SOME CONSEQUENCES OF RESPONSE TIME MODEL MISSPECIFICATION
IN EDUCATIONAL MEASUREMENT

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by
Jeffrey M. Patton

Ying Cheng, Director

Graduate Program in Psychology
Notre Dame, Indiana
August 2014
SOME CONSEQUENCES OF RESPONSE TIME MODEL MISSPECIFICATION
IN EDUCATIONAL MEASUREMENT

Abstract

by

Jeffrey M. Patton

Response times (RTs) on test items are a valuable source of information concerning examinees and the items themselves. As such, they have the potential to improve a wide variety of measurement activities. However, researchers have found that empirical RT distributions can exhibit a variety of shapes among the items within a single test. Though a number of semiparametric and “flexible” parametric models are available, no single model can accommodate all plausible shapes of empirical RT distributions. Thus the goal of this research was to study a few of the potential consequences of RT model misspecification in educational measurement. In particular, two promising applications of RT models were of interest: examinee ability estimation and item selection in computerized adaptive testing (CAT).

By jointly modeling RTs and item responses, RTs can be used as collateral information in the estimation of examinee ability. This can be accomplished by embedding separate models for RTs and item responses in Level 1 of a hierarchical model and allowing their parameters to correlate in Level 2. If the RT model is misspecified, a potential drawback of this hierarchical structure is that any negative impact on estimates of the RT model parameters may, in turn, negatively impact ability estimates. However, a simulation study found that estimates of the RT model parameters were robust to misspecification of the RT model. In turn, ability estimates were also
By considering the time intensity of items during item selection in CAT, test completion times can be reduced without sacrificing the precision of ability estimates. This can be done by choosing items that maximize the ratio of item information to the examinee’s predicted RT. However, an RT model is needed to make the prediction; if the RT model is misspecified, this method may not perform as intended. A simulation study found that whether or not the correct RT model was used to make the prediction had no bearing on test completion times. Additionally, using a simple, average RT as the prediction was just as effective as model-based prediction in reducing test completion times.
Although high school is but a distant memory, I would like to dedicate this work to Greg Bimm and many others on the faculty of Marian Catholic High School who inspired me to become not just a scholar, but a “Scholar with a Soul.”
CONTENTS

FIGURES ................................................................. vi

TABLES ................................................................. vii

ACKNOWLEDGMENTS ..................................................... viii

CHAPTER 1: INTRODUCTION .............................................. 1
  1.1 Approaches to Modeling Response Times in Tests .............. 3
    1.1.1 Models That Incorporate Both Item Responses and Response Time ........................ 5
    1.1.2 A Hierarchical Approach ........................................ 7
  1.2 Response Time Models for Tests .................................. 11
    1.2.1 Box-Cox Normal Model ........................................ 12
    1.2.2 Cox Proportional Hazards Model ............................. 16
    1.2.3 Linear Transformation Model ................................. 19
  1.3 Conclusions ...................................................... 21

CHAPTER 2: USING RESPONSE TIMES AS COLLATERAL INFORMATION FOR CALIBRATION OF ITEM RESPONSE MODELS .... 22
  2.1 Introduction ..................................................... 22
  2.2 Simulation Study ................................................. 25
    2.2.1 Construction of Test Forms ................................. 29
    2.2.2 Person Parameters and Response Time Generation ........ 32
    2.2.3 Model Calibration and Model Fit ........................... 33
    2.2.4 Dependent Measures ......................................... 37
  2.3 Simulation Study Results ......................................... 38
    2.3.1 Response Time Model Fit .................................... 38
    2.3.2 Recovery of Ability Parameters ............................. 44

CHAPTER 3: USING RESPONSE TIMES FOR ITEM SELECTION IN COMPUTERIZED ADAPTIVE TESTING ...................... 51
  3.1 Introduction ..................................................... 51
  3.2 Simulation Study ................................................. 57
    3.2.1 Construction and Calibration of Item Pools ............... 59
    3.2.2 Adaptive Test Procedure .................................... 62
FIGURES

1.1 Examples of several Box-Cox transformations (left subfigure) and several Box-Cox normal densities (right subfigure). On the left, a variable $T$ is on the horizontal axis and the transformed value $T^{(v)}$ is on the vertical axis. On the right, the location and scale parameters of each density are fixed to $\mu = 4.5$ and $\sigma = 0.4$, respectively. In each subfigure, the Box-Cox transformation parameter ranges from $v = 0$ to 0.2 from bottom to top (left subfigure) or right to left (right subfigure). 

1.2 Comparison of the standard normal and standard Gumbel distributions

2.1 Global fit of the log-normal and Cox PH models to each test form. The title of each subfigure indicates the fitted model (log-normal or Cox PH), and below that (in parentheses) is the true response time model (log-normal, Weibull, or Box-Cox). In each subfigure, the solid line indicates the cumulative proportion of posterior predictive probabilities for all observations in the dataset. If the cumulative proportion matches the identity line (dashed line), this indicates good (global) fit of the model. 

2.2 QQ plots of residuals for four items from Test 1 (i.e., true response time model is log-normal). Each column displays QQ plots for the same item fitted with either the log-normal model (top row) or the Cox PH model (bottom row). Residuals based on the log-normal model are plotted against standard normal quantiles, and residuals based on the Cox PH model are plotted against standard Gumbel quantiles. If the residuals fall on the identity line, this indicates good fit of the model to the particular item. 

2.3 QQ plots of residuals for four items from Test 2 (i.e., true response time model is Weibull). Rows and columns are structured in the same way as Figure 2.2. 

2.4 Conditional bias (top row) and average posterior standard deviation (bottom row) of ability estimates for Test 1 (i.e., true response time model is log-normal). Columns 1 and 2 display results for the log-normal and Cox PH models, respectively. Each subfigure displays results for $\rho_{\theta r} = 0$, .6, and .9. (Results for $\rho_{\theta r} = .3$ are very similar to those for $\rho_{\theta r} = 0$ and are thus omitted for clarity.)
2.5 Conditional bias (top row) and average posterior standard deviation (bottom row) of ability estimates for Test 2 (i.e., true response time model is Weibull). Rows and columns are structured in the same way as Figure 2.4.

2.6 Conditional bias (top row) and average posterior standard deviation (bottom row) of ability estimates for Test 3 (i.e., true response time model is Box-Cox). Rows and columns are structured in the same way as Figure 2.4.
ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Ying Cheng, for her guidance throughout the past five years. My strongest suit and my biggest liability are my attention to detail; if not for her constant urgings to focus on the big picture, I would still be revising papers from my first year.

I would also like to thank my dissertation committee members: Drs. Scott Maxwell, Ke-Hai Yuan, and Zhiyong Zhang. That I was able to conduct this research is a testament to their ability as teachers. I have learned from them not only the tools to conduct research, but also how to seek creative solutions to research problems and how to identify problems that are worth solving.
Response times (RTs) on test items are a potentially valuable source of information concerning examinees and the items themselves. Given the prevalence of computer-based tests, RTs are easy to record and widely available. As such, they have the potential to improve a wide variety of measurement activities including calibration of item response models (van der Linden, Entink, & Fox, 2010), test assembly (van der Linden, 2011), item selection in computerized adaptive testing (CAT; Fan, Wang, Chang, & Douglas, 2012; van der Linden, 2008, 2009b), and detection of aberrant response behavior (van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003). However, much of the research on RTs in educational tests is quite young, and many theoretical and practical issues must be addressed before these proposed applications can be widely adopted.

Currently, a fundamental question is how to best model RTs. Several parametric models have been proposed, including models based on the log-normal (van der Linden, 2006), Weibull (Rouder, Sun, Speckman, Lu, & Zhou, 2003), and gamma distributions (Maris, 1993). However, researchers have found that the shape of empirical RT distributions can vary not only among similar types of tests, but also among the items within a single test (Entink, van der Linden, & Fox, 2009; Ranger & Kuhn, 2012). Accordingly, more flexible RT models have been proposed such as the Box-Cox normal model (Entink, van der Linden, et al., 2009) and the semiparametric Cox proportional hazards model (Ranger & Ortner, 2012). These latter models can often fit empirical RT distributions more closely than parametric models, but what
they gain in flexibility, they lose in simplicity. For example, the log-normal model includes a “time intensity” parameter that summarizes how long the item typically takes to complete. In contrast, the Cox proportional hazards model must convey this information with the entire baseline hazard function (Wang, Fan, Chang, & Douglas, 2013).

Given the diversity of shapes exhibited by empirical RT distributions and the current focus of researchers on parametric models versus more flexible alternatives, an important practical concern is the consequences of employing a misspecified RT model. For many of the proposed applications of RT models in educational measurement, RTs are not analyzed in isolation; rather, they are used to augment the information provided by item responses in an effort to improve ability estimates or item selection in CAT, for example. If the RT model does not fit, it has the potential to negatively influence measurement outcomes (e.g., introduce bias into ability estimates). Thus the goal of this research is to explore some of the potential consequences of RT model misfit in educational measurement; in particular, two applications of RTs are of interest: ability estimation and item selection in CAT.

First, researchers have demonstrated that utilizing RTs as collateral information can yield more precise estimates of item response model parameters including examinee ability and item difficulty (e.g., van der Linden et al., 2010). This can be achieved by linking the parameters of separate item response and RT models through a second-order model, which allows information in the RTs to augment that provided by the item responses. The benefits of this approach have been demonstrated under the assumption that the RT model is correctly specified. The goal of the current research was to examine the effects of a misspecified RT model on ability estimates, and compare these estimates with those based on a correctly specified RT model.

The second area of interest is the use of RTs to improve item selection in CAT. Several methods have been proposed, but the focus here is a simple modification
of the popular maximum information criterion called the maximum information per time unit (MIT) criterion (Fan et al., 2012). The MIT criterion is a simple ratio of item information to a model-based prediction of an examinee’s RT. By selecting items with the MIT criterion, a fixed number of items can be completed more quickly, or more items can be completed in a fixed amount of time. The key component is the prediction of an examinee’s RT. How this prediction is made can have far-reaching effects on test completion times, ability estimates, and item pool usage. The goal of this research was to compare these outcomes when RT predictions were made using a misspecified RT model versus outcomes based on a correctly specified model. In addition, these results were compared with a simpler method that does not use a model to predict RTs.

The rest of this chapter is organized into two main sections. The first section reviews some historical and current approaches to modeling RTs in educational measurement. The second section describes and compares three recently-proposed RT models that are flexible alternatives to parametric RT models. Chapter 2 begins with a review of literature concerning the use of RTs as collateral information for the calibration of item response models. This is followed by a simulation study that explores the effects using RTs as collateral information when the RT model is misspecified. Chapter 3 contains a brief overview of adaptive testing, which is followed by a detailed description of the MIT item selection criterion and some alternatives for how the RT prediction can be made. Next is a simulation study that explores the effects of RT model misspecification on CAT outcomes. Lastly, Chapter 4 provides some conclusions and a discussion of the results of the two simulation studies.

1.1 Approaches to Modeling Response Times in Tests

Historically, there have been three major approaches to the analysis of RTs in educational measurement (Entink, Kuhn, Hornke, & Fox, 2009). The first approach
is to ignore item responses and analyze the RTs exclusively. This approach is appropriate for so-called “speed tests,” in which the examinee responds to a large number of easy questions and there is a strict time limit. In this context, it is expected that the examinee will correctly answer all the items that he or she responds to; the only outcome of interest is how quickly the examinee works. The second approach is to analyze item responses and RTs separately. In this case, both the accuracy of responses and the RTs are of interest, but there is an implicit assumption that response accuracy and the speed at which an examinee works are independent. This assumption is unlikely to hold when the examinee is free to choose the pace at which he or she works, for example, when there is a reasonable time limit on the test.

The third approach is to model item responses and RTs jointly in a single model. In contrast with the second approach, joint modeling allows dependencies between responses and RTs to be taken into account. This has become the favored approach in the field of educational measurement; however, there are two very different types of joint models. The first type consists of item response models that incorporate RTs and RT models that incorporate item responses (van der Linden, 2009a). A distinct feature of these models is that they posit a particular functional relationship between responses and RTs. The second type employs a hierarchical framework: separate models for item responses and RTs are linked through a second-order model (van der Linden, 2007). This approach allows responses and RTs to covary, but it does not impose a particular functional relationship between them. In the next section, two examples of the first type of joint model are considered. In the following section, the hierarchical approach, as well as the major differences between the two types of joint models, are discussed in detail.
1.1.1 Models That Incorporate Both Item Responses and Response Times

An example of an item response model that incorporates RTs is the four-parameter logistic response time model (4PL-RTM; Wang & Hanson, 2005). This model inserts RTs into the traditional three-parameter logistic model (3PLM). The 3PLM expresses the probability of a correct response to item $j$ by examinee $i$ as

$$P(u_{ij} = 1) = c_j + \frac{(1 - c_j)}{1 + \exp[-a_j(\theta_i - b_j)]}, \tag{1.1}$$

where $u_{ij}$ is the item response (scored 1 for correct and 0 for incorrect), $\theta_i \in \mathbb{R}$ is the examinee ability parameter, and $a_j \in \mathbb{R}^+, b_j \in \mathbb{R}$, and $c_j \in (0, 1)$ are the item discrimination, difficulty, and pseudo-guessing parameters, respectively. The 4PL-RTM expresses the same probability as

$$P(u_{ij} = 1) = c_j + \frac{(1 - c_j)}{1 + \exp[-a_j(\theta_i - \frac{\rho_i d_j}{t_{ij}} - b_j)]}. \tag{1.2}$$

The only change from the 3PLM is the inclusion of the term $\rho_i d_j/t_{ij}$: $\rho_i \in \mathbb{R}^+$ is the person slowness parameter, $d_j \in \mathbb{R}^+$ is the item slowness parameter, and $t_{ij}$ is the observed RT of examinee $i$ on item $j$. The inclusion of the item slowness parameter and the observed RT explains why this model is referred to as a four-parameter response time model.

As mentioned in the previous section, this type of joint model posits a functional relationship between item responses and RTs. In particular, the 4PL-RTM specifies that $P(u_{ij} = 1)$ is an increasing function of the observed RT, and the rate of increase is determined by the item and person slowness parameters. For fixed values of $\rho_i$ and $d_j$, one can examine the specific functional relationship between the RT and the probability (or log-odds) of a correct response. Because $t_{ij}$ is in the denominator of the additional term and this term is preceded by a subtraction sign, it is clear that
the probability of a correct response increases as the RT increases. For a very high RT, the term will be close to zero, and \( P(u_{ij} = 1) \) will be similar to that given by the 3PLM. That is, by spending a long time on an item, the penalty to the 3PLM probability will be negligible. On the other hand, spending very little time on an item can substantially reduce the 3PLM probability; as the RT approaches zero, the 4PL-RTM probability approaches \( c_j \) (the probability of a correct response due to random guessing). This built-in relationship between the RT and the probability of a correct response is an example of a “speed-accuracy trade-off.” This trade-off is motivated by the finding in experimental psychology that a subject will perform tasks less accurately when forced to work more quickly (and vice versa; van der Linden, 2009a). In this case, the 4PL-RTM predicts that a low RT (i.e., a high rate of speed) will result in less accuracy (i.e., a lower probability of a correct response).

Unlike the 4PL-RTM, which is an item response model that incorporates RTs, Thissen (1979) proposed an RT model that incorporates the log-odds of a correct response. In particular, this model treats an RT as a random variable following a log-normal distribution; because the log of a log-normally distributed variable follows a normal distribution, the model can be expressed as follows:

\[
\log(t_{ij}) = \mu + \beta_j + \tau_i - \rho a_j (\theta_i - b_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \tag{1.3}
\]

where \( t_{ij} \) is the RT of examinee \( i \) on item \( j \), \( \mu \in \mathbb{R} \) is the grand mean of the log RTs (across all combinations of persons and items), \( \beta_j \in \mathbb{R} \) is the item slowness parameter, and \( \tau_i \in \mathbb{R} \) is the person slowness parameter (notation borrowed from van der Linden, 2009a). The item and person slowness parameters are similar to ANOVA effect parameters: as item or person slowness increases, the mean log RT also increases. (Note that these item and person slowness parameters have very different interpretations from those in the 4PL-RTM.)
Thissen’s log-normal RT model, like the 4PL-RTM, specifies a functional relationship between item responses and RTs. The fourth term on the right-hand side of Equation 1.3 is comprised of \(a_j(\theta_i - b_j)\), the log-odds of a correct response under the two-parameter logistic model (or 2PLM, which is equivalent to the 3PLM in Equation 1.1 with \(c_j = 0\)), and \(\rho \in \mathbb{R}\), a slope parameter for the regression of the log RT onto the log-odds. That is, the model specifies that the log RT is a linear function of the log-odds of a correct response. But in contrast with the 4PL-RTM, Thissen’s model does not impose a speed-accuracy trade-off. For \(\rho > 0\), lower RTs are associated with a higher log-odds (i.e., a higher probability of a correct response), whereas \(\rho < 0\) implies the opposite relationship. Interestingly, the subtraction sign in front of \(\rho\) in Equation 1.3 (as well as Thissen’s language in the article) suggest his expectation that \(\rho\) should be positive, implying a positive relationship between speed and accuracy rather than a trade-off. This would imply that higher ability examinees tend to work more quickly than lower ability examinees.

### 1.1.2 A Hierarchical Approach

The previous section presented an example of an item response model that incorporates RTs and an example of an RT model that incorporates item responses. These models acknowledge that item responses and RTs are likely to covary in the context of educational tests, but they do so by imposing a specific functional relationship. The hierarchical approach proposed by van der Linden (2007) offers a more flexible alternative to the joint modeling of item responses and RTs. In this approach, Level 1 consists of separate models for item responses and RTs, and Level 2 specifies a multivariate normal model for the parameters of the Level 1 models. The model for responses can be any item response theory model (such as the 3PLM in Equation 1.1). For the RT model, van der Linden (2006) proposed a log-normal model that is similar to Thissen’s (1979) model (Equation 1.3). van der Linden’s model can be
expressed as follows:

$$\log(t_{ij}) = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_j^{-2}),$$  \hspace{1cm} (1.4)

where $\beta_j \in \mathbb{R}$ is the item time intensity parameter, $\tau_i \in \mathbb{R}$ is the examinee speed parameter, and $\alpha_j \in \mathbb{R}^+$ is the inverse scale parameter. In contrast with Thissen’s model, this model omits the log-odds of a correct response, allows item-specific (inverse) scale parameters, and the examinee speed parameter is the opposite of Thissen’s examinee slowness parameter. Finally, Level 2 consists of two models. The first is a bivariate normal distribution for examinee ability and speed ($\theta$ and $\tau$, respectively), with mean vector

$$\mu_P = (\mu_\theta, \mu_\tau),$$  \hspace{1cm} (1.5)

and covariance matrix

$$\Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta \tau} \\ \sigma_{\tau \theta} & \sigma_\tau^2 \end{pmatrix},$$  \hspace{1cm} (1.6)

where the subscript $P$ refers to “person” parameters. The second model is a separate multivariate normal distribution for all item parameters. If the 3PLM is used, the item parameters include $a$, $b$, and $c$. Along with $\alpha$ and $\beta$ from the log-normal RT model, the mean vector and covariance matrix are

$$\mu_I = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta)$$  \hspace{1cm} (1.7)
\[ \Sigma_I = \begin{pmatrix}
\sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\
\sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\
\sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\
\sigma_{a\alpha} & \sigma_{a\beta} & \sigma_{a\alpha} & \sigma_{a\alpha}^2 & \sigma_{a\beta} \\
\sigma_{b\alpha} & \sigma_{b\beta} & \sigma_{c\alpha} & \sigma_{c\beta} & \sigma_\beta^2
\end{pmatrix}, \tag{1.8} \]

respectively, where the subscript \( I \) refers to “item” parameters.

The hierarchical model addresses the dependence between item responses and RTs in a manner very different from the models considered in the previous section. First, the hierarchical model does not specify a direct relationship between the observed item responses and RTs; rather, the latent ability and speed parameters are allowed to covary in the Level 2 model. This is because observed responses and RTs reflect characteristics of the items as well as those of examinees. For example, an examinee may receive very different scores on two tests, but this may be due to differences in test difficulty rather than a change in ability. Similarly, if an examinee spends three minutes on Item 1 and one minute on Item 2, one cannot conclude that the examinee worked more slowly on Item 1; rather, Item 1 may require more work. Thus it is more appropriate to model the covariance between the ability and speed parameters, which have been adjusted for structural differences among the items (van der Linden, 2009a).

Another feature of the hierarchical model that differs from other joint modeling approaches is that it does not impose a direction or complex functional form on the relationship between ability and speed. The inclusion of a speed-accuracy trade-off in some models (such as the 4PL-RTM) is motivated by findings from reaction-time research in psychology (van der Linden, 2009a). Reaction-time experiments often ask a subject to complete a large number of simple tasks; by manipulating the speed at which the subject works, one can demonstrate that accuracy decreases when speed
increases. During a typical educational test, however, an examinee’s speed is not manipulated by an experimenter (though speed may be influenced by an overall time limit). Further, intuition and some empirical evidence (van der Linden, Breithaupt, Chuah, & Zhang, 2007) suggest that examinees tend to work at a relatively constant speed. Thus, a typical educational test (and the setting in which it is administered) is unlikely to elicit behavior from an examinee that could demonstrate a trade-off between speed and accuracy.

An additional reason why the hierarchical model does not impose a speed-accuracy trade-off is because the Level 2 correlation between ability and speed is a between-person relationship, whereas the speed-accuracy trade-off is a within-person phenomenon (van der Linden, 2009a). That is, two examinees may work at different speeds, but one cannot necessarily expect the slower examinee to receive a higher score. This is because the (unknown) function describing the negative relationship between speed and accuracy (if it were able to be plotted) is unique to each individual (van der Linden, 2009a). For example, one examinee may be able to maintain a high degree of accuracy even when working very quickly, whereas another examinee may have to work much more slowly in order to perform well. Therefore, the hierarchical model does not impose a particular direction or functional relationship on the Level 2 relationship between ability and speed. In fact, among applications of the hierarchical model to empirical data, van der Linden (2009a) reported speed-ability correlations ranging from $-0.65$ to $0.30$. A positive correlation between speed and ability may seem unintuitive, but as explained by van der Linden (2009a), a positive correlation reflects the “better time-management skills among the more able test takers” (p. 267). In other words, when there is a strict time limit, high-ability examinees wisely choose to work quickly (implying a positive correlation), whereas if there is plenty of time, high-ability examinees work more slowly and methodically (implying a negative correlation).
Because of these theoretical improvements over models that specify a functional relationship between item responses and RTs, the hierarchical approach has become a popular joint modeling method in the field of educational measurement. The hierarchical approach is also very flexible: any suitable model for the responses or RTs can be employed in Level 1 (van der Linden, 2007). The literature contains a wealth of item response theory models including the dichotomous 2PL and 3PL models (Hambleton & Swaminathan, 1985), polytomous models such as the graded response and partial credit models (Ostini & Nering, 2006), and multidimensional versions of these models (e.g., Reckase, 2009), among others. The application of item response theory models to assessment data has been well-studied. In contrast, the literature on latent variable models for item RTs is much less developed, and RT models are not routinely applied to assessment data. However, given the increased popularity of computer-based testing, research on RT models has increased in recent years, and many new models have been proposed. In the following sections, three promising item RT models are described in detail.

1.2 Response Time Models for Tests

As stated in the previous section, item response theory models have enjoyed a long history of research and application in educational measurement. RT models, though not new to the field, have not received the same level of attention from researchers or practitioners. As a result, there is currently no consensus on which types of RT models might be most appropriate for typical assessment data. Several parametric models have been proposed; for example, in addition to the log-normal model of van der Linden (2006; Equation 1.4), parametric models based on the Weibull (Rouder et al., 2003) and gamma (Maris, 1993) distributions have also been studied. The latter two models were motivated by their ability to model a particular psychological process that produced the RTs. In contrast, the log-normal model was primarily motivated
by a “distribution-fitting approach” (van der Linden, 2007, p. 295). That is, the log-normal distribution shares several characteristics with empirical RT distributions (e.g., a positive support and right-skewed shape).

In spite of these merits, a common criticism of parametric RT models is that there is no single model (or family of models) that can accommodate the diversity of empirical RT distributions (Ranger & Kuhn, 2012). For each dataset to be analyzed, one must go through the process of fitting several parametric models and choose a best-fitting model according to some criterion. Additionally, the best-fitting model overall may not be the best-fitting model for all items on the test (Entink, van der Linden, et al., 2009). To avoid these issues, several researchers have concluded that the solution is “a flexible model that can routinely be used for response times in tests, because it only makes weak distributional assumptions and subsumes the most successful approaches” (Ranger & Kuhn, 2012, p. 33).

Accordingly, a number of “flexible parametric” and semiparametric RT models have been proposed. In the following sections, three of these models are described: the Box-Cox normal model (Entink, van der Linden, et al., 2009), the Cox proportional hazards model (Ranger & Ortner, 2012; Wang, Fan, et al., 2013), and the linear transformation model (Ranger & Kuhn, 2013; Wang, Chang, & Douglas, 2013). What each of these models has in common is the flexibility to accommodate a wide variety of RT distributions, combined with a degree of parametric simplicity. Also, each of these models can be implemented in the hierarchical model of van der Linden (2007).

1.2.1 Box-Cox Normal Model

An example of a “flexible” parametric model for item RTs is based on the Box-Cox normal distribution (Entink, van der Linden, et al., 2009). Like the log-normal distribution, the Box-Cox normal distribution is often used to model nonnegative, right-skewed data. In fact, the Box-Cox normal model is a generalization of the
log-normal model. Under the Box-Cox model, the Box-Cox transformation of a random variable follows a normal distribution. Because the Box-Cox transformation represents an entire class of transformations (which includes the log transform as a special case), it can be applied to distributions of a variety of shapes. For example, the log-normal model is only appropriate for right-skewed data, but the Box-Cox model can be applied to symmetric or even left-skewed data.

Specifically, the Box-Cox transformation of a random variable $T$ is an order-preserving function that depends on a parameter $v \in \mathbb{R}$:

$$T^{(v)} = \begin{cases} \frac{T^v - 1}{v} & v \neq 0 \\ \log T & v = 0, \end{cases} \quad (1.9)$$

where $T^{(v)}$ is the transformed variable. To make the transformation continuous over the range of $v$, the log transform for $v = 0$ is included as a special case. Then the Box-Cox density is

$$f(t) = t^{v-1} \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{1}{2} \left( \frac{t^{(v)} - \mu}{\sigma} \right)^2 \right], \quad (1.10)$$

where $\mu$ and $\sigma$ are the location and scale parameters, respectively.\textsuperscript{1} Several examples of the Box-Cox transformation are shown in the left half of Figure 1.1; values of a variable $T$ are on the horizontal axis and the transformed value $T^{(v)}$ is on the vertical axis. The transformations range from zero (i.e., the log transform) for the bottommost line to 0.2 for the topmost line. For this range of $v$, the Box-Cox transformation is appropriate for right-skewed data: large values are compressed, and as a result, the

\textsuperscript{1}The Box-Cox transformation of a random variable actually yields a truncated normal distribution, but the truncation usually occurs far from the center of the distribution. Accordingly, the Box-Cox density function includes a normalization constant to ensure that the density integrates to one (Freeman & Modarres, 2002). However, this term was omitted from Equation 1.10 to clarify its presentation.
long right tail of a distribution is also compressed. Also, the degree of compression increases as the value of \( v \) decreases, implying that highly skewed data require small values of \( v \). This fact is reinforced in the right half of Figure 1.1 which displays the Box-Cox normal distribution associated with each value of \( v \). The leftmost density is evaluated with \( v = 0.2 \); in comparison with the other curves, its shape looks nearly symmetric.\(^2\) In contrast, the rightmost curve is a log-normal density (i.e., \( v = 0 \)), and among these curves, it is the most highly skewed.

---

\(^2\)The leftmost curve actually has a positive skew, but it is difficult to discern because of the width of the horizontal axis. A symmetric Box-Cox density is only attained when \( v = 1 \).
The proposed Box-Cox normal model for item RTs (Entink, van der Linden, et al., 2009) can be written as follows:

\[ t_{ij}^{(v)} = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_j^{-2}), \]  

(1.11)

where \( t_{ij}^{(v)} \) is the transformed RT of examinee \( i \) on item \( j \), \( \beta_j \in \mathbb{R} \) is the item time intensity parameter, \( \tau_i \in \mathbb{R} \) is the examinee speed parameter, and \( \alpha_j \in \mathbb{R}^+ \) is the item-specific inverse scale parameter. Notice that the subtraction sign in front of \( \tau_i \) defines it as a “speed” parameter: as examinee speed increases, the transformed RT decreases. Also, with the identifying restriction that \( E(\tau_i) = 0 \), the item time intensity parameter is interpreted as the mean transformed RT. Lastly, \( \alpha_j \) can be interpreted as a discrimination parameter: for larger values of \( \alpha_j \), the RT distributions for different examinees are narrower and share less overlap. When embedded in the hierarchical model of van der Linden (2007), the Level 2 model is identical to that for the log-normal model. That is, examinee ability and speed follow a bivariate normal distribution, and the item parameters follow a separate multivariate normal distribution (see Equations 1.5 - 1.8). Also note that the Box-Cox RT model is equivalent to van der Linden’s (2006) log-normal RT model (Equation 1.4) when \( v = 0 \).

In practice, the Box-Cox model can be applied in two ways. First, one can estimate a single transformation parameter that is common to all items. By choosing an optimal value for \( v \) (rather than fixing it to zero), the Box-Cox model would certainly achieve better fit than the log-normal model. On the other hand, a single value for \( v \) imposes a single parametric form on all items. The alternative is to estimate item-specific transformations. Undoubtedly, this would achieve better fit than a single transformation, but the item-specific approach is not without its drawbacks. The most obvious drawback is that the transformed RTs of different items are on differ-
ent scales. Even if two items have similar ranges of RTs, different values for \( v \) can produce very different scales. For example, if RTs range from 10 to 120 seconds, the range of transformed RTs with \( v = 0.25 \) is (3.1, 9.2), and with \( v = 0.75 \), the range is (6.2, 47.0). As a result, the \( \alpha_j \) and \( \beta_j \) parameters are not comparable across items. Additionally, if the \( v \) estimates are very different across items, it may be necessary to estimate item-specific slopes for the speed parameter.

1.2.2 Cox Proportional Hazards Model

The Box-Cox normal model is an attractive generalization of the log-normal model. Because it employs the entire class of Box-Cox transformations, it can be used to model a wide variety of distributions. Nonetheless, the Box-Cox model has some potential drawbacks. For example, the Box-Cox transformation is parametric; thus it may be unable to model RT distributions that have unusual shapes (e.g., bimodal). Also, the most general version of the model creates item-specific scales, making it difficult to compare parameters across items.

A more flexible model that avoids both of these issues is the semiparametric proportional hazards (PH) model (Ranger & Ortner, 2012; Wang, Fan, et al., 2013). The Cox PH model comes from the field of survival analysis, which is concerned with outcomes that are measured as the time “until an event occurs.” For example, in the context of educational measurement, the event of interest is the completion of a test item, and the time until that event is the response time. Rather than direct analysis of RTs, survival analysis focuses on the hazard rate, or the conditional probability of item completion at time \( t \), given that the item has not yet been completed. Symbolically, if the time-to-event variable is \( T \), its density function is \( f(t) \), and its cumulative distribution function is \( F(t) \), then the hazard function is \( h(t) = f(t)/(1 - F(t)) \). If \( T \) is discrete, the hazard function can be interpreted as \( P(T = t)/P(T > t) \). Psychologically, the hazard rate is the probability of completing an item in the next moment;
it reflects an examinee’s processing capacity, so an examinee with a high hazard rate works more intensely (Ranger & Ortner, 2012; Wenger & Gibson, 2004).

The PH model states that the conditional hazard rate (i.e., conditional on a set of covariates) is equal to the product of the baseline hazard rate and $e^Z$, where $Z$ is a linear combination of the covariates and the baseline hazard rate is the hazard rate when all covariates are equal to zero. Put another way, the conditional hazard function is proportional to the baseline hazard function by a factor of $e^Z$. Specifically, the proposed PH model for RTs can be expressed as

$$h_{ij}(t|\tau_i) = h_{0j}(t) \exp(\gamma_j \tau_i),$$

where $h_{0j}(\cdot)$ is the baseline hazard function for item $j$, $\tau_i \in \mathbb{R}$ is the latent speed parameter for examinee $i$, and $\gamma_j \in \mathbb{R}^+$ is a slope parameter that determines the increase in processing capacity due to an increase in speed. Alternatively, the PH model for RTs can be rewritten as a linear regression model (Wang, Fan, et al., 2013):

$$\log \left[ \int_0^t h_{0j}(s) \, ds \right] = -\gamma_j \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{Gumbel}(0, 1). \quad (1.13)$$

This formulation of the Cox PH model permits easier comparison with other RT models. On the left hand side, $\int_0^t h_{0j}(s) \, ds$ is the (item-specific) cumulative baseline hazard function. This is a monotone increasing function of RTs, and in the Cox PH model, this function is estimated nonparametrically. Thus, after taking the log, the left hand side of Equation 1.13 can simply be thought of as an order-preserving, non-parametric transformation of RTs to the real number line (Wang, Fan, et al., 2013). This is an important advantage over the Box-Cox normal model: both models are capable of item-specific transformations, but the Cox PH model, because it employs nonparametric transformations, is able to capture a wider variety of RT distributions. Another difference between these two models is the error distribution: Box-Cox errors
follow a normal distribution, whereas Cox PH errors follow the standard Gumbel distribution for minimums\(^3\) (0 and 1 are its location and scale parameters, respectively). Figure 1.2 shows a comparison of the standard normal and standard Gumbel distributions. The mode of the Gumbel density is at zero, but unlike the normal curve, it has a clear negative skew; as a result, its mean is slightly below zero (\(\approx -0.5772\)).

![Figure 1.2. Comparison of the standard normal and standard Gumbel distributions](image)

A few additional features of the Cox PH model are important to point out. First, there is no time intensity parameter or item-specific scale parameter. This is because any non-zero intercept or change in the scale of the errors can be absorbed into the

\(^3\)The standard Gumbel distribution for minimums, also known as the standard Type I extreme value distribution for minimums, is the limiting distribution of the minimum of a large number of identically distributed random variables (“Extreme Value Distribution,” n.d.). It is closely related to several distributions employed in survival analysis. For example, the log of an exponential or Weibull distribution yields a Gumbel distribution (Rodríguez, 2010).
nonparametric transformation (Wang, Fan, et al., 2013). Thus, unlike the version of the Box-Cox model with item-specific transformations, there is no concern about being unable to compare parameters across items. Also, the Cox PH model includes an item-specific slope parameter for the effect of speed, which is necessary because the scale of the RT transformation is specific to each item. Finally, when embedded in the hierarchical model of van der Linden (2007), the Level 2 model consists of a bivariate normal distribution for examinee ability and speed. This differs from the Box-Cox and log-normal models, each of which includes in Level 2 a multivariate normal distribution for all item parameters. The Level 2 model for item parameters may be omitted from the hierarchical Cox PH model because there is no reason to expect a significant correlation between the item slope parameter and any of the item response model parameters. A potential concern is the lack of a time intensity parameter in the Cox PH model. Several applications of the hierarchical log-normal model have yielded moderate positive correlations between item difficulty and time intensity (van der Linden, 2009a), suggesting that difficult items also tend to be more time-consuming. With the Cox PH model, it is not possible to model this relationship because time intensity is reflected in the entire baseline hazard function. However, Wang, Fan, et al. (2013) demonstrated that the implicit assumption of independence between item response and RT model item characteristics did not substantially harm the fit of the model.

1.2.3 Linear Transformation Model

The Cox PH model is an extremely flexible model; by applying an item-specific, nonparametric transformation, it can model a wide variety of RT distributions in a single test. However, this model assumes that the errors follow a standard Gumbel distribution, or equivalently, that the RT distribution is closed under proportionality of hazards (at least approximately). This assumption does not necessarily hold for
empirical RT distributions; for example, several applications of the log-normal RT model have achieved excellent fit to empirical data (e.g., van der Linden, 2009a), and the log-normal distribution is not closed under proportionality of hazards.

An even more flexible RT model for test items is the linear transformation model (Ranger & Kuhn, 2013; Wang, Chang, et al., 2013). In its most general form, this model states that after some order-preserving transformation, RTs are the sum of a linear combination of covariates and a random error term. In the case of a single covariate (i.e., examinee speed), the model can be expressed as follows:

\[ H_j(t_{ij}) = -\gamma_j \tau_i + \varepsilon_{ij}, \]  

where \( H_j(\cdot) \) represents the order-preserving transformation for item \( j \). Of course, the model as specified in Equation 1.14 is far too general to be of any practical use, but it does serve as a unifying framework for a variety of RT models, including the models considered in previous sections. For example, substituting the Box-Cox transformation for \( H_j(\cdot) \) and a normal distribution for the errors yields the Box-Cox normal model. Alternatively, a nonparametric transformation and Gumbel errors yields the Cox PH model.

Recently, Wang, Chang, et al. (2013) implemented a semiparametric version of the linear transformation model; specifically, they employed a nonparametric transformation and three different error distributions (normal, logistic, and Gumbel), but to facilitate estimation, a single distribution was assumed for all items on a given test. By employing a nonparametric transformation, this implementation of the model shares the flexibility of the Cox PH model. However, the linear transformation model is even more flexible in that different error distributions can be considered. Wang, Chang, et al. (2013) note that it is possible to allow different error distributions for different items or to estimate the error distributions nonparametrically,
though they note this extreme flexibility would undoubtedly complicate estimation of the model. Lastly, Wang, Chang, et al. (2013) combined their model with van der Linden’s (2007) hierarchical model. Like the hierarchical Cox PH model, Level 2 only consists of a bivariate normal distribution for examinee ability and speed.

1.3 Conclusions

A variety of flexible RT models have been proposed in recent years. Much of this flexibility derives from the hierarchical approach to joint modeling. In contrast with models that include both item responses and RTs, Level 2 of the hierarchical model “takes over” the job of modeling the relationship between ability and speed. As a result, the development of RT models can focus solely on capturing the relevant features of empirical RT distributions.

The linear transformation model provides a useful framework for RT models as combinations of a link function (or monotone transformation), a linear component, and an error distribution. This framework is useful not only because it unifies a variety of existing models, but also because it suggests numerous possibilities for new models.
CHAPTER 2

USING RESPONSE TIMES AS COLLATERAL INFORMATION FOR CALIBRATION OF ITEM RESPONSE MODELS

2.1 Introduction

As stated in the previous chapter, the purpose of analyzing item responses and item RTs jointly is to capture the dependence between them. The hierarchical approach accomplishes this by allowing the parameters of separate Level 1 models to covary in Level 2 (van der Linden, 2007). An obvious practical benefit of this hierarchical structure is that information can be shared between the Level 1 models. As a result, the parameter estimates of each model may be more precise than if each model was fit alone.

For example, van der Linden et al. (2010) demonstrated that application of the hierarchical 3PL/log-normal model yielded more precise 3PLM parameter estimates than did application of the 3PLM alone. Specifically, they generated item responses and RTs for 1000 examinees according to the hierarchical model using different values for the correlation between speed and ability ($\rho_{\theta \tau}$). Using Markov chain Monte Carlo (MCMC) methods to fit the hierarchical model or the 3PLM alone, they found that the hierarchical model yielded a 5% reduction (on average) in the mean squared error of the ability estimates when $\rho_{\theta \tau} = .5$; when $\rho_{\theta \tau} = .75$, the average reduction was 20%. With a smaller sample of 300 examinees, they found that the hierarchical model yielded similar improvements for item discrimination and difficulty estimates.

These results can be understood by considering the posterior distribution of ability under the 3PLM alone versus that under the hierarchical model (van der Linden et al., 2010).
For examinee $i$, the posterior distribution of ability under the item response model can be represented as follows:

$$p(\theta_i | u_i, \mu_\theta, \sigma^2_\theta) \propto f(u_i | \theta_i) \, p(\theta_i | \mu_\theta, \sigma^2_\theta),$$  \hspace{1cm} (2.1)$$

where $u_i$ is the examinee’s vector of observed item responses and $f(u_i | \theta_i)$ is the likelihood of ability given the observed responses. The prior $p(\theta_i | \mu_\theta, \sigma^2_\theta)$ is often a normal distribution with mean and variance estimated from a separate dataset. Notice that $u_i$ provides information that is specific to the examinee, but the prior merely reflects the shape of the population ability distribution. Under the hierarchical model, more information is available: the examinee’s vector of observed RTs ($t_i$) and the population distribution of ability and speed (with mean vector $\mu_P$ and covariance matrix $\Sigma_P$). van der Linden et al. (2010) showed that the new posterior distribution can be expressed as follows:

$$p(\theta_i | u_i, t_i, \mu_P, \Sigma_P) \propto f(u_i | \theta_i) \, p(\theta_i | t_i, \mu_P, \Sigma_P).$$  \hspace{1cm} (2.2)$$

Here, the prior takes into account the examinee’s observed RTs (which contain information about the examinee’s speed) and the population correlation between ability and speed. The result is a prior that is tailored to the individual. Compared with the prior in Equation (2.1), the tailored prior should be narrower and closer to the true ability value, yielding an ability estimate that is more precise and less biased. An analogous argument can be made to show how item difficulty and discrimination parameter estimates can benefit from their Level 2 correlation with the RT model item parameters. Conversely, the speed parameter and item parameters of the RT model can benefit from their Level 2 correlations with the item response model parameters.

Of course, one cannot expect the hierarchical model to always yield improved parameter estimates. For the hierarchical 3PL/log-normal model, Ranger (2013)
studied the contribution of RTs to the test information function under various conditions.\footnote{Test information is simply another name for the Fisher information associated with the likelihood of the ability parameter. The square root of the inverse of test information is commonly used to approximate the standard error of the ability estimate.} Specifically, he found that the potential increase in test information due to RTs is bounded from above, and the upper bound is determined by two primary factors: the correlation between ability and speed and the test length. The increase in test information due to RTs is rather small when $\rho_{\theta r} < .5$, regardless of the test length. For $\rho_{\theta r} > .5$, the increase in test information can be substantial, even for short tests. However, information from the item responses is not bounded, so when the test is long, information from the item responses becomes quite large and the contribution of the RTs becomes negligible.

The research demonstrating the potential benefits of using RTs as collateral information is encouraging. In contrast with traditional examples of collateral variables (e.g., demographic information or survey questions), RTs are collected automatically during the test. Also, no special model is needed; the desired item response model can simply be inserted into Level 1 of the hierarchical model. Thus, one can obtain (potentially) more precise estimates of item response model parameters with minimal effort. However, the results of the few studies to date rest on the assumption that the chosen RT model fits the observed RT distributions. As recent research on RT modeling has found, this assumption should not be taken lightly. It is not unusual for RT distributions to exhibit a variety of shapes in a single test, and even semiparametric models cannot be expected to closely fit all test items (e.g., Wang, Chang, et al., 2013; Wang, Fan, et al., 2013).

Therefore, an obvious question is whether the application of a hierarchical model with the RT model misspecified could have negative effects on item response model parameter estimates. Recalling the prior distribution for ability under the hierarchical model (Equation 2.2), information from the RT model is shared with the ability
estimate in two ways: the examinee’s vector of observed RTs (which contain information about the examinee’s speed) and the population correlation between speed and ability. It seems plausible that fitting an incorrect RT model could negatively affect an individual’s speed estimate or the correlation estimate; in turn, “bad” information from either source could introduce bias into the ability estimate or reduce its precision.

A simulation study was conducted to explore this possibility in greater detail. Specifically, the goal was to investigate the effect of a misspecified RT model on recovery of 3PLM ability parameters. To do so, item responses and RTs were generated according to a particular hierarchical model. The simulated data were then fit with two different hierarchical models: one employed the correct RT model, and the other employed an incorrect RT model. Additionally, the true correlation between speed and ability was manipulated. This study focuses on two RT models in particular: the log-normal model (van der Linden, 2006) and the Cox PH model (Wang, Fan, et al., 2013). The log-normal model was chosen because it is a simple, parametric model that has received much attention in the research literature and is likely to be a common choice among measurement practitioners. The Cox PH model was chosen because it is a more flexible alternative to the log-normal model that has also received substantial attention from researchers.

2.2 Simulation Study

Three tests of 30 items were created. A test length of 30 items was chosen because it represents an educational test of typical length (i.e., neither too short nor too long). For each test, the 3PLM was used to generate the item responses. As a reminder, the 3PLM expresses the probability of a correct response to a multiple-choice item as follows:

\[ P(u_{ij} = 1) = c_j + \frac{(1 - c_j)}{1 + \exp[-a_j(\theta_i - b_j)]}, \]  \hspace{1cm} (2.3)
where $\theta_i \in \mathbb{R}$ is the examinee ability parameter, and $a_j \in \mathbb{R}^+$, $b_j \in \mathbb{R}$, and $c_j \in (0, 1)$ are the item discrimination, difficulty, and pseudo-guessing parameters, respectively.

The same item parameters were used in each of the three test forms. In contrast, the test forms differed in terms of the true RT model: log-normal, Weibull, or a variation of the Box-Cox normal model (each of these models is described in detail in the following paragraphs). After item responses and RTs were generated, the hierarchical model was fit to each dataset using the 3PLM (for the item responses) and either the log-normal or the Cox PH model (for the response times). Thus the model for the item responses was always correctly specified, whereas the RT model was either correctly or incorrectly specified. These six conditions (three true RT models $\times$ two fitted RT models) are summarized in Table 2.1.

### TABLE 2.1

**STUDY 1: SUMMARY OF SIMULATED CONDITIONS**

<table>
<thead>
<tr>
<th>Test</th>
<th>True Response Time Model (&amp; Description)</th>
<th>Fitted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>Log-normal (log transform + normal errors)</td>
<td>Log-normal (correct)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cox PH (misspecified)</td>
</tr>
<tr>
<td>Test 2</td>
<td>Weibull (log transform + Gumbel errors)</td>
<td>Log-normal (misspecified)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cox PH (correct)</td>
</tr>
<tr>
<td>Test 3</td>
<td>Box-Cox transform + normal or Gumbel errors</td>
<td>Log-normal (misspecified)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cox PH (misspecified)</td>
</tr>
</tbody>
</table>
For Test 1, RTs were generated with the log-normal model (van der Linden, 2006); as a reminder, this model can be expressed as

\[ \log(t_{ij}) = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \alpha_j^{-2}), \]  

(2.4)

where \( \beta_j \in \mathbb{R} \) is the item time intensity parameter, \( \tau_i \in \mathbb{R} \) is the examinee speed parameter, and \( \alpha_j \in \mathbb{R}^+ \) is the item-specific inverse scale (or “discrimination”) parameter. In this condition, fitting the log-normal model is the optimal choice. The Cox PH model, though it employs a flexible, nonparametric transformation, has Gumbel-distributed errors and is therefore a “misspecified” model. Thus the purpose of Test 1 was to examine the effects of fitting the hierarchical Cox PH model on recovery of the 3PLM ability parameters. The hierarchical log-normal model was also fit to provide a baseline for comparison.

For Test 2, the true RT model was based on the Weibull distribution.\(^5\) This distribution was chosen because it is commonly used in parametric proportional hazards models. The density function is

\[ f(t) = \alpha \lambda^\alpha t^{\alpha-1} \exp\left[-(\lambda t)^\alpha\right], \]  

(2.5)

where \( \alpha \in \mathbb{R} \) and \( \lambda \in \mathbb{R} \) are the shape and inverse scale parameters, respectively. To analyze the effects of covariates on a Weibull variable, it is common to specify that \( \lambda = \exp(-Z) \), where \( Z \) is a linear combination of the covariates (Dobson & Barnett, 2008; Rodríguez, 2010). To see how this model compares with the log-normal model, it is useful to take the log of a Weibull variable, which yields the following regression model:

\[ \log(t) = -\log(\lambda) + \varepsilon, \quad \varepsilon \sim \text{Gumbel}(0, \alpha_j^{-2}). \]  

(2.6)

\(^5\)Note that this Weibull model is different from the model proposed by Rouder et al. (2003).
Replacing \( \lambda \) with the covariates of interest, the result is a model that looks very similar to the log-normal RT model:

\[
\log(t_{ij}) = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{Gumbel}(0, \alpha_j^{-2}).
\] (2.7)

The only difference between the two models is the distribution of the error term.\(^6\)

Now it is clear that Test 2 reverses the situation in Test 1: the Cox model is now the “true” model\(^7\) and the log-normal model is the misspecified model. Thus the purpose of this condition was to examine the effects of fitting the hierarchical log-normal model on recovery of the 3PLM ability parameters. And as a baseline of comparison, the hierarchical Cox PH model was also fit.

Finally, the RTs for Test 3 were generated from a variation of the Box-Cox normal model (Entink, van der Linden, et al., 2009). Again, the Box-Cox RT model is

\[
t_{ij}^{(v)} = \beta_j - \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \alpha_j^{-2}),
\] (2.8)

where \( t_{ij}^{(v)} \) is the Box-Cox transformed RT. For this test, three different values of the transformation parameter \( v \) were chosen. Also, the errors were drawn from either normal or Gumbel distributions. By using different combinations of transformations and error distributions for different items, the goal was to generate RTs such that the shape of the distribution varied across items, a property often exhibited by empirical RTs (e.g., Wang, Chang, et al., 2013; Wang, Fan, et al., 2013). A drawback of gener-

\(^6\)However, it should be noted that the \( \alpha \) and \( \beta \) parameters have somewhat different interpretations in the two models. When \( \tau_i = 0 \), \( \beta_j \) in the log-normal model is the average log RT; in the Weibull model, the average log RT is equal to \( \beta_j - \gamma/\alpha \), where \( \gamma \approx 0.5772 \) is Euler’s constant. Also in the log-normal model, \( \alpha_j^{-2} \) is the conditional variance of the log RTs, whereas in the Weibull model, the conditional variance is \( \alpha_j^{-2}(\pi^2/6) \).

\(^7\)Strictly speaking, the Cox PH model is not the “correct” model in this case. However, the Weibull model is a parametric proportional hazards model, so any misfit of the (semiparametric) Cox PH model is due solely to sampling error.
ating RTs in this way is that it would be difficult to fit the “true” RT model, because a different model would have to be specified for each item. Thus both the log-normal and Cox PH models are misspecified, and there was no baseline condition. Accordingly, it was not possible to determine the effect of either model on ability recovery (relative to that based on the true model). But it was possible to determine whether ability estimates were more robust to misspecification under one of the models. This situation is analogous to a one-factor ANOVA with two treatment conditions and no control condition: one cannot infer the effect of either treatment relative to the control, but a difference between the two treatments implies that at least one of them is different from the control.

2.2.1 Construction of Test Forms

As stated in the previous section, three test forms were created, each containing 30 items. The 3PLM $a$, $b$, and $c$ parameters were drawn from the retired item pool of a large-scale achievement test. To simulate multiple-choice items with four response options, items in the pool with $c$ values close to .25 were chosen. Among the chosen items, $b$ parameters ranged from $-1.77$ to $1.38$ with a mean of $-0.15$, $a$ parameters ranged from $1.06$ to $2.49$ with a mean of $1.60$, and the correlation between $b$ and $a$ was $.21$. The same set of 3PLM parameters was used for each test form. In contrast, each test used a different model to generate RTs.

For Test 1, the time intensity parameters ($\beta$) of the log-normal model (Equation 2.4) were generated by adding random normal deviates to a linear combination of the $a$ and $b$ parameters. The coefficients for the linear combination and the variance of the normal deviates were chosen so as to achieve a correlation of $0.5$ between $\beta$ and $b$ and a correlation of $0.3$ between $\beta$ and $a$. These correlations imply that difficult items tend to be more time-consuming than easy items; also, highly discriminating items are slightly more time-consuming than less-discriminating items. Additionally, the
β parameters were generated to have a mean and standard deviation (SD) of 4 and 0.33, respectively; these values imply that the RTs are on the scale of seconds (i.e., instead of milliseconds or minutes). Lastly, the inverse scale parameters (α) were generated independently of all other item parameters from a uniform distribution between 1.75 and 3.25. These distributions of the log-normal item parameters and their correlations with the 3PLM item parameters were chosen to reflect the results of empirical applications of the log-normal model (e.g., van der Linden & Xiong, 2013).

For Test 2, the β parameters of the Weibull model (Equation 2.7) were generated in a similar fashion to achieve correlations of .5 and .3 with b and a, respectively. But in contrast with the log-normal parameters, the mean and SD of β were chosen to be 4.5 and 0.25, respectively, and the α parameters were drawn from a uniform distribution between 2.75 and 4. These different ranges for the parameters of the Weibull model were chosen to achieve RT distributions in the same general range as those of the log-normal items.

For Test 3, recall that RTs were generated with different combinations of Box-Cox transformations and either normal or Gumbel errors. Specifically, three different Box-Cox transformation parameters were used: v = 0, 0.25, or 0.5. This range of values was chosen to achieve plausible shapes for the RT distributions. Values of v below zero tend to produce highly right-skewed distributions that are probably more representative of reaction-time studies than of educational test items. When combined with normal errors, v > 0.5 produces distributions that are nearly symmetric; when combined with Gumbel errors, the distributions actually become left-skewed. Thus the range of v was restricted to [0, 0.5]. Each of the three transformations was employed by 10 items. Of each set of 10 items, five items had normal errors and five items had Gumbel errors. Recall that v = 0 corresponds to a log transformation, so the true RT model for five items was the log-normal model, and the true RT model for another five items was the Weibull model. For the remaining items, the true
RT model was either the Box-Cox normal model (Equation 2.8) or a hybrid model combining a Box-Cox transformation with Gumbel errors.

The $\beta$ and $\alpha$ parameters for Test 3 were chosen to achieve RT distributions in the same general range as those of Tests 1 and 2. However, the range of $\beta$ and $\alpha$ parameters was unique to each combination of transformation and error distribution. For example, if RTs range from 10 seconds to 300 seconds, log RTs range from 2.3 to 5.7. This implies, for example, that the log-normal or Weibull $\beta$ parameters should be near the middle of this range, and that their $\alpha$ parameters should not be too small (or else the conditional variance of log RTs at a given value for speed would be too large). When $v = 0.5$, the transformed RTs range from 4.3 to 32.6. Whether this transformation is paired with normal or Gumbel errors, the $\beta$ parameters must be much larger than those for the log-normal or Weibull models. Similarly, the $\alpha$ parameters must be much smaller (to achieve a larger conditional variance on the transformed RT scale).

In addition to different ranges of $\beta$ and $\alpha$ parameters, each group of Test 3 items also required a different slope for the effect of speed. In the log-normal and Weibull models, the slope is the same for all items (−1) because all items use the log transformation. In Test 3, a slope of −1 would be appropriate for items with $v = 0$ (i.e., the log transform), but it would be far too small (in absolute value) for items with $v = 0.5$. To ensure that the effect of speed was similar across items, the following trial-and-error approach was used to select the slope parameters: RTs were generated with a candidate value for the slope; then the log RTs were regressed onto examinee speed and the least-squares slope estimate was recorded. The slope parameter was then modified (as necessary) to achieve a slope estimate (on the log scale) that was close to −1.
2.2.2 Person Parameters and Response Time Generation

The examinee sample size was fixed at \( N = 1000 \), and four values of the correlation between speed and ability were simulated: \( \rho_{\theta \tau} = 0, .3, .6, \) and .9. Though it is possible for the correlation to be negative, only the magnitude of the correlation determines the amount of information that RTs contain about \( \theta \). Thus the sign of the correlation was not manipulated. Regardless of the correlation, examinee ability parameters were fixed at 1000 equally-spaced quantiles of the \( N(0,1) \) distribution. That is, ability values were defined as \( \theta_p = \Phi^{-1}(p) \) for \( p = \frac{1}{1000}, \frac{2}{1000}, \ldots, 1 \), where \( \Phi^{-1} \) is the inverse cumulative distribution function of \( N(0,1) \). This method (borrowed from van der Linden et al., 2010) ensured that when examinees were divided into equally-spaced bins (e.g., 0.5 units wide), each bin contained enough examinees so that the conditional bias and variance of ability estimates could be studied.

For a given value of the correlation \( \rho_{\theta \tau} \), each examinee’s speed parameter was sampled from a normal distribution conditional on their ability parameter. Specifically, the speed parameter for examinee \( i \) was sampled from a normal distribution with mean \( \rho_{\theta \tau} \sigma_{\tau} \theta_i \) and standard deviation \( \sigma_{\tau} \sqrt{1 - \rho_{\theta \tau}^2} \). This yielded a bivariate normal distribution with the desired correlation between speed and ability and a \( N(0, \sigma_{\tau}^2) \) marginal distribution for speed. The SD of speed was chosen to be \( \sigma_{\tau} = 0.25 \).

Each sample of examinees (one for each value of \( \rho_{\theta \tau} \)) was “administered” all three test forms. Item responses and RTs were generated using the true parameter values of the 3PLM and the log-normal, Weibull, or modified Box-Cox model, respectively. Because many of the true RT distributions have non-negligible densities for unrealistically small RTs (e.g., a few seconds) and/or have very long right tails, the generated RTs were truncated from below and above at 10 seconds and 300 seconds, respec-
tively.\textsuperscript{8} This range of RTs implies that each test of 30 items could be completed in anywhere from five minutes to 150 minutes (or 2.5 hours). This is probably an unrealistically wide range of test completion times. However, the vast majority of completion times ranged from about 15 minutes to 90 minutes.

2.2.3 Model Calibration and Model Fit

The hierarchical 3PL/log-normal model was fit using full Bayesian estimation; MCMC sampling was conducted with the computer program JAGS (Plummer, 2013). Details concerning the Level 2 model and the specification of priors are shown in Table 2.2. To identify the 3PLM, the scale of $\theta$ was fixed to $N(0, 1)$. The conditional distribution of $\tau|\theta$ yields a bivariate normal distribution for speed and ability, and it also serves to identify the log-normal model by fixing the (marginal) expectation of $\tau$ to zero. Because the slope parameter in the log-normal model is fixed to $-1$, the SD of $\tau$ can be freely estimated.

Concerning the item parameters, the Level 2 model is slightly different from the full model suggested by van der Linden (2007; see Equations 1.7 and 1.8). First, the log of the 3PLM discrimination parameter ($a$) is modeled instead of the parameter itself. This is because the $a$ parameters in the simulated test forms have an obvious positive skew. Second, the pseudo-guessing parameter $c$ was fixed to its true value rather than estimated. This is a common choice when fitting the 3PLM (Han, 2012). $c$ estimates are often unstable without a strong prior distribution, particularly among easy items, and instability in the $c$ estimate can cause problems in the estimation of the other item parameters. An added benefit of fixing $c$ is that six fewer parameters

\textsuperscript{8}It should be noted that the item parameters of the RT models and the range of speed parameters were chosen (in part) so that the generated RT distributions would not be severely truncated. Truncation of the RTs from below is not strictly necessary, but for a typical educational test item, very small RTs (e.g., below 10 seconds) are probably indicative of a random guess rather than earnest responding behavior. Truncation of the RTs from above, however, is a practical necessity. With a highly skewed distribution like the log-normal distribution, it is not unusual to generate individual RTs that are tens of minutes long.
### TABLE 2.2
LEVEL 2 MODEL AND PRIOR SPECIFICATION FOR THE HIERARCHICAL 3PL/LOG-NORMAL MODEL

<table>
<thead>
<tr>
<th>Level 1 Parameters</th>
<th>Level 2 Model</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_i, \tau_i$</td>
<td>$\theta_i \sim N(0, 1)$</td>
<td>$\sigma_{\theta \tau} \sim N(0, 10)$</td>
</tr>
<tr>
<td></td>
<td>$\tau_i</td>
<td>\theta_i \sim N(\sigma_{\theta \tau} \theta_i, \sigma_{\tau}^2 - \sigma_{\theta \tau}^2)$*</td>
</tr>
<tr>
<td>$\log(a_j), b_j, \alpha_j, \beta_j$</td>
<td>$N_4(\mu_I, \Sigma_I)$</td>
<td>$\mu_I \sim N[(0, 0, 1, 3.5)', \Sigma_I]$</td>
</tr>
<tr>
<td></td>
<td>$\Sigma_I \sim W^{-1}(10 \times I_4, \text{df} = 4)$†</td>
<td></td>
</tr>
</tbody>
</table>

* This conditional distribution implies a bivariate normal distribution for $\theta_i$ and $\tau_i$ in which only $\sigma_{\theta \tau}$ and $\sigma_{\tau}^2$ are unknown. Also, the implied marginal distribution for $\tau_i$ is $N(0, \sigma_{\tau}^2)$.
† “$W^{-1}$” refers to an inverse Wishart distribution.

need to be estimated in the Level 2 model (one mean, one variance, and four covariances). Lastly, for all simulated datasets, it was sufficient to run a single chain of 8000 samples. The first 2000 iterations were discarded and a thinning interval of two was used, yielding a sample size of 3000 for each parameter.

The hierarchical 3PL/Cox PH model was fit using a two-stage estimation method proposed by Wang, Fan, et al. (2013). In the first stage, the parametric part of the model (i.e., everything but the cumulative baseline hazard functions) is fit using full Bayesian estimation, with a Metropolis-Hastings algorithm to conduct MCMC sampling. In the second stage, the cumulative hazard functions are estimated using the nonparametric Breslow estimator (Breslow, 1972), which requires the observed RTs as well as the item slope estimates and person speed estimates from the first stage. Both stages of estimation were implemented in R programs (R Core Team,
A summary of the Level 2 model and specification of priors for the hierarchical Cox model is shown in Table 2.3. Again, the 3PLM is identified by fixing the scale of $\theta$ to $N(0, 1)$. The conditional distribution of $\tau|\theta$ yields a bivariate normal distribution for speed and ability. In contrast with the log-normal model, the Cox model includes item-specific slope parameters, so the RT model is identified by fixing the scale of $\tau$ to $N(0, 1)$. Again, the 3PLM $c$ parameters were fixed to their true values. As explained in the previous chapter, there is no need to estimate the correlation of the slope parameter in the Cox model with the 3PLM item parameters, so all three item parameters were assigned independent priors. Lastly, the Metropolis-Hastings algorithm suggested by Patz and Junker (1999) was used to sample the $a$ and $b$ parameters of the 3PLM. This algorithm utilizes the slope-intercept parameterization of the 3PLM logit, i.e., $(a_j \theta_i + d_j)$ instead of $a_j(\theta_i - b_j)$.

The MCMC algorithm for the entire hierarchical model was exceptionally slow and the chains tended to exhibit a moderate autocorrelation. Thus each dataset was fit by running two parallel chains of 4300 iterations each. The first 1000 iterations of each chain were discarded, and a thinning interval of three was applied to the remaining samples. This resulted in a sample size of 2200 for each parameter.

Finally, two methods were used to examine the fit of each RT model to a given dataset. The first method employs posterior predictive probabilities to examine the global fit of the model. For a given (observed) response time $t_{ij}$, its posterior predictive probability is the left-tailed probability under its posterior predictive density.

---

9Because the Cox PH model utilizes a non-standard partial likelihood function, it cannot be fit using general-purpose MCMC software like JAGS. The author is indebted to Chun Wang for sample R code to implement the stage one Metropolis-Hastings algorithm (personal communication, March 17, 2014). An EM algorithm is available to fit the Cox model in a single stage using maximum likelihood (Ranger & Ortner, 2012), but this approach requires observed RTs to be discretized.

10Each chain took anywhere from 12 to 18 hours to complete. In spite of the relatively short chains, the results of several diagnostic tests were consistent with convergence of the chains.
TABLE 2.3
LEVEL 2 MODEL AND PRIOR SPECIFICATION FOR THE HIERARCHICAL 3PL/COX PROPORTIONAL HAZARDS MODEL

<table>
<thead>
<tr>
<th>Level 1 Parameters</th>
<th>Level 2 Model</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_i, \tau_i )</td>
<td>( \theta_i \sim N(0, 1) )</td>
<td>( \rho_{\theta \tau} \sim N_{[-1,1]}(0, 10) )</td>
</tr>
<tr>
<td>&amp; ( \tau_i</td>
<td>\theta_i \sim N(\rho_{\theta \tau} \theta_i, 1 - \rho_{\theta \tau}^2)^* )</td>
<td></td>
</tr>
<tr>
<td>( a_j, b_j, \gamma_j )</td>
<td>N/A</td>
<td>( d_j = a_j b_j \sim N(0, 10) )</td>
</tr>
<tr>
<td>&amp; ( \gamma_j \sim \ln N(1, 1.25) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* This conditional distribution implies a bivariate normal distribution for \( \theta_i \) and \( \tau_i \) in which only \( \rho_{\theta \tau} \) is unknown. Also, the implied marginal distribution for \( \tau_i \) is \( N(0, 1) \).

† “\( \ln N \)” refers to a log-normal distribution.

Under the log-normal model, this density is obtained by integrating the log-normal density over the joint posterior distribution of \( \tau_i, \alpha_j, \) and \( \beta_j \). In practice, this can be approximated by computing the left-tailed probability of \( t_{ij} \) at each iteration of the Markov chain and then averaging over the iterations. This can be represented symbolically as:

\[
\frac{1}{R} \sum_{r=1}^{R} F\left( t_{ij} | \tau_i^{(r)}, \alpha_j^{(r)}, \beta_j^{(r)} \right)
\]

(2.9)

where \( F(\cdot) \) is the CDF of the log-normal distribution, \( \tau_i^{(r)}, \alpha_j^{(r)}, \) and \( \beta_j^{(r)} \) are the parameter values at iteration \( r \) of the Markov chain, and \( R \) is the total number of iterations. The posterior predictive probability of an RT can be computed in an analogous way under the Cox PH model. Ideally, the distribution of these probabilities over all observations will be uniform. (Equivalently, the empirical CDF of the probabilities should fall on the identity line.) But if, for example, there is an
abundance of probabilities near zero or one, this suggests that the model does not provide adequate fit for a large proportion of the data.

The second method compares the distribution of item-level residuals with their theoretical distribution under the model (e.g., using QQ plots). These plots provide more detailed information than the posterior predictive probabilities because they can be examined for individual items. It is straightforward to compute residuals under the log-normal model. The model states that log RTs are normally distributed, so a residual is obtained by standardizing the log RT:

$$\hat{\varepsilon}_{ij} = \frac{\log(t_{ij}) - (\hat{\beta}_j - \hat{\tau}_i)}{\hat{\alpha}_j^{-1}}. \quad (2.10)$$

Under the log-normal model, all $N$ residuals associated with a given item should follow a $N(0, 1)$ distribution. It is also straightforward to compute residuals under the Cox PH model:

$$\hat{\varepsilon}_{ij} = \log \left[ \hat{H}_{0j}(t_{ij}) \right] - (-\hat{\gamma}_j \hat{\tau}_i). \quad (2.11)$$

Here, $\log \left[ \hat{H}_{0j}(t_{ij}) \right]$ is the transformed RT, and it is standardized by subtracting its location parameter $(-\hat{\gamma}_j \hat{\tau}_i)$. Under the Cox model, the $N$ residuals associated with a given item should follow a standard Gumbel distribution (for minimums).

2.2.4 Dependent Measures

To summarize, three different test forms were created, and each test was “administered” to four samples of examinees that differed in terms of the correlation between speed and ability ($\rho_{\theta r} = 0, .3, .6, \text{ or }.9$). Each of the resulting datasets was fit with the hierarchical 3PL/log-normal model and the hierarchical 3PL/Cox PH
model, yielding eight conditions per test form.\textsuperscript{11} To ensure that the outcomes were not specific to any particular sample, 10 different examinee samples were generated for each correlation, and the outcomes were averaged over the 10 replications (when appropriate).

The global and item-level fit of each RT model was investigated using the methods described in the previous section. Recovery of the ability parameters was investigated by dividing examinees into equally-spaced groups based on their true ability value; within each group, the bias of the ability estimates and the average of their posterior standard deviations was computed. These results were then averaged over the 10 replications. Finally, estimates of the correlation between speed and ability were obtained and then averaged over the 10 replications.

\section*{2.3 Simulation Study Results}

\subsection*{2.3.1 Response Time Model Fit}

First, the fit of each RT model to the simulated datasets is considered. Figure \ref{fig:empiricalCDF} displays the empirical CDF of the posterior predictive probabilities for each combination of test form (log-normal, Weibull, or modified Box-Cox) and fitted model (log-normal or Cox PH). Each plot is based on a single replicated dataset in the $\rho = 0$ condition; however, the correlation had no effect on the fit of the RT models. In Row 1, the true RT model is log-normal. The left column displays the fit of the log-normal model, and as expected, the fit is essentially perfect. In the right column, the fit of the Cox PH model is actually quite good. Given that the Cox model employs a misspecified error distribution, the good (global) fit of the model might be attributed to the flexibility of the nonparametric RT transformation.

\textsuperscript{11}The (non-hierarchical) 3PLM was also fit to each generated dataset. However, the 3PLM-only results were indistinguishable from those based on the hierarchical models in the $\rho = 0$ condition, so the 3PLM-only results are not reported.
Figure 2.1. Global fit of the log-normal and Cox PH models to each test form. The title of each subfigure indicates the fitted model (log-normal or Cox PH), and below that (in parentheses) is the true response time model (log-normal, Weibull, or Box-Cox). In each subfigure, the solid line indicates the cumulative proportion of posterior predictive probabilities for all observations in the dataset. If the cumulative proportion matches the identity line (dashed line), this indicates good (global) fit of the model.
In Row 2, the true RT model is Weibull. In this case, the Cox PH model is the “correct” model, and accordingly, the fit is very good. In contrast, the log-normal model shows clear evidence of misfit; specifically, there are too many probabilities close to one. This is because the Weibull distribution tends to be less skewed than the log-normal distribution (at least for the range of item parameters employed in this study). So for a fixed mean and variance, the highest density region of a Weibull distribution will tend to be shifted to the right, relative to that of a log-normal distribution. The result is that when a log-normal model is applied to RTs generated from a Weibull model, many RTs will be larger than expected.

In Row 3 of Figure 2.1, the true RT model is a combination of different Box-Cox transformations and either normal or Gumbel errors. As might be expected, the Cox PH model fits quite well: the nonparametric transformation can capture the Box-Cox transformation, and the error distribution is misspecified for only half of the items. Again, the log-normal model shows evidence of misfit, but the magnitude of misfit appears to be less than that for the Weibull test. This is because the log-normal model employs an incorrect error distribution for only half of the items. Apparently, the misspecified transformation of the log-normal model does not have a large effect on the global fit of the model.

Next, Figure 2.2 displays QQ plots of residuals for four items selected from the log-normal test. In Row 1, residuals based on the log-normal model are plotted against the corresponding \( N(0,1) \) quantiles; in Row 2, residuals based on the Cox PH model (for the same four items) are plotted against the corresponding Gumbel(0,1) quantiles. As the true RT model is log-normal, it is unsurprising that the log-normal residuals closely match their theoretical distribution. In contrast, the Cox residuals are clearly misfitting. Specifically, the largest negative residuals are consistently smaller (i.e., closer to zero) than what is expected. This is because the distribution of residuals based on log-normal RTs will more closely resemble a nor-
mal distribution than a Gumbel distribution. As a result, there are fewer very large negative residuals than what is expected under the Gumbel distribution (recall that the standard Gumbel density has a long left tail relative to the standard normal density; see Figure 1.2). Also, note that the items shown in Figure 2.2 were chosen to represent the range of misfit exhibited by the Cox PH model. Thus Items 1 and 8 are representative of the best- and worst-fitting items on the test, respectively. It is perhaps surprising that the consistent misfit exhibited by the Cox PH model did not yield greater evidence of global misfit in Figure 2.1. However, this is probably because the misfitting residuals comprise a relatively small proportion of the data.

Figure 2.3 displays QQ plots of residuals for four items selected from the Weibull test. In this case, the Cox PH model is the correct model, and the residuals closely match their corresponding Gumbel quantiles. In contrast, the log-normal model clearly does not fit any of the items. (Again, these particular items were chosen to represent the range of misfit exhibited by the misfitting model.) Specifically, the largest negative residuals are larger than expected, and the largest positive residuals are smaller than expected. This pattern can again be understood by referring to the difference between the standard Gumbel and standard normal distributions (Figure 1.2). The distribution of residuals based on Weibull RTs will more closely resemble a Gumbel density than a normal density; thus the right tail is too short and the left tail is too long (relative to what is expected under the normal density). For this test, the pattern of item-level misfit exhibited by the log-normal model is consistent across items and affects a large proportion of the data; as a result, the global misfit of the model is apparent in Figure 2.1.
Figure 2.2. QQ plots of residuals for four items from Test 1 (i.e., true response time model is log-normal). Each column displays QQ plots for the same item fitted with either the log-normal model (top row) or the Cox PH model (bottom row). Residuals based on the log-normal model are plotted against standard normal quantiles, and residuals based on the Cox PH model are plotted against standard Gumbel quantiles. If the residuals fall on the identity line, this indicates good fit of the model to the particular item.
Figure 2.3. QQ plots of residuals for four items from Test 2 (i.e., true response time model is Weibull). Rows and columns are structured in the same way as Figure 2.2.
Item-level QQ plots are not shown for Test 3 because they are very similar to those shown in Figures 2.2 and 2.3. Specifically, RTs generated with a normal error distribution and fit with the Cox PH model produced residuals very similar to those shown in the bottom row of Figure 2.2. And RTs generated with a Gumbel error distribution and fit with the log-normal model produced residuals very similar to those shown in the top row of Figure 2.3. Interestingly, when RTs were generated with a Box-Cox transformation using \( v = 0.25 \) or 0.5, there was no perceptible effect on the fit of the log-normal model. (The Box-Cox transformation had no effect on the fit of the Cox PH model, either, but this result was expected.) Thus, among the items in Test 3, the primary factor that determined whether the log-normal or Cox PH model fit a given item was the true error distribution, whereas the true RT transformation had no effect.

2.3.2 Recovery of Ability Parameters

For Test 1 (i.e., the true RT model is log-normal), Figure 2.4 presents a summary of ability estimates based on the hierarchical log-normal model (Column 1) and the hierarchical Cox PH model (Column 2). Row 1 presents the conditional bias of ability estimates, and Row 2 presents the conditional (average) posterior standard deviations (PSDs) of the ability estimates. And each subfigure presents results for \( \rho_{\theta_r} = 0, .6, \) and .9. (Results for \( \rho_{\theta_r} = .3 \) are very similar to those for \( \rho_{\theta_r} = 0 \) and are thus omitted for clarity.) First, the pattern of bias is very similar for the two models when \( \rho_{\theta_r} = 0.12 \) Even though the Cox PH model is misspecified, the correlation between speed and ability is zero. As a result, no information is “shared” between the RT and item response models, and there is no opportunity for a misfitting RT model to affect ability estimation. Under the log-normal model, bias is reduced for

\[12\] Though the magnitude of bias is rather large (almost half of a standard deviation on the ability scale), a strong \( N(0, 1) \) prior on ability is quite common for item response models.
examinees with extreme ability values, but only in the $\rho_{\theta \tau} = .9$ condition. It is somewhat surprising that the smaller correlations do not yield a reduction in bias. However, previous studies have found that RTs are not likely to yield practically large improvements in ability estimates unless the correlation between speed and ability is at least .5 (Ranger, 2013; van der Linden et al., 2010). Under the Cox PH model, the results are very similar: $\rho_{\theta \tau} = .6$ does not yield a reduction in bias, but the $\rho_{\theta \tau} = .9$ condition shows a reduction in bias for examinees at the extremes. The reduction for low-ability examinees is smaller than that under the log-normal model, but the reduction for high-ability examinees is actually greater than that under the log-normal model.

Next, the bottom row of Figure 2.4 displays the average PSDs of ability estimates for each model. Again, the results are very similar for the two models in the baseline condition (i.e., when $\rho_{\theta \tau} = 0$). As $\rho_{\theta \tau}$ increases, the log-normal PSDs decrease for all levels of ability. In the $\rho_{\theta \tau} = .6$ condition, PSDs are reduced by about 5%, and in the $\rho_{\theta \tau} = .9$ condition, PSDs are reduced by about 20%. The Cox PH results are very similar: increases in $\rho_{\theta \tau}$ yield similar decreases in the PSDs.

To summarize, even though the Cox PH model is the incorrect RT model, increases in the correlation between speed and ability yield reductions in the bias and PSDs of ability estimates; and these reductions are very similar in magnitude to those under the log-normal model. Put more simply, ability estimates based on the hierarchical model were robust to misspecification of the RT model. Of course, this conclusion is specific to Test 1. One might not expect the Cox PH model to have adverse affects on ability estimation given that the global fit of the model is quite good (see Figure 2.1, top-right panel). Test 2 might yield more interesting results, given that the misfit of the log-normal model is more severe (Figure 2.1, middle-left panel). A summary of the ability estimates from Test 2 is presented in Figure 2.5. However, the pattern of results is strikingly similar to that for Test 1. The only notable difference is that Test
Figure 2.4. Conditional bias (top row) and average posterior standard deviation (bottom row) of ability estimates for Test 1 (i.e., true response time model is log-normal). Columns 1 and 2 display results for the log-normal and Cox PH models, respectively. Each subfigure displays results for $\rho_{\theta \tau} = 0$, .6, and .9. (Results for $\rho_{\theta \tau} = .3$ are very similar to those for $\rho_{\theta \tau} = 0$ and are thus omitted for clarity.)
2 yields larger reductions in bias for examinees at the extremes, and the reduction appears to be larger for the Cox PH model. Otherwise, there is no evidence to suggest that the misfitting log-normal model has adverse effects on ability estimates. For Test 3, both models are misfitting, but again, there are no practically significant differences between the ability estimates from each model (Figure 2.6).

Figure 2.5. Conditional bias (top row) and average posterior standard deviation (bottom row) of ability estimates for Test 2 (i.e., true response time model is Weibull). Rows and columns are structured in the same way as Figure 2.4.
Recall that information from the RT model can be shared with ability estimates in two different ways: the examinee’s speed parameter and the population correlation between speed and ability. Because the choice of fitted model (i.e., correct versus misspecified) had no effect on ability estimates, this suggest that the choice of model also had no effect on speed estimates or the correlation estimate. This is indeed
what occurred. The bias and precision of the speed estimates (not shown) were examined in each condition, and there was little difference between estimates based on the log-normal model and those based on the Cox PH model. The correlation estimates were also examined, and these are presented in Table 2.4. Again, there is no evidence to suggest that the correlation estimates are more accurate when the log-normal or Cox PH model is the correctly specified model. However, there are two interesting trends to note. First, for each combination of $\rho_\theta \tau$ and true RT model, the Cox PH correlation estimate is equal to or lower than the log-normal estimate (though the difference is not large). There is no obvious explanation for this, but one (uninteresting) possibility is that the MCMC chains of the Cox PH model were not run long enough. This seems plausible because the Cox PH estimates nearly always underestimate the true correlation. A second notable trend in Table 2.4 is that for each true correlation value, the lowest correlation estimates nearly always occur when the true RT model is log-normal. Again, the effect is small and there is no obvious explanation, but this suggests an (unintended) systematic difference between the properties of the log-normal distributions that were used to generate RTs and the properties of the distributions in the other two tests.
TABLE 2.4

AVERAGE ESTIMATE OF THE CORRELATION BETWEEN SPEED AND ABILITY FOR EACH COMBINATION OF TRUE CORRELATION, TRUE MODEL, AND FITTED MODEL

<table>
<thead>
<tr>
<th>$\rho_{\theta_r}$</th>
<th>Log-normal Model</th>
<th>Cox PH Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log-normal</td>
<td>Weibull</td>
</tr>
<tr>
<td>0</td>
<td>-.007</td>
<td>-.002</td>
</tr>
<tr>
<td>.3</td>
<td>.300</td>
<td>.306</td>
</tr>
<tr>
<td>.6</td>
<td>.585</td>
<td>.593</td>
</tr>
<tr>
<td>.9</td>
<td>.895</td>
<td>.891</td>
</tr>
</tbody>
</table>
CHAPTER 3

USING RESPONSE TIMES FOR ITEM SELECTION IN COMPUTERIZED ADAPTIVE TESTING

3.1 Introduction

A prominent area of educational measurement that could benefit greatly from the information provided by response times is computerized adaptive testing (CAT). In particular, the literature contains a number of proposals for ways to use RTs to improve item selection in CAT. For example, van der Linden (2008) proposed a method that takes advantage of the correlation between ability and speed in Level 2 of the hierarchical model for item responses and RTs (van der Linden, 2007). Specifically, the method uses an examinee’s RTs, which are collected during the CAT, and the (previously estimated) correlation between ability and speed to construct an informative prior for the examinee’s ability parameter (see Equation 2.2). When an additional RT is collected, the prior is updated, and the ability estimate is updated as well. Because items are chosen to match (a function of) the current ability estimate, this method indirectly improves item selection by producing an ability estimate that more closely reflects the examinee’s true ability.

Alternatively, RTs can improve item selection in a more direct way by considering the time intensity of the items. For example, if the item selection criterion prefers items that are less time intensive, the amount of time required to complete the test will be reduced. This is the goal of the maximum information per time unit (MIT) criterion (Fan et al., 2012). In contrast with the method proposed by van der
Linden (2008), the MIT criterion does not explicitly employ the hierarchical model. Nonetheless, its effect on CAT outcomes (e.g., test completion times or item pool usage) does depend on second-order relationships such as the correlation between ability and speed and the correlation between item difficulty and item time intensity. Before describing the MIT criterion in greater detail, it is useful to cover the basic structure of adaptive tests.

In general, the goal of item selection in CAT is to estimate an examinee’s ability parameter as precisely as possible. This is accomplished by adapting or “tailoring” the choice of items to each examinee. Historically, the most popular item selection method has been the maximum-information (MI) criterion. The justification for this criterion is as follows. After an examinee has been administered \( m \) dichotomously-scored items, the likelihood of the ability parameter (assuming local independence) is given by

\[
L(\theta|\mathbf{u}) = \prod_{j=1}^{m} P_{j}^{u_j} Q_{j}^{1-u_j},
\]

where \( \mathbf{u} \) is the vector of \( m \) item responses, \( P_{j} = P(u_j = 1|\theta) \), \( Q_{j} = 1 - P_{j} \), and the functional form of \( P(u_j = 1|\theta) \) is determined by the chosen item response model (e.g., the 3PLM; see Equation 1.1).\(^{13}\) The corresponding test (or Fisher) information function is given by

\[
I_m = -E \left[ \frac{d^2 \log L(\theta|\mathbf{u})}{d\theta^2} \right] = \sum_{j=1}^{m} \frac{(dP_j/d\theta)^2}{P_{j} Q_{j}},
\]

and the standard error of the maximum likelihood ability estimate is approximated by the square root of the inverse of test information. According to Equation 3.2, the contribution of each item to test information is additive; thus the motivation behind choosing the item with maximum information is to produce the largest possible

\(^{13}\)The functions \( P_{j} \) and \( Q_{j} \) depend on item parameters as well as the ability parameter; however, dependence on the item parameters is omitted here for clarity.
reduction in the standard error of the ability estimate.

Clearly, the goal of the MI criterion is efficiency: accrue as much information as possible with a fixed number of items, or alternatively, achieve a fixed amount of information using as few items as possible. But this goal overlooks a potentially important feature of items, namely, the amount of time that they require. For example, a highly-informative item may not look so attractive if it also takes a long time to complete. It may be a more efficient use of time to administer a few less-informative items that can each be completed more quickly. The maximum information per time unit (MIT) criterion (Fan et al., 2012) seeks to address this shortcoming of the MI criterion: instead of simply maximizing information, the goal is to maximize the ratio of information to the examinee’s predicted RT. The result of choosing items in this way is that a fixed number of items can be completed more quickly, or alternatively, more items can be completed in a fixed amount of time. The MIT criterion does not try to explicitly control the time required to complete the test, but nonetheless, this simple modification of the widely used MI criterion has the potential to dramatically reduce test completion times.

The key feature of the MIT criterion is the examinee’s predicted RT. Fan et al. (2012) proposed using the expected RT based on van der Linden’s (2006) log-normal model. For an item \( k \) that is being considered for administration, the expected RT of examinee \( i \) has a simple formula:

\[
E(T_{ik} | \hat{\tau}_i^{(m)}) = \exp \left( \beta_k - \hat{\tau}_i^{(m)} + \frac{1}{2\alpha_k^2} \right),
\]  

(3.3)

where \( \hat{\tau}_i^{(m)} \) is the maximum likelihood estimate of the examinee’s speed parameter after \( m \) items have been administered. Based on the \( m \) observed RTs, the speed

\[14\]For examples of item selection methods that put explicit constraints on test completion times, see van der Linden (2009b) and van der Linden and Xiong (2013).
estimate also has a simple solution:

\[
\hat{\tau}_i^{(m)} = \frac{\sum_{j=1}^{m} \alpha_j^2 (\beta_j - \log t_{ij})}{\sum_{j=1}^{m} \alpha_j^2}.
\]  \hspace{1cm} (3.4)

When an item is administered, both the item response and the RT are recorded. Just as the ability estimate is updated after each additional response, the speed estimate is updated after each additional RT. To select the next item, the item information function of each candidate item is evaluated with the updated ability estimate, and the expected RT for each candidate item is estimated using the updated speed estimate. It is interesting to note that although a precise speed estimate is required to obtain a precise estimate of the expected RT, the MIT criterion does not take into account the information that items contain about speed. (That is, the goal of the MIT criterion is to obtain a precise ability estimate; the speed parameter is incidental.) However, it is encouraging to note that the information-based standard error (SE) of the maximum likelihood speed estimate does not depend on the speed estimate itself:

\[
\widehat{\text{SE}}(\hat{\tau}_i^{(m)}) = \left[ \sum_{j=1}^{m} \alpha_j^2 \right]^{-1/2}.
\]  \hspace{1cm} (3.5)

Thus, assuming that examinees of high versus low speed (or high versus low ability) are not administered items with systematically different ranges of \(\alpha_j\) parameters, one can expect the precision of \(\hat{\tau}_i\) to be relatively uniform for all examinees.

As the previous paragraph shows, the MIT criterion can be implemented in a very straightforward manner using the log-normal RT model. However, the accuracy of an examinee’s predicted RT depends on how well the model-implied RT distribution resembles the examinee’s (hypothetical) distribution of RTs for the item. In the context of CAT, an item pool might consist of hundreds or even thousands of items, and it is very unlikely that a single parametric model would fit every item. A much more flexible alternative to the log-normal model is the semiparametric Cox PH
model. Under the Cox model, an examinee’s expected RT for a candidate item $k$ can be computed as follows (Wang, Fan, et al., 2013):

$$E(T_{ik}|\hat{\tau}_i^{(m)}) = \int_0^\infty \exp \left[ -H_{0k}(t) \cdot \exp \left( \gamma_k \hat{\tau}_i^{(m)} \right) \right] \, dt, \quad (3.6)$$

where $H_{0k}(\cdot)$ is the cumulative baseline hazard function for item $k$. Two features of this formula should be pointed out. First, it is common to estimate $H_{0k}(\cdot)$ using the nonparametric Breslow estimator (Breslow, 1972), which yields a step function and requires the RTs and speed estimates of the original calibration sample. However, Wang, Fan, et al. (2013) recommend smoothing the estimated Breslow function with a B-spline function so that $H_{0k}(\cdot)$ can be evaluated quickly and without the calibration data. Second, because $H_{0k}(\cdot)$ is a nonparametric function, the integral in Equation 3.6 must be evaluated numerically.

The expected RT based on the Cox PH model also requires an estimate of the examinee’s speed parameter. Given the vector of $m$ observed RTs ($t_i$), the likelihood function for $\tau_i$ under the Cox model is:

$$f(t_i|\tau) = \prod_{j=1}^{m} h_{0j}(t_{ij}) e^{\gamma_j \tau_i} \cdot \exp \left[ -H_{0j}(t_{ij}) e^{\gamma_j \tau_i} \right]. \quad (3.7)$$

Taking the first derivative of the log-likelihood and setting it equal to zero yields the following estimating equation:

$$0 = \sum_{j=1}^{m} \gamma_j [1 - H_{0j}(t_{ij}) e^{\gamma_j \tau_i}]. \quad (3.8)$$

No closed-form solution is available, but $\hat{\tau}_i$ can be found iteratively using Newton’s method. It is also useful to consider the factors that affect the precision of $\hat{\tau}_i$. Under
the Cox model, the information-based SE is:

\[
\widehat{SE}\left(\hat{\tau}_i^{(m)}\right) = \left[\sum_{j=1}^{m} \gamma_j^2 H_0(t_{ij}) \cdot \exp\left(\gamma_j \hat{\tau}_i^{(m)}\right)\right]^{-1/2}.
\]  

(3.9)

In contrast with the SE under the log-normal model, this SE depends on the examinee’s speed; specifically, the SE decreases as speed increases. However, this relationship is counterbalanced by the cumulative baseline hazard, which is a monotone increasing function of RTs. Thus the cumulative hazard is large for slower examinees (which reduces the SE) and small for faster examinees (which increases the SE).

Given the diversity of RT distributions that the items in a large, operational item pool might exhibit, it was of interest to study the consequences of using a misspecified model for RT prediction when employing the MIT item selection criterion in CAT. Similar to the study in Section 2.2, two item pools were created, and the pools differed in terms of the “true” model used to generate RTs. Using each of the item pools, two adaptive tests were simulated: one CAT employed the MIT criterion with predicted RTs based on the true model, and the other CAT employed the MIT criterion with predicted RTs based on a misspecified model. By comparing the results of the two CATs (for a given item pool), it could be determined whether using an incorrect model for RT prediction had an effect on practical CAT outcomes including test completion times, recovery of ability parameters, and item exposure rates.

The simulation study was conducted with two additional goals in mind. First, the MIT criterion has only been implemented with the log-normal RT model (Fan et al., 2012). Thus it was of interest to illustrate how the more flexible Cox PH model could be used instead. A second goal was motivated by the fact that the predicted RT in the denominator of the MIT criterion can be obtained in any number of ways. For example, the prediction need not be based on the expected RT; it could instead be the mode, median, or some other quantile of the RT distribution. In contrast,
the prediction could be made without an RT model; for example, it may suffice to use an estimate of the average RT in the examinee population. The latter method (i.e., using an average RT as the prediction) was included in the study because it is far simpler than model-based prediction: it does not require calibration of an RT model, nor does it require computation of an examinee-specific RT prediction during the adaptive test. Also, if using an incorrect model for RT prediction has negative consequences on CAT outcomes, this simpler method may be an attractive alternative because it does not require specification of an RT model.

3.2 Simulation Study

Two item pools were created, each containing 450 items. The 3PLM was chosen as the true model for item responses, and the same set of 3PLM item parameters was used in each pool. In contrast, the pools differed in terms of the true RT model: log-normal (Equation 2.4) or Weibull (Equation 2.7). Regardless of the true RT model, each item pool was used to administer an adaptive test that employed one of four item selection methods. Three of these methods were based on the MIT criterion, but each obtained an examinee’s predicted RT in a different way:

1. expected RT based on the log-normal model (MIT-LN method),
2. expected RT based on the Cox PH model (MIT-PH method), or
3. mean RT obtained from a separate calibration sample (MIT-RT method).

Note that the first two methods use an RT model to make an examinee-specific prediction. As such, these methods require the examinee’s speed estimate to be updated throughout the test. The third method is much simpler: the “predicted” RT is the average RT for that item based on a separate calibration sample. (How these average RTs were obtained will be explained in the next section.) Unlike the model-based methods, this prediction is not tailored to the examinee. However, the average
RTs can simply be included in the item pool as an additional item “parameter” (i.e., there is no need to compute an expected RT during the test), and the examinee’s speed parameter is not required. The fourth method employed the traditional MI criterion. In contrast with the other three methods, this method does not consider the time intensity of the items or the speed of the examinee. A summary of these eight conditions (two item pools × four item selection methods) is provided in Table 3.1.

TABLE 3.1

STUDY 2: SUMMARY OF SIMULATED CONDITIONS

<table>
<thead>
<tr>
<th>Pool 1</th>
<th>True RT Model</th>
<th>Item Selection Criterion + RT Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-normal</td>
<td>MIT + Log-normal model (correct)</td>
<td>MIT + Cox PH model (misspecified)</td>
</tr>
<tr>
<td></td>
<td>MIT + mean RTs from calibration sample</td>
<td>MIT only</td>
</tr>
<tr>
<td>Pool 2</td>
<td>Weibull</td>
<td>MIT + Log-normal model (misspecified)</td>
</tr>
<tr>
<td></td>
<td>MIT + Cox PH model (correct)</td>
<td>MIT + mean RTs from calibration sample</td>
</tr>
<tr>
<td></td>
<td>MIT only</td>
<td></td>
</tr>
</tbody>
</table>

For the log-normal item pool, the MIT-PH method represents a “misfit” condition: during the adaptive test, an examinee’s RTs are generated using the true log-normal
item parameters, whereas RT prediction and speed estimation are based on estimated Cox PH item parameters. (How these parameter estimates were obtained will be explained in the next section.) To determine the effects of using an incorrect RT model to predict RTs and estimate speed, the MIT-PH results were compared with the MIT-LN results. In this latter condition, RT generation, RT prediction, and speed estimation are all based on the true log-normal item parameters.

Conversely, for the Weibull item pool, the MIT-LN method represents a misfit condition: RTs are generated using the true Weibull item parameters, whereas RT prediction and speed estimation are based on estimated log-normal item parameters. To determine the effects of using an incorrect RT model to predict RTs and estimate speed, these results were compared with the MIT-PH results. In this latter condition, RT generation is based on the true Weibull item parameters, and RT prediction and speed estimation are based on estimated Cox PH item parameters.\textsuperscript{15}

For both item pools, the MIT-RT method was investigated as a possible alternative to model-based RT prediction. If the MIT-RT method performs as well as (or nearly as well as) RT prediction based on the correct model, the former would be a very attractive choice because it does not require calibration of an RT model, nor does it require RT prediction or speed estimation during the adaptive test. If it also happens that RT prediction based on an incorrect model has negative effects on CAT outcomes, an argument for the MIT-RT method would be even stronger given that the true RT model is never known in practice.

3.2.1 Construction and Calibration of Item Pools

As stated in the previous section, two item pools were created, each containing 450 items. The 3PLM $a$, $b$, and $c$ parameters were drawn from the retired item

\textsuperscript{15}As in the previous study, the Cox PH model is not the “correct” model when RTs are generated from a Weibull distribution. However, the Weibull model is a parametric proportional hazards model, so any misfit of the (semiparametric) Cox PH model is due solely to sampling error.
bank of a large-scale achievement test. $b$ parameters ranged from $-3.11$ to 3.28 with a mean of 0.12, $a$ parameters ranged from 0.47 to 3.12 with a mean of 1.72, and $c$ parameters ranged from 0.09 to 0.31 with a mean of 0.19. Also, the correlation between $b$ and $a$ was $0.48$, the correlation between $b$ and $c$ was $-0.23$, and the correlation between $a$ and $c$ was $-0.17$. The same set of parameters was used for each of the pools. In contrast, each item pool used a different model to generate RTs.

For the first item pool, the time intensity parameters ($\beta$) of the log-normal model (Equation 2.4) were generated by adding random normal deviates to a linear combination of the $a$ and $b$ parameters. The coefficients for the linear combination and the variance of the normal deviates were chosen so as to achieve a correlation of $0.5$ between $\beta$ and $b$ and a correlation of $0.3$ between $\beta$ and $a$. These correlations imply that difficult items tend to be more time-consuming than easy items; also, highly discriminating items are slightly more time-consuming than less-discriminating items. Additionally, the $\beta$ parameters were generated to have a mean and standard deviation (SD) of 4 and 0.33, respectively; these values imply that the RTs are on the scale of seconds (i.e., instead of milliseconds or minutes). Lastly, the inverse scale parameters ($\alpha$) were generated independently of all other item parameters from a uniform distribution between 1.75 and 3.25. These distributions of the log-normal item parameters and their correlations with the 3PLM item parameters were chosen to reflect the results of empirical applications of the log-normal model (e.g., van der Linden & Xiong, 2013).

For the second item pool, the $\beta$ parameters of the Weibull model (Equation 2.7) were generated in a similar fashion to achieve correlations of $0.5$ and $0.3$ with $b$ and $a$, respectively. But in contrast with the log-normal parameters, the mean and SD of $\beta$ were chosen to be 4.5 and 0.25, respectively, and the $\alpha$ parameters were drawn from a uniform distribution between 2.75 and 4. These different ranges for the parameters of the Weibull model were chosen to achieve RT distributions in the same general
range as those of the log-normal items.

Again, each item pool was used to administer a CAT that employed one of four item selection methods. For two of these methods, either the log-normal model or the Cox PH model was used to predict examinee RTs and estimate examinee speed parameters. When the log-normal or Cox PH model differed from the true RT model, it was necessary to use the parameters of the true RT model to generate RTs for a hypothetical “calibration sample”; these data were then used to estimate the parameters of the log-normal or Cox PH model.

For each of the item pools, 15 randomly-equivalent calibration samples were created, each of size \(N = 1000\) with speed parameters sampled from \(N(0, 0.25^2)\). Then each calibration sample was “administered” a random set of 30 items from the pool (15 samples \(\times\) 30 items = 450 items). Based on these simulated RTs, the average RT was recorded for each item, and these averages were used as the predicted RTs in the MIT-RT item selection method. For the log-normal pool, the true log-normal item parameters were used for the MIT-LN method. Thus, only the Cox PH model was fit to the log-normal calibration data. This yielded \(\gamma\) (slope) estimates for each item as well as the estimated cumulative baseline hazard (CBH) function for each item.

In contrast, both the log-normal and Cox PH models were fit to the Weibull calibration data. The fitted log-normal model yielded \(\alpha\) and \(\beta\) estimates, and again, the fitted Cox PH model yielded \(\gamma\) estimates and the estimated CBH functions. Finally, full Bayesian estimation of each RT model was conducted using MCMC methods as described in Subsection 2.2.3. In contrast with the previous study, the RT models were fit individually (i.e., not as part of a hierarchical model with the 3PLM), but otherwise, the details are the same.
3.2.2 Adaptive Test Procedure

Each simulated CAT administered 45 items to four different samples of 5000 examinees. These samples differed in terms of the correlation between speed and ability: $\rho_{\theta\tau} = 0, .3, .6, \text{ or } .9$. Note that in contrast with the previous study, there was no expectation that larger correlations would improve ability estimates. This is because RTs were (in the current study) used only to reduce test completion times. Nonetheless, the correlation was manipulated in order to investigate what effects it might have on outcomes of the CATs. For example, if $\rho_{\theta\tau} = 0$, high- and low-ability examinees work at the same speed (on average). Because the correlation between item difficulty and time intensity is positive (.5), high-ability examinees will receive more time-intensive items and will have longer test completion times. If $\rho_{\theta\tau} > 0$, high-ability examinees will still receive more time-intensive items, but now they will work more quickly than low-ability examinees (on average) and will not necessarily have longer test completion times. Thus it was of interest to see what effects $\rho_{\theta\tau}$ might have on test completion times and how it might (indirectly) affect ability estimates.

It is possible for $\rho_{\theta\tau}$ to be negative, and unlike the previous study, the sign of $\rho_{\theta\tau}$ could have an impact on the results of the current study. However, it was desirable to keep the number of simulated conditions manageable, and because a positive $\rho_{\theta\tau}$ seems to be more common among empirical applications of the hierarchical model, only positive correlations were simulated.

Regardless of the correlation, examinee ability parameters were fixed at 5000 equally-spaced quantiles of the $N(0,1)$ distribution. That is, ability values were defined as $\theta_p = \Phi^{-1}(p)$ for $p = \frac{1}{5000}, \frac{2}{5000}, \ldots, 1$, where $\Phi^{-1}$ is the inverse cumulative distribution function of $N(0,1)$. For a given value of the correlation $\rho_{\theta\tau}$, each examinee’s speed parameter was sampled from a normal distribution conditional on their ability parameter. Specifically, the speed parameter for examinee $i$ was sampled from a normal distribution with mean $\rho_{\theta\tau}\sigma_{\tau}\theta_i$ and standard deviation $\sigma_{\tau}\sqrt{1 - \rho_{\theta\tau}^2}$. This
yielded a bivariate normal distribution with the desired correlation between speed and ability and a $N(0, \sigma^2_\tau)$ marginal distribution for speed. The SD of speed was chosen to be $\sigma_\tau = 0.25$.

At the beginning of the test, a random “pretest” of five items was drawn from the item pool. This was done to obtain an initial maximum likelihood estimate (MLE) of ability and (for the MIT-LN and MIT-PH conditions only) an initial MLE of speed. If an examinee provided all correct or all incorrect answers, $\hat{\theta}_{ML}$ does not exist, so instead, ability was estimated by its posterior mean with a uniform prior on $[-4,4]$ until a finite $\hat{\theta}_{ML}$ could be obtained. Thereafter, items were selected by maximizing one of the four item selection criteria: MIT-LN, MIT-PH, MIT-RT, or MI. Note that even though the MIT-RT and MI methods did not require an examinee’s speed estimate, RTs were still generated for examinees in these conditions (using the true RT model) so that their test completion times could be compared with those in the other two conditions. Also, because many of the true RT distributions have non-negligible densities for unrealistically small RTs (e.g., a few seconds) and/or have very long right tails, the generated RTs were truncated from below and above at 10 seconds and 300 seconds, respectively. (For a more thorough explanation, refer to Footnote 8.)

3.2.3 Dependent Measures

To summarize, each of eight different adaptive tests (two item pools × four item selection methods) was administered to four different samples of examinees, and these samples differed in terms of the correlation between speed and ability ($\rho_{\theta \tau} = 0, .3, .6, \text{ or } .9$). The outcomes of interest were test completion times, recovery of the ability parameters, and item pool usage. An examinee’s test completion time was obtained by summing his or her RTs over the 45 items, and the completion times were averaged over the 5000 examinees in each condition. Recovery of the
ability parameters in each condition was summarized by the root mean squared error (RMSE) of ability estimates versus their true values. Additionally, item pool usage was measured by comparing the observed and ideal item exposure rates. An item’s exposure rate is defined as the number of times that the item is administered, divided by the number of examinees. Ideally, every item in the pool should have the same exposure rate. This ideal rate is found by dividing the test length by the size of the item pool (in this case, the ideal is 45/450 = .1). In each condition, the following statistic was used to summarize the discrepancy between the observed exposure rates and the ideal rate (Chang & Ying, 1999):

$$\chi^2 = \sum_{j=1}^{M} \frac{(r_j - r)^2}{r},$$

(3.10)

where $r_j$ is the observed item exposure rate for item $j$, $r$ is the ideal rate, and $M$ is the number of items in the pool. Lastly, test completion times, ability recovery, and item pool usage were examined at test lengths of 15 to 45 items in increments of 10. Note that this did not require any additional CAT simulations; rather, the dependent measures were computed after 15, 25, 35, or 45 items had been administered. However, there was no meaningful change in the pattern of results for different test lengths, so only those results based on the full test length of 45 items are presented.

3.3 Simulation Study Results

Table 3.2 displays the average test completion time (in minutes) for each combination of item pool, item selection method, and correlation between speed and ability. First, completion times in the MI condition are noticeably longer than those in the MIT conditions, regardless of the item pool. This is because the MI criterion simply chooses the most informative item, regardless of how long the item might take to complete. Next, for the log-normal pool, the MIT-PH method produced slightly shorter
tests than the MIT-LN method. This might seem to suggest that using the incorrect model to predict RTs caused the test to be shorter; however, the MIT-PH method tended to produce shorter tests than the MIT-LN method for the Weibull pool as well. So instead, these results suggest that, compared with the MIT-PH method, the MIT-LN method tends to select slightly more time-consuming items, regardless of the true RT model. More importantly, these results suggest that whether or not the true model is used to predict RTs has no bearing on test completion times.

A surprising result in Table 3.2 is that the MIT-RT method produced completion times that are remarkably similar to those produced by the model-based MIT methods. This suggests that when employing the MIT method, it is sufficient to predict
examinee RTs with only a rough measure of how long an item takes to complete in the examinee population; going a step further by using a model to tailor the prediction to each examinee has no added benefit. Finally, for each item selection method, the average completion time decreased slightly as the correlation between speed and ability increased. But the important conclusion is that there was no obvious interaction between the correlation and the item selection method.

Next, Table 3.3 displays the RMSE of ability estimates for each of the simulated CATs. As expected, the MI method invariably yields the most precise ability estimates, though the difference relative to the MIT methods is not large. This difference occurred because the MIT method does not necessarily choose items that provide a lot of information; rather, what matters is how much information the item can provide in a unit of time. Similar to the pattern of test completion times, there are no systematic differences among the three MIT methods in terms of ability recovery, and there is no obvious interaction between the item selection method and the correlation between speed and ability.

Lastly, Table 3.4 displays the $\chi^2$ statistic (Equation 3.10) that was used to summarize item pool usage in each condition. Note that higher $\chi^2$ values are indicative of larger discrepancies between the observed item exposure rates and the ideal exposure rate. First, the MI method clearly uses the item pool more efficiently than the MIT methods. This is because the MI method favors highly informative items, whereas the MIT method favors a smaller subset of the pool: items that are both highly informative and can be completed quickly. Again, there are no systematic differences among the MIT methods or among the correlations.
### TABLE 3.3

RMSE OF ABILITY ESTIMATES FOR EACH COMBINATION OF ITEM POOL, ITEM SELECTION METHOD, AND ABILITY-SPEED CORRELATION

<table>
<thead>
<tr>
<th></th>
<th>Log-normal Pool</th>
<th>Weibull Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MIT-LN</td>
<td>MIT-PH</td>
</tr>
<tr>
<td>$\rho_{\theta\tau}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.198</td>
<td>0.201</td>
</tr>
<tr>
<td>.3</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>.6</td>
<td>0.197</td>
<td>0.200</td>
</tr>
<tr>
<td>.9</td>
<td>0.199</td>
<td>0.200</td>
</tr>
</tbody>
</table>

NOTE: “RMSE” refers to “root mean squared error.” Also, abbreviations for the item selection methods are defined in Section 3.2.

### TABLE 3.4

SUMMARY OF ITEM POOL USAGE FOR EACH COMBINATION OF ITEM POOL, ITEM SELECTION METHOD, AND ABILITY-SPEED CORRELATION

<table>
<thead>
<tr>
<th></th>
<th>Log-normal Pool</th>
<th>Weibull Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MIT-LN</td>
<td>MIT-PH</td>
</tr>
<tr>
<td>$\rho_{\theta\tau}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>98.27</td>
<td>98.12</td>
</tr>
<tr>
<td>.3</td>
<td>98.44</td>
<td>97.76</td>
</tr>
<tr>
<td>.6</td>
<td>98.38</td>
<td>98.07</td>
</tr>
<tr>
<td>.9</td>
<td>98.17</td>
<td>97.90</td>
</tr>
</tbody>
</table>

NOTE: Item pool usage is summarized by the $\chi^2$ statistic defined in Equation 3.10. Also, abbreviations for the item selection methods are defined in Section 3.2.
Chapter 2 described a study to investigate the effects of RT model misspecification on 3PLM ability estimates in the context of van der Linden’s (2007) hierarchical model. When the correct RT model was fit, increases in the correlation between speed and ability yielded reductions in the bias and variance of ability estimates, replicating the findings of van der Linden et al. (2010). Surprisingly, when the incorrect RT model was used, increases in the correlation produced nearly identical reductions in the bias and variance of ability estimates. That is, whether or not the true RT model was used, ability estimates benefited from the added information from RTs. Upon closer inspection, the robustness of the ability estimates was due to the robustness of the correlation and speed estimates. This robustness may be due to the type and/or magnitude of model misspecification that was simulated in this study. It may be fruitful to investigate other forms of misspecification. For example, the RT models considered in this study assume that examinees work at a constant speed; however, it is not uncommon for examinees to exhibit a “warm-up” effect in which they work more slowly at the beginning of the test (e.g., van der Linden et al., 2007). If the assumption of constant speed is violated, a single speed parameter does not adequately describe an examinee’s behavior, and its relationship with ability may be misleading.

Chapter 3 described a study to investigate the effects of RT model misspecification on the performance of the MIT item selection criterion. This study replicated earlier findings by Fan et al. (2012) concerning differences between the MI and MIT
criteria. Specifically, the MI criterion tends to produce slightly more accurate ability estimates and use the item pool somewhat more efficiently, but this comes at the cost of noticeably longer test completion times. Concerning the MIT criterion, using the true or a misspecified model for RT prediction yielded only trivial differences in test completion times, ability recovery, and item pool usage. This suggests that the log-normal and Cox PH models provided very similar predictions, regardless of the true RT model. Though surprising, this result is also encouraging because it provides some evidence that the MIT criterion can successfully reduce test completion times even when the RT model is misspecified.

A perhaps more surprising finding was that using a simple average RT as the prediction produced test completion times strikingly similar to those produced by model-based prediction. For someone who works very quickly, the average RT will be higher than the model-based prediction. And for someone who works very slowly, the average RT will be lower than the model-based prediction. Why did these discrepancies not matter? For the items in the simulated pools, it turns out that the average RTs were very similar to the expected RTs at $\tau = 0$ (the mean value of speed), and this was true whether the expectation was based on the log-normal model or the Cox PH model. So a more specific question is: why would it not matter if every examinee’s expected RT was evaluated at $\tau = 0$ instead of his or her own speed estimate? For the log-normal model, the reason is that speed has a multiplicative effect on the expected value (see Equation 3.3). For example, the expected RT at $\tau = -0.5$ is $e^{0.5}$ times larger than the expected RT at $\tau = 0$, and this is true for any given values of the item parameters $\alpha$ and $\beta$. Thus when computing the MIT criterion for all candidate items in the pool, it does not matter what value of speed is used. Shifting $\tau$ by a constant $\delta$ has the effect of multiplying the information/RT ratios by the constant $e^{-\delta}$; accordingly, the rank-ordering of the items (with respect to the MIT criterion) does not change. For the Cox PH model, it is not strictly true
that speed has a multiplicative effect on the expected value; however, this can be expected to hold at least approximately when the Cox model is fit to RTs that were generated from a model possessing this property. Indeed, this was the case for both of the item pools in this study: like the log-normal model, the Weibull model also has this property.
REFERENCES


71


