 plots the relationship between the expected RT and the boundary parameter in the D-diffusion model.

Item response Figure 4.3 shows the relationship between the probability of choosing the first response category and the drift rate component ($\theta_i - b_j$). Each of the three lines corresponds

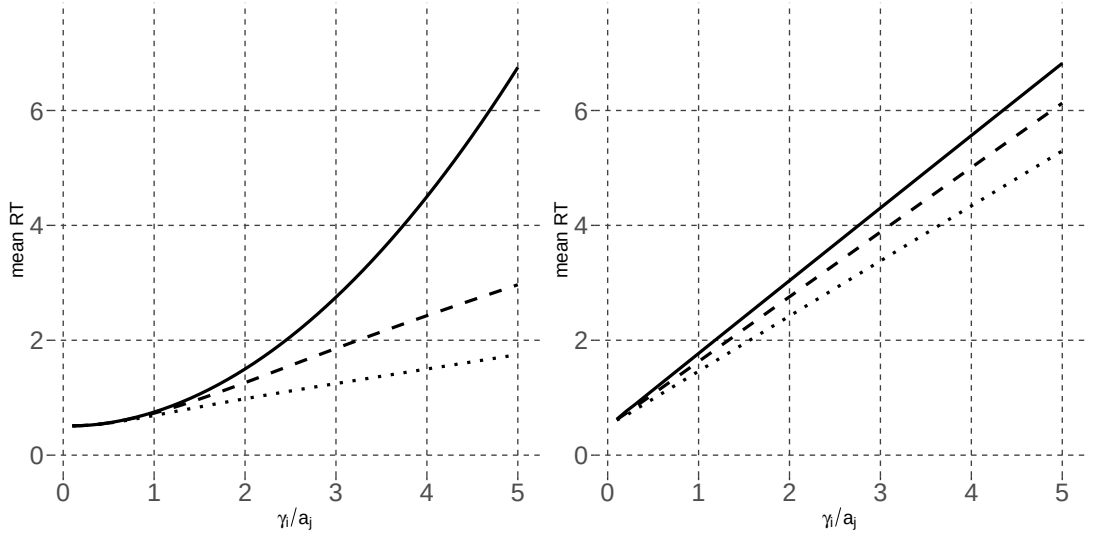

Figure 4.2: Relationship between the boundary (γ_i/ξ_j) and the expected RT. Left panel: D-diffusion model; Right panel: D-LBA model; Solid line: $|\theta_i - b_j| = 0$; Dashed line: $|\theta_i - b_j| = 1$; Dotted line: $|\theta_i - b_j| = 2$.

to different boundary parameter values. In the D-diffusion model, the discriminability differs in conjunction with the boundary parameters. In the D-LBA model, on the other hand, the discriminability does not differ even when the boundary parameters differ. This suggests that, in the D-LBA model, the boundary parameters only affect the expected RT. Here, Figure 4.4 indicates how the between-trial variability in the slope, η_{ij}, affects the discriminability and expected RT. The results suggest that η_{ij} affects both the discriminability and expected RT.

The relationship shown in Figure 4.4 shows the advantage of the proposed D-LBA model. The D-LBA model is free from the strong and unrealistic assumption of linearity between the discriminability and the expected RT by incorporating two different parameters B_{ij} and η_{ij}. In contrast, the D-diffusion model has this linearity. This may limit the applicability of the D-diffusion model to empirical data that has a longer RT than the typical context of the diffusion model.

4.3.3 Prior Distribution

The proposed D-LBA model can be numerically estimated using the MCMC estimation method. For this, the prior distributions for each parameter need to be set. In this section, the following

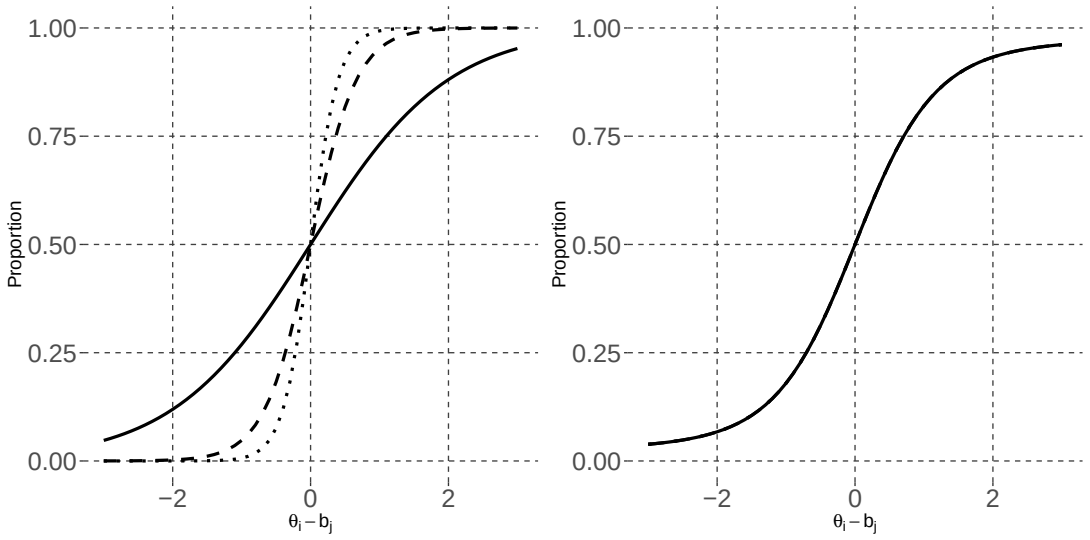

Figure 4.3: Relationship between the drift rate component $(\theta_i - b_j)$ and the probability of choosing the first category. Left panel: D-diffusion model; Right panel: D-LBA model; Solid line: $\gamma_i/\xi_j = 1$; Dashed line: $\gamma_i/\xi_j = 3$; Dotted line: $\gamma_i/\xi_j = 5$.

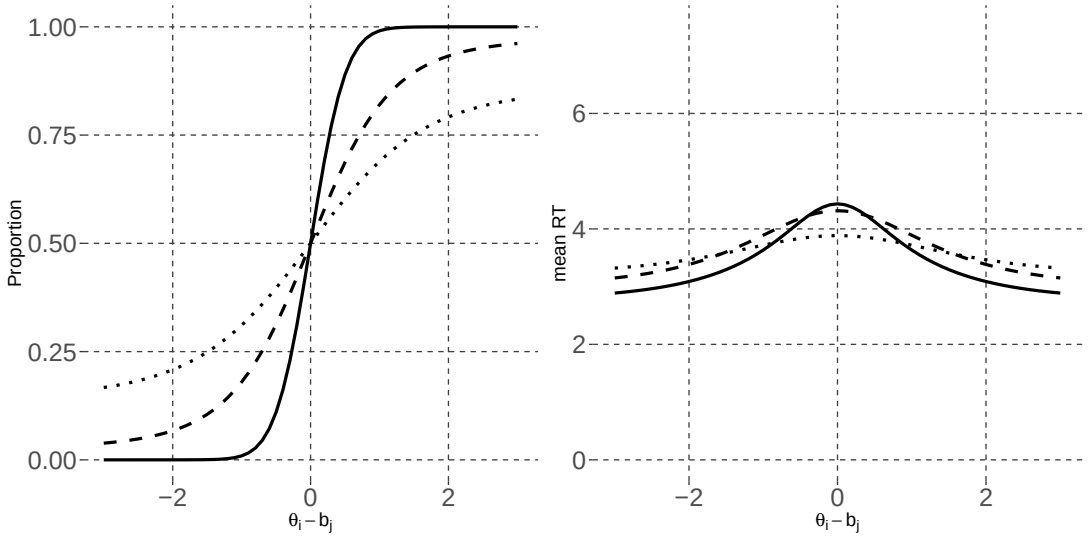

Figure 4.4: Relationships between the parameters and the observed quantities in the D-LBA model with different intertrial variabilities η_{ij} in the D-LBA model. Solid line: $\eta_{ij} = 0.1$; Dashed line: $\eta_{ij} = 0.3$; Dotted line: $\eta_{ij} = 0.5$. Left panel: relationship between the drift rate component $(\theta_i - b_j)$ and the probability of choosing the correct response. Right panel: relationship between the expected RT and the drift rate component $(\theta_i - b_j)$.

priors were used:

$$
\begin{aligned}
&\xi_j \sim Cauchy_{[0,\infty)}(0,5), && \gamma_i \sim LN(0,1),\\
&b_j \sim N(0,2.5^2), && \theta_i \sim N(0,1^2),\\
&\psi_j \sim Cauchy_{[0,\infty)}(0,5), && \sigma_i \sim LN(0,1),\\
&\tau_j \sim Cauchy_{[0,\infty)}(0,5),
\end{aligned}
\tag{4.8}
$$

where $Cauchy_{[0,\infty)}(\cdot)$ denotes truncated Cauchy distribution. Here, θ_i follows the standard normal distribution for identification purposes; this is a popular assumption in IRT models. Other respondent parameters, γ_i and σ_i, follow the standard lognormal distribution for the same reason. On the other hand, we set weakly informative priors (Gelman, Jakulin, Pittau, & Su, 2008) to all item parameters. For determining the hyperparameters, a sensitivity analysis was conducted, and the results showed that even when $Cauchy_{[0,\infty)}(0,5)$ was set for ξ_j, ψ_j, and τ_j, each posterior mean negligibly differs from a very weakly informative prior ($Cauchy_{[0,\infty)}(0,100)$). Thus, the abovementioned priors were adopted for computationally efficient estimation.

In the original LBA model, some parameters (viz., B, A, and η) are permitted to differ between different response categories. As pointed out by Heathcote and Love (2012), allowing this flexibility can make the model fit better. However, in this study, these parameters are assumed to be equal between response categories for the following reasons. First, the proposed D-LBA model is meant to be a simpler alternative to the D-diffusion model; thus, we would not want to increase the number of parameters unless they are of substantial importance. Second, this constraint facilitates faster and stable estimation.

4.4 Simulation Study

In this simulation study, several issues were investigated. First, parameter recovery was assessed. If the proposed model cannot properly recover parameters with the simulation data, it is pointless to examine other issues. Second, the similarity between the D-diffusion and D-LBA models was investigated. Because the parameters of these two models do not analytically map

to each other, the empirical correspondence between the parameters needs to be investigated. Although the D-LBA and D-diffusion models may have similar parameter interpretations, they have different parameter scales. Hence, the similarity was evaluated by correlations rather than absolute differences. Third, both models were compared in terms of the number of iterations to convergence. Previous studies have shown that the LBA model is simpler than the diffusion model. If so, it may be natural to expect the D-LBA model to converge faster than D-diffusion, even if new parameters are added to the D-LBA model. Finally, the model performance was examined according to the information criteria.

Conditions In this simulation, artificial data were generated from both the D-diffusion and D-LBA models, and the parameters were estimated using both models. In addition, $3 \times 3 = 9$ conditions were set by combining the following factors: (1) the number of respondents $I = (100, 300, 500)$ and (2) the number of items $J = (10, 20, 30)$. Overall, $3 \times 3 \times 2 \times 2 = 36$ conditions were simulated with 20 replications for each condition.

Data generation To generate simulation data from the D-LBA IRT model, the following distributions were used:

$$\begin{aligned} \xi_j &\sim U(0.5, 3), & \gamma_i &\sim LN(0, 1), \\ b_j &\sim U(-3, 3), & \theta_i &\sim N(0, 1^2), \\ \psi_j &\sim U(1, 4), & \sigma_i &\sim LN(0, 1), \\ \tau_j &\sim U(0.1, 0.5). \end{aligned} \tag{4.9}$$

When generating data from D-diffusion, the same distributions were used except $\xi_j \sim U(0.3, 0.7)$ owing to the difference in the parameter scale. We set the range of parameters in Equation 4.9 following Donkin et al. (2011). Furthermore, s was fixed to 1 in the diffusion model.

Using the true parameter values generated from Equation 4.9, simulation data were generated with the `rdiffusion` function in the `rtdists` R package. In this process, we found that a very small proportion of the generated RTs were greater than 120 s. However, such data make estimation (calculation of the log-probability) difficult and unstable for technical reasons,

Table 4.1: Descriptive statistics of the RTs generated by the simulation.

	>120 s	mean	2.5%	50%	97.5%
D-diffusion	0.054%	2.111s	0.254s	1.066s	10.587s
D-LBA	0.007%	1.539s	0.282s	0.907s	6.635s

especially in the case of diffusion IRT[1]. Moreover, in practice, the observation of such large RT data is unlikely; if they exist, they are usually excluded from the analysis as "lazy" responses. For these reasons, we used only data for which the RT was less than 120 s.

Table 4.1 summarizes the descriptive statistics of the RTs used in this simulation. As seen in the first column, the proportion of excluded data is very slight. The following columns indicate the mean, 2.5% quantile, median, and 97.5% quantile of the data that were used after exclusion. The range of the RT can be considered adequate and relevant.

Prior distributions For estimation by the proposed D-LBA model, our priors are those in Equation 4.8. The same prior was used for the D-diffusion model, except ψ_j and σ_i.

For all of the estimation results presented in this chapter, we used R (3.4.0) and `rstan` 2.15.0 on a Windows 10 PC. Three MCMC chains were run for each dataset. The number of MCMC iterations per chain was 10,000, 9,000 of which were discarded as warmup. The Stan code for LBA was obtained from the work of Annis, Miller, and Palmeri (2017) and extended to the D-LBA model. The Stan codes used in this study is provided in Appendices C and D, or the Open Science Framework website (`https://osf.io/ck7fr/`). The posterior means of the parameters were chosen as their point estimates.

Results

Parameter recovery (RMSE and bias) Table 4.2 lists the mean RMSE values for each condition when both the generation and estimation models are the same. Since γ_i and σ_i are log-normally distributed, the RMSE between the log-transformed estimates and the log-transformed true values for γ_i and σ_i are shown for these parameters. A comparison of the absolute values

[1]When estimated with Stan, the built-in `wiener_lpdf` function sometimes produces log(0) and returns an error. This error occurs more often when the response time is longer.

between the two models may not be meaningful because the scales of the parameters are different. Nevertheless, it is evident that as the number of respondents I increases, the RMSEs for all item parameters decrease in both the D-diffusion and D-LBA models. On the other hand, the RMSEs for the respondent parameters did not decrease as a function of I. This can be attributed to the well-known "Neyman–Scott paradox" (Neyman & Scott, 1948). Specifically, in some IRT models, the estimates of the respondent parameters do not converge to the true value even in the large-sample limit because the number of parameters also increases as the number of respondents increases.

Table 4.2: Mean RMSE for each model. The SD is stated within parentheses.

	I	J	ξ_j	$\log(\gamma_i)$	b_j	θ_i	τ_j	ψ_j	$\log(\sigma_i)$
D-diffusion	100	10	0.062(0.047)	0.175(0.075)	0.144(0.050)	0.465(0.046)	0.003(0.002)	-	-
		20	0.119(0.054)	0.224(0.079)	0.152(0.050)	0.391(0.051)	0.003(0.001)	-	-
		30	0.213(0.080)	0.348(0.103)	0.144(0.040)	0.364(0.052)	0.003(0.001)	-	-
	300	10	0.032(0.022)	0.150(0.024)	0.093(0.031)	0.482(0.032)	0.001(0.000)	-	-
		20	0.039(0.026)	0.119(0.034)	0.094(0.027)	0.390(0.028)	0.001(0.000)	-	-
		30	0.055(0.033)	0.128(0.048)	0.085(0.028)	0.338(0.028)	0.001(0.000)	-	-
	500	10	0.029(0.016)	0.145(0.017)	0.070(0.026)	0.471(0.023)	0.001(0.000)	-	-
		20	0.035(0.018)	0.115(0.021)	0.072(0.021)	0.389(0.027)	0.001(0.000)	-	-
		30	0.037(0.016)	0.108(0.024)	0.065(0.012)	0.340(0.018)	0.001(0.000)	-	-
D-LBA	100	10	0.230(0.128)	0.302(0.056)	0.435(0.158)	0.670(0.071)	0.010(0.004)	0.525(0.208)	0.509(0.046)
		20	0.323(0.167)	0.276(0.054)	0.429(0.138)	0.548(0.058)	0.010(0.004)	0.616(0.219)	0.401(0.047)
		30	0.456(0.171)	0.288(0.051)	0.381(0.096)	0.523(0.058)	0.010(0.003)	0.674(0.228)	0.348(0.045)
	300	10	0.120(0.057)	0.281(0.034)	0.285(0.133)	0.684(0.055)	0.005(0.003)	0.330(0.180)	0.485(0.020)
		20	0.176(0.099)	0.252(0.028)	0.233(0.089)	0.564(0.032)	0.005(0.001)	0.317(0.105)	0.373(0.026)
		30	0.199(0.097)	0.224(0.035)	0.228(0.067)	0.517(0.037)	0.004(0.001)	0.308(0.103)	0.321(0.024)
	500	10	0.094(0.056)	0.284(0.020)	0.175(0.106)	0.650(0.053)	0.004(0.002)	0.224(0.121)	0.478(0.018)
		20	0.106(0.060)	0.233(0.018)	0.176(0.099)	0.553(0.031)	0.003(0.001)	0.212(0.083)	0.360(0.017)
		30	0.105(0.062)	0.209(0.017)	0.167(0.072)	0.495(0.027)	0.003(0.001)	0.209(0.069)	0.311(0.020)

Table 4.3: Mean bias for each model.

	I	J	ξ_j	$\log(\gamma_i)$	b_j	θ_i	τ_j	ψ_j	$\log(\sigma_i)$
D-diffusion	100	10	-0.044	-0.084	0.020	0.021	0.000	-	-
		20	-0.109	-0.194	-0.030	-0.032	0.001	-	-
		30	-0.205	-0.338	0.022	0.008	0.000	-	-
	300	10	-0.023	-0.056	-0.012	-0.020	0.000	-	-
		20	-0.030	-0.061	0.001	-0.001	0.001	-	-
		30	-0.049	-0.095	0.009	0.007	0.000	-	-
	500	10	-0.013	-0.035	-0.019	-0.010	0.000	-	-
		20	-0.028	-0.059	-0.013	-0.014	0.000	-	-
		30	-0.033	-0.070	-0.008	-0.002	0.001	-	-
D-LBA	100	10	-0.129	-0.090	0.008	-0.028	0.000	-0.244	-0.145
		20	-0.269	-0.163	0.066	0.006	0.001	-0.428	-0.189
		30	-0.405	-0.215	0.007	0.001	0.000	-0.490	-0.172
	300	10	-0.054	-0.067	0.023	-0.005	0.001	-0.134	-0.137
		20	-0.142	-0.100	0.005	-0.003	0.000	-0.134	-0.098
		30	-0.163	-0.101	0.006	0.010	0.000	-0.186	-0.099
	500	10	-0.050	-0.064	0.001	0.005	0.001	-0.050	-0.110
		20	-0.057	-0.052	0.006	0.000	0.000	-0.064	-0.074
		30	-0.076	-0.063	-0.011	-0.006	0.000	-0.106	-0.077

Table 4.3 summarizes the mean biases for each condition. From the table, the biases for all parameters were close to zero. This suggests that the proposed D-LBA model and D-diffusion model produce good parameter estimates on average.

From the viewpoint of the RMSEs and biases, it can be concluded that the proposed model parameters are properly estimated and that parameter recovery is as good as that of the D-diffusion model.

Correspondence (correlations) Tables 4.4–4.8 summarize the correlation of each parameter for each condition. With regard to the correlations for the conditions in which data were generated with the D-diffusion model and estimated with the D-LBA model, all of the parameters except ξ_j took sufficient values (greater than 0.9). Even though ξ_j had lower correlations than the others, this is consistent with the results of Donkin et al. (2011). In addition, the proposed model expresses the relationship between the discriminability and the RT with two parameters, B_{ij} and η_{ij}, while the D-diffusion model uses only α_{ij} to express this relationship. This would explain the result that the correlation of ξ_j is lower, particularly when the data were generated

Table 4.4: Correlation between the true parameter value of ξ_j in the data generating model and its estimate in the estimation model.

Data generating model		D-diffusion		D-LBA	
Estimation model		D-diffusion	D-LBA	D-diffusion	D-LBA
I	J				
100	10	0.973	0.407	0.920	0.990
	20	0.982	0.598	0.905	0.991
	30	0.981	0.665	0.890	0.989
300	10	0.994	0.519	0.942	0.996
	20	0.993	0.547	0.939	0.996
	30	0.993	0.553	0.936	0.996
500	10	0.996	0.530	0.937	0.998
	20	0.996	0.582	0.932	0.998
	30	0.996	0.603	0.948	0.998

from the D-diffusion model and estimated with the D-LBA model. These results, especially those in Table 4.7, show an interesting point. When the data generation and estimation models were the same, the D-diffusion model exhibited a higher correlation than the D-LBA model. On the other hand, when data generation and estimation models were different, the D-LBA model exhibited a higher correlation than the D-diffusion model. This suggests that when the data generation model is known to be the D-diffusion model, the true D-diffusion model seems to indicate higher performance than the D-LBA model. However, when the data generation model is unknown, which is natural in a practical situation, the D-LBA model indicates more stable performance regardless of the true data generation model.

Estimation efficiency ($\hat{R}$ and effective sample size) We computed the Gelman–Rubin diagnostic statistic ($\hat{R}$; Gelman et al., 2014) as a convergence diagnostic measure. When the number of respondents is small, both models converge quite fast (fewer than 1,000 iterations when $I = 100$). Therefore, we will show the results when the number of respondents is the largest under the conditions that were considered.

Figure 4.5 shows the results for the condition $I = 500$ and $J = 30$ with the D-LBA model as the data generation model. The x axis represents the warmup iterations (1,000 iterations after this 9,000 warmup iterations were used for posterior estimation). The y axis represents

Table 4.5: Correlation between the true parameter value of $\log(\gamma_i)$ in the data generating model and its estimate in the estimation model.

Data generating model		D-diffusion		D-LBA	
Estimation model		D-diffusion	D-LBA	D-diffusion	D-LBA
I	J				
100	10	0.992	0.947	0.917	0.960
	20	0.996	0.968	0.939	0.976
	30	0.997	0.977	0.942	0.982
300	10	0.992	0.936	0.921	0.964
	20	0.996	0.964	0.932	0.973
	30	0.997	0.974	0.940	0.980
500	10	0.991	0.938	0.918	0.962
	20	0.996	0.966	0.931	0.974
	30	0.997	0.974	0.941	0.980

Table 4.6: Correlation between the true parameter value of b_j in the data generating model and its estimate in the estimation model.

Data generating model		D-diffusion		D-LBA	
Estimation model		D-diffusion	D-LBA	D-diffusion	D-LBA
I	J				
100	10	0.998	0.958	0.944	0.981
	20	0.998	0.971	0.947	0.984
	30	0.998	0.976	0.934	0.982
300	10	0.999	0.964	0.946	0.990
	20	0.999	0.978	0.963	0.993
	30	0.999	0.980	0.952	0.993
500	10	>0.999	0.964	0.970	0.996
	20	>0.999	0.980	0.961	0.996
	30	>0.999	0.980	0.955	0.996

Table 4.7: Correlation between the true parameter value of θ_i in the data generating model and its estimate in the estimation model.

Data generation model		D-diffusion		D-LBA	
Estimation model		D-diffusion	D-LBA	D-diffusion	D-LBA
I	J				
100	10	0.882	0.739	0.612	0.736
	20	0.921	0.853	0.695	0.836
	30	0.935	0.894	0.738	0.856
300	10	0.879	0.721	0.596	0.727
	20	0.924	0.830	0.708	0.826
	30	0.941	0.892	0.741	0.861
500	10	0.880	0.716	0.652	0.754
	20	0.922	0.838	0.719	0.831
	30	0.941	0.887	0.745	0.870

Table 4.8: Correlation between the true parameter value of τ_j in the data generating model and its estimate in the estimation model.

Data generation model		D-diffusion		D-LBA	
Estimation model		D-diffusion	D-LBA	D-diffusion	D-LBA
I	J				
100	10	>0.999	0.999	0.986	0.995
	20	>0.999	0.999	0.988	0.996
	30	>0.999	0.999	0.985	0.997
300	10	>0.999	>0.999	0.997	0.999
	20	>0.999	>0.999	0.995	0.999
	30	>0.999	>0.999	0.995	0.999
500	10	>0.999	>0.999	0.997	0.999
	20	>0.999	>0.999	0.998	>0.999
	30	>0.999	>0.999	0.997	>0.999

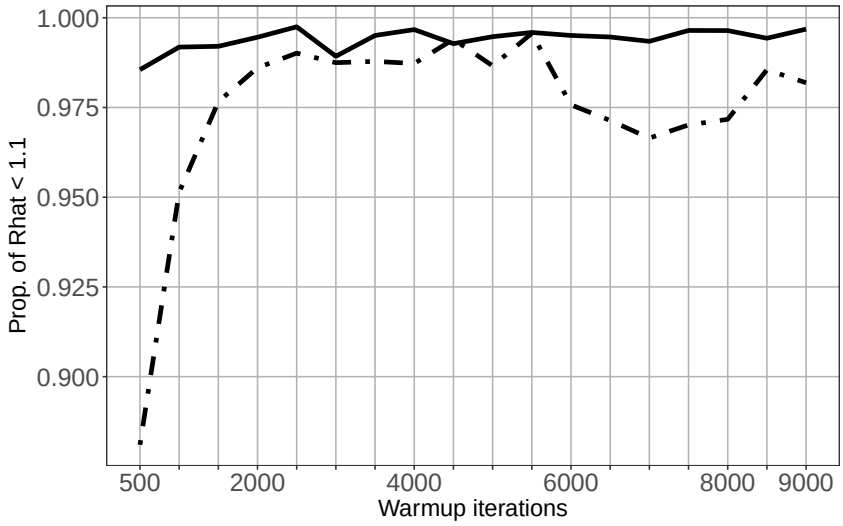

Figure 4.5: Proportion of parameters for which $\hat{R}$ is lower than 1.1 when the data generation model is the D-LBA model. The solid line represents the results estimated by the D-LBA model. The dashed line represents the results estimated by D-diffusion.

the proportion for which $\hat{R}$ is below the threshold. Gelman et al. (2014) suggested a threshold of 1.1 for $\hat{R}$; therefore, the same threshold was adopted. These proportions were computed based on the MCMC samples from zero to the x-axis value iterations in increments of 500. The solid and dashed lines represent the results estimated by the D-LBA and D-diffusion models, respectively.

The results indicate that the D-LBA model converges much faster than D-diffusion, despite the fact that the number of parameters in the D-LBA model is larger. The D-LBA model took less than $1,000$ warmups to converge for more than 99% of the parameters, whereas D-diffusion seems to be more unstable. Figure 4.6, which shows the results when the data generation model was the D-diffusion model, also indicates that the D-LBA model converges several times faster than the D-diffusion model.

In addition, the average effective sample sizes were checked as a measure of the estimation efficiency. Table 4.9 summarizes the effective sample sizes under the condition $I = 500$ and $J = 30$. It is evident that the D-LBA model obtained larger effective sample sizes than the D-diffusion model for all parameters regardless of the true data generation model. This result was consistent under all conditions. The results suggest that the proposed model can estimate

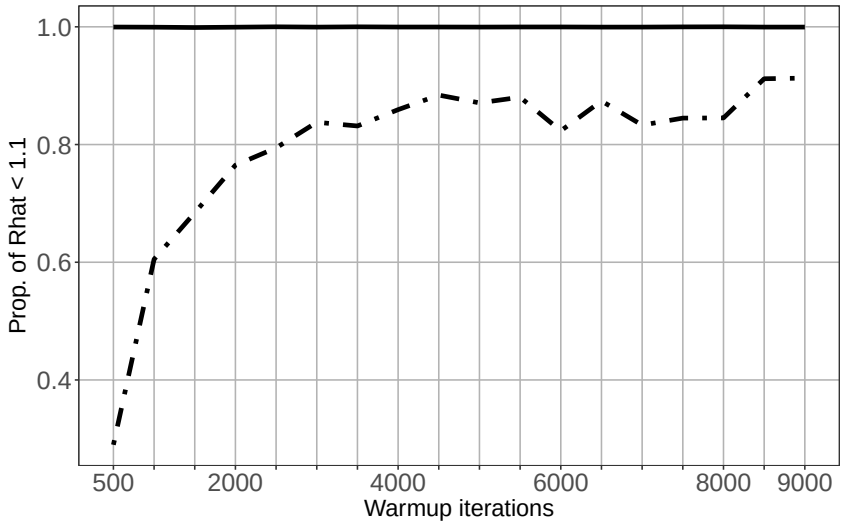

Figure 4.6: Proportion of parameters for which $\hat{R}$ is lower than 1.1 when the data generation model is the D-diffusion model. The solid and dashed lines have the same meanings as those in Figure 4.5.

Table 4.9: Average effective sample sizes for each parameter under the conditions $I = 500$ and $J = 30$.

Generation model	D-diffusion		D-LBA	
Estimation model	D-diffusion	D-LBA	D-diffusion	D-LBA
ξ_j	48.25	146.09	49.03	63.41
γ_i	333.72	938.00	320.33	781.89
b_j	188.29	2101.40	202.00	1164.00
θ_i	969.67	3615.56	2411.72	4153.28
τ_j	1253.42	2291.05	2739.38	3153.62
ψ_j	-	508.43	-	430.78
σ_i	-	1516.56	-	2171.01

parameters more efficiently than the D-diffusion model. Note that with regard to the empirical computational time per iteration, given the same computational environment, there exist no systematic differences between the two models. This means that the proposed D-LBA model provides a higher efficiency for estimation per iteration and per actual time.

Information criteria In addition to the results presented above, the fit of the models was assessed in terms of information criteria (WAIC: widely applicable information criterion and WBIC: widely applicable Bayesian information criterion; Vehtari, Gelman, & Gabry, 2017; Watanabe, 2010, 2013). Figure 4.7 shows the results under the largest data conditions ($I =$

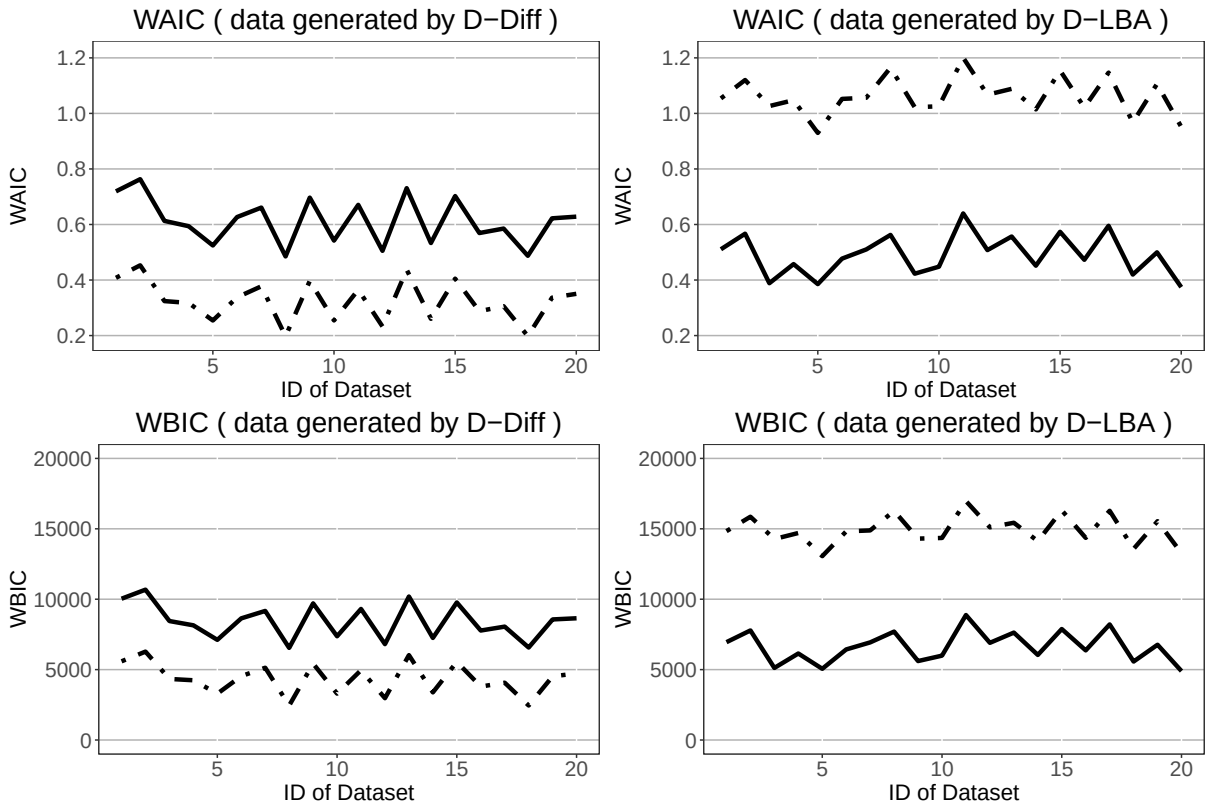

Figure 4.7: Results under the condition $I = 500$ and $J = 30$. Solid lines: estimated with D-LBA; Dot–dash lines: estimated with D-diffusion; Upper half: WAIC, Lower half: WBIC; Left side: data generated by D-diffusion, Right side: data generated by D-LBA.

$500, J = 30$). In all of these graphs, the solid lines represent the results estimated by the D-LBA model, and the dot–dash lines represent the results estimated by the D-diffusion model. The upper half represents the WAIC, and the lower half represents the WBIC. The graphs on the left are for data generated by D-diffusion, whereas those on the right are for data generated by the D-LBA model. For all datasets generated by the D-LBA model, both indices become lower when estimated by the D-LBA model. On the other hand, for all datasets generated by the D-diffusion model, the D-LBA model shows worse values than D-diffusion. These results are expected, and the same results were confirmed under all conditions.

4.5 Real Data Application: Extraversion Data

In this section, we consider a more realistic situation by using real data in order to examine the applicability of the proposed D-LBA model.

Data We used the `extraversion` data in the `diffIRT` R package. These data, obtained by Molenaar et al. (2015), comprise 146 respondents for 10 items. Each item is a particular word or phrase related to extraversion behavior (e.g., *"active"* or *"noisy"*). Respondents were asked whether each item is appropriate to their personalities. For all respondents and all items, the actual response (yes/no) and RT were recorded, some of which are missing.

Results

Table 4.10 summarizes the estimates of the item parameters obtained by both models. The correlations of the ξ_j, b_j, and τ_j parameters between the D-diffusion and D-LBA estimates were 0.707, 0.806, and 0.851, respectively. Although they are slightly lower than the results of the simulation study, given the relatively small number of respondents ($I = 143$), it can be said that the D-diffusion and D-LBA models are substantially similar models.

In addition, Figure 4.8 shows the posterior density for each item parameter. In the figure, the left-hand side shows the plot for ξ_j, and the right-hand side shows the plot for b_j. From this figure, we can see that each parameter estimate was properly obtained because each density has only one peak. Moreover, the parameter estimates for each model were highly correlated with the summary statistics. For items having a high proportion of "yes" responses (e.g., "*viable*" and "*eupeptic*"), b_j become lower. As for ξ_j, it corresponds with the mean response time (MRT). This can be considered evidence for the validity of the estimates.

Table 4.11 summarizes the mean effective sample sizes for both models. All but τ_j indicated higher values in the D-LBA model than in the D-diffusion model. This result suggests that the D-LBA model can estimate parameters more efficiently than the D-diffusion model, even for real data.

Furthermore, a posterior predictive check was conducted to validate the model. Specifically,

Table 4.10: Item parameters obtained by D-LBA and D-diffusion. Prop: proportion that answers "yes"; MRT: mean response time.

				ξ_j		b_j		τ_j		ψ_j
	Item	Prop.	MRT	D-LBA	D-diff	D-LBA	D-diff	D-LBA	D-diff	D-LBA
1	active	0.741	1.486	0.966	0.520	-2.019	-0.704	0.451	0.575	1.503
2	noisy	0.538	1.357	1.091	0.540	-0.087	-0.117	0.327	0.475	2.513
3	energetic	0.846	1.120	1.573	0.597	-2.032	-1.322	0.394	0.502	2.813
4	enthusiastic	0.916	1.000	2.090	0.579	-2.440	-1.854	0.398	0.458	2.957
5	impulsive	0.539	1.298	1.240	0.551	-0.240	-0.213	0.339	0.464	2.751
6	jovial	0.902	1.262	1.187	0.507	-4.338	-1.380	0.412	0.501	1.937
7	viable	0.937	1.142	1.306	0.490	-2.694	-1.788	0.372	0.511	3.393
8	eupeptic	0.958	1.090	1.419	0.454	-3.448	-2.065	0.352	0.434	2.955
9	communicative	0.824	1.728	0.786	0.408	-2.056	-0.904	0.472	0.609	2.088
10	spontaneous	0.860	0.986	1.848	0.632	-2.488	-1.527	0.370	0.462	2.750

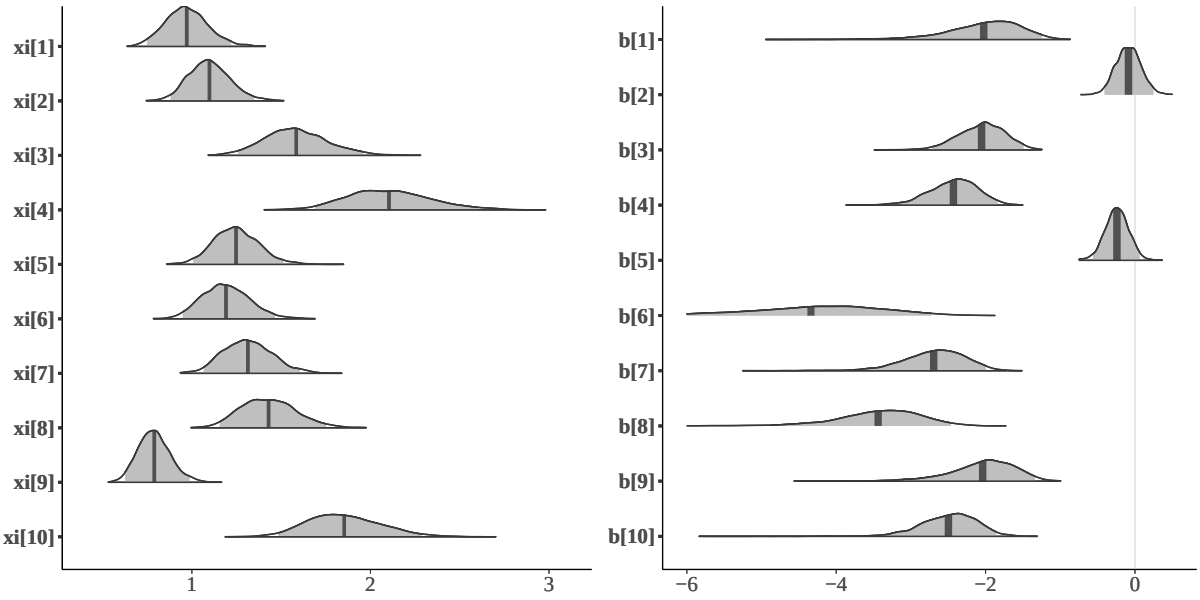

Figure 4.8: Posterior densities of the item parameters obtained by the D-LBA model. The posteriors of ξ_j are shown on the left-hand side, and the posteriors of b_j are shown on the right-hand side. The item indices correspond to those of Table 4.10.

Table 4.11: Mean effective sample size for each parameter in the D-diffusion and D-LBA models.

Estimation model	D-diffusion	D-LBA
ξ_j	101.84	260.60
γ_i	370.02	843.82
b_j	684.45	1255.58
θ_i	2465.66	3201.51
τ_j	2035.36	1706.55
ψ_j	-	720.89
σ_i	-	1414.07

15,000 random samples of responses and RTs were generated from the posterior predictive distribution. Then, the proportion of responses that were the same as the observed data for each respondent was calculated. Figures 4.9 and 4.10 show the histograms of this proportion for all 143 respondents for the D-LBA and D-diffusion models, respectively. Similarly, Figure 4.11 shows the posterior predictive distributions of the RT for the first respondent. Each line corresponds to the posterior predictive distribution for each item, and the black vertical line represents the probability density at the point of the observed RT. In Figure 4.11, longer black lines mean that the model performs better in terms of posterior prediction of the RT. These results indicate that the proposed D-LBA model adequately explains the observed data and that its predictive performance is at least as good as that of the D-diffusion model.

One of the major advantages of using the model-based parameters instead of much simpler descriptive statistics such as the MRT or the proportion of choosing the first category is that, while the theory of psychological measurement suggests that the observed data contain random fluctuations or errors, the substantially informed model parameters directly reflect the underlying psychological process that elicited the observed responses (Molenaar et al., 2015). The model also decomposes the observed information into several different meaningful sources of variability. For instance, the MRT is used as a property of an item, although it may be influenced by some respondents' traits, which were represented by the parameter γ_i. If more "deliberate" respondents were unintentionally collected, the observed MRT may be longer even if they answered the same items. However, the respondents' traits cannot be distinguished from the item traits as long as the simple MRT is used. By estimating both γ_i and ξ_j at the same time, ξ_j can

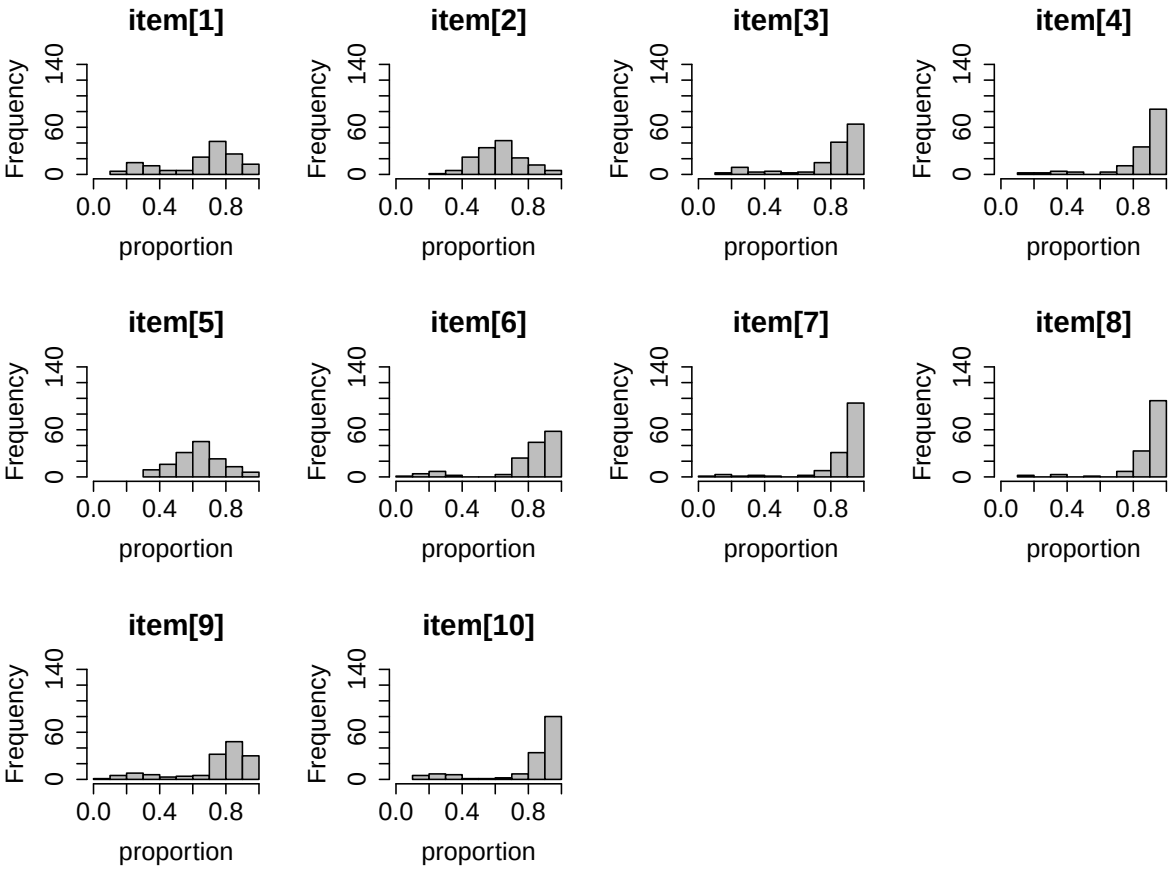

Figure 4.9: Histograms of the posterior predictive samples that correspond to the observed response in the D-LBA model (sample size: 143).

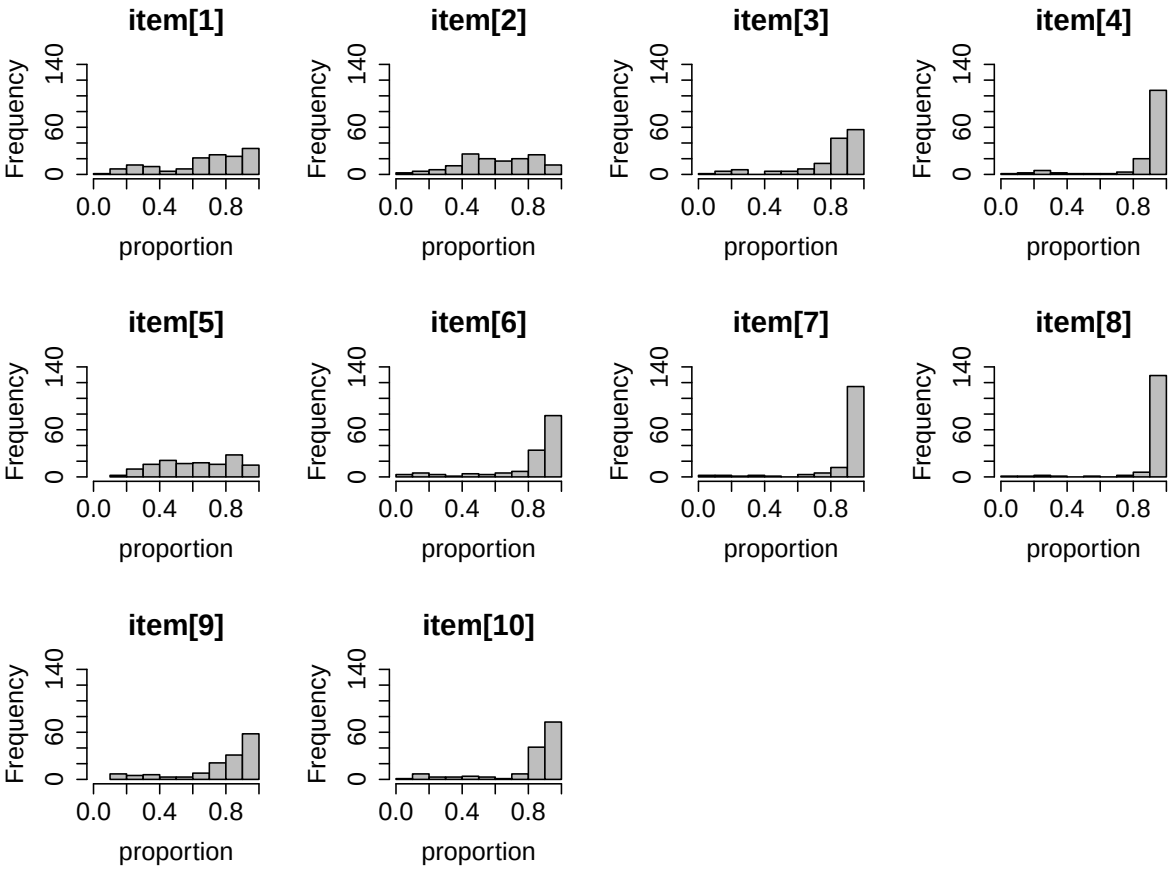

Figure 4.10: Histograms of the posterior predictive samples that correspond to the observed response in the D-diffusion model (sample size: 143).

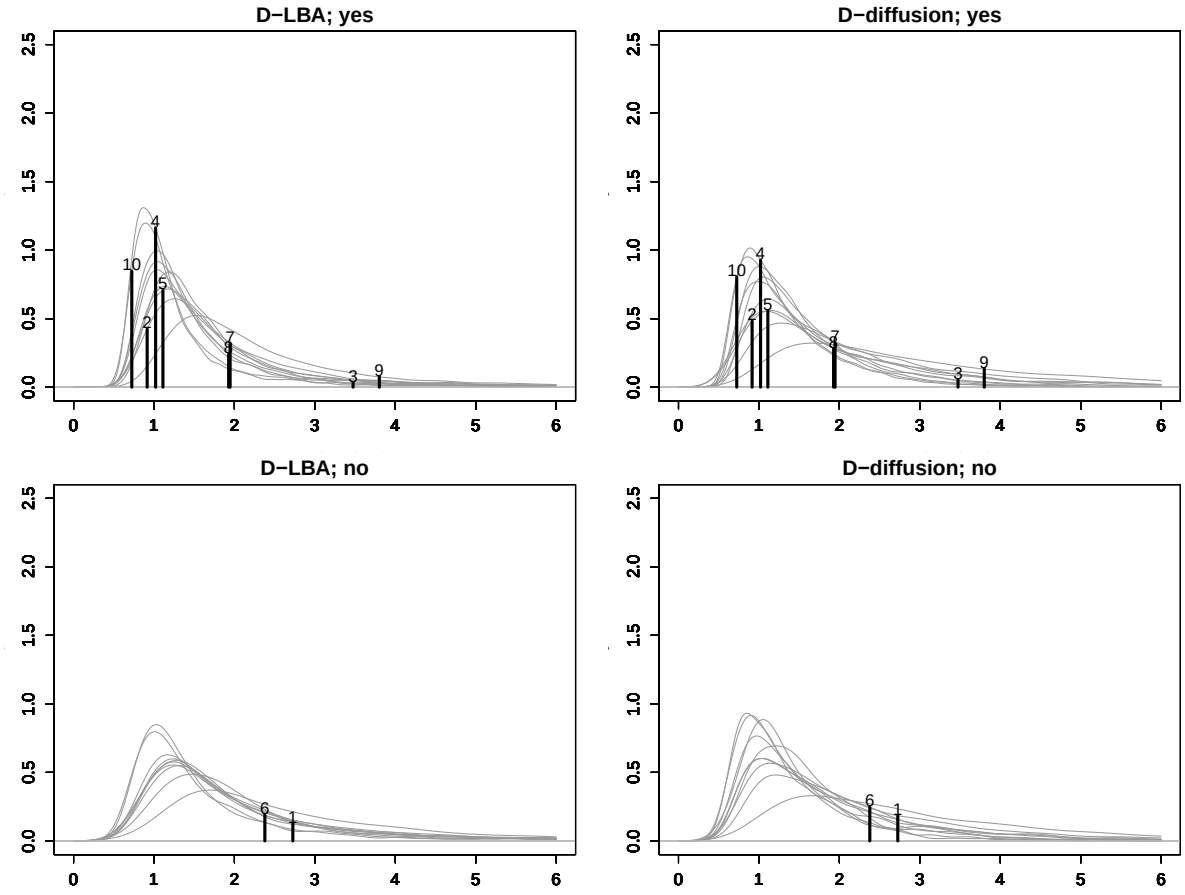

Figure 4.11: Posterior predictive distributions of the RT in the D-LBA (left side) and D-diffusion (right side) models along with the observed RT (black vertical line) for the first respondent. The upper half shows the distributions when the response is "yes," and the lower half shows the distributions when the response is "no." This respondent only answered "no" for items one and six.

be seen as that not influenced by respondents' traits. This also makes it reasonable to examine the item parameter estimates from a qualitative perspective without considering respondents' traits. For example, it may be difficult to provide an answer quickly and intuitively for an item having a higher ξ_j (e.g., "communicative"). In other words, such items seem to indicate more complicated meanings than those having a lower ξ_j. In addition, more respondents seem to answer "yes" for items having a lower b_j. From a qualitative viewpoint, items with a higher b_j tend to have negative meanings. For example, in the Oxford Advanced Learner's Dictionary (Bradbery, Deuter, & Turnbull, 2015), the word "noisy" is defined as "making a lot of noise," and the word "impulsive" is defined as "acting suddenly without thinking carefully about what might happen because of what you are doing." Obviously, these two words have more negative connotations than positive ones. Therefore, respondents may be reluctant to answer "yes" to these items.

One of the main interests in this application is comparative model fitting. Our primary objective here is to comparatively evaluate the proposed D-LBA model with existing D-diffusion model. However, in order to check whether or not more parsimonious model formulations show better performance than the full models, several more parsimonious model formulations were also examined. The conditions examined are shown in Table 4.12. With regard to boundary, four conditions were considered: when the boundary parameter depends both on the respondents and items, when it depends on the items but is common across respondents, when it depends on respondents but is common across items, and when it is common across both respondents and items. The specific parameterizations are shown in Table 4.12. Similarly, these four conditions were considered for the between-trial variance η_{ij} in the D-LBA model. As a result, $4 \times 4 = 16$ coniditions were conducted for the D-LBA model and 4 conditions for the D-diffusion model. Table 4.12 summarizes the obtained WAIC and WBIC for these 20 models. Among all models, the *full* D-LBA models has shown the best value in terms of both WAIC and WBIC. Particularly, the obtained WAIC values of the *full* D-diffusion and *full* D-LBA models were 0.829 and 0.757, whereas the WBIC values were 1024.98 and 898.71, respectively. Therefore, the results indicate that the proposed D-LBA model is better fitted than the D-diffusion model in

Table 4.12: Information criteria values of the full model as well as more parsimonious sub-models for the D-LBA and D-diffusion models.

Model	Boundary	Between-trial variance	Number of parameters	WAIC	WBIC
D-LBA	$B_{ij} = \gamma_i/\xi_j$	$\eta_{ij} = \sigma_i/\psi_j$	$4J + 3I$	0.757	898.71
		$\eta_{ij} = \sigma_i$	$3J + 3I$	0.777	968.52
		$\eta_{ij} = \psi_j$	$4J + 2I$	0.828	1023.02
		$\eta_{ij} = \eta$	$3J + 2I + 1$	0.784	982.39
	$B_{ij} = \xi_j$	$\eta_{ij} = \sigma_i/\psi_j$	$4J + 2I$	0.868	1123.83
		$\eta_{ij} = \sigma_i$	$3J + 2I$	0.909	1234.57
		$\eta_{ij} = \psi_j$	$4J + I$	0.939	1234.60
		$\eta_{ij} = \eta$	$3J + I + 1$	0.915	1241.96
	$B_{ij} = \gamma_i$	$\eta_{ij} = \sigma_i/\psi_j$	$3J + 3I$	0.816	989.34
		$\eta_{ij} = \sigma_i$	$2J + 3I$	0.818	1030.26
		$\eta_{ij} = \psi_j$	$3J + 2I$	0.818	1030.26
		$\eta_{ij} = \eta$	$2J + 2I + 1$	0.820	1040.59
	$B_{ij} = B$	$\eta_{ij} = \sigma_i/\psi_j$	$3J + 2I + 1$	0.880	1144.93
		$\eta_{ij} = \sigma_i$	$2J + 2I + 1$	0.931	1269.21
		$\eta_{ij} = \psi_j$	$3J + I + 1$	0.962	1273.85
		$\eta_{ij} = \eta$	$2J + I + 2$	0.944	1286.09
D-diffusion	$\alpha_{ij} = \gamma_i/\xi_j$	-	$3J + 2I$	0.829	1024.98
	$\alpha_{ij} = \xi_j$	-	$3J + I$	0.902	1184.12
	$\alpha_{ij} = \gamma_i$	-	$2J + 2I$	0.849	1052.88
	$\alpha_{ij} = \alpha$	-	$2J + I + 1$	0.914	1204.62

Note: When $\alpha_{ij} = \alpha$, $B_{ij} = B$ or $\eta_{ij} = \eta$, these parameters do not depend on respondent nor item. Also, when $\alpha_{ij}, = \gamma_i$, $B_{ij} = \gamma_i$ or $\eta_{ij} = \sigma_i$, the prior distributions were set to $Cauchy_{[0,\infty)}(0, 5)$ instead of $LN(0, 1)$ because identification constraints were not essential in these conditions.

terms of these information criteria with this dataset. In other words, the assumption that the item discriminability and expected RTs are completely correlated is unlikely in real data. In addition, the results would indicate that both α_{ij} and η_{ij} should be decomposed into item factors and person factors to show better fit to real data.

4.6 Discussion

In this study, a new cognitively-based IRT model was proposed that can explain the flexible relationship between the item discriminability and the expected RTs for personality assessment using RT information. The likelihood function of the proposed D-LBA IRT model can be essentially seen as a reparameterization of the LBA model. Our argument is that this reparameterization is the point: the LBA framework for modeling the response time data, which has been proved to be useful in the field of cognitive and mathematical psychology, has not been

applied to IRT models in psychometrics. Our contributions of this study is to clearly reveal the relationship between these two models, which are both popular in different fields, and to combine the strengths of both models to propose the D-LBA IRT model.

From the simulation results, four advantageous properties of the proposed D-LBA model were identified. First, the proposed model can recover parameters as sufficiently as the D-diffusion model. Second, each parameter in the proposed and D-diffusion models can be interpreted in nearly the same way. Third, the correlations between the true values and the estimates obtained from the proposed model are higher than those from the D-diffusion model when the true data generation model is different. Fourth, the proposed model converges much faster and estimates more effectively than the D-diffusion model. These findings suggest that the proposed D-LBA model is a more realistic, efficient, and practical yet simpler alternative to the D-diffusion model of the item response and RT.

In addition, the D-LBA and D-diffusion models were applied to a real personality measurement dataset. Consequently, from the viewpoint of the information criterion, the D-LBA model was found to fit this dataset better than the D-diffusion model.

By introducing a new parameter and extending the simple LBA model, the proposed model can mitigate the problems that originate from the diffusion IRT model. Nevertheless, in empirical applications of the proposed D-LBA model, three potentially significant issues might persist.

First, the time required for the MCMC estimation algorithm of the proposed D-LBA model might be substantial. In our simulation study, the proposed model took around $7,000$ iterations to achieve full convergence, which was judged on the basis of $\hat{R}$ when all parameters were less than 1.1. This corresponds to a few hours (with $I = 500$ and $J = 30$) in our computational environment (CPU: Intel Core i7-7700K; Memory: 64 GB; Operating system: Windows 10). Note that a greater computational time was needed for the existing D-diffusion model; at times, even $40,000$ iterations were not sufficient for convergence. Nevertheless, for researchers who want to analyze real data, the estimation time of the proposed model might not be sufficiently fast. In this case, one may be able to use variational Bayes (VB) inference instead of MCMC estimation

for the proposed D-LBA model. Using `rstan`, it is easy to apply automatic differentiation variational inference (ADVI) without the need to specify the approximating variational distribution. With this approach, researchers need to satisfy only one condition: each parameter should be approximately transformed to a normal distribution.

Second, our proposed-approach is model based, and therefore, strictly speaking, the advantages of the model only apply when the underlying model is correct (Tuerlinckx et al., 2016). However, as is often said, "all models are wrong, but some are useful." We believe, in line with Box (1979), that the relevant question is not whether the model assumptions are met exactly, but rather whether the model is illuminating and useful enough as an approximation of reality. Based on the model comparison and posterior predictive check that were present in this study, we are particularly positive about empirical applicability of the model. That being said, more thorough investigation regarding the fit and prediction of the model, such as the evaluation of person fit (e.g., Ferrando, 2007), would be desirable in future studies.

Finally, one of the major advantages of the IRT framework is that thanks to the decomposition of observed data variability into item and respondent parameters, items can be scaled independently of respondents, and likewise respondents can be scaled independently of items. This characteristic would be particularly advantageous in situations when the observed data are accumulated from different sets of samples and items over time, as in the typical application of IRT for educational measurement. On the other hand, in typical personality assessment, items do not change across respondents. However, recent technological development allows us to administer large-scale personality assessment in which different items are presented to different respondents, and to perform modeling of such data (e.g., Condon & Revelle, 2015; Okada, Vandekerckhove, & Lee, 2018). Therefore, the proposed D-LBA IRT model may be applicable in such modern personality measurement studies in which response time is also corrected. This can be a fruitful direction of future research.

One of the major advantages of the D-LBA model over the D-diffusion model is that the model is conceptually applicable to the MAFC tasks. However, in this chapter, we applied the D-LBA model to a unidimensional binary personality scale in order to compare the results with

those of the existing D-diffusion model. In the next chapter, we will show an extension of the D-LBA model to the multidimensional MAFC tasks and also give some real data examples.

Chapter 5

Study 3: Multidimensional MAFC D-LBA IRT Model

5.1 Objective of the Study

For the MAFC personality measurement data, the author is not aware of any models that consider the RT. This can be due to the following two difficulties in combining the TIRT model and the inverted-U relationship. First, the TIRT model internally codes an MAFC item response by a set of implied pairwise comparisons (A. Brown & Maydeu-Olivares, 2013). For instance, when a respondent chooses the best option in a four-alternative forced-choice (4AFC) item, the TIRT model codes this single polytomous item response as an implied set of three implied pairwise comparisons among the chosen and unchosen alternatives, which are mutually dependent by design. However, we can only observe a single RT value that corresponds to these three implied pairwise comparisons. Due to this mismatch in the number of data points between the binary-coded item response and RT, developing a joint model involving both of them is not a straightforward but requires a substantial modeling effort. Second, in the MAFC format, respondents need to judge multiple statements in a presented item as opposed to just one in Likert and binary formats. In fact, Sass, Frick, Reips, and Wetzel (n.d.) found empirical evidence that respondents have several distinct underlying decision processes in responding to the MAFC

item, and that many respondents switch between the processes even within a single questionnaire. Due to the above reasons, the single inverted-U relationship would not be appropriate for the MAFC response and its RT. Instead, we need a different approach for this type of data.

In Chapter 4, the unidimensional binary D-LBA IRT model has been proposed. Actually, one of the key characteristics of the LBA model as opposed to the diffusion model is that it can represent the process underlying forced-choice tasks with more than two alternatives. This is because whereas the diffusion model assumes a single accumulator, the LBA model assumes multiple accumulators, each of which corresponds to a choice alternative.

Nevertheless, an extension of the D-LBA IRT model for MAFC personality measurement has not been developed. As outlined above, such an extension is not straightforward, but requires a careful combination of the two different foundations. More specifically, in the unidimensional binary D-LBA IRT model, both accumulators can be represented based on the relative position of the respondent to the item on a single psychological dimension. However, in the MAFC format, each alternative corresponds to different psychological dimensions. In addition, all choice alternatives are simultaneously compared with each other. These properties make it difficult to define the accumulators of the LBA process in maintaining theoretical soundness and enabling estimation.

The main objective of this study is to propose a model that solves this problem. To tackle the problem, we apply the Thurstone's (1927) random utility theory to the LBA framework, and reformulate an LBA parameter as the function of the latent utility of each choice alternative. After formulating the proposed D-LBA IRT model for MAFC personality questionnaire, we study its performance through a simulation study and empirical examples.

The remainder of this chapter consists of four sections. The proposed model is derived and formulated in the next section. In Section 5.3, we report a simulation study to check parameter recovery. Section 5.4 presents two empirical examples of Big-Five questionnaire data. Finally, concluding remarks and possible directions for further research are discussed in Section 5.5.

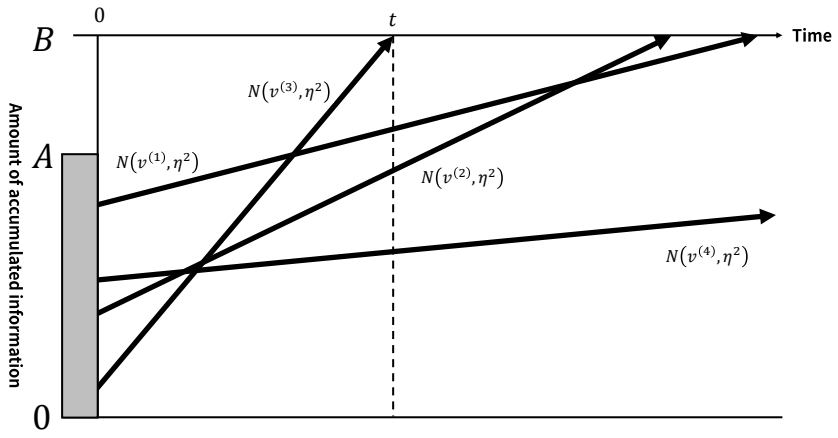

Figure 5.1: Schematic of the LBA model for four-alternative forced-choice task.

5.2 Multidimensional MAFC D-LBA IRT model

The proposed model is based on the LBA model, which was mathematically introduced in Section 2.3. Figure 5.1 shows the schematic of the LBA model for 4AFC format. The likelihood function and the prior distributions of the proposed model are described in Sections 5.2.1 and 5.2.2, respectively.

5.2.1 Proposed model likelihood

The key development in the proposed model is in the formulation of the drift rate parameter $v_{ij}^{(k)}$. In unidimensional binary D-LBA IRT, this can be simply modeled as a function of the relative position of the respondent to the item's position. However, as noted in Section 5.1, things become complicated when we consider the MAFC responses because each alternative corresponds to different psychological dimension to be measured. The drift rate of each accumulator, which corresponds to one of the alternatives, should reflect its own underlying dimensional property.

To tackle this issue and derive the appropriate form of $v_{ij}^{(k)}$ for the MAFC format, let us consider the relationship between accumulators in the LBA framework and the latent utilities. The TIRT model succeeded in modeling the forced-choice response through the latent utility of each alternative based on Thurstone's (1927) random utility theory. The proposed model

applies the same concept to derive the reparametrization of $v_{ij}^{(k)}$ in the LBA framework.

In the random utility theory, each compared alternative elicits its latent utility, and respondents are to choose the one with the largest utility. Thus, the observed response depends on the latent utility of each (k-th) alternative, $\delta_j^{(k)}$, which is given as the sum of the option-specific component and random effect (Böckenholt, 2006):

$$\delta_j^{(k)} \sim N\left(p_j^{(k)}, \Psi_j^{2(k)}\right), \tag{5.1}$$

where $p_j^{(k)}$ is the component unique to the k-th alternative and $\Psi_j^{2(k)}$ represents the variance of the random component. Later, Bock (1958) proposed the extended three-component model. In this model, respondent i's latent utility to k-th choice alternative in item j is represented by:

$$\delta_{ij}^{(k)} \sim N\left(\mu_j^{(k)} + p_{ij}^{(k)}, \Psi_{ij}^{2\,(k)}\right), \tag{5.2}$$

where $\mu_j^{(k)}$ is the component specific to that object and common to all respondents, and both $p_{ij}^{(k)}$ and $\Psi_{ij}^{2\,(k)}$ now differ among respondents. Several further extensions of these basic models have been proposed (A. Brown, 2016). Among them, we applied the linear factor analysis model (Brady, 1989; Harman, 1960) for its compatibility with personality measurement. In this approach, the latent utility is represented as the linear combination of factor loadings $\beta_{jd}^{(k)}$ and factor score θ_{id} as

$$\delta_{ij}^{(k)} \sim N\left(\mu_j^{(k)} + \sum_{d=1}^{D} \beta_{jd}^{(k)} \theta_{id}, \Psi_{ij}^{2\,(k)}\right), \tag{5.3}$$

where $d(= 1, \ldots, D)$ denote the d-th psychological dimension measured in the questionnaire.

Here, the slope of the accumulator in the LBA process ($\Delta_{ij}^{(k)}$) and the utility ($\delta_{ij}^{(k)}$) in Equation 5.3 have the common property that (a) both follow normal distributions and (b) the probability of choosing each alternative increases with the mean of the normal distribution. Therefore, it is reasonable to substitute $\delta_{ij}^{(k)}$ for $\Delta_{ij}^{(k)}$ in the LBA framework. Then, the mean parameter of Equation 5.3 corresponds to the drift rate parameter $v_{ij}^{(k)}$. To uniquely determine its formulation, we further assumed the following in this study:

- The MAFC questionnaire has the simple structure in its factor analytic part $\sum_{d=1}^{D} \beta_{jd}^{(k)} \theta_{id}$. In other words, each statement measures one of the D psychological dimensions, and thus it has only one non-zero factor loading $\beta_{jd}^{(k)}$ to that dimension,
- Variances of the random component $\Psi_{ij}^{2\ (k)} (k = 1, \cdots, K)$ are the same within each item and respondent. They are represented by η_{ij}^2.
- In order to suffice the model identification, the sum of $v_{ij}^{(k)} (k = 1, \cdots, K)$ is fixed as one for all i and j.

In order to restrict the sum of $v_{ij}^{(k)}$ to one, we apply the softmax function, which can be seen as the generalization of the logistic response function for $K > 2$ (A. Brown, 2016). The resulting formulation of the parameter $v_{ij}^{(k)}$ $(k = 1, \cdots, K)$ becomes

$$v_{ij}^{(k)} = \frac{\exp\left(\mu_j^{(k)} + \beta_{jd_j^{(k)}} \theta_{id_j^{(k)}}\right)}{\sum_{c=1}^{K} \exp\left(\mu_j^{(c)} + \beta_{jd_j^{(c)}} \theta_{id_j^{(c)}}\right)}, \tag{5.4}$$

where $d_j^{(k)}$ is the factor that is measured with k-th alternative in item j. As a result, the drift rate $v_{ij}^{(k)}$ is proportional to *response strength*, which is any mapping of the mean latent utility to a numerical ratio scale (Luce, 1977). This condition is analogous to Luce's (1959, 1977) model based on the choice axiom.

Equation 5.4 can be seen as the generalization of the unidimensional binary D-LBA IRT model formulation in which the drift rate is given as a logistic function of single trait parameter. Additionally, Equation 5.4 is equivalent to the response probability of the multidimensional nominal response model (MNRM; Revuelta, 2014; Revuelta & Ximénez, 2017). The same parameter constraints are thus needed in the proposed model. In the MNRM, at least one $\mu_j^{(k)}$ and D element of $\beta_{jd_j^{(k)}}$ need to be fixed to obtain model identification (Revuelta & Ximénez, 2017). Under the assumption of the simple structure, $(D-1)K$ elements of $\beta_{jd_j^{(k)}}$ are automatically fixed

to zero, which suffices its required constraint. Regarding $\mu_j^{(k)}$, the sum-to-zero constraint,

$$\sum_{k=1}^{K} \mu_j^{(k)} = 0, \tag{5.5}$$

is adapted. This type of constraint is preferable when there are no reference alternatives, which is the case in typical MAFC measurements.

For other parameters, the parameters B and η in the LBA model are decomposed into both respondent and item parameters in the same way as that in the unidimensional binary D-LBA IRT model, as follows:

$$B_{ij} = \frac{\gamma_i}{\xi_j} \text{ with } \gamma_i, \xi_j \in \mathbb{R}_{>0}, \qquad \eta_{ij} = \frac{\sigma_i}{\psi_j} \text{ with } \sigma_i, \psi_j \in \mathbb{R}_{>0}. \tag{5.6}$$

Following the unidimensional binary D-LBA IRT model, the upper bound of the start point (A_{ij}) is fixed to the half of B_{ij}.

The boundary parameter B_{ij} is represented as the quotient of γ_i and ξ_j. This indicates that the expected RT becomes longer when γ_i is large or ξ_j is small. Therefore, γ_i can be interpreted as the cautiousness or deliberateness of the respondent. ξ_j represents the item property that affects the response latency such as the length of the sentence and difficulty in reading. The between-trial standard deviation of the slope, η_{ij}, functions as a discrimination parameter of the IRT model. Furthermore, η_{ij} is also decomposed into σ_i and ψ_j in this model. ψ_j is interpreted in the same manner as item discrimination in the IRT model, and σ_i can be interpreted as the response consistency of the respondent. Section 4.3.2 depicts the relationship between these D-LBA IRT parameter values and observed quantities.

Note that the proposed model has two types of item discrimination-related parameters: $\beta_{jd_j^{(k)}}$ and ψ_j. While $\beta_{jd_j^{(k)}}$ is a *statement* parameter that is unique to each statement in an item, ψ_j is an *item* parameter that reflects the overall property of the statements in each item.

The resulting joint cumulative distribution function and probability density function for

choosing the k-th alternative at RT t is given as

$$F(k,t) = LBA_{CDF}\left(k,t \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{\exp\left(\mu_j^{(k)} + \beta_{jd_j^{(k)}}\theta_{id_j^{(k)}}\right)}{\sum_{c=1}^{K}\exp\left(\mu_j^{(c)} + \beta_{jd_j^{(c)}}\theta_{id_j^{(c)}}\right)}, \frac{\sigma_i}{\psi_j}\right) \tag{5.7}$$

and

$$f(k,t) = LBA_{PDF}\left(k,t \,\middle|\, \frac{\gamma_i}{\xi_j}, \frac{\gamma_i}{2\xi_j}, \frac{\exp\left(\mu_j^{(k)} + \beta_{jd_j^{(k)}}\theta_{id_j^{(k)}}\right)}{\sum_{c=1}^{K}\exp\left(\mu_j^{(c)} + \beta_{jd_j^{(c)}}\theta_{id_j^{(c)}}\right)}, \frac{\sigma_i}{\psi_j}\right), \tag{5.8}$$

respectively.

In the proposed multidimensional MAFC D-LBA IRT model, the non-decision time parameter (τ) of the LBA model is fixed to zero. The expected RT for responding to an MAFC questionnaire item is typically much longer than that for a cognitive psychological task. Consequently, the portion of the non-decision time in observed RT should be much smaller in the kind of item response in which we are interested in this study. This makes its estimation more difficult. In fact, Cassey, Gaut, Steyvers, and Brown (2016) reported that the LBA model tends to estimate the non-decision time as being smaller than its actual duration. We have also tried to estimate this parameter using real data in Section 5.4.1 and found that its estimates were similarly quite small (results not shown in this thesis).Additionally, considering the fact that this parameter is generally not of interest in personality measurement, we decided not to incorporate the non-decision time parameter.

5.2.2 Prior distribution

We set the following prior distributions for the parameters:

$$
\begin{aligned}
\xi_j &\sim Cauchy_{[0,\infty)}(0, \lambda_\xi), & \gamma_i &\sim LN(0,1), \\
\mu_j^{(k)} &\sim N(0, \lambda_\mu^2), & \beta_{jd_j^{(k)}} &\sim Cauchy_{[0,\infty)}(0, \lambda_\beta), \\
\boldsymbol{\theta}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), & \boldsymbol{\Sigma} &\sim LKJCorr(1), \\
\psi_j &\sim Cauchy_{[0,\infty)}(0, \lambda_\psi), & \sigma_i &\sim LN(0,1), \\
\lambda_\xi, \lambda_\mu, \lambda_\beta, \lambda_\psi, &\sim Cauchy_{[0,\infty)}(0, 5), &&
\end{aligned}
\tag{5.9}
$$

where $Cauchy_{[0,\infty)}(.)$ represents the half Cauchy distribution. As shown above, we set standard (log-)normal distributions to respondent parameters for identification purposes, which is a common setting in IRT literature. For the correlation matrix of the latent traits, we used the LKJ correlation distribution (Lewandowski et al., 2009) with shape parameter one. This distribution corresponds to a uniform prior over the space of $D \times D$ correlation matrices.

Our preliminary simulation study revealed that, when we set a specific value, such as five or ten, to the scale parameter of the prior distribution of item parameters, the resultant estimates were not close to the true values. To obtain more accurate parameter estimates, for item parameters, we adopted hierarchical Bayesian approach to achieve adequate flexibility as well as shrinkage. The prior distribution for hyperparameter $\boldsymbol{\lambda}$ is set to be the half-Cauchy distribution as a weakly informative prior, which is recommended by Gelman (2006).

Following A. Brown and Maydeu-Olivares (2012) and Bürkner, Schulte, and Holling (2019), the present study assumed that the keyed direction of each statement in the questionnaire is already known. This means that the researchers know whether each statement is in the regular or reversed direction in terms of the psychological dimension to be measured. This would be a natural assumption in most applications. Under this assumption, the factor loading $\beta_{jd_j^{(k)}}$ is estimated within non-negative constraint at first, then the signs of estimates on negatively keyed statements are manually reversed.

The proposed model is numerically estimated using the Markov chain Monte Carlo (MCMC)

method. All estimation procedures in this chapter were conducted with R 3.5.3 and `stan` 2.17.3 on Windows 10 PCs. On the proposed D-LBA IRT model extension, three MCMC chains with the length of 2,000 iterations (half of them were discarded as warmup period) were run for each dataset. The stan code of the proposed model can be found at Appendix E.

5.3 Simulation study

Simulation design In this section, we conducted a simulation study to check parameter recovery of the proposed model. We considered the following two scenarios:

- The first scenario considers the situation when all items contain statements that reflect all latent traits, that is, the number of statements in an item (K) is the same as the number of latent traits (D). Here, total number of items J and the number of latent traits D are manipulated as $J = (8, 16, 24)$ and $D = K = (2, 3, 4)$ to check the effect of these factors. In this scenario, the total number of statements is $J \times D$. For example, when $J = 8$ and $D = 4$, each item consists of four statements and the total number of statements is therefore $4 \times 8 = 32$. The number of statements that belong to each trait is also the same as J. The number of respondents is fixed to $I = 500$.

- The second scenario examines the effect of the number of statements K in an item. In this scenario, the total numbers of latent traits D and statements are fixed to five and 60, respectively. Thus, each latent trait consists of 12 statements. The following two valuables are manipulated: (a) the number of statements in an item $K = (2, 3, 4)$; (b) the number of respondents $I = (100, 300, 500)$. The number of items J varies according to K ($J = 60/K$).

We repeated data generation and parameter estimation 30 times for each of the above-mentioned 18 conditions.

Data generation First, true values are randomly drawn from the following distributions: $\xi_j \sim U(0.15, 0.4)$, $\gamma_i \sim LN(0, 1)$, $\mu_j^{(k)} \sim U(-1.5, 1.5)$, $\left|\beta_{jd_j^{(k)}}\right| \sim U(0.1, 2)$, $\boldsymbol{\theta}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, $\psi_j \sim$

$U(2.5,6)$, $\sigma_i \sim LN(0,1)$. These distributions are chosen on the basis of the results of our real data example, which is summarized in the next section, such that the distributions cover all obtained estimates except outliers. For each j, $\mu_j^{(k)}$s are centered so that the sum-to-zero constraint (Equation 5.5) is satisfied. Regarding the elements of $\boldsymbol{\Sigma}$, which are represented by $\rho_{dd'}$, the true values are respectively drawn from $U(-.5,5)$ with the restriction that the resulting factor correlation matrix $\boldsymbol{\Sigma}$ is a positive definite. For the second scenario, items are constructed so that the numbers of *implied paired-comparison* (this will be elaborated in the next paragraph) on all pairs of traits are the same as far as possible. In addition, for both scenarios, negatively keyed statements are arranged so that each pair of latent traits has both, same- and opposite-direction implied paired-comparison.

To facilitate the understanding on how items are constructed in the present simulation, let us consider a 4AFC item that consists of statements that measure traits A, B, D, and E, respectively. The former two statements are positively keyed, and the latter two are negatively keyed. This item corresponds to item 6 in Table 5.1. From this item, six implied paired-comparisons shown in the right column in Table 5.2 are derived. This binarization step is the same as the TIRT model. When total number of items is $J = 15$, the total number of implied paired-comparisons becomes $6 \times 15 = 90$. In contrast, the number of the patterns of trait pairs is ${}_5C_2 = 10$ when $D = 5$. Thus, each pattern should appear in nine out of 15 items. In addition, each trait pair consists of four (or five) same-direction comparisons and five (or four) opposite-direction comparisons. Table 5.1 shows an example set of 15 items, which corresponds to the second scenario. The item set is also generated at randomly in each repetition.

After true values are determined, the artificial item response and RT data are randomly generated based on the proposed model. Before parameter estimation, three observations that have an RT longer than 120 sec are omitted as irregular responses.

Results Tables 5.3 and 5.4 show the means of RMSE and bias values for each parameter obtained from each of the 18 conditions, respectively. As γ_i and σ_i are lognormally distributed, the RMSEs and biases of the log-transformed estimates for γ_i and σ_i are presented. Regarding

Table 5.1: Example set of 15 items for 4AFC measurement

	Statement (1)		Statement (2)		Statement (3)		Statement (4)	
Item	Trait	Direction	Trait	Direction	Trait	Direction	Trait	Direction
1	A	Positive	B	Negative	C	Positive	D	Positive
2	A	Positive	B	Positive	C	Positive	E	Positive
3	A	Negative	B	Positive	D	Positive	E	Negative
4	A	Negative	C	Negative	D	Positive	E	Positive
5	B	Positive	C	Positive	D	Positive	E	Positive
6	A	Positive	B	Positive	D	Negative	E	Negative
7	A	Positive	B	Positive	C	Negative	E	Positive
8	A	Positive	B	Positive	C	Negative	D	Positive
9	A	Positive	C	Negative	D	Negative	E	Positive
10	B	Positive	C	Negative	D	Positive	E	Negative
11	A	Positive	B	Positive	C	Positive	D	Positive
12	A	Negative	B	Positive	C	Positive	E	Positive
13	A	Positive	B	Negative	D	Positive	E	Negative
14	A	Positive	C	Negative	D	Positive	E	Positive
15	B	Negative	C	Negative	D	Positive	E	Positive

Table 5.2: Six possible implied paired-comparisons from the item 6 in Table 5.1 that is constructed by the following four statements: (1) trait A, positive; (2) trait B, positive; (3) trait D, negative; (4) trait E, negative.

Statement (a)		Statement (b)		Paired-Comparison	
Trait	Direction	Trait	Direction	Traits	Direction
A	Positive	B	Positive	A–B	Same
A	Positive	D	Negative	A–D	Opposite
A	Positive	E	Negative	A–E	Opposite
B	Positive	D	Negative	B–D	Opposite
B	Positive	E	Negative	B–E	Opposite
D	Negative	E	Negative	D–E	Same

β, for negatively keyed statements, sign inversion of both $\beta_{jd_j^{(k)}}$ and $\hat{\beta}_{jd_j^{(k)}}$ are applied before calculating the RMSE and bias.

In the first scenario, the RMSEs for all parameters are acceptably small in all conditions, even when $D = K = 2$. Note that under this condition, the TIRT and TDIRT models encounter the problem of rotation indeterminacy (A. Brown & Maydeu-Olivares, 2011). This is because the loadings of these models become equivalent to the case of exploratory factor analysis model. To see why the parameters of the proposed model can be estimated under this condition, let us take a look at the factor loading matrix $\boldsymbol{\beta}$ in the TIRT model. When $D = K = 2$, the matrix is given as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1d_1^{(1)}} & \beta_{1d_1^{(2)}} \\ \beta_{2d_2^{(1)}} & \beta_{2d_2^{(2)}} \\ \vdots & \vdots \\ \beta_{Kd_K^{(1)}} & \beta_{Kd_K^{(2)}} \end{bmatrix}. \tag{5.10}$$

This matrix suffers from the rotational indeterminacy, because the matrix does not contain zeros in its elements. To resolve the rotational indeterminacy, A. Brown and Maydeu-Olivares (2011, 2012) arbitrarily recommends fixing the values of the first item $\left(\beta_{1d_1^{(1)}} \text{ and } \beta_{1d_1^{(2)}}\right)$ to their true values. Though the approach may be used in the simulation study, it is not applicable in real data analysis. This is because researchers do not know the true parameter values, and the resultant estimates depend on the item for which the parameters are fixed (Fontanella, Fontanella, Valentini, & Trendafilov, 2019; Lopes & West, 2004).

On the other hand, the drift rate of the proposed D-LBA IRT approach is mathematically equivalent to the choice probability in the MNRM model. Therefore, the factor loading matrix

Table 5.3: Mean RMSEs of each parameter for the simulated conditions.

I	D	J	K	ξ	$\log(\gamma)$	μ	β	θ	ψ	$\log(\sigma)$	ρ
500	2	8	2	0.016	0.231	0.144	0.393	0.609	0.394	0.575	0.171
		16		0.012	0.185	0.095	0.276	0.481	0.353	0.410	0.125
		24		0.012	0.160	0.089	0.236	0.422	0.360	0.351	0.095
	3	8	3	0.016	0.260	0.176	0.289	0.613	0.424	0.552	0.072
		16		0.017	0.205	0.141	0.217	0.490	0.385	0.402	0.041
		24		0.014	0.183	0.123	0.176	0.436	0.378	0.338	0.035
	4	8	4	0.019	0.273	0.229	0.287	0.631	0.468	0.554	0.064
		16		0.016	0.215	0.189	0.211	0.521	0.409	0.393	0.044
		24		0.018	0.195	0.175	0.201	0.470	0.339	0.334	0.035
100	5	15	4	0.036	0.245	0.323	0.817	0.576	0.956	0.451	0.104
		20	3	0.033	0.216	0.266	0.853	0.550	0.910	0.409	0.090
		30	2	0.035	0.189	0.186	0.869	0.506	0.841	0.345	0.080
300		15	4	0.023	0.237	0.237	0.317	0.563	0.524	0.429	0.062
		20	3	0.021	0.190	0.163	0.263	0.516	0.506	0.393	0.054
		30	2	0.019	0.158	0.124	0.286	0.482	0.470	0.333	0.044
500		15	4	0.020	0.237	0.187	0.228	0.564	0.372	0.412	0.046
		20	3	0.018	0.199	0.138	0.207	0.519	0.377	0.382	0.040
		30	2	0.016	0.162	0.093	0.225	0.485	0.328	0.328	0.034

Table 5.4: Mean biases of each parameter for the simulated conditions.

I	D	J	K	ξ	$\log(\gamma)$	μ	β	θ	ψ	$\log(\sigma)$	ρ
500	2	8	2	-0.006	0.008	–	0.053	0.001	0.094	0.159	-0.010
		16		0.001	0.012	–	0.008	-0.009	0.101	0.081	-0.037
		24		0.002	0.015	–	0.004	0.004	0.097	0.060	-0.026
	3	8	3	0.003	0.042	–	0.029	0.006	0.071	0.137	0.004
		16		0.004	0.024	–	0.003	0.001	0.119	0.080	0.000
		24		0.003	0.016	–	-0.004	0.003	0.102	0.053	-0.005
	4	8	4	0.001	0.041	–	0.017	0.005	0.195	0.164	-0.001
		16		0.001	0.024	–	0.006	-0.003	0.087	0.078	0.003
		24		0.002	0.019	–	0.008	0.006	0.059	0.053	0.003
100	5	15	4	0.005	0.040	–	0.131	-0.007	0.237	0.108	-0.005
		20	3	0.005	0.025	–	0.108	0.001	0.214	0.073	-0.003
		30	2	0.006	0.024	–	0.144	-0.001	0.190	0.046	-0.009
300		15	4	0.006	0.032	–	0.035	0.001	0.153	0.093	0.001
		20	3	0.005	0.029	–	0.019	-0.007	0.214	0.101	-0.002
		30	2	0.004	0.017	–	0.042	0.000	0.132	0.060	0.004
500		15	4	0.005	0.034	–	0.012	-0.005	0.004	0.065	-0.005
		20	3	0.008	0.037	–	0.016	0.005	0.095	0.067	0.000
		30	2	0.002	0.009	–	0.018	-0.003	0.053	0.044	-0.002

Note: biases for μ cannot be calculated due to the sum-to-zero constraint (Equiation. 5.5).

$\boldsymbol{\beta}$ can also be written in the same form as the MNRM model:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1d_1^{(1)}} & 0 \\ 0 & \beta_{1d_1^{(2)}} \\ \hdashline \beta_{2d_2^{(1)}} & 0 \\ 0 & \beta_{2d_2^{(2)}} \\ \hdashline \vdots & \vdots \\ \hdashline \beta_{Kd_K^{(1)}} & 0 \\ 0 & \beta_{Kd_K^{(2)}} \end{bmatrix}. \tag{5.11}$$

As noted in section 5.2.1, the MNRM requires at least D elements in the matrix $\boldsymbol{\beta}$ to be fixed such that the model is sufficiently identified. Thus, the proposed model extension is always identified even when the number of latent traits D and statements in item K is the same, as long as the simple structure is assumed. This is the reason why the proposed D-LBA IRT approach can recover parameters properly even when $D = K = 2$ while the TIRT and TDIRT models cannot unless some extra constraints are imposed. Still, the simulation results shown in Table 5.3 indicate that when $D = 2$, the RMSEs for β and ρ are larger than when $D > 2$. On the other hand, other parameters tend to have larger RMSEs as D increases. Either way, the RMSEs for all parameters but ξ_j become smaller as J increases. Concerning ξ_j, this can be estimated sufficiently with a small number of items. These would be preferable results.

From the results of the second scenario, it is evident that all parameters but θ_{id} obtain smaller RMSEs as the number of respondents I increases. The reason why the precision of θ_{id} does not improve is attributed to the "Neyman–Scott paradox" (Neyman & Scott, 1948). This indicates that in general IRT models, the estimates of respondent parameter do not satisfy the consistency when respondent parameters and item parameters are simultaneously estimated.

The results of the second scenario also show that when the number of statements in an item (K) increases, the RMSEs become larger. Evidently, this result is because we can obtain just the same number of response data as the number of items (J). However, the 4AFC format is expected to shorten the total RT than the two-alternative forced-choice (2AFC) format (Vancleef

et al., 2018). This result therefore suggests that the 4AFC format would be useful when the researcher wants to obtain estimates as accurately as possible, whereas the 4AFC format is good for reducing total time at the expense of a little accuracy.

The biases for all parameters are acceptably small when the number of respondents (I) and the number of items (J) are not too small. However, both ψ_j and σ_i tend to be slightly overestimated. This is inconsistent with the simulation result of study 2 that showed small negative biases for corresponding parameters. Nevertheless, as long as biases for these parameters are in the same direction, these biases would not affect other parameters because they are given in the quotient form. In this study we chose standard log-normal distribution as the prior of σ_i. While there might be a different class of prior that further reduces the biases of ψ_j and σ_i, the current prior choice would be practical and acceptable in terms of its simulation performance.

Again, the results demonstrated that the biases for all parameters are small when there exist enough J and I. This evidences that the proposed model can recover parameters properly with adequate numbers of items and respondents.

5.4 Real Data Application

In this section, we report two applications of the proposed model to actual 4AFC and 2AFC questionnaire data with RT information.

5.4.1 Example 1: Application to 4AFC data

In the first application, we apply the proposed method to the 4AFC ($K = 4$) personality questionnaire data. The stan code of the proposed D-LBA IRT, TIRT, and MNRM models used in this application example can be found in Appendix E, A, and F, respectively.

Item construction Eighteen 4AFC questionnaire items were constructed using the statements from the Mini-IPIP scale (Donnellan, Oswald, Baird, & Lucas, 2006) and the Japanese version of the Balanced Inventory of Desired Responding (BIDR-J; Tani, 2008). The Mini-IPIP scale

was developed to measure Big-Five personality traits. It consists of four statements for each latent trait. The Mini-IPIP scale therefore contains 20 statements. Though the Japanese version of the Mini-IPIP scale has not been developed, this scale is a part of the Big-Five factor questionnaire (Goldberg, 1992, 1999), of which the Japanese translation has been developed by Apple and Neff (2012). Thus, we used its corresponding translations. The BIDR-J, which is designed to measure social desirability, originally consists of 24 statements. For this study, four of 24 statements were selected based on the reported factor loadings, considering that both positively and negatively keyed statements are included.

Using the $20+4=24$ statements in total, 18 4AFC items were carefully constructed so that all patterns of trait pairs appear almost the same number of times and both same- and opposite-direction comparisons are included in each pattern, as noted in Section 5.3. Note that each statement appears thrice in the item set for the purpose of the studies other than the current one. The resulting item set used in this example is summarized in Tables 5.5–5.6. Table 5.5 summarizes the block of statements in all items, and Table 5.6 gives the 24 statements used in this example. For instance, item 1 consists of the following four statements: "*Am the life of the party* (Ext-P1)," "*Am not interested in other people's problems* (Agr-N1)," "*Get upset easily* (Emo-N1)," and "*Have excellent ideas* (Int-P1)."

Data Four hundred and eighty-four Japanese participants were recruited via an online crowdsourcing platform. The sample consists of 184 male and 296 female participants. Four of the 484 participants did not report a specific gender. The distribution of participants' ages is as following (three did not answer): 20's=94; 30's=168; 40's=150; 50's=59; 60's=8; and 70's=2. The total number of item responses was $8,674$.

During the survey, an item that consists of four statements appeared on the screen at a time. Participants were instructed to choose the most appropriate one that described themselves by pressing the corresponding number on the keyboard. Moreover, participants were asked to press the button as soon as they made the decision. The time limit to answer each item was 30 seconds. When 30 seconds passed before giving an answer, the item response was

recorded as unanswered. The orders of items and statements in each item were randomized across participants.

Procedure Before analysis, one out of 8,674 item response recorded shorter RT than 300 milliseconds. The data was therefore listwise deleted as an ineligible response. Note that the eligibility check was conducted at the item-level. This means that the respondent answered faster than 300 ms in only one of 18 items, the item response to that item was removed, while the remaining 17 item responses were used in the following analysis procedure. The dataset was analyzed using the three models: the proposed D-LBA IRT model, the TIRT model, and the MNRM model.

In this study, the priors of the TIRT model were set as

$$\begin{aligned} \mu_j^{(k)} &\sim N(0, \lambda_\mu^2), & \beta_{jd_j^{(k)}} &\sim Cauchy_{[0,\infty)}(0, \lambda_\beta), \\ \boldsymbol{\theta}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), & \boldsymbol{\Sigma} &\sim LKJCorr(1), \\ \Psi_j^{(k)} &\sim Cauchy_{[0,\infty)}(0, \lambda_\Psi), & \lambda_\mu, \lambda_\beta, \lambda_\Psi &\sim Cauchy_{[0,\infty)}(0, 5). \end{aligned} \tag{5.12}$$

Note that the TIRT model requires some parameter constraints for model identification. We therefore set the following constraints: $\sum_{k=1}^{K} \mu_j^{(k)} = 0 \ \forall \ j$, $\Psi_j^{(K)} = 1 \ \forall \ j$.

For the MNRM model, the probability that respondent i chooses statement k from an MAFC item j is

$$P(x_{ij} = k) = \frac{\exp\left(\mu_j^{(k)} + \beta_{jd_j^{(k)}} \theta_{id_j^{(k)}}\right)}{\sum_{c=1}^{K} \exp\left(\mu_j^{(c)} + \beta_{jd_j^{(c)}} \theta_{id_j^{(c)}}\right)}, \tag{5.13}$$

which is the same as the drift rate parameter ($v_{ij}^{(k)}$). The priors for the MNRM were

$$\begin{aligned} \mu_j^{(k)} &\sim N(0, \lambda_\mu^2), & \beta_{jd_j^{(k)}} &\sim Cauchy_{[0,\infty)}(0, \lambda_\beta), \\ \boldsymbol{\theta}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), & \boldsymbol{\Sigma} &\sim LKJCorr(1), \\ \lambda_\mu, \lambda_\beta, &\sim Cauchy_{[0,\infty)}(0, 5). \end{aligned} \tag{5.14}$$

Though the TIRT model is based on Thurstone's random utility approach, the MNRM model is based on Luce's choice axiom. This can be regarded as a special case of Thurstone's random utility approach when the variances of the unique factor ($\Psi_j^{2(k)}$) are independent, normally distributed, and equivalent among all choice alternatives (A. Brown, 2016; Luce, 1977).

Note that we applied the hierarchical Bayesian approach to item parameters of the TIRT and MNRM models because otherwise the parameter estimates are affected by the prior distribution selection.

As mentioned in "*Data*" paragraph, each statement appears thrice in the 4AFC questionnaire used in this study. Nevertheless, we assumed that the statement has different parameters when it appeared in different items. For instance, $\mu_1^{(1)}$, $\mu_2^{(1)}$, and $\mu_5^{(1)}$ are treated as different parameters, although statement (1) in items 1, 2, and 5 are the same (Ext-P1: *Am the life of the party*).

Results To begin with, we checked the histogram of the RT of the obtained data (Figure 5.2). The mean RT was 5.431 seconds (min=0.528 sec, max=28.823 sec). Obviously, this distribution is similar to the frequently used distributions for the RT, such as log-normal, Gamma, or ex-Gaussian (De Boeck & Jeon, 2019). We considered this as an evidence that the RT was collected properly and proceeded to the following analysis.

Table 5.7 summarizes the estimated item parameters by the proposed D-LBA IRT model for the MAFC personality measurement data.

The parameters ξ_j and γ_i are the item and respondent elements of the boundary, respectively. When the boundary is expected to be larger, the RT is expected to be longer. In other words, the expected RT becomes longer when ξ_j is small or γ_j is large. To check if these parameter estimates are related to the RT, we checked the correlation with the mean RT. Figure 5.3 shows two scatter plots between the mean RTs and the parameter estimates. The left panel shows the scatter plot between *itemwise* mean RT and ξ_j. As we anticipated, there existed a negative correlation ($r = -.740$) between them. Regarding item 17, mean RT is relatively shorter for its value of ξ_j. This is possibly because the estimates of β_{17} are extremely larger than that for other items. We are uncertain about why this happened, thus these parameters should be interpreted

Table 5.5: Structure of 4AFC items used in Section 5.4.1. Names of statements correspond to the "ID" column in Table 5.6.

Item	Statement (1)	Statement (2)	Statement (3)	Statement (4)
1	Ext-P1	Agr-N1	Emo-N1	Int-P1
2	Ext-P1	Agr-P1	Emo-N2	Soc-P1
3	Ext-N1	Agr-N2	Emo-N2	Int-P1
4	Agr-N1	Con-N1	Emo-P1	Int-N1
5	Ext-P1	Agr-N2	Con-P1	Int-N2
6	Ext-P2	Agr-P2	Int-N2	Soc-P2
7	Ext-P2	Emo-P1	Int-P2	Soc-P1
8	Ext-P2	Con-N1	Emo-N2	Soc-N1
9	Ext-N1	Con-N2	Int-P2	Soc-N2
10	Agr-N1	Emo-P2	Int-P2	Soc-P2
11	Agr-P2	Con-N2	Int-P1	Soc-N1
12	Agr-P1	Con-P1	Emo-P1	Soc-N2
13	Ext-N1	Agr-P2	Con-P2	Emo-N1
14	Con-P1	Emo-N1	Int-N1	Soc-P1
15	Ext-N2	Con-N2	Emo-P2	Int-N1
16	Ext-N2	Con-N1	Int-N2	Soc-N2
17	Ext-N2	Agr-N2	Con-P2	Soc-P2
18	Agr-P1	Con-P2	Emo-P2	Soc-N1

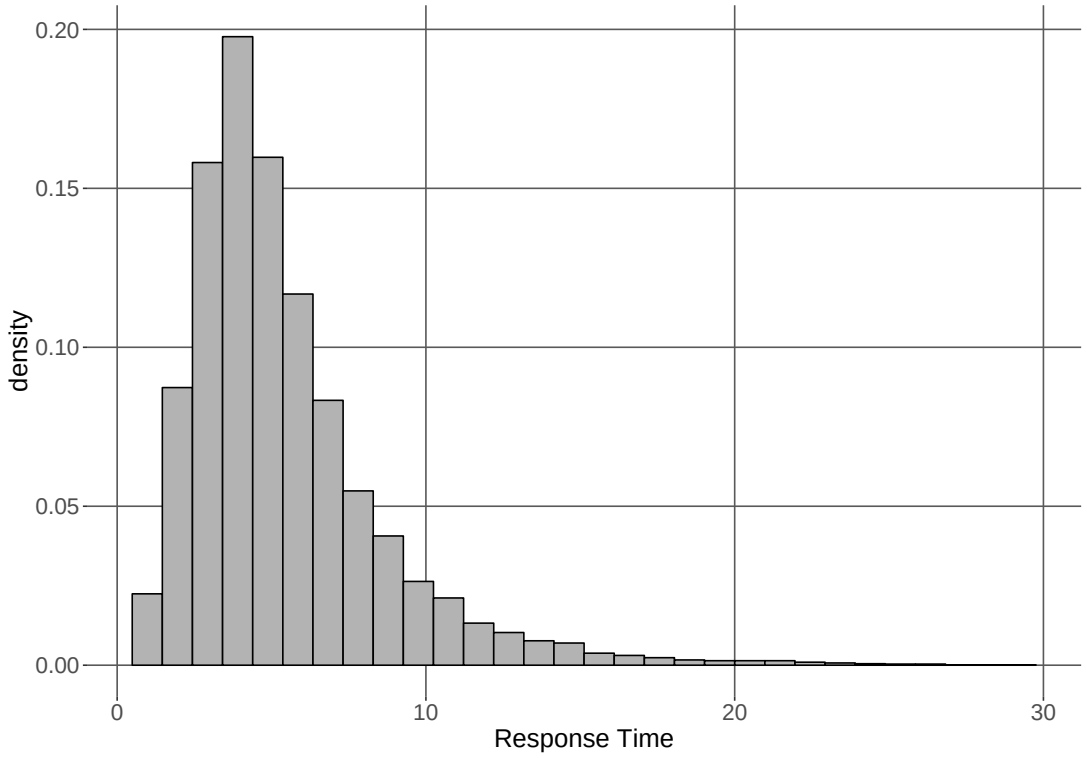

Figure 5.2: Histogram of the RT

Table 5.6: List of statements used in Section 5.4.1.

Trait	Direction	ID	Statement
Extraversion	positive	Ext-P1	Am the life of the party.
		Ext-P2	Talk to a lot of different people at parties.
	negative	Ext-N1	Don't talk a lot.
		Ext-N2	Keep in the background.
Agreeableness	positive	Agr-P1	Feel others' emotions.
		Agr-P2	Sympathize with others' feelings.
	negative	Agr-N1	Am not interested in other people's problems.
		Agr-N2	Am not really interested in others.
Conscientiousness	positive	Con-P1	Get chores done right away.
		Con-P2	Like order.
	negative	Con-N1	Often forget to put things back in their proper place.
		Con-N2	Make a mess of things.
Emotional Stability	positive	Emo-P1	Am relaxed most of the time.
		Emo-P2	Seldom feel blue.
	negative	Emo-N1	Get upset easily.
		Emo-N2	Have frequent mood swings.
Intellect-Imagination	positive	Int-P1	Have excellent ideas.
		Int-P2	Have a vivid imagination.
	negative	Int-N1	Am not interested in abstract ideas.
		Int-N2	Do not have a good imagination.
Social Desirability	positive	Soc-P1	Never swear.
		Soc-P2	Never regret my decisions.
	negative	Soc-N1	Sometimes lose out on things because I can't make up my mind soon enough.
		Soc-N2	Sometimes tell lies if I have to.

Note: "ID" column corresponds to Table (a). English versions of the statements for Social Desirability are retrieved from Paulhus (1988). Japanese versions of the statements are available at `https://ipip.ori.org/JapaneseBig-FiveFactorMarkers.htm`. and Tani (2008)

Table 5.7: Item Parameter Estimates obtained by the proposed D-LBA IRT model.

Item	MRT	ξ_j	ψ_j	$\mu_j^{(1)}$	$\mu_j^{(2)}$	$\mu_j^{(3)}$	$\mu_j^{(4)}$	$\beta_j^{(1)}$	$\beta_j^{(2)}$	$\beta_j^{(3)}$	$\beta_j^{(4)}$
1	5.155	0.261	4.185	0.386	-0.050	-0.621	0.285	1.075	0.811	0.714	1.613
2	5.182	0.284	4.641	0.920	-0.558	-0.187	-0.176	1.549	0.753	0.729	0.126
3	5.307	0.247	4.006	-0.647	0.566	-0.248	0.329	0.757	1.256	1.356	1.585
4	6.134	0.226	3.737	0.384	-0.004	-0.429	0.049	1.365	0.858	0.804	0.088
5	5.430	0.247	3.816	0.371	0.185	-0.809	0.252	1.277	0.949	0.821	0.782
6	5.450	0.260	4.191	1.218	-0.817	0.134	-0.534	1.692	0.630	0.580	0.690
7	5.295	0.268	4.054	0.706	-0.149	-0.557	0.001	1.411	0.830	0.996	0.436
8	5.985	0.233	3.963	0.586	0.130	-0.078	-0.638	1.316	0.818	1.069	0.537
9	4.963	0.322	5.239	-0.197	0.600	-0.335	-0.068	0.487	0.102	0.747	0.273
10	5.732	0.235	4.171	0.292	0.835	-0.661	-0.466	1.054	0.912	1.120	0.764
11	5.444	0.259	4.213	-0.619	0.455	0.589	-0.426	0.915	0.233	1.747	0.478
12	5.275	0.302	4.664	-0.019	-0.199	0.171	0.047	0.830	0.582	0.763	0.295
13	4.655	0.352	5.358	-0.177	-0.177	0.359	-0.005	0.335	0.530	1.043	0.640
14	5.296	0.272	3.677	-0.432	-0.291	0.369	0.354	0.687	0.693	0.086	0.442
15	5.539	0.255	4.438	-0.445	0.086	0.399	-0.040	1.442	0.195	0.913	0.119
16	5.646	0.236	3.688	-0.636	0.362	0.543	-0.269	1.192	1.341	0.463	0.166
17	5.424	0.189	3.134	0.042	0.784	0.294	-1.120	8.291	2.272	4.812	1.955
18	5.855	0.258	5.532	-0.270	0.182	0.353	-0.265	0.903	1.210	0.488	0.548
		λ_ξ	λ_ψ				λ_μ				λ_β
		0.282	4.350				0.520				0.768

Note: MRT = mean RT for each item.

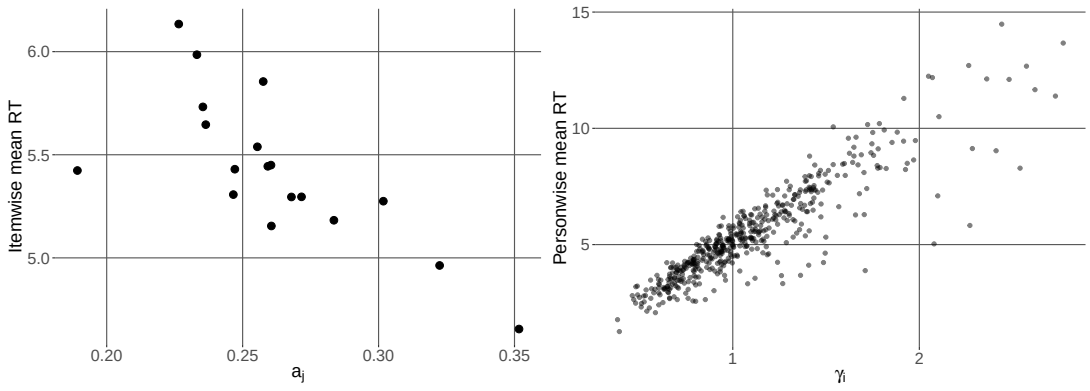

Figure 5.3: Left panel: scatter plot between the itemwise MRT and ξ_j; Right panel: scatter plot between the personwise MRT and γ_i.

carefully. The right panel is the scatter plot between the *personwise* mean RT and γ_i. There also exist strong correlations between them (r = .901).

Now that we ensured that the value of ξ_j has a strong relationship with the itemwise mean RT, the parameter ξ_j is considered to reflect some factors of the statements that affect the mean RT, such as the sentence length and the readability. To check the correlation between the readability and the estimates of ξ_j, we calculated the readability score of each sentence by jReadability (Hasebe & Lee, 2015). The readability score is obtained by the length of the sentence and the percentage of some components (e.g., verbs, particles). The correlation between the mean of readability score per item and ξ_j was .392. This result suggests that the more difficult the sentences in the item, the longer is the expected RT.

As shown in Equation 5.3, the parameter $\mu_j^{(k)}$ is the mean of the latent utility of each statement. Therefore, it is expected that the choice proportion of the statement increases as $\mu_j^{(k)}$ increases. Table 5.4 shows the scatter plot between the estimates of $\mu_j^{(k)}$ and the choice proportion of each statement. Note that both values are centered per item. The correlation was .835, implying the strong relationship between them, as expected.

Next, we compared the estimated latent trait scores $\boldsymbol{\theta}_i$ among the D-LBA IRT, TIRT, and MNRM models. For item parameter estimates, The values obtained by the TIRT and MNRM models are therefore shown in the Appendix Table 5.8 although we did not go into the details in this study. The correlation matrix of the estimated latent trait scores $\boldsymbol{\theta}_i$ is summarized in

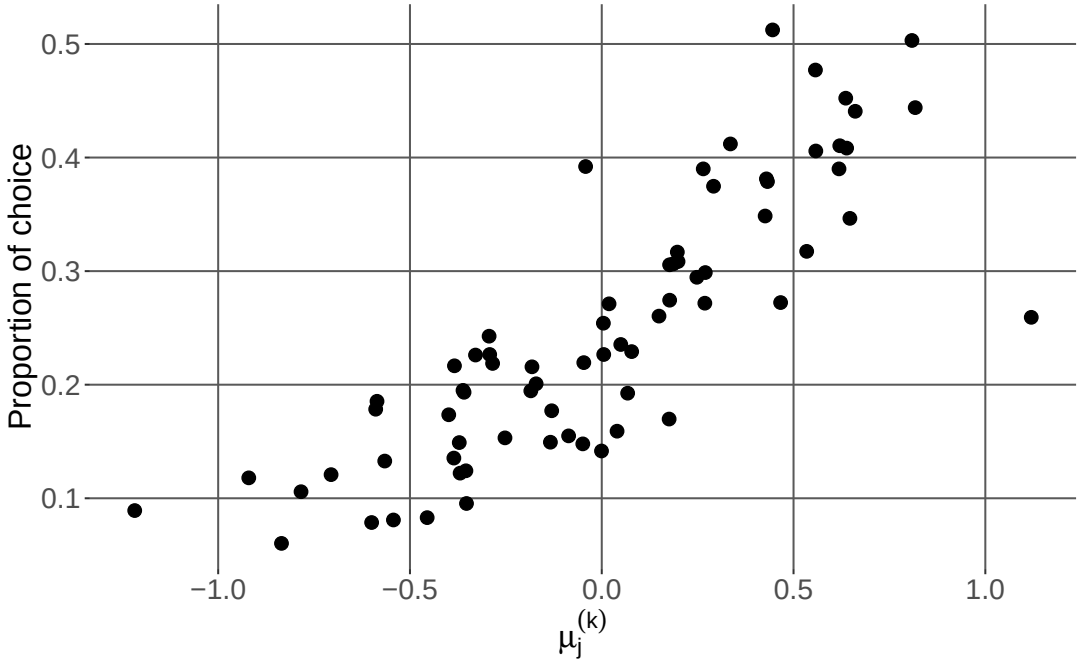

Figure 5.4: Scatter plot between the choice proportion of each sentence in the item and $\mu_j^{(k)}$.

Table 5.9. For simplicity, this table, along with Table 5.11 in example 2, only shows the correlations among trait scores within each model, and the correlations between the same trait scores obtained by different models.

Table 5.8: Item parameter estimates obtained by the TIRT and MNRM models.

	TIRT											MNRM							
Item	$\mu_j^{(1)}$	$\mu_j^{(2)}$	$\mu_j^{(3)}$	$\mu_j^{(4)}$	$\beta_j^{(1)}$	$\beta_j^{(2)}$	$\beta_j^{(3)}$	$\beta_j^{(4)}$	$\psi_j^{(1)}$	$\psi_j^{(2)}$	$\psi_j^{(3)}$	$\mu_j^{(1)}$	$\mu_j^{(2)}$	$\mu_j^{(3)}$	$\mu_j^{(4)}$	$\beta_j^{(1)}$	$\beta_j^{(2)}$	$\beta_j^{(3)}$	$\beta_j^{(4)}$
1	0.799	-0.165	-1.332	0.698	2.463	1.934	1.390	3.356	0.706	0.971	1.479	0.634	-0.059	-0.968	0.393	1.878	1.656	1.605	2.438
2	1.602	-1.089	-0.616	0.104	2.973	1.178	0.905	1.425	1.149	0.689	0.778	1.160	-0.872	-0.225	-0.062	2.132	1.487	1.652	0.169
3	-0.901	0.578	-0.670	0.993	1.619	1.930	2.002	3.307	0.975	0.953	1.102	-0.711	0.546	-0.414	0.579	1.865	1.582	2.366	2.518
4	0.469	-0.028	-0.545	0.104	1.920	1.268	1.163	0.064	1.214	0.965	0.885	0.268	0.071	-0.594	0.254	1.639	1.431	1.119	0.115
5	0.860	0.167	-1.287	0.260	2.122	1.431	1.091	1.178	0.866	1.031	1.286	0.685	0.126	-1.171	0.360	1.805	1.225	1.297	1.088
6	1.979	-1.262	0.137	-0.855	2.836	0.719	0.899	0.924	1.156	1.283	0.506	1.392	-1.035	0.414	-0.771	2.205	1.075	0.847	0.964
7	1.586	-0.777	-1.592	0.783	3.857	1.807	2.506	2.582	0.394	1.644	1.483	0.972	-0.260	-0.961	0.249	2.171	0.926	1.744	0.522
8	1.127	0.114	-0.291	-0.949	2.593	1.174	0.970	0.554	0.671	1.363	0.923	0.876	0.256	-0.152	-0.980	2.093	1.652	1.781	1.391
9	-0.158	0.607	-0.427	-0.021	1.156	0.089	2.599	0.530	0.797	0.787	1.486	-0.149	0.744	-0.518	-0.077	1.658	0.233	2.317	0.720
10	0.242	1.207	-0.811	-0.639	1.631	1.606	1.202	0.746	1.103	0.884	0.607	0.220	1.292	-0.876	-0.636	1.524	1.408	1.870	0.900
11	-0.812	0.595	0.809	-0.591	1.228	0.115	2.377	0.523	1.132	0.995	0.482	-0.818	0.769	0.494	-0.445	1.443	0.381	2.016	1.405
12	-0.056	-0.460	0.297	0.219	1.876	1.446	1.643	0.815	0.901	1.643	2.154	0.062	-0.326	0.136	0.129	1.505	0.956	0.777	0.532
13	-0.296	-0.600	0.997	-0.102	0.971	0.983	2.624	1.170	0.730	1.295	0.393	-0.255	-0.456	0.575	0.136	1.095	1.147	1.964	1.514
14	-2.314	-2.183	-1.467	5.964	0.926	0.851	0.091	8.452	1.041	0.985	0.762	-0.569	-0.430	0.418	0.581	1.086	1.191	0.147	0.633
15	-1.507	-0.381	2.323	-0.434	1.434	0.123	4.421	0.172	1.478	0.473	0.312	-0.965	0.129	0.792	0.044	1.912	0.273	1.858	0.273
16	-0.774	0.287	0.812	-0.325	1.045	1.249	0.859	0.089	1.158	0.926	0.811	-0.808	0.278	0.928	-0.398	1.369	1.370	0.943	0.172
17	-0.560	0.606	0.287	-0.333	1.143	1.115	1.804	0.818	0.987	0.316	0.304	-0.541	0.740	0.183	-0.382	1.785	1.191	2.041	0.762
18	-0.552	0.482	0.831	-0.761	1.378	2.400	1.244	0.845	0.957	0.530	1.099	-0.502	0.285	0.791	-0.574	1.581	2.099	0.861	1.988
				λ_μ				λ_β			λ_ψ				λ_μ				λ_β
				1.046				1.266			0.885				0.699				1.257

Note: $\psi_j^{(4)}$ are all fixed to one because of a parameter constraint.

Among the three models, the correlations between the TIRT and MNRM models were the highest. Comparatively, the correlations between the proposed D-LBA IRT model and each of the TIRT and MNRM models were slightly lower. This result may indicate that the D-LBA IRT model estimates slightly different aspect of the latent traits, because only the model uses the RT information in estimation. Nevertheless, the D-LBA IRT model estimates correlate higher with the MNRM model ones than with the TIRT model ones. This would be partly because both the proposed model and the MNRM model adopt the softmax function in the model formulation (Equations 5.4 and 5.13). In a more detailed review, the proposed D-LBA IRT model tended to obtain higher between-trait correlations than the other two models. Especially, correlations with social desirability (θ_{i6}) were much higher than the TIRT model. The same tendency can be found in the MNRM model. This is possibly related to the result that correlations of θ_{i6} between the TIRT model and each of the D-LBA IRT and MNRM models are smaller than those of other latent traits (.739 and .737, respectively).

Table 5.9: Correlation matrix of the estimated latent trait scores $\boldsymbol{\theta}_i$ among the D-LBA IRT, TIRT, and MNRM models.

	D-LBA IRT						TIRT						MNRM					
	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i6}	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i6}	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i6}
θ_{i1}	-																	
θ_{i2}	0.316	-																
θ_{i3}	0.324	0.211	-															
θ_{i4}	0.415	0.160	0.293	-														
θ_{i5}	0.588	0.609	0.216	0.213	-													
θ_{i6}	0.613	0.181	0.406	0.772	0.425	-												
θ_{i1}	0.882						-											
θ_{i2}		0.905					0.324	-										
θ_{i3}			0.927				0.143	0.281	-									
θ_{i4}				0.922			0.330	0.141	0.252	-								
θ_{i5}					0.940		0.439	0.507	0.181	0.169	-							
θ_{i6}						0.739	0.152	0.018	0.122	0.583	0.072	-						
θ_{i1}	0.939						0.973						-					
θ_{i2}		0.943						0.976					0.249	-				
θ_{i3}			0.955						0.975				0.163	0.247	-			
θ_{i4}				0.953						0.965			0.339	0.081	0.337	-		
θ_{i5}					0.950						0.978		0.477	0.605	0.174	0.065	-	
θ_{i6}						0.924						0.737	0.626	0.177	0.357	0.787	0.290	-

Note: θ_{i1} = emotional stability; θ_{i2} = extraversion; θ_{i3} = agreeableness; θ_{i4} = conscientiousness; θ_{i5} = intellect/imagination; θ_{i6} = social desirability

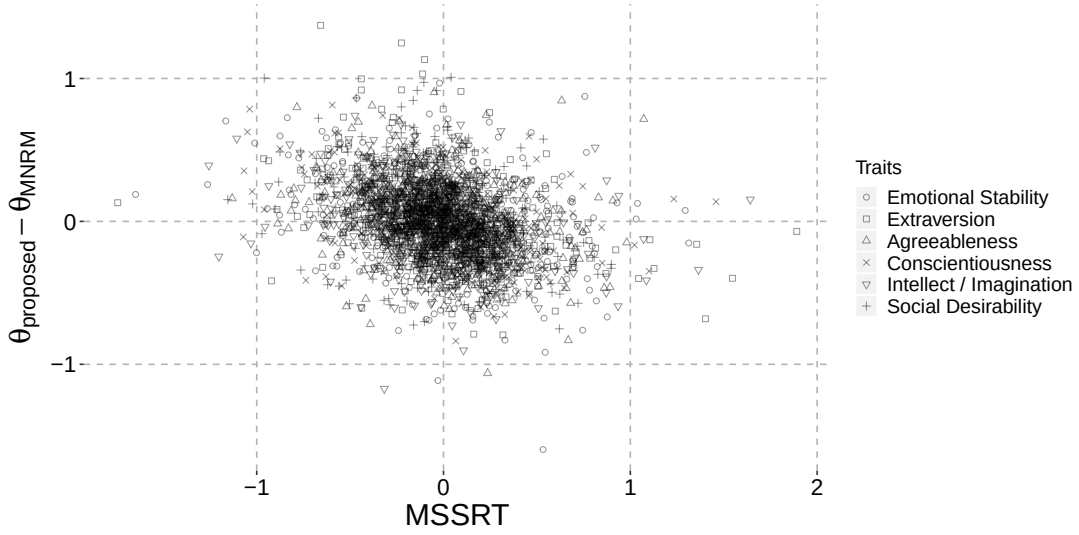

Figure 5.5: Scatter plot between the mean signed standardized RT and the difference of the estimates (Proposed model minus the MNRM model).

The above-mentioned result suggests that the proposed D-LBA IRT approach estimates slightly different latent traits due to its use of RT information. In order to examine how much the RT affected the parameter estimation, we examined the relationship between the difference in estimates and the mean signed standardized RT (MSSRT). The MSSRT indicates whether the respondent takes less time to choose positively-keyed statement and avoid negatively-keyed statement regarding each latent trait. Thus, the MSSRT is expected to correlate negatively with the difference in estimates between the proposed model, which uses RT information, and the MNRM model, which does not use it. For more detailed procedure of calculating the MSSRT, refer to Section 3.4.

Figure 5.5 is the scatter plot between the MSSRT and the difference in estimated trait scores. The sample correlation was $-.32$ when all traits are intermingled. On each trait, the sample correlation ranged from $-.41$ to $-.27$. This result indicates that the proposed D-LBA IRT model did use the RT information in a systematic manner in estimating $\boldsymbol{\theta}_i$.

5.4.2 Example 2: Application to 2AFC data

Next, we applied the proposed model to the 2AFC ($K = 2$) personality questionnaire in order to compare the model performance with the TDIRT model. The data used in this application were collected in Study 1. The data consist of 499 Japanese participants collected through an online crowdsourcing service platform. Participants were instructed to answer 25 2AFC personality questionnaire items which were designed to measure Big-Five traits. Each item consists of a pair of statements from the Japanese version of the Big-Five factor marker questionnaire Apple and Neff (2012). For more details of the data collection procedure, see Section 3.4. The data and sample R code can be obtained from the author's Open Science Framework website (`https://osf.io/jswqg/`).

Procedure Before analysis, some responses with RTs shorter than 300 ms were deleted list-wise as ineligible ones. We applied the proposed model and the TDIRT model to the dataset. The proposed model was estimated with the priors in Equation 5.9. For the TDIRT model, the priors used in this study were

$$\begin{aligned} \xi_j &\sim Cauchy_{[0,\infty)}(0, \lambda_\xi), & \gamma_i &\sim LN(0,1), \\ \mu_j^{(1-2)} &\sim N(0, \lambda_\mu^2), & \beta_j^{(k)} &\sim Cauchy_{[0,\infty)}(0, \lambda_\beta), \\ \boldsymbol{\theta}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}), & \boldsymbol{\Sigma} &\sim LKJCorr(1), \\ \tau_j &\sim U(0, \min(\mathrm{RT})_j), & \lambda_\xi, \lambda_\mu, \lambda_\beta, &\sim Cauchy_{[0,\infty)}(0, 5), \end{aligned} \tag{5.15}$$

where ξ_j is the item component of the boundary parameter in the diffusion process, which is denoted by γ_i/ξ_j. The intercept parameter $\mu_j^{(1-2)}$ is the same as that of $\mu_j^{(1)} - \mu_j^{(2)}$ in the TIRT and proposed D-LBA IRT models, which can be interpreted as the extent to which the mean utility of the first statement is larger than that of the second.

Note that in this example, as well as the example 1, we applied the hierarchical Bayesian approach to item parameters of the TDIRT model, whereas half-t distributions were adopted to ξ_j and $\beta_{jd_j^{(k)}}$ in Study 1.

Results Table 5.10 summarizes the parameter estimates obtained from the proposed D-LBA IRT and TDIRT models. The correlations between corresponding parameter estimates were .796 for boundary parameters (ξ), .974 for intercept parameters (μ), and .816 for factor loading parameters (β), respectively. Regarding the respondents' latent trait scores shown in Table 5.11, the correlations between estimates obtained by both models are .949, the smallest for all five traits. These results indicate that both models estimate similar parameters. Meanwhile, compared to the result of the TDIRT model, the proposed model tended to show higher correlations among the Big-five traits. Furthermore, the sample standard deviations were smaller in the D-LBA IRT model than the TDIRT model, though the prior distribution of population standard deviation was set to be one in both models.

Next, we compared the model performance using the two information criteria, the WAIC and WBIC. Both were calculated using the full MCMC sample. The WAIC were 2.00 and 2.24 for the D-LBA IRT model and the TDIRT model, respectively. The difference evidently indicates that the proposed model performs better than the TDIRT model. The same applies to the WBIC, as the values were 23607.02 for the D-LBA IRT model and 26648.41 for the TDIRT model. These results appear to show that the parameters of the D-LBA IRT corresponds to those of the TDIRT model to a certain extent. Hence, these models are somewhat compatible in terms of mere parameter estimation. Nevertheless, it would be important to choose the appropriate model based on the underlying cognitive processes. Still, researchers are usually not sure about the cognitive processes in many real-world situations. Thus, using the D-LBA IRT model rather than the TDIRT could be recommended because the D-LBA IRT model shows better performance in terms of information criteria and is expected to be more robust against model misspecification.

5.5 Discussion

In this study, we proposed a new cognitive process model for MAFC data by extending the unidimensional binary D-LBA IRT model. The proposed model is based on the LBA model

Table 5.10: Item parameters obtained by the D-LBA IRT model and the TDIRT model. Item numbers correspond with Table 3.4.

	Boundary		Intercept		Loadings			
	D-LBA IRT	TDIRT	D-LBA IRT	TDIRT	LBA IRT		TDIRT	
Item	ξ_j	ξ_j	$\mu_j^{(1)}$	$\mu_j^{(1-2)}$	$\beta_{jd_j^{(1)}}$	$\beta_{jd_j^{(2)}}$	$\beta_j^{(1)}$	$\beta_j^{(2)}$
1	0.454	0.296	-0.137	-0.213	0.688	0.483	0.722	0.473
2	0.314	0.248	0.383	0.363	0.609	0.607	0.287	0.333
3	0.442	0.284	0.207	0.406	0.236	0.385	0.258	0.383
4	0.422	0.294	-0.336	-0.536	0.396	0.536	0.256	0.441
5	0.384	0.276	-0.159	-0.258	0.423	0.564	0.366	0.411
6	0.481	0.301	-0.349	-0.653	0.176	0.649	0.252	0.560
7	0.323	0.233	0.174	0.226	0.235	0.145	0.161	0.088
8	0.362	0.239	-0.298	-0.451	0.450	0.072	0.332	0.075
9	0.466	0.294	0.052	0.109	0.716	0.100	0.710	0.115
10	0.408	0.265	0.088	0.137	0.737	0.076	0.502	0.161
11	0.401	0.267	-0.034	-0.055	0.347	0.422	0.329	0.410
12	0.370	0.256	0.168	0.137	1.143	1.724	0.460	0.730
13	0.323	0.247	-0.085	-0.102	0.460	0.440	0.380	0.246
14	0.416	0.250	0.328	0.558	0.274	0.309	0.204	0.251
15	0.414	0.291	0.034	0.026	0.096	0.058	0.095	0.044
16	0.299	0.221	0.230	0.214	1.003	1.129	0.401	0.460
17	0.340	0.236	0.201	0.319	0.437	0.171	0.343	0.142
18	0.354	0.243	-0.093	-0.160	0.270	0.482	0.221	0.433
19	0.424	0.236	0.354	0.673	0.345	0.225	0.316	0.241
20	0.306	0.236	0.026	0.022	0.519	0.627	0.312	0.355
21	0.421	0.279	-0.302	-0.412	0.622	0.518	0.441	0.372
22	0.357	0.275	-0.063	-0.084	0.119	0.136	0.086	0.113
23	0.317	0.234	-0.059	-0.048	0.513	0.242	0.250	0.186
24	0.479	0.276	-0.324	-0.395	0.906	0.113	0.576	0.070
25	0.333	0.259	0.100	0.165	0.205	0.494	0.175	0.366

Note: $\mu_j^{(2)}$ in the D-LBA IRT model always becomes $-\mu_j^{(1)}$ due to the sum-to-zero constraint and is therefore not shown in the table. Regarding to $\beta_{jd_j^{(k)}}$, the absolute values are shown.

and the random utility theory (Thurstone, 1927), both of which have firm scientific and mathematical foundations. The proposed model makes it possible for the first time to estimate the multidimensional psychological traits jointly based on the MAFC item response and its RT. The results of the simulation study revealed that the proposed model can recover parameters properly. In addition, it was demonstrated that the proposed D-LBA IRT model can be applied when the number of latent traits and statements per item are both two ($D = K = 2$), which is the case when the existing TIRT and TDIRT models cannot be applied.

We have shown two real data applications. The first one was for a 4AFC questionnaire dataset. Each parameter of the proposed model was considered to reflect specific psychometric properties. Correlations between some parameters and corresponding summary statistics were found, which supports the assumed meanings of the parameters. From the comparison of the estimates of the latent trait scores $\boldsymbol{\theta}_i$, we found that the proposed model obtains parameter estimates that are slightly different from the TIRT and MNRM models. As the correlation between the difference in estimates and the MSSRT suggests, this difference is due to the use of RT information in the proposed model. In the second application, we comparatively applied the proposed D-LBA IRT model and the TDIRT model to a 2AFC questionnaire dataset. The result indicated these two models obtain similar parameter estimates. Nevertheless, information criteria found the proposed model to be better than the TDIRT model. As indicated by the results in study 2, the D-LBA IRT model tends to be more robust than the TDIRT model when the true data-generation model is unknown.

The proposed model is expected to take over the advantages of both the forced-choice format and cognitive process models. As a future direction, it would be fruitful to empirically investigate whether and how much applications of the proposed model benefit from these advantages. We elaborate on both of these points.

First, as discussed in Section 1, the proposed approach is expected to remove systematic response biases such as acquiescence, extreme, and socially desirable responding. Due to the use of forced-choice measurement format, it is reasonable to believe that the acquiescence and extreme responding biases can be prevented. On the other hand, concerning socially desirable

responding, careful item construction would be a key to preventing this bias (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006). However, from the simulation results by A. Brown and Maydeu-Olivares (2011) and Bürkner et al. (2019), it is anticipated that the estimation becomes unstable if statements with similar social desirability are combined. Further detailed investigation would therefore be needed on the prevention of social desirable responding.

Second, the proposed model is a class of cognitive process models. The parameters in the proposed model are thus considered to have psychological meanings. One main advantage of the accumulator models is that they take both the boundary and the amount of information accumulation into account. As a result, researchers can investigate psychological phenomena in detail (e.g., Neville, Raaijmakers, & Maanen, 2019; Palada et al., 2016; Ratcliff et al., 2007). For instance, Ballard, Sewell, Cosgrove, and Neal (2019) investigated the influence of reward and punishment on decision making. From the viewpoint of the simple RT, the mean RT was longer under punishment than neutral condition, while the correct response rate was lower. Based on the LBA parameter estimates, they found that respondents set higher boundary to respond when reward is provided, and lowers the boundary when punishment is provided. In addition, the sum of drift rates was smaller under punishment condition. These results indicate that people tend to be less cautious while the rate of information accumulation also decreases when they are threatened with punishment. In this way, the breakdown of observed RT into several psychologically meaningful parameters makes it possible to provide insights into the cognitive processes involved in the task.

In addition, the proposed D-LBA IRT model decomposes LBA parameters into both item and respondent components. The proposed model can therefore be used for more in-depth analyses. For example, researchers can investigate which (item or respondent) factor is the main cause of the longer RT by observing the estimates of boundary parameters (ξ_j and γ_i). In our real data application, we have confirmed the correlation between the item boundary parameter ξ_j and the empirical readability score. The result suggests that the required amount of information accumulation increases as the statements in an item are more difficult to process. More such external validation of the model parameters would be hoped through empirical applications.

Table 5.11: The correlation matrix among the latent trait scores obtained by the D-LBA IRT model and the TDIRT model.

	D-LBA IRT					TDIRT						
	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	Mean	SD
θ_{i1}	-										-0.016	0.820
θ_{i2}	0.652	-									-0.011	0.865
θ_{i3}	0.327	0.627	-								0.004	0.766
θ_{i4}	0.421	0.208	0.359	-							0.009	0.798
θ_{i5}	0.474	0.684	0.288	0.181	-						-0.003	0.766
θ_{i1}	0.950					-					-0.020	0.859
θ_{i2}		0.966				0.588	-				-0.013	0.887
θ_{i3}			0.949			0.215	0.513	-			0.007	0.784
θ_{i4}				0.961		0.267	0.122	0.365	-		0.008	0.838
θ_{i5}					0.967	0.352	0.605	0.224	0.168	-	-0.003	0.800

Note: θ_{i1} = emotional stability; θ_{i2} = extraversion; θ_{i3} = agreeableness; θ_{i4} = conscientiousness; θ_{i5} = intellect/imagination

Chapter 6

General Discussion

6.1 Summary of the Series of Studies

The major objective of the series of studies was to propose new cognitive process IRT models for forced-choice personality measurement using RT information.

Three studies were conducted in this thesis. In Study 1, the combination of the Thurstonian IRT model and the D-diffusion IRT model was proposed. The Thurstonian IRT model has been successfully used to analyze forced-choice personality data by means of Thurstone's Law of Comparative Judgment. In contrast, the D-diffusion IRT model is a novel cognitive process model that can naturally use RT information, and estimate concurrently item and respondent parameters. As a result, a combination of the models, namely the Thurstonian D-diffusion IRT model, can analyze two-alternative forced-choice personality measurement data with RT information. In addition, the success of this model would indicate that the diffusion model is applicable to personality measurement data, even though the original model did not focus on these types of data.

Study 2 investigated a combination of the IRT model and the linear ballistic accumulator (LBA) model. The LBA model is considered simpler than the diffusion model, yet the parameters in the LBA model can be interpreted as the corresponding parameters in the diffusion model. We therefore proposed parameter decomposition of the LBA model, similar to that of

the D-diffusion IRT model. The proposed model was therefore named the (unidimensional binary) D-LBA IRT model. Several simulation results revealed that the proposed D-LBA model can estimate parameters more quickly and efficiently than the D-diffusion IRT model. Moreover, simulation and empirical data revealed that the D-LBA IRT model is expected to be more robust than the D-diffusion IRT model, when the true model is unknown.

Study 3 examined an extension of the D-LBA IRT model, namely the multidimensional MAFC D-LBA IRT model. To date, there exist no models that can be used to analyze MAFC personality measurement data with RT information. Empirical data showed that the proposed model can generate slightly different trait scores in comparison with other models that do not use RT information. In addition, the proposed model appears more robust than the Thurstonian D-diffusion IRT model, even when applied to two-alternative forced-choice data.

We proposed three new cognitive process IRT models. Each model applies to different situations: dichotomous or polytomous, unidimensional or multidimensional. Nevertheless, the success of these three models suggests that cognitive process models, such as the diffusion and LBA models, can be used in the field of psychometrics, particularly in personality measurement.

6.2 Future Orientations

There exist numerous directions for future studies. In this thesis, two different types of D-LBA IRT models were proposed. However, these models are not suitable for polytomous unidimensional measurement data, i.e., Likert scale data. Several psychometric models have been proposed that are applicable to Likert-type measurements with RT data (Ferrando & Lorenzo-Seva, 2007b; Ranger, 2013; Ranger & Ortner, 2011). Nevertheless, consideration of cognitive process models would be beneficial for investigating psychologically meaningful parameters, as noted in Section 1.3. A D-LBA IRT model for Likert-type data would therefore represent a promising direction for future study.

This thesis only considered personality measurement data. In terms of the use of RT data,

there is a fundamental difference between ability and personality measurement. In ability measurement, more able respondents generally find correct answers faster (van der Linden, 2016). This is obviously different from the inverted-U relationship. Therefore, the proposed models in the series of studies are not suitable for ability measurements as the drift rate is expressed as the difference between the respondent and item parameters. These model the respondent as answering more quickly when he or she has an extremely low ability; this is unlikely in ability measurement unless the respondent is simply guessing. Therefore, another interesting direction for future research would be to develop a Q-version of the LBA IRT model for ability measurements.

Throughout the series of studies, new models based on cognitive process models were proposed. These models include parameters such as the boundary or the drift rate. The likelihood function of the diffusion and the LBA models were derived so that the models can explain several empirical phenomena, e.g., the inverted-U relationship. The parameter values based on these models should therefore correlate with the corresponding summary statistics, as the results of the studies have shown. However, these results actually do not demonstrate the construct validity of the parameters. Therefore one promising direction for future research would be investigating the construct validity, or empirical meanings, of the parameters.

Even though the cognitive process models are expected to have advantages over psychometric models, there also exist two disadvantages. First, as mentioned in the discussion of study 2, these models take time to obtain parameter estimates. One major reason for the long estimation time is that these models need to be estimated through the MCMC approach. To reduce the calculation time, one solution would be to derive the posterior distributions analytically. However, the likelihood functions of the diffusion and the LBA models are quite complex. In addition, the number of parameters becomes large as the number of items and respondent increases. These issues render it impossible to derive the posterior distributions analytically. Second, the likelihood functions of these models are considered to non-differentiable. One of the major advantages of IRT is the item information. In IRT, the standard error of the estimate of trait score can be calculated using the item information function. In addition, the item information is

frequently used for adaptive testing. However, the item information function is equivalent to the Fisher information, which is the second derivative of the likelihood function with respect to the respondent parameter (θ_i). The proposed model contains two or three respondent parameters. For adaptive testing, the Fisher information matrix must be derived from the likelihood function of the proposed models. However, as mentioned above, the likelihood functions are quite complex and therefore not differentiable. These disadvantages appear to support the utility of psychometric models for e.g., fast estimation or adaptive testing. For instance, the log-normal model (van der Linden, 2006) is one well-known psychometric model that uses RT information. van der Linden (2008) demonstrated how to use RT information to improve item selection in adaptive testing. If researchers are interested in adaptive testing, this type of approach would be more beneficial than a cognitive process model approach.

Although there still exist difficulties with the use of cognitive process models, we believe these models provide information that is beneficial to several research fields, and the models proposed in this thesis lay a new path for RT modeling in psychometrics.

References

Akrami, N., Hedlund, L. E., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-U effect revisited. *Personality and Individual Differences*, *43*(3), 611–618. doi: 10.1016/j.paid.2006.12.005

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J., & Roberts, R. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment*, *33*(1), 83-97. doi: 10.1177/0734282914550387

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods*, *49*(3), 863–886. doi: 10.3758/s13428-016-0746-9

Apple, M. T., & Neff, P. (2012). Using Rasch measurement to validate the Big Five factor marker questionnaire for a Japanese university population. *Journal of Applied Measurement*, *13*(3), 276–296.

Ballard, T., Sewell, D. K., Cosgrove, D., & Neal, A. (2019). Information processing under reward versus under punishment. *Psychological Science*, *30*(5), 757–764. doi: 10.1177/0956797619835462

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56. doi: 10.1111/j.2044-8325.1996.tb00599.x

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*(4), 331–344. doi: 10.1037/1040-3590.10.4.331

Bertling, M., & Weeks, J. P. (2018). Using response time data to reduce testing time in cognitive

tests. *Psychological Assessment*, *30*(3), 328–338. doi: 10.1037/pas0000466

Bock, R. D. (1958). Remarks on the test of significance for the method of paired comparisons. *Psychometrika*, *23*(4), 323–334. doi: 10.1007/BF02289782

Böckenholt, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, *71*(4), 615–629. doi: 10.1007/s11336-006-1598-5

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press. doi: 10.1016/C2013-0-11050-1

Bradbery, J., Deuter, M., & Turnbull, J. (2015). *Oxford advanced learner's dictionary*. London: Oxford University Press.

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, *54*(2), 181–202. doi: 10.1007/BF02294514

Brown, A. (2010). *How item response theory can solve problems of ipsative data (Doctoral dissertation)* (Unpublished doctoral dissertation). University of Barcelona, Spain.

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. doi: 10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. doi: 10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147. doi: 10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How irt can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. doi: 10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2018). Modelling forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 523–569). Hobo-

ken: NJ: Wiley. doi: 10.1002/9781118489772.ch18

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of thurstonian irt models. *Educational and Psychological Measurement*. doi: 10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? a meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*. doi: 10.1037/apl0000414

Cassey, P. J., Gaut, G., Steyvers, M., & Brown, S. D. (2016). A generative joint model for spike trains and saccades during perceptual decision-making. *Psychonomic Bulletin and Review*, *23*(6), 1757–1778. doi: 10.3758/s13423-016-1056-z

Cattell, R. B. (1944). Psychological measurement: normative, ipsative, interactive. *Psychological Review*, *51*(5), 292–303. doi: 10.1037/h0057299

Chen, C., Lee, S.-y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among east asian and north american students. *Psychological Science*, *6*(3), 170–175. doi: 10.1111/j.1467-9280.1995.tb00327.x

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi: 10.1207/S15328007SEM0902

Christiansen, N., Burns, G., & Montgomery, G. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance*, *18*(3), 267–307.

Condon, D. M. (2018). *The SAPA personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model.* Retrieved from `https://psyarxiv.com/sc4p9/`

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3*(1), e6.

doi: 10.5334/jopd.al

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational and Organizational Psychology*, *67*(2), 89–100. doi: 10.1111/j.2044-8325.1994.tb00553.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi: 10.1007/BF02310555

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*(FEB). doi: 10.3389/fpsyg.2019.00102

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E. J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin and Review*, *18*(1), 61–69. doi: 10.3758/s13423-010-0022-4

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. doi: 10.1037/1040-3590.18.2.192

Feller, W. (1968). Random Walk and Ruin Problems. In W. Feller (Ed.), *An introduction to probability theory and its applications (vol. 1).* (3rd ed., pp. 342–371). New York: John Wiley and Sons, Inc.

Ferrando, P. J. (2007). A Pearson-Type-VII item response model for assessing person fluctuation. *Psychometrika*, *72*(1), 25–41. doi: 10.1007/s11336-004-1170-0

Ferrando, P. J., & Lorenzo-Seva, U. (2007a). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543. doi: 10.1177/0146621606295197

Ferrando, P. J., & Lorenzo-Seva, U. (2007b). A Measurement Model for Likert Responses That Incorporates Response Time. *Multivariate Behavioral Research*, *42*(4), 675–706.

doi: 10.1080/00273170701710247

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 427–448. doi: 10.1348/000711009X470740

Fontanella, L., Fontanella, S., Valentini, P., & Trendafilov, N. (2019). Simple structure detection through Bayesian exploratory multidimensional IRT models. *Multivariate Behavioral Research*, *54*(1), 100–112. doi: 10.1080/00273171.2018.1496317

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi: 10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed. ed.). New York: Chapman and Hall/CRC. doi: 10.1201/b16018

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. doi: 10.1214/08-AOAS191

Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology*, *7*, 201–208. doi: 10.1111/j.1744-6570.1954.tb01593.x

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. doi: 10.1037/1040-3590.4.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facts of several Five-Factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Grice, G. R. (1968). Stimulus intensity and response evocation. *Psychological Review*, *75*(5), 359–373.

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian

item response modeling. *Assessment*, *25*(4), 513–526. doi: 10.1177/1073191116641181

Harman, H. H. (1960). *Modern factor analysis.* Oxford, England: Univ. of Chicago Press.

Hasebe, Y., & Lee, J.-H. (2015). Introducing a readability evaluation system for Japanese language education. In *Proceedings of the 6th international conference on computer assisted systems for teaching & learning japanese (castel/j)* (pp. 19–22).

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology*, *45*(7), 1028–1045. doi: 10.1177/0022022114534773

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, NY: Springer New York. doi: 10.1007/978-1-4939-2236-9_2

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: Different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology*, *66*(2), 125–136. doi: 10.1037/a0028189

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*(AUG), 1–19. doi: 10.3389/fpsyg.2012.00292

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, *91*(1), 9–24. doi: 10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167–184. doi: 10.1037/h0029780

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "Fake-Proof" measure of the Big Five. *Journal of Research in Personality*, *42*(5), 1323–1333. doi: 10.1016/j.jrp.2008.04.006

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. doi: 10.1177/0146621615585851

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, *13*(4), 371–388. doi: 10.1207/S15327043HUP1304_3

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The ues of ipsative personality tests. *Journal of Occupational Psychology*, *61*, 153–162.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and Likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, *23*(1), 92–97. doi: 10 .1111/ijsa.12098

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY, US: Farrar, Straus and Giroux.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. doi: 10.1177/1094428115571894

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi: 10.1177/0013164406294779

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The "inverted-U RT effect" for self and other judgments. *Personality and Social Psychology Bulletin*, *7*(3), 438–443. doi: 10.1177/014616728173012

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* Oxford, England: Academic Press.

Lee, H., & Smith, W. Z. (2019). A Bayesian random block item response theory model for forced-choice formats. *Educational and Psychological Measurement*, *Advance online publication*. doi: 10.1177/0013164419871659

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Lee, P., Joo, S. H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Indi-*

vidual Differences, *142*(August 2018), 13–20. doi: 10.1016/j.paid.2019.01.022

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*(November 2017), 229–235. doi: 10.1016/j.paid.2017.11.031

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651–687.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 5–55.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41–67.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luce, R. D. (1959). *Individual choice behavior.* Oxford, England: John Wiley.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233. doi: 10.1016/0022-2496(77)90032-3

Luce, R. D. (1986). *Response times*. New York: Oxford University Press. doi: 10.1093/acprof:oso/9780195070019.001.0001

Masuda, S., & Sakagami, T. (2017). Respondents with low motivation tend to choose middle category : survey questions on happiness in Japan. *Behaviormetrika*, *44*(2), 593–605. doi: 10.1007/s41237-017-0026-8

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531–552. doi: 10.1348/0963179042596504

Meng, X. B., Tao, J., & Shi, N. Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computa-*

tion and Simulation, *84*(1), 1–21. doi: 10.1080/00949655.2012.692368

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437–449. doi: 10.2139/ssrn.1901533

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). Fitting diffusion item response theory models for responses and response times using the R package diffirt. *Journal of statistical software*, *66*(4), 1–34. doi: 10.18637/jss.v066.i04

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. doi: 10.1016/j.jmp.2009.02.003

Neville, D. A., Raaijmakers, J. G. W., & Maanen, L. V. (2019). Modulation of the word frequency effect in recognition memory after an unrelated lexical decision task. *Journal of Memory and Language*, *108*, 104026. doi: 10.1016/j.jml.2019.05.004

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32. doi: 10.2307/1914288

Okada, K., Vandekerckhove, J., & Lee, M. D. (2018). Modeling when people quit: Bayesian censored geometric models with hierarchical and latent-mixture extensions. *Behavior Research Methods*, *50*, 406–415. doi: 10.3758/s13428-017-0879-5

Palada, H., Neal, A., Vuckovic, A., Martin, R., Samuels, K., & Heathcote, A. (2016). Evidence accumulation in a complex task: Making choices about concurrent multiattribute stimuli under time pressure. *Journal of Experimental Psychology: Applied*, *22*(1), 1–23. doi: 10.1037/xap0000074

Paulhus, D. L. (1988). Balanced inventory of desirable responding (BIDR). *Acceptance and Commitment Therapy. Measures Package*, *41*, 79586–79587.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego: Academic Press, Inc. doi: 10.1016/

B978-0-12-590241-0.50006-X

Phelps, L., Schmitz, C. D., & Boatright, B. (1986). The effects of halo and leniency on cooperating teacher reports using Likert-type rating scales. *Journal of Educational Research*, *79*(3), 151–154.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. doi: 10.1037/0021-9010.88.5.879

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*(4), 1137–1145. doi: 10.1037/a0033323

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, *55*(4), 361–382.

Ranger, J., Kuhn, J. T., & Szardenings, C. (2016). Limited information estimation of the diffusion-based item response theory model for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *69*(2), 122–138. doi: 10.1111/bmsp.12064

Ranger, J., Kuhn, J. T., & Szardenings, C. (2017). Analysing model fit of psychometric process models: An overview, a new test and an application to the diffusion model. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 209–224. doi: 10.1111/bmsp.12082

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406. doi: 10.1177/0013164410382895

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., Thapar, A., & Mckoon, G. (2007). Application of the diffusion model to two-choice

tasks for adults 75 − 90 years old. *Psychology and Aging*, *22*(1), 56–66.

Ratcliff, R., Thompson, C. A., & Mckoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, 115–136. doi: 10.1016/j.cognition .2014.12.004.Modeling

Raz, S., Bar-Haim, Y., Sadeh, A., & Dan, O. (2014). Reliability and validity of the online continuous performance test among young adults. *Assessment*, *21*(1), 108–118. doi: 10.1177/1073191112443409

Reddi, B. A. J., & Carpenter, R. H. S. (2000). The influence of urgency time on performance. *Nature neuroscience*, *3*(8), 827–830.

Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, *38*(7), 549–562. doi: 10.1177/0146621614536272

Revuelta, J., & Ximénez, C. (2017). Bayesian dimensionality assessment for the multidimensional nominal response model. *Frontiers in Psychology*, *8*(JUN). doi: 10.3389/ fpsyg.2017.00961

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*(2), 184–201. doi: 10.1037/1040-3590.14.2.184

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In *Progress in mathematical psychology, 1.* (pp. 151–174). New York, NY, US: Elsevier Science.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York, NY: Springer New York. doi: 10.1007/978-1-4757-2691-6_11

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*(1), 3–30. doi: 10.1080/1359432X.2012.716198

Sass, R., Frick, S., Reips, U. D., & Wetzel, E. (n.d.). Taking the test taker's perspective:

Response process and test motivation in multidimensional forced-choice versus rating scale instrument. *Assessment*. doi: 10.1177/1073191118762049

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442–450. doi: 10.1177/0963721417708229

Stan Development Team. (2018). *Stan modeling language users guide and reference manual, version 2.18.0.* Retrieved from `http://mc-stan.org`

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184–203. doi: 10.1177/0146621604273988

Stone, C. A. (1992). Recovery of marginal maximum likelihood response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1–16. doi: 10.1177/014662169201600101

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. doi: 10.1007/BF02289729

Tani, I. (2008). Baransu gata syakaiteki nozomasisa hannou syakudo (BIDR-J) no sakusei to sinraisei, datousei no kentou [Development of Japanese version of balanced inventory of desirable responding (BIDR-J)]. *The Japanese Journal of Personality*, *17*(1), 18–28. doi: 10.2132/personality.17.18

Thissen, D. (1983). Timed testing: An approach using item response theory. In *New horizons in testing: Latent trait test theory and computerized adapting testing* (pp. 179–203). New York: Academic Press. doi: 10.1016/B978-0-12-742780-5.50019-6

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. doi: 10.1037/h0070288

Thurstone, L. L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology*, *14*(3), 187–201.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650. doi: 10.1007/s11336-000-0810-3

Tuerlinckx, F., Molenaar, D., & van der Maas, H. L. J. (2016). Diffusion-based response-time models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 283–300). Boca Raton: Chapman & Hall/CRC Press.

Vancleef, K., Read, J. C., Herbert, W., Goodship, N., Woodhouse, M., & Serrano-Pedraza, I. (2018). Two choices good, four choices better: For measuring stereoacuity in children, a four-alternative forced-choice paradigm is more efficient than two. *PLoS ONE*, *13*(7), 1–15. doi: 10.1371/journal.pone.0201366

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. doi: 10.1016/j.jmp.2014.06.004

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. doi: 10.3102/10769986031002181

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. doi: 10.3102/1076998607302626

van der Linden, W. J. (2016). Lognormal response-time model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 261–282). Boca Raton: Chapman & Hall/CRC Press. doi: 10.1201/9781315374512-17

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356. doi: 10.1037/a0022749

van der Maas, H. L. J., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*(1), 29–60. doi: 10.2307/30039042

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter re-

covery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*(6), 463–473. doi: 10.1016/j.jmp.2009.09.004

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197–210. doi: 10.1177/00131649921969802

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. doi: 10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220. doi: 10.3758/BF03196893

Wagenmakers, E. J., van der Maas, H. L., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. doi: 10.3758/BF03194023

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*(1), 223–254. doi: 10.1007/s11336-016-9525-x

Wang, W.-C., Qiu, X.-l., Chen, C.-w., Ro, S., & Jin, K.-y. (2017). Item Response Theory Models for Ipsative Tests With Multidimensional Pairwise Comparison Items. *Applied Measurement in Education*, *41*(8), 600–613. doi: 10.1177/0146621617703183

Watanabe, S. (2010). Asymptotic equivalence of Bayescross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi: 10 .1207/s15324818ame1802_2

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental Psychology*. NY: Holt.

Appendix A: stan code for the Thurstonian IRT model

```
data{
     int <lower=1> I;
     int <lower=1> J;
     int <lower=1> K;
     int <lower=2> D; //number of factors
     int <lower=1,upper=D> D1[J]; //trait of choice 1
     int <lower=1,upper=D> D2[J]; //trait of choice 2
     int <lower=1,upper=I> II[K];
     int <lower=1,upper=J> JJ[K];
     int <lower=0,upper=1> X[K];
     int <lower=-1,upper=1> R1[J  ];// indicator whether the statement 1 is reversed
     int <lower=-1,upper=1> R2[J]; // indicator whether the statement 2 is reversed
}

parameters {
  matrix[D,I] z_trait;
  real mu[J]; //threshold
  real <lower=0>beta1[J]; // discriminaton1
  real <lower=0>beta2[J]; // discriminaton2
  cholesky_factor_corr[D] Sigma_lower; //correlation
}

transformed parameters{
  matrix[I,D] theta;
  theta = (Sigma_lower * z_trait)';
}


model {
  mu ~ normal(0,2.5);
  beta1 ~ student_t(4,0,2.5);
  beta2 ~ student_t(4,0,2.5);
  Sigma_lower ~ lkj_corr_cholesky(1);
  to_vector(z_trait) ~ normal(0,1);

  for(k in 1:K){
```

```
37     X[k] ~ bernoulli_logit(((R1[JJ[k]]*beta1[JJ[k]]*theta[II[k]][D1[JJ[
           k]]]-R2[JJ[k]]*beta2[JJ[k]]*theta[II[k]][D2[JJ[k]]])-mu[JJ[k
           ]]));
38   }
39 }
40
41 generated quantities{
42   corr_matrix[D] Sigma;
43   Sigma = multiply_lower_tri_self_transpose(Sigma_lower);
44 }
```

Appendix B: stan code for the Thurstonian D-diffusion IRT model

```
    data{
     int <lower=1> I;
     int <lower=1> J;
     int <lower=1> K;
     int <lower=2> D; //number of factors
     int <lower=1,upper=D> M1[J]; //trait of choice 1
     int <lower=1,upper=D> M2[J]; //trait of choice 2
     int <lower=1,upper=I> II[K];
     int <lower=1,upper=J> JJ[K];
     int <lower=1,upper=2> X[K];
     real<lower=0> T[K];
     int <lower=-1,upper=1> R1[J];
     int <lower=-1,upper=1> R2[J];
     vector<lower=0>[J] minRT;
}

parameters {
  matrix[D,I] z_trait;
  real mu[J];
  real <lower=0>beta1[J];
  real <lower=0>beta2[J];
  real <lower=0>xi[J];
  real <lower=0>gamma[I];
  // real <lower=0> s;
  vector<lower=0,upper=1>[J] tau_raw;
  cholesky_factor_corr[D] Sigma_lower;
}

transformed parameters{
  matrix[I,D] theta;
  vector[J] tau;
  real <lower=0> alpha[K];
  real v[K];
  real taus[K];
  theta = (Sigma_lower * z_trait)';
  tau = minRT .* tau_raw;
```

```
37
38   for (k in 1:K){
39     alpha[k] = (gamma[II[k]]/xi[JJ[k]]);
40     taus[k] = tau[JJ[k]];
41     if(X[k]==1){
42       v[k] = ((R1[JJ[k]]*beta1[JJ[k]]*theta[II[k]][D1[JJ[k]]]-R2[JJ[k
           ]]*beta2[JJ[k]]*theta[II[k]][D2[JJ[k]]])-mu[JJ[k]]);
43     } else {
44       v[k] = -((R1[JJ[k]]*beta1[JJ[k]]*theta[II[k]][D1[JJ[k]]]-R2[JJ[k
           ]]*beta2[JJ[k]]*theta[II[k]][D2[JJ[k]]])-mu[JJ[k]]);
45     }
46   }
47 }
48 model {
49   xi ~ student_t(4,0,2.5);
50   gamma ~ lognormal(0,1);
51   mu ~ normal(0,2.5);
52   beta1 ~ student_t(4,0,2.5);
53   beta2 ~ student_t(4,0,2.5);
54   to_vector(z_trait) ~ normal(0,1);
55   tau_raw ~ beta(1,1);
56   Sigma_lower ~ lkj_corr_cholesky(1);
57
58   T ~ wiener(alpha,taus,0.5,v);
59 }
60
61 generated quantities{
62   corr_matrix[D] Sigma;
63   Sigma = multiply_lower_tri_self_transpose(Sigma_lower);
64 }
```

Appendix C: stan code for the unidimensional binary D-LBA IRT model

```
functions{

 real lba_pdf(real t, real B, real A, real v, real s){
      // PDF of the LBA model

      real B_A_tv_ts;
      real B_tv_ts;
      real term_1;
      real term_2;
      real term_3;
      real term_4;
      real pdf;

      B_A_tv_ts = (B - A - t*v)/(t*s);
      B_tv_ts = (B - t*v)/(t*s);
      term_1 = v*Phi_approx(B_A_tv_ts);
      term_2 = s*exp(normal_lpdf(B_A_tv_ts|0,1));
      term_3 = v*Phi_approx(B_tv_ts);
      term_4 = s*exp(normal_lpdf(B_tv_ts|0,1));
      pdf = (1/A)*(-term_1 + term_2 + term_3 - term_4);

      return pdf;
 }

 real lba_cdf(real t, real B, real A, real v, real s){
      // CDF of the LBA model

      real B_A_tv;
      real B_tv;
      real ts;
      real term_1;
      real term_2;
      real term_3;
      real term_4;
      real cdf;

```

```
37          B_A_tv = B - A - t*v;
38          B_tv = B - t*v;
39          ts = t*s;
40          term_1 = B_A_tv/A * Phi_approx(B_A_tv/ts);
41          term_2 = B_tv/A * Phi_approx(B_tv/ts);
42          term_3 = ts/A * exp(normal_lpdf(B_A_tv/ts|0,1));
43          term_4 = ts/A * exp(normal_lpdf(B_tv/ts|0,1));
44          cdf = 1 + term_1 - term_2 + term_3 - term_4;
45
46          return cdf;
47
48      }
49
50      real lba_lpmf(int response, real RT, real B, real A, row_vector v,
            real s,real tau){
51
52          real t;
53          real cdf;
54          real pdf;
55          real prob;
56          real out;
57          real prob_neg;
58
59              t = RT - tau;
60              if(t > 0){
61                  cdf = 1;
62
63                  for(j in 1:num_elements(v)){
64                      if(response == j){
65                          pdf = lba_pdf(t, B, A, v[j], s);
66                      }else{
67                          cdf = (1-lba_cdf(t, B, A, v[j], s)) * cdf;
68                      }
69                  }
70                  prob_neg = 1;
71                  for(j in 1:num_elements(v)){
72                      prob_neg = Phi(-v[j]/s) * prob_neg;
73                  }
74                  prob = pdf*cdf;
75                  prob = prob/(1-prob_neg);
76                  if(prob < 1e-10){
77                      prob = 1e-10;
78                  }
79
80              }else{
81                  prob = 1e-10;
82              }
83          return log(prob);
84      }
85
86  }
87
```

```
data{
     int <lower=1> I;
     int <lower=1> J;
     int <lower=1> K;
     int <lower=1,upper=I> II[K];
     int <lower=1,upper=J> JJ[K];
     int<lower=0,upper=2> X[K];
     real<lower=0> T[K];
}

parameters {
     real<lower=0> xi [J];
     real<lower=0> gamma [I];
     real theta [I];
     real b [J];
     real<lower=0> sigma[I];
     real<lower=0> psi[J];
     real<lower=0> tau [J];
}

transformed parameters {
     matrix <lower=0,upper=1>[K,2] v ;
     for (k in 1:K){
         v[k,1] = inv_logit(theta[II[k]]-b[JJ[k]]);
         v[k,2] = inv_logit(-theta[II[k]]+b[JJ[k]]);
     }
}

model {
     xi ~ cauchy(0,5);
     gamma ~ lognormal(0,1);
     b ~ normal(0,2.5);
     theta ~ normal(0,1);
     psi ~ cauchy(0,5);
     sigma ~ lognormal(0,1);
     tau ~ cauchy(0,5);

     for (k in 1:K){
         target += (lba_lpmf(X[k]|T[k],gamma[II[k]]/xi[JJ[k]],0.5*gamma
             [II[k]]/xi[JJ[k]],v[k,],sigma[II[k]]/psi[JJ[k]],tau[JJ[k
             ]]));
     }
}

generated quantities{
  vector[K] log_lik;
  for(k in 1:K){
    log_lik[k] = lba_lpmf(X[k]|T[k],gamma[NN[k]]/xi[JJ[k]],0.5*gamma[
        II[k]]/xi[JJ[k]],v[k,],sigma[II[k]]/psi[JJ[k]],tau[JJ[k]]);
  }
}
```

Appendix D: stan code for the D-diffusion IRT model

```
data{
     int <lower=1> I;
     int <lower=1> J;
     int <lower=1> K;
     int <lower=1,upper=I> II[K];
     int <lower=1,upper=J> JJ[K];
     int <lower=0,upper=2> X[K];
     real<lower=0> T[K];
}

parameters {
  real theta[I];
  real b[J];
  real <lower=0>xi[J];
  real <lower=0>gamma[I];
  real <lower=0.0000000001>tau[J];
  // An error occurs if tau == 0 in wiener;
}

transformed parameters{
  real <lower=0> alpha[K];
  real v[K];
  real taus[K];
  for (k in 1:K){
    alpha[k] = (gamma[II[k]]/xi[JJ[k]]);
    taus[k] = tau[JJ[k]];
    if(X[k]==1){
      v[k] = (theta[II[k]]-b[JJ[k]]);
    } else {
      v[k] = (-theta[II[k]]+b[JJ[k]]);
    }
  }
}
model {
  xi ~ cauchy(0,5);
  gamma ~ lognormal(0,1);
```

```
37   b ~ normal(0,2.5);
38   theta ~ normal(0,1);
39   tau ~ cauchy(0,5);
40
41   T ~ wiener(alpha,taus,0.5,v);
42 }
43
44 generated quantities{
45   vector[K] log_lik;
46   for(k in 1:K){
47     log_lik[k] = (
48       wiener_lpdf(T[k]|(gamma[II[k]]/xi[JJ[k]]),tau[JJ[k]],0.5,( theta
           [II[k]]-b[JJ[k]]))*(2-X[k])+
49       wiener_lpdf(T[k]|(gamma[II[k]]/xi[JJ[k]]),tau[JJ[k]],0.5,(-theta
           [II[k]]+b[JJ[k]]))*(X[k]-1)
50       );
51   }
52 }
```

Appendix E: stan code for the multidimensional MAFC LBA IRT model

```
functions{

    real lba_pdf(real t, real B, real A, real v, real s){
        // PDF of the LBA model

        real B_A_tv_ts;
        real B_tv_ts;
        real term_1;
        real term_2;
        real term_3;
        real term_4;
        real pdf;

        B_A_tv_ts = (B - A - t*v)/(t*s);
        B_tv_ts = (B - t*v)/(t*s);
        term_1 = v*Phi_approx(B_A_tv_ts);
        term_2 = s*exp(normal_lpdf(B_A_tv_ts|0,1));
        term_3 = v*Phi_approx(B_tv_ts);
        term_4 = s*exp(normal_lpdf(B_tv_ts|0,1));
        pdf = (1/A)*(-term_1 + term_2 + term_3 - term_4);

        return pdf;
    }

    real lba_cdf(real t, real B, real A, real v, real s){
        // CDF of the LBA model

        real B_A_tv;
        real B_tv;
        real ts;
        real term_1;
        real term_2;
        real term_3;
        real term_4;
        real cdf;

```

```
37          B_A_tv = B - A - t*v;
38          B_tv = B - t*v;
39          ts = t*s;
40          term_1 = B_A_tv/A * Phi_approx(B_A_tv/ts);
41          term_2 = B_tv/A * Phi_approx(B_tv/ts);
42          term_3 = ts/A * exp(normal_lpdf(B_A_tv/ts|0,1));
43          term_4 = ts/A * exp(normal_lpdf(B_tv/ts|0,1));
44          cdf = 1 + term_1 - term_2 + term_3 - term_4;
45
46          return cdf;
47
48      }
49
50      real lba_lpmf(int response, real RT, real B, real A, row_vector v,
            real s){
51
52          real t;
53          real cdf;
54          real pdf;
55          real prob;
56          real out;
57          real prob_neg;
58
59              t = RT;
60                  cdf = 1;
61
62                  for(j in 1:num_elements(v)){
63                      if(response == j){
64                          pdf = lba_pdf(t, B, A, v[j], s);
65                      }else{
66                          cdf = (1-lba_cdf(t, B, A, v[j], s)) * cdf;
67                      }
68                  }
69                  prob_neg = 1;
70                  for(j in 1:num_elements(v)){
71                      prob_neg = Phi(-v[j]/s) * prob_neg;
72                  }
73                  prob = pdf*cdf;
74                  prob = prob/(1-prob_neg);
75                  if(prob < 1e-10){
76                      prob = 1e-10;
77                  }
78
79          return log(prob);
80      }
81  }
82
83  data{
84      int <lower=1> I;
85      int <lower=1> J;
86      int <lower=1> K;
87      int <lower=2> D; //number of factors
```

```
88      int <lower=2> S; //number of statements in each item
89      int <lower=1,upper=D> D_each[J*S]; //trait of each statement
90      int <lower=1,upper=I> II[K];
91      int <lower=1,upper=J> JJ[K];
92      int <lower=0> X[K];
93      real<lower=0> T[K];
94      vector<lower=-1,upper=1>[J*S] R; // indicator whether each statement is
            reversed
95
96      // indicator of each statement
97      // when all statement have different parameters:
98      // R syntax > ITEM <- matrix(1:(J*S), nrow=J, ncol=S, byrow=T)
99      int <lower=1,upper=J*S> ITEM[J,S];
100 }
101
102 parameters {
103   matrix[D,I] z_trait;
104   matrix[S-1,J] mu; //threshold
105   vector<lower=0>[J*S] beta; //factor loadings
106   real <lower=0>xi[J];
107   real <lower=0>gamma[I];
108   cholesky_factor_corr[D] rho;
109   real<lower=0> sigma[I];
110   real<lower=0> psi[J];
111   real<lower=0> lambda_xi;
112   real<lower=0> lambda_beta;
113   real<lower=0> lambda_mu;
114   real<lower=0> lambda_psi;
115 }
116
117 transformed parameters {
118   matrix[S,J] mu_center; // centered mu
119   matrix[D,I] theta;
120   matrix[J*S,I] pref;
121
122   theta = (rho * z_trait);
123   for(j in 1:J) mu_center[,j] = append_row(mu[,j],-sum(mu[,j]));
124
125   pref = rep_matrix(R .* beta,I) .* theta[D_each,] - rep_matrix(
          to_vector(mu_center),I);
126 }
127
128 model {
129      xi ~ cauchy(0,lambda_xi);
130      gamma ~ lognormal(0,1);
131      to_vector(mu) ~ normal(0,lambda_b);
132      to_vector(z_trait) ~ normal(0,1);
133      rho ~ lkj_corr_cholesky(1);
134      beta ~ cauchy(0,lambda_beta);
135      psi ~ cauchy(0,lambda_psi);
136      sigma ~ lognormal(0,1);
137
```

```
138     lambda_xi ~ cauchy(0,5);
139     lambda_mu ~ cauchy(0,5);
140     lambda_beta ~ cauchy(0,5);
141     lambda_psi ~cauchy(0,5);
142
143     for (k in 1:K){
144         target += lba_lpmf(X[k]|T[k],gamma[II[k]]/xi[JJ[k]],0.5*gamma[
                II[k]]/xi[JJ[k]],
145         softmax(pref[ITEM[JJ[k]],II[k]]),sigma[II[k]]/psi[JJ[k]]);
146     }
147 }
148
149 generated quantities{
150   vector[K] log_lik;
151   corr_matrix[D] Cor_theta;
152   Cor_theta = multiply_lower_tri_self_transpose(rho);
153   for(k in 1:K){
154     log_lik[k] = lba_lpmf(X[k]|T[k],gamma[II[k]]/xi[JJ[k]],0.5*gamma[
          II[k]]/xi[JJ[k]],
155          softmax(pref[ITEM[JJ[k]],II[k]]),sigma[II[k]]/psi[JJ[k]]);
156   }
157 }
```

Appendix F: stan code for the multidimensional nominal response model

```
data{
     int <lower=1> I;
     int <lower=1> J;
     int <lower=1> K;
     int <lower=2> D; //number of factors
     int <lower=2> S; //number of statements in each item
     int <lower=1,upper=D> D_each[J*S]; //trait of each statement
     int <lower=1,upper=I> II[K];
     int <lower=1,upper=J> JJ[K];
     int <lower=0> X[K];
     vector<lower=-1,upper=1>[J*S] R; // indicator whether each statement is
         reversed

     // indicator of each statement
     // when all statement have different parameters:
     // R syntax > ITEM <- matrix(1:(J*S), nrow=J, ncol=S, byrow=T)
     int <lower=1,upper=J*S> ITEM[J,S];
}

parameters {
  matrix[D,I] z_trait;
  matrix[S-1,J] mu; //threshold
  vector<lower=0>[J*S] beta; //factor loadings
  cholesky_factor_corr[D] rho;
  real<lower=0> beta_scale;
  real<lower=0> b_scale;
}

transformed parameters {
  matrix[S,J] mu_center; // centered b
  matrix[D,I] theta;
  matrix[J*S,I] pref;

  theta = (rho * z_trait);
  for(j in 1:J) mu_center[,j] = append_row(mu[,j],-sum(mu[,j]));

```

```
36   pref = rep_matrix(R .* beta,I) .* theta[D_each,] - rep_matrix(
         to_vector(mu_center),I);
37 }
38
39 model {
40     to_vector(b) ~ normal(0,b_scale);
41     to_vector(z_trait) ~ normal(0,1);
42     rho ~ lkj_corr_cholesky(1);
43     beta ~ cauchy(0,beta_scale);
44
45     b_scale ~ cauchy(0,5);
46     beta_scale ~ cauchy(0,5);
47
48   for(k in 1:K){
49     X[k] ~ categorical_logit(pref[ITEM[JJ[k]],II[k]]);
50   }
51 }
52
53 generated quantities{
54   corr_matrix[D] Sigma = multiply_lower_tri_self_transpose(rho);
55 }
```

謝辞

本学位論文は、著者が東京大学大学院教育学研究科教育心理学コース博士課程に在学中の研究成果をまとめたものです。論文の執筆にあたって、著者の指導教員である東京大学大学院教育学研究科の岡田謙介准教授には論文の構成から表現の内容に至るまで、様々なアドバイスを頂戴いたしました。また、様々な学会や研究会へお誘い頂いたことで、引きこもりがちな私も研究者としての視野を多少広く持つことができたのではないかと感じております。深く感謝いたします。

お忙しい中学位論文の審査をお引き受けいただいた、東京大学大学院教育学研究科の影浦峡教授、針生悦子教授、早稲田大学人間科学学術院の杉澤武俊准教授、東京大学高大接続研究開発センターの宇佐美慧准教授にも、この場を借りてお礼申し上げます。数々のコメントの視点の広さと深さには敬服いたしました。博士論文に限らず、一研究者としての姿勢を考えるきっかけを頂きました。これからも考え続けていきたいと思います。

博士論文執筆の外に目をやると、私の大学院生活は他にも様々な方々のお陰で成り立っていたと感じます。特に東京大学名誉教授の南風原朝和先生、慶應義塾大学経済学研究科の星野崇宏教授には学部生時代からときに厳しく指導いただきながらも暖かく見守っていただきました。星野先生のご指導により学振特別研究員に採択されたことで、博士課程では研究に集中することができ、博士論文を早々に書き上げることができました。また、南風原先生には最後の口述試験にも足をお運びいただきました。緊張感が漲るとともに非常に心強かったことを覚えています。心よりお礼申し上げます。

同じ研究室の学生については、ゼミの時間を通して私の研究に関して様々なコメントを頂いただけでなく、皆さんの研究からも多くの刺激を受けました。研究発表を聞いているうちに新しいネタを思いつくこともありました。皆さんの研究発表へコメントするうちに自身の研究の方法についても再度考えさせられることもありました。振り返ってみると、切磋琢磨しあえる仲間がいるという環境は非常に幸せなことであったと感じています。

最後に、両親に感謝いたします。何やらよくわからないことをやり続けて、当初の予定より５年も社会進出が遅くなってしまった、それでも自由にやらせてくれて、何があっても見捨てることなく見守り続けてくれました。本当にありがとうございます。改めて、これまで多くの方々に支えていただいて今の自分があるという事実に深謝いたします。そして、今後ともご指導ご鞭撻のほど宜しくお願い申し上げます。

2020 年 3 月
分寺杏介