16

Lognormal Response-Time Model

Wim J. van der Linden

CONTENTS

16.1 Introduction

Raw responses to the items in an ability test typically have the physical form of bubbles on an answer sheet, clicks with a computer mouse, or a few lines of text. Each of these responses has two important aspects: (i) its degree of correctness and (ii) the time used to produce it. The former is usually recorded on a numerical scale, typically a 0–1 scale for incorrect–correct responses or a more-refined scale for the evaluation of polytomous items. Until recently, response times (RTs) on items could be recorded only for individually proctored tests. But since the arrival of computers and handheld electronic devices in testing, they have become automatically available for any type of test.

It may be tempting to view the correctness of a response and the time required to produce it as the result of two highly dependent processes, or even one single process. In fact, this is exactly what two different traditions of descriptive analyses of responses and RTs seem to suggest. The first tradition is that of item analysis in educational and psychological testing, where predominantly negative correlations between the correctness of responses and RTs have been found for tests ranging from adaptive matrices (e.g., Hornke, 2000) to medical-licensing exams (e.g., Swanson et al., 2001). It seems easy to explain the sign of these correlations: Test takers with lower abilities struggle for a longer time to find answers to the items, and when they find one, it is likely to be wrong. On the other hand, more able test takers usually find correct answers faster.

The other tradition is that of psychological research on speed–accuracy trade-offs (SATs), typically conducted as experiments with subjects replicating a standardized task multiple times under conditions inducing different levels of speed. These experiments invariably result in positive correlations between the numbers of correct responses and the average RT (for a review, see Luce, 1986, Section 6.5). The standard explanation of these results seems convincing as well: Subjects who work faster do so at the expense of their accuracy and consequently produce fewer correct responses.

To add to the confusion, consider the case of the RT patterns of different test takers responding to the same items. Table 16.1 shows the RTs by two test takers on the first 10 items in a 65-item cognitive ability test. As the test takers were arbitrarily selected from a much larger set, it seems safe to assume that they worked entirely independently. Nevertheless, the full patterns with the RTs for the two students correlated to 0.89! In fact, positive correlations can be found for nearly any pair of students on any kind of ability test.

How is it possible for these three different types of correlations to be so conflicting with each other, or with reality? The answer is: aggregation of data across hidden (or latent) covariates. For instance, for the RT patterns in Table 16.1, the pertinent covariate is the labor intensity of the items in the test. The solution for some of the test items just required more labor than for others. As a consequence, even when they work completely independently, the expected RTs of any pair of test takers simply go up and down jointly with the amount of labor required by the items. An example of a factor that contributes to the difference in labor between the items in an arithmetic test is the length of the computations. The longer-addition item in Figure 16.1 obviously demands more time than the shorter item (even though the latter may be more difficult because of the presence of three-digit numbers). In several other data sets analyzed by this author, the expected RTs required by the items easily differed by a factor of more than 5; for the data set for a large-scale adaptive test analyzed in van der Linden and Guo (2008, Table 16.1), they even ranged from some 14 to 812 s per item.

The same type of spuriousness due to hidden covariates explains the difference between the negative and positive correlations in the two earlier traditions of item analysis and

TABLE 16.1

RTs by Two Test Takers on the First 10 Items in an Ability Test

	Test Taker	
Item	$p = 1$	$p = 2$
1	22	26
2	19	38
3	40	101
4	43	57
5	27	37
6	21	27
7	45	116
8	23	44
9	14	10
10	47	117
r	0.89	

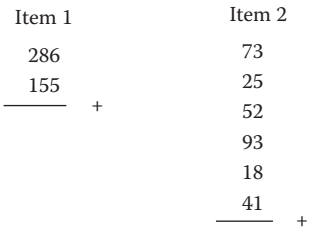

FIGURE 16.1
Two addition items of unequal length differing in the amount of cognitive labor they require.

SAT research. Typical covariates associated with test items are their difficulty and time intensity. If we aggregate responses and RTs across items, depending on the sign of the correlation between these covariates, spurious positive or negative correlations between the responses and RTs may be observed. The same holds for aggregation across test takers who vary in their ability and speed of work. If we aggregate both across items and test takers, the correlations between their covariates interact and, depending on their signs and relative strengths, the results are difficult to predict without any further formal analysis.

An effective solution to the problem of spurious correlation is to explicitly model the relationship between the observed variables and all covariates. If the model does fit the empirical data satisfactorily, we are able to neutralize the effects of the covariates by conditioning on them. This principle, which underlies all latent variable modeling, has found applications in diverse areas such as physics and psychometrics; for a discussion of some of the parallels between them, see van der Linden (2011a).

Several attempts at modeling RTs have been made. The main differences between them exist in how they treat the relationship between the responses and RTs on the items. One category consists of attempts to incorporate the observed RTs as an explanatory variable in the model for the responses. For instance, Roskam's (1987, 1997) model that was

$$\Pr\{U_{pi} = 1 \mid \theta_p\} = \frac{\exp[(\theta_p + \ln t_{pi} - b_i)]}{1 + \exp[(\theta_p + \ln t_{pi} - b_i)]}, \tag{16.1}$$

basically an extension of the well-known Rasch model (Volume One, Chapter 3), was motivated by the success of psychological research on the SAT. The presence of $\theta_j + \ln t_{ij}$ (instead of just θ_j) relative to the difficulty parameter b_i in its parameter structure represents the assumption that more time spent by the test taker on the item should have the same effect on his probability of success as a higher ability in the regular Rasch model. Other models in this category (e.g., Verhelst et al., 1997; Wang and Hanson, 2005) used RT parameters instead of the observed RTs but were motivated similarly.

The models in the second category are the other way around; they incorporate response parameters in a model for the distribution of the RTs. A well-known example is Thissen's (1983) model, which is the product of a normal model for the logtimes

$$\ln T_{pi} \sim N(\mu + \tau_p + \beta_i - \rho(a_i\theta_p - b_i), \sigma^2) \tag{16.2}$$

with a regular 2PL or 3PL model for the responses. Parameters τ_p and β_i in Equation 16.2 are slowness parameters for test taker p and item i, respectively, $a_i\theta_p - b_i$ is the parameter

structure of the 2PL or 3PL model assumed to hold for the responses, and ρ serves as a slope parameter in the regression of this structure on the expected RT. The minus sign between $\mu + \tau_p + \beta_i$ and $\rho(a_i\theta_p - b_i)$ represents a trade-off between the probability of success on the items and the slowness of the response, with ρ controlling the strength of the trade-off. Other models in this category were designed to mimic the same effect (e.g., Gaviria, 2005). Ferrando and Lorenzo-Seva (2007) proposed a variation that was claimed to be more appropriate for personality tests.

The third category consists of completely distinct models for RTs. One of the first was Rasch's (1960) model for reading speed, which postulates a density for the time required to read an item of *m* words equal to

$$p(t_{pi} \mid \xi_p, \delta_i) \equiv \frac{(\xi_p / \delta_i)^m}{\Gamma(m)} t^{m-1} e^{-\xi_p t/\delta_i}, \tag{16.3}$$

where ξ_p is the ability of the test taker, δ_i is the difficulty of the text, and $\Gamma(m) \equiv (m - 1)!$ is the gamma function. The density is a standard gamma density with its intensity parameter modeled as $\lambda_{pi} \equiv \xi_p/\delta_i$; that is, the number of words expected to be read in a given time unit is assumed to be a function of the test taker's speed and the difficulty of the text only. More details on this model and its counterpart in the form of a Poisson model for reading errors can be found in Jansen (Volume One, Chapter 15). Similar models in this category are the gamma models by Maris (1993), an exponential model with an additive version of the basic parameterization in Equation 16.3 by Oosterloo (1975) and Scheiblechner (1979), and a Weibull model by Tatsuoka and Tatsuoka (1980). The lognormal model in this chapter also belongs to this category.

A category of models not reviewed here are the reaction-time models used in mathematical psychology. These models have been developed mainly to explain the stochastic processes leading to reaction times observed in psychological experiments. As nearly all of them assume a single standardized task replicated by groups of subjects considered as exchangeable, they miss a fundamental feature shared by all item response theory (IRT) models—separation of person and item effects. These reaction-time models are therefore less relevant to educational and psychological testing, which needs test taker parameters to adjust for their effects when calibrating the items and, conversely, item parameters to adjust the scores of test takers. An exception is the version of the diffusion model with item and person parameters developed by Tuerlinckx, Molenaar, and van der Maas (Volume One, Chapter 17).

16.2 Presentation of the Model

16.2.1 Assumptions

One of the most careless practices in the educational and psychological literature is the equating of the time on an item or task with the test taker's speed. For instance, nearly all the literature on the SAT (e.g., Luce, 1986, Chapter 6) measures speed as the average time recorded for its experimental subjects (as well as accuracy simply as the proportion of correct responses). However, time and speed are definitely distinct variables. If they were not, it would be unnecessary for our cars to have a speedometer in addition to a clock.

The first assumption underlying the RT model in this chapter is just a definition of speed. We follow the general practice of defining speed as the rate of change of a substantive measure with respect to time. A prime example of this format is the definition of the average speed of a moving object in physics:

$$\text{Average speed} = \frac{\text{Distance traveled}}{\text{Time elapsed}}. \tag{16.4}$$

Two other examples are speed of inflation as the amount of inflation over time in economics and speed of the spread of a bacterial infection measured as the increase in the number of patients infected over time in epidemiology. In this context, it seems natural to define the speed by a test taker on an item as the amount of cognitive labor performed on it over time (van der Linden, 2009a).

Generally, cognitive labor cannot be measured directly. Even for the two simple addition items in Figure 16.1, although their length clearly is a factor with an impact on the amount of labor they require, it would be wrong to equate the two; other factors do have an impact as well, for instance, the combinations of the numbers that have to be added at each step. However, although cognitive labor cannot be observed directly, we do observe its effect on the RT and have the option to introduce a latent item parameter for it in our model. We will refer to this parameter as a time-intensity parameter, just to remind ourselves that it represents the amount of labor required by the items indirectly, through its effect on the observed time spent on them.

The second assumption is that of constancy of speed during testing. This assumption has both a between- and within-item aspect. The assumption of constant speed between items not only has the technical advantage of avoiding overparameterization (i.e., the introduction of a separate speed parameter for each item) but is also consistent with the assumption of constant ability that underlies all the mainstream IRT models used in educational and psychological testing. The history of a satisfactory fit of these models would have been impossible when ability did vary substantially during testing. Given the strong evidence on the existence of SATs in performance tasks collected in experimental psychology, it seems inconsistent to entertain the possibility of constant ability but varying speed during testing.

Strictly speaking, an assumption of constant speed within items is not required. The only thing typically observed is the total time spent on each of them, and whether we should consider the speed parameter to represent the test taker's average or instantaneous speed on each item is beyond possible verification.

Of course, test takers will always vary somewhat in their speed during real-world testing. But rather than complicating the RT model by trying to allow for such changes, it is more practical to assume constant speed and use the actual RTs to check on the seriousness of the violations of the assumption. An example of this type of residual analysis is offered later in this chapter. As for possible larger systematic trends in speed, for instance, in the form of an increase in speed toward the end of the test due to a tight time limit, they typically have nothing to do with the intended ability measured by the test and should be avoided when designing the test. In fact, the RT model introduced in this chapter can be used to assemble test forms or set time limits on them to guarantee an acceptable level of speededness; the methodology required to do so is reviewed in van der Linden (Volume Three, Chapter 13).

The third assumption is conditional independence of RTs given the test taker's speed. The assumption is entirely analogous to that of conditional (or "local") independence

adopted throughout IRT. It may seem to be at odds with the (marginal) correlations found between the RTs on items in the earlier descriptive studies. But, as already argued, these studies necessarily aggregate data across test takers and/or items and tend to create spurious results. To model their results, we need to take the step from conditional, single-level modeling of fixed effects in this chapter to the hierarchical modeling in van der Linden and Fox (Volume One, Chapter 29).

16.2.2 Formal Model

The final assumption is that of a family of densities for the RT distributions of the test takers on the items. Let τ_p^* denote the speed of test takers $p = 1, \ldots, P$ on items $i = 1, \ldots, I$, which are assumed to have parameters β_i^* for their time intensities. Entirely analogous to Equation 16.4, we define speed as

$$\tau_p^* \equiv \frac{\beta_i^*}{t_{pi}}, \tag{16.5}$$

where t_{pi} is the observed RT by test taker p on item i. Equivalently,

$$t_{pi} = \frac{\tau_p^*}{\beta_i^*}, \tag{16.6}$$

which more explicitly demonstrates the assumed separation of the observed RT into an unknown test taker and item effect.

As RT distributions tend to be positively skewed, it is customary to work with logtimes, which leads to

$$\ln t_{pi} = \tau_p - \beta_i. \tag{16.7}$$

Acknowledging the random nature of RTs, the addition of a normally distributed random term gives us

$$\ln T_{pi} = \tau_p - \beta_i + \epsilon_i, \epsilon_i \sim N(0, \alpha_i^{-2}). \tag{16.8}$$

The result is a normal distribution of $\ln T_{pi}$ with mean $\beta_i - \tau_p$ and variance α_i^{-2}. We refer to the a_i parameter as the item discrimination parameter in the model.

The same distribution can be presented by the lognormal density

$$f(t_{pi}; \tau_p, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{pi}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{pi} - (\beta_i - \tau_p))]^2\right\} \tag{16.9}$$

with $\tau_p \in \mathbb{R}$, $\beta_i \in \mathbb{R}$, and $\alpha_i > 0$.

A third representation is

$$T_{pi} = \exp\left(\beta_i - \tau_p + \frac{Z}{\alpha_i}\right), \; Z \sim N(0,1), \tag{16.10}$$

which reveals that Equation 16.9 actually is the density of an exponential rather than the log of a normally distributed variable. The use of "lognormal" for the name of this distribution is thus a misnomer. But since it has been accepted without much practical confusion, we follow the tradition. As demonstrated throughout the rest of this chapter, transformations of random variables with an assumed normal distribution have obvious advantages. Other choices than the log of RTs are possible, though. If the interest is in flexible curve fitting and less in the applications reviewed later in this chapter, the class of Box–Cox transformations could be considered (Klein Entink et al., 2009).

16.2.3 Parameter Interpretation

We need to check if the behavior of each of the model parameters is consistent with the interpretation for it claimed so far. As a starting point, observe that Equations 16.8 through 16.10 do not have the standard parameterization of a lognormal distribution (Volume Two, Chapter 3, Equation 16.30), which would have implied the choice of

$$f(\ln t_{pi}; \mu_{pi}, \sigma_{pi}^2) \equiv \frac{1}{\sigma_{pi}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln t_{pi} - \mu_{pi}}{\sigma_{pi}}\right)^2\right\} \tag{16.11}$$

with μ_{pi} and σ_{pi}^2 parameters representing the mean and variance of the logtime by the test taker on the item. The current model differs because of our assumption of $\mu_{pi} \equiv \beta_i - \tau_p$ as a function of a separate item and test taker effects and the choice of $\sigma_{pi}^2 \equiv \alpha_i^{-2}$ as a function of only the items.

The interpretation of τ_p as a speed parameter was primarily motivated by the agreement between its definition in Equation 16.5 and the notion of speed generally adopted throughout science. But it is also consistent with the behavior of the model. For test takers with a higher value of τ_p, the model returns a lower-expected RT on each item. Likewise, a higher value of β_i results in a higher-expected RT for every test taker. Both are exactly what we would expect from parameters introduced to represent the opposite effects on RTs by the test taker's speed and the laboriousness of the items.

The interpretation of α_i as a discrimination parameter can be motivated by its contribution to the precision of the estimate of the test taker's speed produced by the test. The key quantity is the standard error of estimation for the speed parameter. Consider a test from an item bank calibrated with enough data to treat all item parameters as known. The Fisher information in the logtimes on any τ parameter for an item with parameters α_i and β_i is

$$\begin{aligned}
\mathbf{I}_i(\tau) &\equiv \mathcal{E}\left(\frac{\partial}{\partial \tau_p} \ln f(\ln T; \tau, \alpha_i, \beta_i)\right)^2 \\
&= \mathcal{E}\left(-a_i^2(\ln T - (\beta_i - \tau)\right)^2 \\
&= \alpha_i^4 \mathcal{E}(\ln T - (\beta_i - \tau))^2 \\
&= \alpha_i^2.
\end{aligned} \tag{16.12}$$

Hence, because of conditional independence, for an n-item test, the information is $\sum_{i=1}^{n} \alpha_i^2$, and the maximum-likelihood estimation (MLE) of speed parameter τ has an asymptotic standard error

$$\mathrm{SE}(\hat{\tau} \mid \alpha_1, \ldots, \alpha_n) = \left(\sum_{i=1}^{n} \alpha_i^2 \right)^{-1/2}. \tag{16.13}$$

The discrimination parameter is thus the only parameter in the model that controls how much each item contributes to the accuracy of the estimation of τ. Remarkably, the true value of the τ parameter does not play any role. Neither do the time-intensity parameters of the items. It thus holds that no matter the speed at which a test taker operates, it can be estimated equally well with more or less laborious items—a conclusion with a potentially large practical value.

The reason we have opted for α_i instead of its reciprocal in Equation 16.11 is to obtain three model parameters with a similar interpretation as the θ_p, a_i, and b_i parameters in the 2PL and 3PL response models with their well-established history in educational and psychological testing. The θ_p and b_i parameters in these two models have an effect on the success probability of the item compared to that of τ_p and β_i on the RT. As for the discrimination parameters, entirely analogous to Equation 16.12, the 2PL and 3PL models have a_i^2 as the main factor in their expressions for the item and test information functions (Volume Three, Chapter 9, Equations 16.2 and 16.3).

The similarity of interpretation does *not* extend to the formal structure of these response models, though. The lognormal model specifies a density directly for the distribution of the RTs T_{pi}, whereas the two response models are for the success parameters in the Bernoulli distributions of response variables U_{pi}. Consequently, the former has t_{pi} in its argument but the latter do not contain u_{pi}. Nevertheless, to enhance the similarities between the two types of models, it has been suggested to include an additional discrimination parameter in the lognormal model, adopting

$$\alpha_i(\ln t_{pi} - \phi_i(\beta_i - \tau_p)) \tag{16.14}$$

as the core structure of it (e.g., Fox, 2010, Chapter 8). Although the product of $\phi_i(\beta_i - \tau_p)$ does remind us of the parameter structure of a 2PL or 3PL model, the addition of ϕ_i is unnecessary. In fact, the adoption of two parameters in one model with an entirely similar impact on the distribution of the RT seems an obvious case of overparameterization, which is likely to result in poor identifiability of both.

When test takers guess on an item, their behavior does not necessarily imply anything for their RTs—the guess may have been immediate or the final result of a long uphill battle. Therefore, the RT model does not require any "guessing parameter" to constrain the RTs from below (beyond their natural lower bound at zero).

16.2.4 Parameter Identifiability and Linking

Models with latent parameters generally lack identifiability and it may be difficult to find restrictions that make them identifiable (Volume Two, Chapter 8). However, the presence of the manifest variable $\ln t_{pi}$ in the current RT model makes the job much easier. As a

starting point, observe that if we had chosen Equation 16.11 with parameters μ_{pi} and σ^2_{pi} as our density of $\ln T_{pi}$, the model would have been fully identified (provided we have enough test takers and items). Also, the substitution of $\sigma_{pi} = \alpha_i^{-1}$ for all p, which for $\alpha_i > 0$ amounts to a monotone transformation, would not have led to any loss of identifiability. The possible lack of identifiability of Equation 16.9 can thus only be a consequence of its further parameterization of μ_{pi} as $\beta_i - \tau_p$. Indeed, for any value of ε, the transformations $\beta_i - \varepsilon$ and $\tau_p - \varepsilon$ yield identical RT distributions.

A convenient restriction to establish full identifiability of the model is setting the τ_p parameters for the test takers equal to zero; that is

$$\mu_\tau \equiv 0. \tag{16.15}$$

The choice implies speed parameters with values that are deviations from their average. As for the time-intensity parameters, averaging μ_{pi} across all test takers and items gives an overall mean equal to $\mu \equiv \mu_\beta - \mu_\tau$. Hence, the choice of Equation 16.15 also implies that

$$\mu_\beta = \mu, \tag{16.16}$$

which implies that the individual time-intensity parameters can now be viewed as deviations from the expected logtime across all test takers and items.

As an alternative to Equation 16.15, we could set the sum of the β_i parameters equal to zero. This choice results in time intensity and speed parameters that are deviations from their average and the expected logtime for all test takers and items, respectively.

The necessity to introduce an additional identifiability restriction creates a linking problem for parameters estimated from different samples. For samples that include different test takers and/or items, Equation 16.15 implies true parameter values that are deviations from different overall means. Linking functions that adjust for these differences are

$$\tau_{p_2} = \tau_{p_1} + \nu \tag{16.17}$$

and

$$\beta_{i_2} = \beta_{i_1} + \nu, \tag{16.18}$$

where the extra subscripts have been added to indicate two separate calibration studies.

For a common-item design, the unknown linking parameter ν is equal to

$$\nu = \mu_{\beta_1} - \mu_{\beta_2} \tag{16.19}$$

with the averages taken across the common items in the two calibrations. Similarly, for a common-person design linking, parameter ν is equal to the difference between its average speed parameters for the two test forms.

Obviously, in practice, ν has to be calculated from an estimated item or person parameters. Standard errors of linking due to parameter estimation error for a variety of linking designs were presented in van der Linden (2010). In fact, since this reference, it has become clear that a statistically more accurate form of linking is possible using precision-weighted

averaging rather than plain averages as in Equation 16.19. For a more general treatment of the linking problem in IRT models, including the use of this improved-type estimator and its optimization through the application of optimal design principles, see van der Linden and Barrett (Volume Three, Chapter 2).

16.2.5 Moments of RT Distributions

For the standard family of lognormal densities, with parameters μ and σ^2 for the mean and variance of the natural log of the variate (Volume Two, Chapter 3), the kth moment about the origin is equal to

$$m_k \equiv \mathcal{E}(X^k) = \exp(k\mu + k^2\sigma^2/2) \tag{16.20}$$

(Kotz and Johnson, 1985, 134–136; note that one of the squares is accidentally missing in this reference). From Equation 16.20, the first two moments for the current RT model can be derived as

$$\mathcal{E}(T_{pi}) = \exp(-\tau_p)\exp(\beta_i + \alpha_i^{-2}/2), \tag{16.21}$$

$$\mathcal{E}(T_{pi}^2) = \exp(-2\tau_p)\exp(2\beta_i + 2\alpha_i^{-2}) \tag{16.22}$$

(van der Linden, 2011b,c). Consequently, its second cumulant can be written as

$$\mathcal{E}[T_{pi} - \mathcal{E}(T_{pi})]^2 = \exp(-2\tau_p)\exp(2\beta_i + \alpha_i^{-2})[\exp(\alpha_i^{-2}) - 1]. \tag{16.23}$$

Observe how these expressions factor into separate components for the test takers and the items, a feature that suggests the following reparameterization:

$$q_i \equiv \exp(\beta_i + \alpha_i^{-2}/2), \tag{16.24}$$

$$r_i \equiv \exp(2\beta_i + \alpha_i^{-2})[\exp(\alpha_i^{-2}) - 1] \tag{16.25}$$

without any change of the speed parameters. Two remarkably simple expressions are resulting for the mean and variance of the RT of a test taker on an item:

$$\text{mean}(T_{pi}) = \exp(-\tau_p)q_i, \tag{16.26}$$

$$\text{var}(T_{pi}) = \exp(-2\tau_p)r_i. \tag{16.27}$$

Because of conditional independence, the distribution of the total time of a test taker on a test of n items has mean

$$\text{mean}(T_p) = \exp(-\tau_p)\sum_{i=1}^{n} q_i \tag{16.28}$$

and variance

$$\text{var}(T_p) = \exp(-2\tau)\sum_{i=1}^{n} r_i. \tag{16.29}$$

Although the exact shape of the total-time distribution of a test taker is not known to be lognormal, it can be approximated surprisingly accurately by a standard lognormal with μ and σ^2 matching the mean and variance in Equations 16.28 and 16.29, though. Further treatment of the approximation is given in van der Linden (Volume Two, Chapter 6).

16.2.6 Relationships with Other Models

As pointed out by Finger and Chee (2009), the lognormal model in Equation 16.8 can be reparameterized as a one-factor model with item-dependent intercepts

$$\ln t_{pi} \equiv \upsilon_i + \lambda_i \xi_p + \varepsilon_i. \tag{16.30}$$

That is, under its standard assumptions of $\varepsilon_i \sim N(0, \psi_i^2)$, $\mathcal{E}(\xi_p) \equiv 0$, and $\mathcal{E}(\xi_p \varepsilon_i) \equiv 0$, substitution of $\psi_i \equiv \alpha_i^{-1}$, and setting $\lambda_i \equiv -1$ for all i, the time-intensity parameter β_i corresponds to intercept υ_i and speed parameters τ_p with the factor score ξ_p in Equation 16.30. A more general treatment of the correspondence between IRT models for continuous response variables and factor analysis models is offered by Mellenbergh (Volume One, Chapter 10).

For the analysis of possible dependencies between the RTs of pairs of test takers without any confounding due to effects of the latent covariates associated with the items or test takers, a bivariate version of the model is available. Let p and q denote two different test takers with RTs that fit the lognormal model. Their joint RTs (T_{pi}, T_{qi}) are then distributed with density

$$\begin{aligned} &f(t_{pi}, t_{qi}; \tau_p, \tau_q, \alpha_i, \beta_i, \rho_{pq}) \\ &\equiv \frac{\alpha_i^2}{t_{pi} t_{qi} 2\pi\sqrt{1-\rho_{pq}^2}} \exp\left\{\frac{-1}{2(1-\rho_{pq}^2)}(\psi_{pi}^2 - 2\rho_{pq}\psi_{pi}\psi_{qi} + \psi_{qi}^2)\right\}, \end{aligned} \tag{16.31}$$

where

$$\psi_{pi} \equiv \alpha_i[\ln t_{pi} - (\beta_i - \tau_p)] \tag{16.32}$$

is the residual RT after adjustment for the item and person effects and ρ_{pq} is the correlation between ψ_{pi} and ψ_{qi} (van der Linden, 2009b).

This bivariate version of the model can be used, for instance, for the detection of collusion between test takers during testing. Because all regular item and person effects have been removed, no matter the properties of the items, independently working test takers should have $\rho_{pq} = 0$. Deviations from this null value could thus point at a potentially suspicious behavior (Volume Three, Chapter 13).

16.3 Parameter Estimation

Our primary estimation method is Bayesian with Gibbs sampling of the joint posterior distribution of all model parameters (Fox, 2010, Chapter 4; Junker, Patz, and Vanhoudnos, Volume Two, Chapter 15). It is convenient to use the version of the model in Equation 16.8; that is, estimate the parameters from the logtimes. The sampler proposed in van der Linden (2006) alternates conveniently between slight modifications of the two standard cases of (i) normal data with known variances and a conjugate normal prior for the unknown speed parameters and (ii) normal data with a conjugate normal-gamma gamma prior for the unknown time-intensity and discrimination parameters (e.g., Gelman et al., 2014, Sections 2.5 and 3.3). The two modifications are: First, because of our choice of $\alpha_i \equiv \sigma^{-1}$, the inverse-$\chi^2$ in the second case has to be replaced by a gamma distribution. Second, when sampling the item parameters, the data need to be adjusted for the speed parameters, and vice versa.

16.3.1 Estimating Both Item and Person Parameters

The common prior distribution for the speed parameters τ_p is

$$\tau_p \sim N(\mu_\tau, \sigma_\tau^2), \tag{16.33}$$

while the one for the item parameters (α_i, β_i) is specified to be a normal gamma:

$$\beta_i \,|\, \alpha_i \sim N(\mu_\beta, (\alpha_i^2 \kappa)^{-1}); \tag{16.34}$$

$$\alpha_i^2 \sim G(\nu/2, \nu/(2\lambda)). \tag{16.35}$$

Thus, μ_τ is our prior guess of the value of the speed parameters with σ_τ^2 expressing the strength of our evidence for it. Likewise, μ_β is our guess of the value of β_i given α_i, now with $\alpha_i^2\kappa$ representing the strength of our evidence, while λ represents our prior guess of α_i^2, this time with ν expressing the strength of our evidence. The identifiability restriction in Equation 16.15 allows us to set the prior means equal to $\mu_\tau = 0$ and $\mu_\beta = \overline{\ln t}$, where $\overline{\ln t}$ is the average observed logtime across all test takers and items in the sample. Our choices for σ_τ^2 and κ should reflect the expected variability of the speed and time-intensity parameters. Similarly, it makes sense to set λ relative to the observed variance of the logtimes.

The Gibbs sampler alternates between the blocks of speed parameters $\tau_1, \ldots, \tau_P$, and item parameters $(\alpha_1, \beta_1), \ldots, (\alpha_I, \beta_I)$. At step k, the conditional posterior distributions that need to be sampled for these blocks have the following two forms:

1. When sampling each $\tau_p^{(k)}$, the item parameters have fixed values of $\alpha_i^{(k-1)}$ and $\beta_i^{(k-1)}$. Using the latter to redefine the data on each test taker as $\beta_i^{(k-1)} - \ln t_{pi}$, $i = 1, \ldots, I$, the posterior distribution of τ_p is the one for normal data with known variance $(\alpha_i^{(k-1)})^{-2}$ but an unknown mean τ_p with a conjugate normal prior. Consequently, assuming a prior mean $\mu_\tau \equiv 0$, $\tau_j^{(k)}$ has to be drawn from a normal distribution with mean

$$\frac{\sum_{i=1}^{I} (\alpha_i^{(k-1)})^2 (\beta_i^{(k-1)} - \ln t_{pi})}{\sigma_\tau^{-2} + \sum_{i=1}^{n} (\alpha_i^{(k-1)})^2} \tag{16.36}$$

and variance

$$\left(\sigma_\tau^{-2} + \sum_{i=1}^{I} (\alpha_i^{(k-1)})^2\right)^{-1}. \tag{16.37}$$

2. When sampling each $(\alpha_i^{(k)}, \beta_i^{(k)})$, the speed parameters have fixed values of $\tau_j^{(k)}$. Using them to redefine the data on each item as $\ln t_{pi} + \tau_p^{(k)}$, $p = 1, \ldots, P$, the posterior distribution of (α_i, β_i) is now the one for normal data with an unknown mean β_i and unknown variance α_i^{-2}. Given their conjugate prior distribution, the values for these parameters have to be drawn from a normal-gamma posterior. That is, assuming a prior mean $\mu_\beta \equiv \overline{\ln t}$, $\alpha_i^{(k)}$ is drawn from a gamma distribution $G(\phi/2, \omega/2)$ with parameters

$$\phi = \nu + P; \tag{16.38}$$

$$\omega = \nu\lambda^{-1} + \sum_{p=1}^{P} (\ln t_{pi} - \overline{\ln t}_i + \tau_p^{(k)})^2 + \frac{\kappa P\left[\overline{\ln t}_i - \overline{\ln t}\right]^2}{\kappa + P}, \tag{16.39}$$

while $\beta_i^{(k)}$ is drawn from a normal distribution with mean

$$\frac{\kappa\overline{\ln t} + \sum_{p=1}^{P} (\ln t_{pi} + \tau_p^{(k)})}{\kappa + P} \tag{16.40}$$

and variance

$$(\kappa + P)^{-1}. \tag{16.41}$$

As a frequentist alternative to this Bayesian method, the item and person parameters can be estimated by confirmatory factor analysis of the logtimes with an appropriate back transformation of the discrimination parameters (Section 16.2.6), for instance, using *MPlus* (Volume Three, Chapter 28). A comparative study with empirical RTs by Finger and Chee (2009) showed correlations between the factor analysis and Gibbs sampler estimates uniformly greater than 0.99 for all parameters.

16.3.2 Estimating the Speed Parameters

When speed parameters τ_p have to be reported for examinees on a test with items already calibrated under the RT model, an attractive option is to sample their posterior distributions using an adjusted version of the above Gibbs sampler. The average of the sampled values for each test taker is their expected a posterior (EAP) estimate, which is the standard deviation of their posterior standard error.

The adjusted sampler has the following two alternating steps:

1. Drawing the τ_p parameters from the normal posterior distribution with the mean and variance in Equations 16.36 and 16.37.
2. Resampling of vectors of draws from the stationary posterior distributions of the α_i and β_i parameters for the items in the test saved during their calibration.

The estimation is extremely fast because only the draws for the τ_p parameters need to converge; the draws for the item parameters are already from extremely narrow posterior distributions centered at their true values. The vectors with the saved draws for the item parameter do not need to contain more than 1000 or so well-spaced draws by the stationary Gibbs sampler during their calibration. For a proof of the convergence of this type of Gibbs sampler, see van der Linden and Ren (2015).

Again, as a frequentist alternative to this Bayesian method, if the items have been calibrated with enough precision to treat the remaining uncertainty as negligible, the speed parameters can be estimated using the following MLE with point estimates substituted for the item parameters:

$$\hat{\tau}_p = \frac{\sum_{i=1}^{n} \hat{\alpha}_i^2 (\hat{\beta}_i - \ln t_{pi})}{\sum_{i=1}^{n} \hat{\alpha}_i^2}. \tag{16.42}$$

As the expression reveals, the speed of an individual test taker is then estimated as the precision-weighted average of $\hat{\beta}_i - \ln t_{pi}$ (i.e., the logtime on each item adjusted for its estimated time intensity). Returning to our earlier discussion of the interpretation of the α_i parameters as a discrimination parameter, observe how they serve as precision weights in this estimator. The same could already have been observed for the similar role played by these parameters in the expressions for the posterior mean of τ_p in Equations 16.36 and 16.37.

16.4 Model Fit

It is convenient to combine these two Gibbs samplers with posterior predictive checks of the fit of the model to the items and test takers. As we have one observation per combination of test taker and item, an obvious choice is marginal checks with aggregation of the results across test takers or items to evaluate their fit more specifically.

Let $\widetilde{\ln t}_{pi}$ denote the predicted logtime for test taker p on item i. The (lower-tail) posterior predictive p-values of the observed logtimes are defined as

$$\begin{aligned} \pi_{pi} &\equiv \Pr\{\widetilde{\ln t}_{pi} < \ln t_{pi}\} \\ &= F_{\widetilde{\ln T}_{pi}}(\ln t_{pi}) \\ &= \int\int\int \Phi(\ln t_{pi}; \tau_p, \alpha_i, \beta_i) f(\tau_p, \alpha_i, \beta_i \mid \mathbf{t}) d\tau_p \, d\alpha_i \, d\beta_i, \end{aligned} \tag{16.43}$$

where $\Phi(\ln t_{pi}; \tau_p, \alpha_i, \beta_i)$ is the cdf for the normal distribution of $\ln T_{pi}$ in Equation 16.8 and $f(\tau_p, \alpha_i, \beta_i | \mathbf{t})$ is the posterior density of its parameters given the observed data $\mathbf{t} \equiv (\mathbf{t}_{ij})$.

With increasing sample size and test length, the posterior density in Equation 16.43 converges to that of a degenerate distribution at the true parameter values, and consequently $F_{\widehat{\ln T_{pi}}}(\ln t_{pi}) \to \Phi(\ln t_{pi}; \tau_p, \alpha_i, \beta_i)$. The probability integral transform theorem (e.g., Casella and Berger, 2002, Section 2.1) tells us that $F_{\widehat{\ln T_{pi}}}(X)$ has the $U(0,1)$ distribution. Combining the two observations, it follows that each π_{pi}-value has the same asymptotic uniform distribution on [0,1]. For posterior distributions with noticeable variance, the predictive distribution of $\widehat{\ln t_{pi}}$ is still wider than that of $\ln T_{pi}$. Depending on the true value of π_{pi}, the observed distribution will then tend to be concentrated more toward a point somewhere in the middle of its scale.

Upon stationarity of the Gibbs sampler, for $k = 1, \ldots, K$ additional draws from the posterior distribution, each of these π_{pi}-values can be approximated as

$$\pi_{pi} \approx K^{-1} \sum_{k=1}^{K} \Phi(\ln t_{pi}; \tau_p^{(k)}, \alpha_i^{(k)}, \beta_i^{(k)}). \tag{16.44}$$

The proposed tool for checking on the fit of the model are plots of the cumulative distributions of these Bayesian π_{pi}-values across items or test takers. The use of cumulative distributions allows us to visually establish the lack of uniformity as deviations from the identity line. However, as just discussed, due to less than full convergence of the posterior distributions, the curves may be somewhat lower than the identity line at the lower end of the scale with the corresponding compensation at the upper end.

Lagrange multiplier (LM) tests (or score tests) of the assumption of conditional independence of RTs have been presented for the two cases of marginal MLE of all model parameters (Glas and van der Linden, 2010) and MLE of the speed parameters for known item parameters (van der Linden and Glas, 2010). Both statistical tests are based on the assumption of the bivariate lognormal distribution in Equation 16.31, this time posited for the case of the RTs of a single test taker on pairs of items i and i'. More specifically, they evaluate $H_0: \rho_{ii'} = 0$ against $H_1: \rho_{ii'} \neq 0$. LM tests are known to be uniformly most powerful.

When the items are taken from a bank gone through a process of item calibration careful enough to allow treatment of their parameters as known, the use of the second LM test has the advantage of a test statistic in a closed form

$$LM(\rho_{ii'}) = \frac{\left(\sum_{p=1}^{P} \hat{\psi}_{pi} \hat{\psi}_{pi'} \right)^2}{\sum_{p=1}^{P} \left[\hat{\psi}_{pi}^2 + \hat{\psi}_{pi'}^2 - 1 - \left((\alpha_i \hat{\psi}_{pi} + \alpha_{i'} \hat{\psi}_{pi'})^2 \Big/ \sum_{i=1}^{n} \alpha_i^2 \right) \right]}, \tag{16.45}$$

where

$$\hat{\psi}_{pi} = \alpha_i [\ln t_{pi} - (\beta_i - \hat{\tau}_p)] \tag{16.46}$$

is the estimate of the standardized residual RT for test taker p on item i in Equation 16.32, $\hat{\psi}_{pi'}$ is defined analogously for item i', and $\hat{\tau}_p$ is the MLE in Equation 16.42. Under H_0, the statistic has a χ^2 distribution with one degree of freedom.

Observe that, since the model posits a normal distribution of each ln T_{pi}, their standardized versions will have $N(0,1)$ as an asymptotic distribution. The fact has been used by van der Linden et al. (2007) to check observed distributions of residuals for changes in speed during test administration. The procedure is illustrated in the empirical example in the next section.

16.5 Empirical Example

The data set is from a computerized high-stakes test existing of the RTs of 396 test takers on 65 multiple-choice items. The lognormal model was fitted using the Gibbs sampler in Equations 16.33 through 16.41 with a burn-in of 1500 iterations and 4500 additional iterations for the calculation of the EAP estimates of the parameters. These numbers have been shown to suffice in extensive explorations of the model, provided the following implementation choices are made: As an identifiability restriction, we used $\mu_\tau = 0$ in Equation 16.15, which was implemented simply by recentering the draws for all τ_p parameters after each iteration. All prior distributions were set to have parameter values according to our suggestions above. As variance of the common prior distribution of the τ_p parameters, $\sigma_\tau^2 = 100$ was chosen. Because of the identifiability restriction, we were able to use Equation 16.16 and set the prior mean of the β_i parameters equal to $\mu_\beta = \overline{\ln t} = 3.88$. In addition, we set $\kappa = 1$. As for the common prior distribution for the α_i parameters, $\lambda = \text{var}(\ln t) = 0.68$ was used in combination with the choice of $\upsilon = 1$. The choice of values for σ_τ^2, κ, and υ were all low informative. The Gibbs sampler was initialized using their prior mean as starting value for each of its parameters.

The estimates of the β_i parameters ranged from 2.75 to 5.32 with an average of 3.89. Observe that, for test takers operating at the average speed of $\tau = 0$, the endpoints of this range correspond to the expected RTs of exp(2.75) = 15.64 and exp(5.32) = 204.38 s, respectively. The estimates of the τ_p parameters ranged from −0.43 to 0.70 around their mean of zero with a standard deviation equal to 0.23. The variation was much smaller than for the time-intensity parameters of the items—a result typical of high-stakes tests with a well-chosen time limit. Finally, all estimates of the α_i parameters ranged from 1.33 to 2.55 with a mean of 1.96.

The fit of the model to the dataset was evaluated using the tools described earlier. Figure 16.2 shows the plot of the cumulative distribution of the Bayesian π-values in Equation 16.44 across all test takers and items in the set. The curve shows a minor deviation from the identity line at the lowest π-values due to the remaining posterior variance of the parameter estimates discussed earlier; otherwise, its shape is entirely according to expectation. It is important to note, however, that this result only represents a necessary condition for the overall model fit. It is still possible for the individual items or test takers to have opposite types of misfit canceling out at this higher level of aggregation.

Figure 16.3 shows the same cumulative curves for each of the 65 items. As each of them is based on some 2.5% of the data only, the results are less stable. However, some of the variation may also reflect a true misfit of some of the items. For the three items with the lowest curves at the lower end of the scale, there certainly was some misfit. For instance, each of them had an observed proportion of π-values of some 0.07 lower than expected at the nominal value of 0.30.

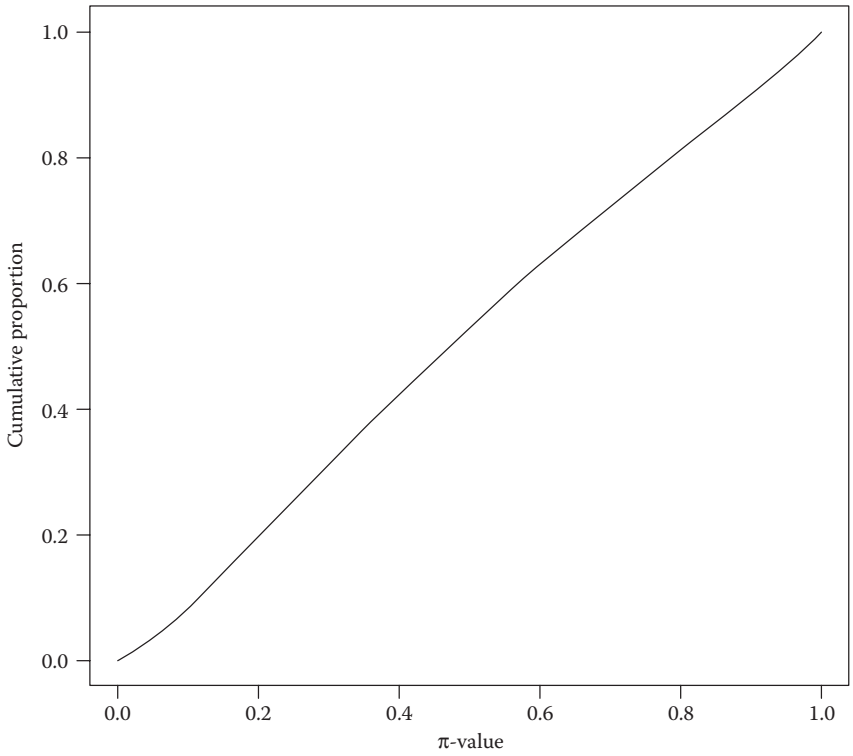

FIGURE 16.2
Cumulative distribution of Bayesian π-values across all 65 items and 396 test takers in the dataset.

Obviously, the same type of plot with the curves for the 396 test takers in Figure 16.4 shows even more variation as each of the curves is now based on the observed RTs on only 65 items. Nevertheless, the plot is still useful to identify test takers with aberrant behavior. For instance, the outlying curve at the higher end of the scale was for a test taker whose observed proportion of π-values below 0.75 was equal to 0.48, a result clearly below expectation. Inspection of his record revealed some six items RTs in the range from 0 to 10 s. Apparently, this test taker had inspected only these items quickly entering random responses to move to the next item.

As for the assumption of constancy of speed, it is possible to check on systematic violations by plotting the mean-standardized residual logtimes on the items in Equation 16.46 as a function of their position in the test. The results in Figure 16.5 reveal generally small mean residuals, varying about their expected value of zero in the range from −0.05 to 0.05. The slight tendency for the residuals to be more negative toward the end of the test may point at a corresponding increase in speed. However, as a mean-standardized residual logtime of 0.05 corresponds to just exp(0.05) = 1.1 s, the observed violations hardly have any practical value. The standard deviations of the residuals were close to their expected value of 1. Their values ranged from 0.94 to 1.02, with a mean equal to 0.99.

As a check on the assumption of local independence, we calculated the correlations between the residual RTs in Equation 16.32 on the pairs of adjacent items in the test beginning with the first item. The results are presented in Table 16.2. As expected, the correlations

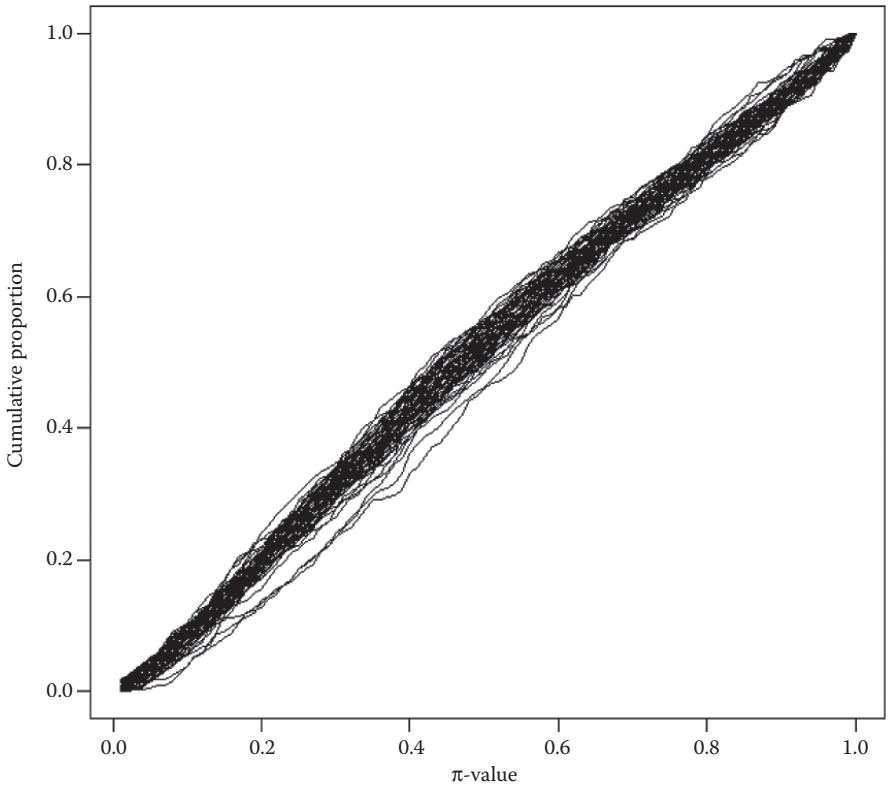

FIGURE 16.3
Cumulative distribution of Bayesian π-values for each of the 65 items across all 396 test takers in the dataset.

varied around zero and, with a few exceptions, were generally low. We also calculated the LM test in Equation 16.45, which returned all correlations with an absolute value greater than 0.10 as significant at $\alpha = 0.05$. Of course, these results were greatly influenced by the high power of the test for a sample size of $P = 369$ test takers. Violations of the assumption of local independence of this size do not necessarily have a practical meaning. Generally, they lead to standard errors for the speed parameters based on a somewhat lower-than-anticipated effective number of observations but do not imply any unnecessary bias or even inconsistency of the parameter estimates. The only violation with a noticeable impact on the standard error might be Item Pair (45,46). If its correlation of 0.21 is deemed too high, replacement of one of its items in the test would solve the problem.

16.6 Discussion

We began this chapter by observing that every response to a test item has both an aspect of correctness and time. During its first century, test theory exclusively focused on responses scored for their correctness. It had to do so because, except for an individually proctored

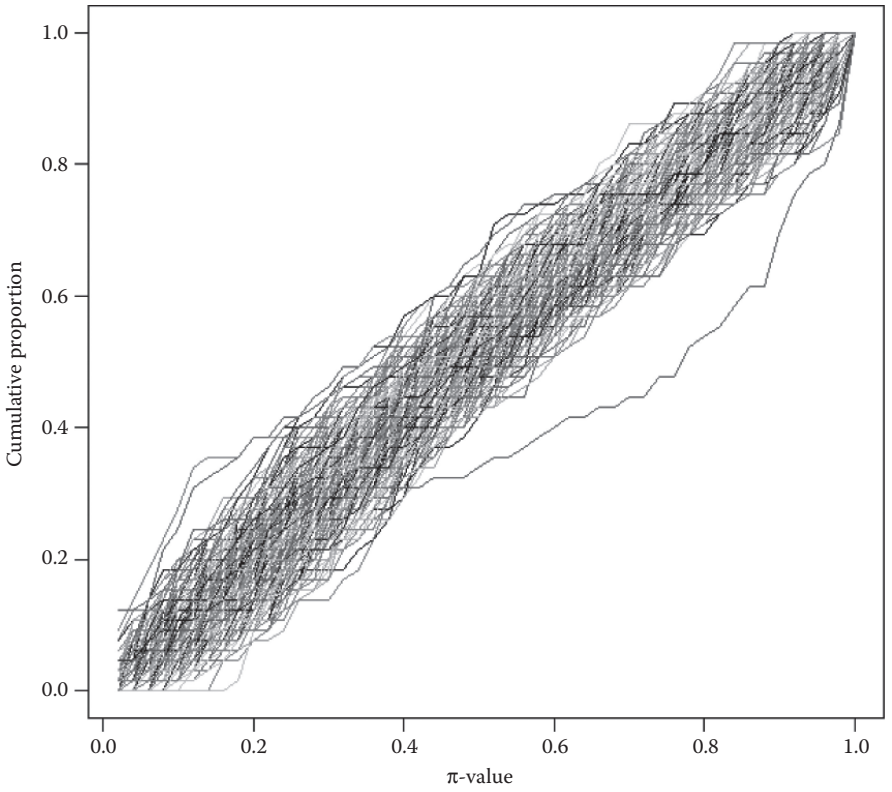

FIGURE 16.4
Cumulative distribution of Bayesian π-values for each of the 396 test takers across all 65 items in the test.

test, accurate recording of the RTs at the level of the individual test takers and items was practically infeasible. The arrival of computers and handheld electronic devices in test administration has removed this limitation. Consequently, test theorists will have to rethink their basic notions and applications. Examples already present in the literature focus on the notion of test speededness (e.g., Bolt et al., 2002; Goegebeur et al., 2008; Lu and Sireci, 2007; Schmitt et al., 2010; Volume Three, Chapter 12), the role of motivation in test performance (e.g., Wise and DeMars, 2009; Wise and Kong, 2005), and the nature of parameter drift (e.g., Li and Shen, 2010; Wollack et al., 2003). More in-depth psychological analyses of the role time limits and the time spent on test items can be found in Goldhammer and Kroehne (2014), Goldhammer et al. (2014), and Ranger and Kuhn (2015). Besides, it has become possible to improve such practices as test item calibration (van der Linden et al., 2010) and test accommodations (e.g., Stretch and Osborne, 2005). Adaptive testing has already profited from the use of RTs. Fan et al. (2012), in the spirit of Woodbury and Novick's (1968) pioneering work on the assembly of fixed forms, worked out the details of maximization of its information about a test takers' ability parameter for a given time interval while van der Linden (2008) showed how RTs can be used as collateral information to optimize item selection and ability estimation in adaptive testing. More complete reviews of all these changes and new promises of RT analyses can be found in Lee and Chen (2011) and van der Linden (2011a).

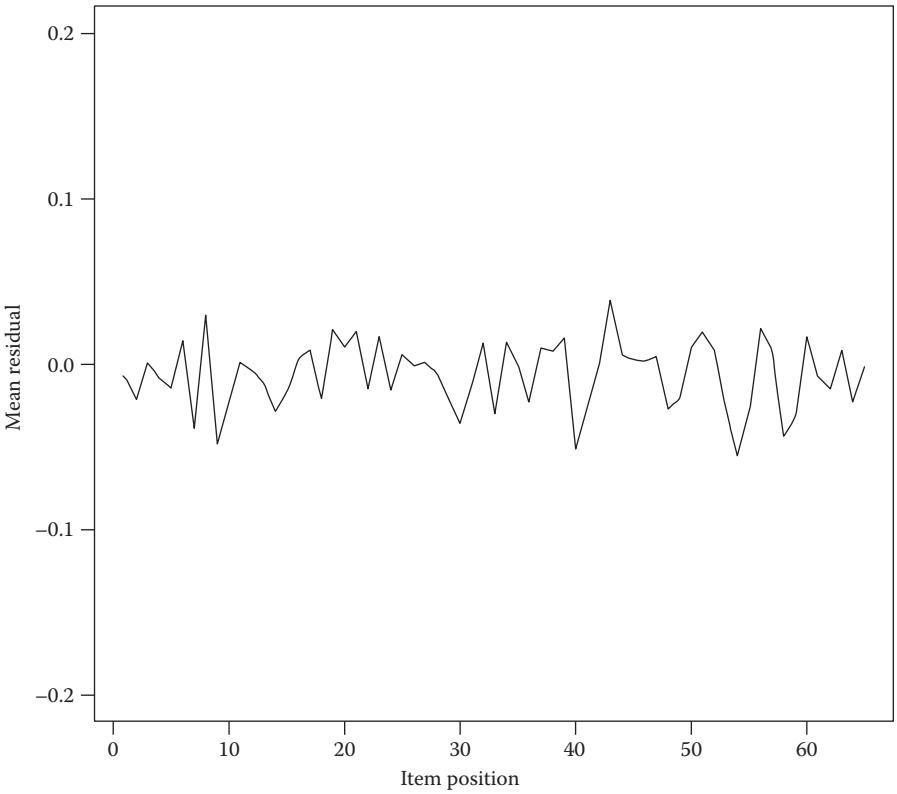

FIGURE 16.5
Mean residual logtimes on the 65 items as a function of their position in the test.

TABLE 16.2

Correlations between Residual RTs on the Items

Items	ρ	Items	ρ	Items	ρ	Items	ρ
1,2	0.00	17,18	–0.01	33,34	0.06	49,50	–0.07
3,4	0.03	19,20	0.06	35,36	–0.12	51,52	0.00
5,6	0.00	21,22	–0.02	37,38	–0.12	53,54	0.00
7,8	0.15	23,24	0.01	39,40	–0.05	55,56	–0.06
9,10	–0.11	25,26	–0.03	41,42	0.03	57,58	0.07
11,12	0.04	27,28	–0.02	43,44	–0.15	59,60	–0.07
13,14	0.01	29,30	0.09	45,56	0.21	61,62	–0.04
15,16	0.08	31,32	–0.02	47,48	0.08	63,64	–0.10

References

Bolt, D. M., Cohen, A. S., and Wollack, J. A. 2002. Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Casella, G. and Berger, R. L. 2002. *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Fan, Z., Wang, C., Chang, H.-H., and Douglas, J. 2012. Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37,* 655–670.

Ferrando, P. J. and Lorenzo-Seva, U. 2007. An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543.

Finger, M. S. and Chee, C. S. 2009. Response-time model estimation via confirmatory factor analysis. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, San Diego, CA, April.

Fox, J.-P. 2010. *Bayesian Item Response Modeling*. New York: Springer.

Gaviria, J.-L. 2005. Increase in precision when estimating parameters in computer assisted testing using response times. *Quality and Quantity, 39*, 45–69.

Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2014. *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Glas, C. A. W. and van der Linden, W. J. 2010. Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology, 63,* 603–626.

Goegebeur, Y., De Boeck, P., Wollack, J. A., and Cohen, A. S. 2008. A speeded response model with gradual process change. *Psychometrika*, *73*, 65–87.

Goldhammer, F. and Kroehne, U. 2014. Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response-time modeling. *Applied Psychological Measurement*, *38*, 255–267.

Goldhammer, F., Naumann, J., Stelter, A., Toth, K., Roelle, H., and Klieme, E. 2014. The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large scale assessment. *Journal of Educational Psychology*, *106*, 608–626.

Hornke, L. F. 2000. Item response times in computerized adaptive testing. *Psicológica*, *21*, 175–189.

Klein Entink, R. H., van der Linden, W. J., and Fox, J.-P. 2009. A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Kotz, S. and Johnson, N. L. 1985. *Encyclopedia of Statistical Science* (Volume 5). New York: Wiley.

Lee, Y.-H. and Chen, H. 2011. A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Li, F. and Shen, L. 2010. Detecting item parameter drift by item response and item response time in a computer-based exam. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, Denver, CO, April 29–May 3.

Lu, Y. and Sireci, S. G. 2007. Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37.

Luce, R. D. 1986. *Response Times: Their Roles in Inferring Elementary Mental Organization*. Oxford, UK: Oxford University Press.

Maris, E. 1993. Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

Oosterloo, S. J. 1975. *Modellen voor reactie-tijden* [Models for reaction times]. Unpublished master's thesis, Faculty of Psychology, University of Groningen, The Netherlands.

Ranger, J. and Kuhn, J.-T. 2015. Modeling information accumulation in psychological tests using item response times. *Journal of Educational and Behavioral Statistics*, *40*, 274–306.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Roskam, E. E. 1987. Toward a psychometric theory of intelligence. In E. E. Roskam and R. Suck (Eds.). *Progress in Mathematical Psychology* (pp. 151–171). Amsterdam: North-Holland.

Roskam, E. E. 1997. Models for speed and time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.). *Handbook of Modern Item Response Theory* (pp. 187–208). New York: Springer.

Scheiblechner, H. 1979. Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19,* 18–38.

Schmitt, T. A., Sass, D. A., Sullivan, J. R., and Walker, C. M. 2010. A Monte Carlo simulation investigating the validity and reliability of ability estimation in item response theory with speeded computer adaptive tests. *International Journal of Testing*, *10*, 230–261.

Stretch, L. S. and Osborne, J. W. 2005. Extended time test accommodation: Directions for future research and practice. *Practical Assessment, Research and Evaluation*, *10*, 1–8.

Swanson, D. B., Case, S. E., Ripkey, D. R., Clauser, B. E., and Holtman, M. C. 2001. Relationships among item characteristics, examinee characteristics, and response times on USMLE, Step 1. *Academic Medicine*, *76*, 114–116.

Tatsuoka, K. K. and Tatsuoka, M. M. 1980. A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.). *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236–256). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D. 1983. Timed testing: An approach using item response theory. In D. J. Weiss (Ed.). *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 179–203). New York: Academic Press.

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

van der Linden, W. J. 2008. Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J. 2009a. Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J. 2009b. A bivariate lognormal model response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*, 378–394.

van der Linden, W. J. 2010. Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, *47*, 92–114.

van der Linden, W. J. 2011a. Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, *53*, 334–358.

van der Linden, W. J. 2011b. Setting time limits on tests. *Applied Psychological Measurement*, *35*, 183–199.

van der Linden, W. J. 2011c. Test design and speededness. *Journal of Educational Measurement*, *48*, 44–60.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., and Zhang, Y. 2007. Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

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

van der Linden, W. J. and Guo, F. 2008. Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., and Klein Entink, R. H., and Fox, J.-P. 2010. Item parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

van der Linden, W. J., and Ren, H. 2015. Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, *80*, 263–288.

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. 1997. A logistic model for time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.). *Handbook of Modern Item Response Theory* (pp. 169–185). New York: Springer.

Wang, T. and Hanson, B. A. 2005. Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wise, S. L. and DeMars, C. E. 2009. An application of item-response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L. and Kong, C. 2005. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Wollack, J. A., Cohen, A. S., and Wells, C. S. 2003. A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*, 307–330.

Woodbury, M. A. and Novick, M. R. 1968. Maximizing the validity of a test battery as a function of relative test length for a fixed total testing time. *Journal of Mathematical Psychology*, *5*, 242–259.