Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 20

A Model for the Joint Distribution of Item Response and Response Time using a One-Parameter Weibull Distribution

Tianyou Wang [†]

April 2006

[†]Tianyou Wang is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: tianyou-wang@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

This paper presents a model for the joint distribution of item response and response time. This model extends the four parameter logistic response time (4PLRT) model (Wang & Hanson, 2005) that treats response time only as a conditioning variable. This extended model for the joint distribution includes both the conditional probability of correct response and the marginal distribution of response time. A one parameter Weibull distribution is used for the marginal distribution of response time. By modelling the joint distribution, it is not necessary to assume that response time is independent of person parameters as required in Wang and Hanson (2005) and thus expands the applicability of the model. An EM algorithm for parameter estimation is developed and programmed. Real test data were used to fit the model and calibrate item parameters. A simulation study was undertaken to evaluate the precision of the item parameter estimation procedure.

1 Introduction

One of the benefits of computerized testing is that it makes response time data available at individual item level. This new availability of a different dimension of observable data opens up new opportunities for the theory and practice of educational measurement. As a result, there has been an increased research interest in how to make use of the response time data. Schnipke and Scrams (1998) provided a comprehensive review of the history and present state of research on response time. Most of the previous modelling approaches on response time have focused response time as a dependent variable. Except for a few cases, response accuracy and response time are not modelled simultaneously. Those a few exceptions include Thissen (1983), Verhelst, Verstralen, and Jansen (1997), and Roskam (1997). The approaches by Verhelst et al and Roskam are similar in that they both use Rasch type of models and as response time goes to infinity, the probability of correct response will approach one. For that reason, they all applied only to speed tests where unlimited response time almost always guarantee correct response. Thissen's model applies to power tests and is most relevant to the model proposed in this paper. Thissen (1983) used the commonly used three parameter logistic model for the marginal distribution of response accuracy, and used a lognormal model for the marginal distribution of response time. One limitation of Thissen's model is that it assumes that response accuracy and response time are independent, so that the joint distribution is the product of the marginal distributions of two variable, although the two marginal distributions shared some common parameters. There is ample reason to suspect that this assumption does not generally hold.

More recently, Wang and Hanson (2005) proposed a four parameter logistic response time (4PLRT) model. In their model, response accuracy and response time are modelled simultaneously, and the response time is an independent variable that affects the probability of a correct response. In their formulation, as response time goes to infinity, the probability of correct response approach the probability of a regular three parameter logistic (3PL) model. For that reason, their model applies to power test where unlimited response time does not guarantee correct response. In formulating the EM algorithm for calibrating the item parameters, their procedure requires an assumption that the item response time is independent of person parameters. This assumption is necessary in order to avoid modelling response time. This assumption, however, poses a severe limitation to the applicability of their model. In order to avoid this limitation, it is needed to model the joint distribution of response accuracy and response time.

Bloxom (1985) pointed out different ways in modelling the joint distribution of response accuracy and response time. One way is to model the conditional distribution of response accuracy given response time and then multiply it with the marginal distribution of response time. The other is to model the conditional distribution of response time given response accuracy and then multiply it with the marginal distribution of response accuracy. With either approach, there are a variety of models that can be used in modelling the conditional and marginal

distributions. Most of these possibilities have not been thoroughly explored in the literature.

The present paper will explore the former way of modelling the joint distribution. It adopts a slightly revised version of the Wang and Hanson's (2005) 4PLRT model for the conditional distribution of response accuracy given response time. An one parameter Weibull distribution is used for the marginal distribution of response time. The joint distribution is, of course, the product of the conditional and the marginal distributions.

2 The model for the Joint Distribution of Response Accuracy and Response Time

Let y_{ij} be the dichotomous item response variable, with 1 for a correct response, 0 for an incorrect response. Let t_{ij} be the response time variable. In this extended 4PLRT model, the joint distribution of y_{ij} and t_{ij} is expressed as:

$$f(y_{ij}, t_{ij}|\theta_i, \rho_i, \boldsymbol{\delta}_j) = f(y_{ij}|t_{ij}, \theta_i, \boldsymbol{\delta}_j) f(t_{ij}|\theta_i, \rho_i, \boldsymbol{\delta}_j), \tag{1}$$

where examinee with ability and speed parameters θ_i and ρ_i, $\boldsymbol{\delta}_j = (a_j, b_j, c_j, d_j)$ are item parameters for item j. The person ability parameter θ_i determines the probability of examinee i answering item j correctly. The latent variable ρ_i will be called the speed paramter for reasons that will be discussed later.

The conditional distribution of y_{ij} conditioned on t_{ij} can be expressed as

$$f(y_{ij}|t_{ij}, \theta_i, \boldsymbol{\delta}_j) = P(t_{ij}, \theta_i, \boldsymbol{\delta}_j)^{y_{ij}} [1 - P(t_{ij}, \theta_i, \boldsymbol{\delta}_j)]^{1-y_{ij}}, \tag{2}$$

where

$$P(t_{ij}, \theta_i, \boldsymbol{\delta}_j) = Prob(y_{ij} = 1|t_{ij}, \theta_i, \boldsymbol{\delta}_j) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j[\theta_i - (d_j/t_{ij}) - b_j]}}. \tag{3}$$

Note this is somewhat different as the conditional distribution in Wang and Hanson (2005). The difference is that ρ_i is not in the exponent part of the logistic function in Equation 3.

There are a variety of distribution forms that have been used to model response time. The most common ones are the log-normal distribution, the Gamma distribution, and the Weibull distribution. Van Breukelen (1989) and Roskam (1997) used a one parameter Weibull distribution to model the response time distribution. In this paper, we adopt one parameter Weibull distribution for the marginal distribution of response time, which is expressed as:

$$f(t_{ij}|\theta_i, \rho_i, \boldsymbol{\delta}_j) = \lambda t_{ij} e^{-\lambda t_{ij}^2/2}. \tag{4}$$

The one parameter Weibull distribution is a special case of the more general three-parameter Weibull distribution with the location parameter set to zero and shape parameter set to 2. The parameter λ is called a scale parameter that determines both the mean and variance of the distribution. The smaller the λ, the larger the mean and variance. This one parameter has a hazard function:

$$h(t_{ij}) = \lambda t_{ij}, \tag{5}$$

which is a linear function of t_{ij}. That means, given that a response has not been given at time t_{ij}, the likelihood of it being given in the next moment increases as t_{ij} increases. The λ parameter is the slope of this linear function. The larger the λ, the more likely the response will be given in the next moment, which means that the response time distribution will have smaller mean and variance.

The difficult issue remains is how to model the λ parameter in relation to the other person and item parameters. In Roskam's (1997) model, λ is a linear function of $(\theta - b)$ divided by some person persistence parameter. That means, the hard the item, the more response time, the higher the person ability, the less the response time. This is probably a reasonable way to model response time when there is basically no time restriction on the whole test and the examinees can spend as much time as they want to on the test. However, this paper mainly concerns a situation where there is strict time limit and the test is a power test rather than a speed test. In real test situation, most often there is a time restriction on standardized tests and classroom assessment. Under that situation, Wang and Zhang (2006) used analytical approach to show that it is to the examinees' best advantage to spent more time on items that match their ability level and spend less time on items either too easy or too hard. Other empirical studies show that test-wise examinees actually follow this pacing strategy when there is time restriction. For this reason, we propose a simple model for the λ parameter:

$$\lambda = \rho_i(\theta_i - b_j)^2. \tag{6}$$

This expression means that the higher the ρ_i, the higher the λ, and the less the mean and variance for the response time distribution. In that sense, the examinee parameter ρ_i can be viewed as a speed parameter. Also, the more the θ_i is different from b_j, the larger the λ, and the less the mean and variance for the response time distribution. That means that examinees will generally spend more time on items that match their ability level. Thus, the marginal distribution of response time can be expressed as:

$$f(t_{ij}|\theta_i, \rho_i, \boldsymbol{\delta}_j) = \rho_i(\theta_i - b_j)^2 t_{ij} e^{-\rho_i(\theta_i - b_j)^2 t_{ij}^2/2}. \tag{7}$$

It should be pointed out that there are a number of varieties of modelling the response time even under the general category of Weibull distribution. Even with the one-parameter Weibull distribution, there are different possibilities of modelling the λ parameter in relationship with other item and person parameters. The present paper only explore a very simple model. There is possibility that the model is oversimplified and thus does not sufficiently capture the realistic relationship. Thus this paper can only be viewed as an initial step in exploring the appropriate models for response time.

3 The EM Algorithm for Parameter Calibration

The EM algorithm for the extended 4PLRT model described below is an extension of the EM algorithm for the 4PLRT model described in Wang and Hanson

(2005), which was a special application of the EM algorithm for finite mixture model described in Woodruff and Hanson (1996). The main difference is that here we work on the joint likelihood of the item responses and the response times, which includes the response time distribution in the formulation.

Data

The latent variables are assumed to be discrete. It is assumed that θ_i can be one of K known discrete values q_k, $k = 1, \ldots, K$, and that ρ_i can take on L known discrete values u_l, $l = 1, \ldots, L$. The probability that a randomly chosen examinee is in category k of the ability latent variable and in category l of the speed latent variable in π_{kl}. With this assumption the joint distribution of the latent variables has a multinomial distribution with probabilities π_{kl}, $k = 1, \ldots, K, l = 1, \ldots, L$ (the set of all π_{kl} is denoted $\boldsymbol{\pi}$). The notational convention used in this paper is that q_k, $k = 1, \ldots, K$ and u_l, $l = 1, \ldots, L$ are the possible values of the two latent variables, whereas θ_i and ρ_i are unspecified values of the latent variables for examinee i which can equal any of the q_k and u_l.

The model treats both item responses and response times as observed realizations of random variables in the observed and complete data likelihoods.

Observed Data. The observed data are the item responses and response times of a sample of N examinees to J dichotomous items. The item responses are contained in a $N \times J$ matrix $\boldsymbol{Y}$, where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)^t$, $\boldsymbol{y}_i$ is a vector given by $(y_{i1}, y_{i2}, \ldots, y_{iJ})$, and y_{ij} is one if examinee i answered item j correctly and zero if examinee i answered item j incorrectly. The response times are contained in a $N \times J$ matrix $\boldsymbol{T}$, where $\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_N)^t$, $\boldsymbol{t}_i$ is a vector given by $(t_{i1}, t_{i2}, \ldots, t_{iJ})$, and t_{ij} is the response time of examinee i used to answer item j.

Missing Data. The missing data are values of the unobserved ability and speed latent variables for each examinee. The missing data are $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$ and $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_N)$, where θ_i and ρ_i are the values of the ability and speed latent variables for examinee i.

Complete Data. The complete data are the observed data plus the missing data for each examinee. The complete data are $[(\boldsymbol{y}_1, t_1, \theta_1, \rho_1), (\boldsymbol{y}_2, t_2, \theta_2, \rho_2), \ldots, (\boldsymbol{y}_N, t_N, \theta_N, \rho_N)]$.

Algorithm

The EM algorithm can be used to find parameter estimates that maximize the likelihood of the observed data based on a sequence of calculations that involve finding parameter estimates that maximize a conditional expectation of the complete data likelihood. In the current model the maximum likelihood estimates are found for the conditional observed joint likelihood of the item response and the response times. Parameter estimates are found that maximize the following observed data likelihood:

$$L(\boldsymbol{Y}, \boldsymbol{T} | \boldsymbol{\Delta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} \prod_{j=1}^{J} f(y_{ij}, t_{ij} | q_k, u_l, \boldsymbol{\delta}_j) \right), \tag{8}$$

where $\boldsymbol{\Delta}$ is the set of item parameters for all items $(\boldsymbol{\delta}_j, j = 1, \ldots, J)$.

The corresponding likelihood for the complete data is:

$$
\begin{aligned}
&L(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\rho} | \boldsymbol{\Delta}, \boldsymbol{\pi}) \\
&\quad = \prod_{i=1}^{N} \left(\prod_{j=1}^{J} f(y_{ij}, t_{ij} | \theta_i, \rho_i, \boldsymbol{\delta}_j, \boldsymbol{\pi}) \right) f(\theta_i, \rho_i | \Delta, \boldsymbol{\pi}) \\
&\quad = \prod_{i=1}^{N} \left(\prod_{j=1}^{J} f(y_{ij}, t_{ij} | \theta_i, \rho_i, \boldsymbol{\delta}_j) \right) f(\theta_i, \rho_i | \boldsymbol{\pi}),
\end{aligned} \tag{9}
$$

where $f(\theta_i, \rho_i|\boldsymbol{\pi}) = \pi_{kl}$ if $\theta_i = q_k$ and $\rho_i = u_l$. Note that here we do not need to make the simplifying assumption that the joint distribution of θ_i and ρ_i are not independent of response time; that is $f(\theta_i, \rho_i|t_{ij}, \boldsymbol{\pi}) = f(\theta_i, \rho_i|\boldsymbol{\pi})$. This assumption was needed in Wang and Hanson (2005) because it that paper t_{ij} was treated as a conditioning variable, and without this assumption the response time distribution would eventually need to be specified. In the current model, t_{ij} is not treated only as a conditioning variable, but rather, the joint distribution of y_{ij} and t_{ij} is treated in the first place, and the the response time distribution is specified.

The log-likelihood corresponding to Equation 9 is

$$
\begin{aligned}
&\log[L(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\rho} | \boldsymbol{T}, \boldsymbol{\Delta}, \boldsymbol{\pi})] \\
&\quad = \sum_{i=1}^{N} \{ \sum_{j=1}^{J} log[f(y_{ij}, t_{ij} | \theta_i, \rho_i, \boldsymbol{\delta}_j)] + \log[f(\theta_i, \rho_i | \boldsymbol{\pi})] \} \\
&\quad = \sum_{j=1}^{J} \sum_{i=1}^{N} log[f(y_{ij}, t_{ij} | \theta_i, \rho_i, \boldsymbol{\delta}_j)] + \sum_{i=1}^{N} \log[f(\theta_i, \rho_i | \boldsymbol{\pi})].
\end{aligned} \tag{10}
$$

The computations to be performed in the E and M steps of the EM algorithm are described in the next two sections.

E Step

The E step at iteration $s (s = 0, 1, \ldots)$ consists of computing the expected value of the log-likelihood given in Equation 10 over the conditional distribution of the missing data $(\boldsymbol{\theta}, \boldsymbol{\rho})$ given the observed data $(\boldsymbol{Y}, \boldsymbol{T})$, and fixed values of the parameters $\boldsymbol{\Delta}^{(s)}$ and $\boldsymbol{\pi}^{(s)}$ obtained in the M step of iteration $s - 1$ (with some type of starting values for the parameters are used for $\boldsymbol{\Delta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$). The expected complete data log-likelihood is given by (Woodruff and Hanson, 1996):

$$
\phi(\boldsymbol{\Delta}) + \psi(\boldsymbol{\pi}) \tag{11}
$$

where

$$
\phi(\boldsymbol{\Delta}) = \sum_{j=1}^{J} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} log[f(y_{ij}, t_{ij} | q_k, u_l, \boldsymbol{\delta}_j)] f(q_k, u_l | \boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}}) \tag{12}
$$

and

$$\psi(\boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\log[f(q_k,u_l|\boldsymbol{\pi})]f(q_k,u_l|\boldsymbol{y_i},\boldsymbol{t_i},\boldsymbol{\Delta^{(s)}},\boldsymbol{\pi^{(s)}})$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\log\pi_{kl}f(q_k,u_l|\boldsymbol{y_i},\boldsymbol{t_i},\boldsymbol{\Delta^{(s)}},\boldsymbol{\pi^{(s)}}) \quad (13)$$

The conditional probability of the ability latent variable being equal to q_k and the speed latent variable being equal to u_l for examinee i given observed item response $\boldsymbol{y}_i$, observed response times $\boldsymbol{t}_i$ and parameter values of $\boldsymbol{\Delta}^{(s)}$ and $\boldsymbol{\pi}^{(s)}$ is (Woodruff and Hanson, 1996):

$$f(q_k,u_l|\boldsymbol{y_i},\boldsymbol{t_i},\boldsymbol{\Delta^{(s)}},\boldsymbol{\pi^{(s)}})$$
$$= \frac{f(\boldsymbol{y}_i,\boldsymbol{t}_i|q_k,u_l,\boldsymbol{\Delta}^{(s)})\pi_{kl}^{(s)}}{\sum_{k'=1}^{K}\sum_{l'=1}^{L}f(\boldsymbol{y}_i,\boldsymbol{t}_i|q_k,u_l,\boldsymbol{\Delta}^{(s)})\pi_{k'l'}^{(s)}}$$
$$= \frac{\pi_{kl}^{(s)}\prod_{j=1}^{J}f(y_{ij},t_{ij}|q_k,u_l,\boldsymbol{\delta}_j)}{\sum_{k'=1}^{K}\sum_{l'=1}^{L}\pi_{k'l'}^{(s)}\prod_{j=1}^{J}f(y_{ij},t_{ij}|q_k,u_l,\boldsymbol{\delta}_j)} \quad (14)$$

The E step consists of computing the conditional probabilities in Equation 14 which are used to compute the derivatives of $\phi(\boldsymbol{\Delta})$ and $\psi(\boldsymbol{\pi})$ in the M step.

The previous equations used the joint distribution $f(y_{ij},t_{ij}|q_k,u_l,\boldsymbol{\delta}_j)$ which is given by Equations 1, 2, and 7.

M Step

Estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\Delta}$ can be computed independently in the M step by finding values of $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ that separately maximize $\phi(\boldsymbol{\Delta})$ and $\psi(\boldsymbol{\pi})$. The values of $\pi_{kl}^{(s+1)}$ computed in the M step at iteration $s+1$ are (Equation 30 of Woodruff and Hanson, 1996):

$$\pi_{kl}^{(s+1)} = \frac{n_{kl}^{(s)}}{N} \quad (15)$$

where

$$n_{kl}^{(s)} = \sum_{i=1}^{N} f(q_k,u_l|\boldsymbol{y_i},\boldsymbol{t_i},\boldsymbol{\Delta^{(s)}},\boldsymbol{\pi^{(s)}}), \quad (16)$$

and $f(q_k,u_l|\boldsymbol{y_i},\boldsymbol{t_i},\boldsymbol{\Delta^{(s)}},\boldsymbol{\pi^{(s)}})$ is given by Equation 14.

The values of $\boldsymbol{\delta}_j^{(s+1)}$ computed in the M step at iteration s are the solution of the system of four equations:

$$\begin{aligned}\frac{\partial\phi(\boldsymbol{\Delta})}{\partial a_j} &= 0\\ \frac{\partial\phi(\boldsymbol{\Delta})}{\partial b_j} &= 0\\ \frac{\partial\phi(\boldsymbol{\Delta})}{\partial c_j} &= 0\\ \frac{\partial\phi(\boldsymbol{\Delta})}{\partial d_j} &= 0,\end{aligned} \quad (17)$$

where $\phi(\boldsymbol{\Delta})$ is expressed in Equation 12 using $f(q_k, u_l|\boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}})$ computed in the E step at iteration s.

This system of equations can be simultaneously solved using Newton-Ralphson algorithm which involved a Hessian matrix of second order derivatives. These derivatives are too messy to express here and are thus included in the appendix.

There is an issue of scale indeterminacy as with other IRT models. This issue is resolved by enforcing the distribution of θ to have a mean of zero and standard deviation of one.

The above described EM algorithm was programmed in C++.

4 Application to a Real Test Data Set

Table 1: Item Parameter Estimates for the 20-Item Math Test under the Extended 4PLRT Model.

Item	a	b	c	d
1	0.7838	-0.5753	0.2612	0.5205
2	0.4534	-1.3934	0.1552	0.5964
3	0.5900	-0.6755	0.2061	1.3362
4	0.5441	-0.6405	0.1332	0.6704
5	0.5233	-0.3829	0.1666	0.6975
6	0.4490	-0.8186	0.2898	0.8678
7	0.3527	-0.5600	0.1283	2.2162
8	0.5411	-1.3981	0.1276	20.2259
9	1.1087	-0.7049	0.1856	21.7185
10	1.1077	-0.5410	0.1880	26.4511
11	0.4604	-0.4511	0.1342	15.4335
12	0.9209	0.8645	0.4174	9.6417
13	1.0841	0.5990	0.3320	28.6968
14	0.6431	1.0253	0.1996	8.9783
15	0.6934	0.7043	0.3390	12.5383
16	0.4642	0.9014	0.2198	22.7151
17	0.5555	0.9570	0.1438	13.6889
18	0.5206	0.7926	0.1623	11.0413
19	0.6895	0.7914	0.1518	19.8005
20	0.8910	0.5602	0.2160	12.9700

The same real test data set used in Wang and Hanson (2005) was input to the calibration program. The data contains the item responses and response times for 1161 examinee who answered 20 Math Items on computers. Table 1 contains the calibrated item parameter estimates. These parameter estimates are somewhat different compared to the item parameter estimated in Wang and Hanson (2005). It is understandable that the d parameter estimates are different

because in the current model, d is the only scale parameter in the term that contains response time, whereas in the Wang and Hanson model, there are two scale parameters d and ρ. The differences in the a, b, and c parameters are not readily explained. Wang and Hanson's (2005) parameter estimates for a, b, and c based on the 4PLRT model were quite close to the parameter estimates obtained from a regular three parameter logistic (3PL) model calibrated using BILOG3 (Mislevy & Bock, 1990). With a different modelling approach, we always expect these parameter estimates will somehow be affected, but how much differences are reasonable differences is a question that needs to be further explored. One way to explore this question is to try out different models for the marginal distributions of response time and see if the item parameter values are greatly affected by these different models.

5 Simulation Study to Evaluate the Precision of the Calibration Procedure

5.1 Simulation Procedure

In order to evaluate the precision of the calibration procedure, a simulation study was carried out. Item responses and response times are generated under the assumption that the joint distribution described in Equations 1 through 7 was the correct model. The item parameters estimated in Wang and Hanson (2005) was assumed to the true item parameters. Random samples of examinees were generated with a standard normal θ distribution and a uniform ρ distribution ($U(0.002, 0.02)$). Because there is a lack of prior knowledge of the relationship between θ and ρ, they are basically drawn independently. Item responses and response times were generated with the joint probability distributions given in Equations 1, 2. Data set for each of the random samples were input into the calibration program. Because of the extreme long CPU time for running the simulations, it is not possible to include a larger test length in this simulation study. Three different sample sizes were used: 1000, 2000, and 4000. For each sample size, 100 random samples were repeated drawn. The estimated item parameters were used to compute the bias, standard error of estimates (SEE), and root mean square errors (RMSE). Bias is defined as the difference between mean estimated item parameter values across 100 replications and the true item parameter value for a particular item parameter. SEE is the standard deviation of the estimated parameter values across 100 replications. RMSE is computed from bias and SEE by the following equation:

$$RMSE^2 = Bias^2 + SEE^2. \tag{18}$$

Correlations between estimated and true item parameter values across items are computed for each item parameter and for each replication. The mean and standard deviation of the correlation coefficients across replications are also computed.

5.2 Results

The results are presented in Tables 2, 3, and 4. Note that the error indexes for different parameters are on the different scales and thus are not comparable. For example, the errors for the c parameter are on a much smaller scale than the errors for the other parameters. The error for the d parameter is on a larger scale than for other parameters. The mean bias, SEE, and RMSE across items reported in Tables 2, 3, and 4 are comparable to the errors in Wang and Hanson (2005). The bias tends to decrease as sample size increases, but not as rapidly as in Wang and Hanson (2005). The SEE also consistently decreases as sample size increases, so are the RMSE. If we look at the errors for each item, we can see there are a few items that have unusually larger errors than other item, especially items 8 and 9. The large biases for the b parameter for these items are especially disturbing in light of the fact that the b parameters are usually estimated with high accuracy in other IRT models. The reason for such high bias needs to be further explored.

Table 5 contains the mean correlations of true and estimated item parameter averaged across replications. Because correlation coefficients are scale free and thus are comparable across different parameters. This table shows that the d parameter has the highest correlation, followed by b, a, and c, in that order. These mean correlations indicate that the item parameter estimates recover the true parameters reasonably well. As expected, the correlations increase as sample size increases.

6 Conclusion and Discussion

This paper explores a model for the joint distribution of item response accuracy and response time. It adopts a slightly revised version of the conditional distribution response accuracy given response time in Wang and Hanson (2005), and adds a one-parameter Weibull distribution for the marginal distribution of response time. The EM algorithm is used for the parameter estimation procedure. The application of this procedure to a real test data showed that the item parameters are somewhat different from the values reported in Wang and Hanson (2005). The simulation study showed that the calibration procedure can recover true item parameters with very well except for a few items. Overall, we can say the model and the calibration procedure are moderately successful. Additional studies with real test data are needed to test the model.

Modelling response accuracy and response time simultaneously is an intriguing and challenging enterprise. There are a number of ways to model the joint distribution of response accuracy and response time. Modelling distribution of response time relating to other person and item parameters is also a tricky undertaking. Even within the one-parameter Weibull distribution family, there are a number of ways to model the λ parameter. This paper should be viewed as a initial step in a sequence of explorations.

Table 2: Item Parameter Estimates for the 20-Item Math Test under the Extended 4PLRT Model for Sample Size 1000.

	a			b		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.143	0.115	0.183	-0.015	0.083	0.084
2	-0.201	0.122	0.235	-0.125	0.102	0.161
3	-0.121	0.112	0.165	-0.038	0.081	0.089
4	-0.239	0.117	0.266	-0.174	0.091	0.196
5	-0.160	0.132	0.207	-0.247	0.163	0.296
6	-0.001	0.082	0.082	0.014	0.072	0.073
7	-0.075	0.089	0.116	-0.137	0.095	0.167
8	-0.176	0.157	0.236	0.658	0.217	0.693
9	-0.375	0.179	0.415	0.539	0.136	0.556
10	-0.166	0.153	0.226	0.400	0.128	0.420
11	-0.059	0.098	0.114	-0.058	0.082	0.100
12	-0.252	0.245	0.351	-0.004	0.117	0.117
13	-0.237	0.480	0.536	-0.085	0.108	0.137
14	-0.135	0.295	0.324	-0.126	0.108	0.166
15	-0.140	0.257	0.293	0.031	0.102	0.107
16	0.024	0.099	0.102	0.184	0.128	0.224
17	-0.286	0.226	0.365	0.165	0.069	0.179
18	-0.008	0.110	0.110	0.158	0.111	0.193
19	-0.248	0.256	0.357	0.103	0.109	0.150
20	-0.141	0.153	0.208	-0.294	0.117	0.316
Mean	0.159	0.174	0.245	0.178	0.111	0.221
SD	0.097	0.097	0.120	0.176	0.034	0.164
	c			d		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.002	0.033	0.033	-0.557	0.458	0.721
2	-0.059	0.040	0.071	-0.246	0.309	0.395
3	-0.014	0.027	0.030	-0.382	0.387	0.543
4	-0.047	0.032	0.057	0.279	0.273	0.390
5	-0.075	0.051	0.090	0.470	0.300	0.557
6	0.034	0.028	0.044	0.211	0.227	0.310
7	-0.043	0.035	0.055	0.266	0.362	0.449
8	0.121	0.056	0.133	-2.878	1.013	3.051
9	0.096	0.034	0.102	-2.995	0.807	3.101
10	0.062	0.029	0.069	-2.461	0.840	2.601
11	-0.018	0.029	0.034	-0.181	0.914	0.932
12	-0.027	0.035	0.044	-1.360	1.831	2.281
13	-0.036	0.061	0.071	-1.784	3.080	3.559
14	-0.015	0.025	0.029	0.706	1.499	1.657
15	-0.016	0.029	0.033	-1.170	1.248	1.711
16	0.022	0.021	0.030	-1.520	1.387	2.058
17	0.015	0.012	0.020	-1.754	0.858	1.953
18	0.017	0.020	0.027	-1.303	1.163	1.747
19	0.005	0.015	0.016	-1.439	1.119	1.823
20	-0.050	0.039	0.063	1.471	1.602	2.175
Mean	0.039	0.033	0.053	1.172	0.984	1.601
SD	0.031	0.012	0.030	0.890	0.693	1.019

Table 3: Item Parameter Estimates for the 20-Item Math Test under the Extended 4PLRT Model for Sample Size 2000.

	a			b		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.121	0.088	0.150	-0.026	0.061	0.067
2	-0.213	0.078	0.226	-0.122	0.096	0.155
3	-0.152	0.084	0.174	-0.035	0.081	0.088
4	-0.249	0.102	0.269	-0.149	0.082	0.170
5	-0.150	0.110	0.186	-0.209	0.149	0.256
6	-0.004	0.079	0.079	0.010	0.056	0.057
7	-0.084	0.081	0.117	-0.137	0.118	0.181
8	-0.182	0.135	0.226	0.702	0.196	0.729
9	-0.360	0.160	0.394	0.521	0.127	0.536
10	-0.131	0.115	0.175	0.366	0.068	0.372
11	-0.063	0.092	0.111	-0.053	0.059	0.080
12	-0.164	0.194	0.254	0.020	0.087	0.090
13	-0.070	0.316	0.324	-0.057	0.084	0.102
14	-0.112	0.176	0.209	-0.085	0.057	0.103
15	-0.156	0.161	0.224	0.053	0.061	0.081
16	0.003	0.070	0.070	0.139	0.086	0.164
17	-0.332	0.131	0.357	0.180	0.040	0.184
18	-0.011	0.076	0.077	0.118	0.047	0.127
19	-0.268	0.181	0.323	0.107	0.113	0.156
20	-0.127	0.125	0.179	-0.300	0.068	0.307
Mean	0.148	0.128	0.206	0.169	0.087	0.200
SD	0.100	0.059	0.094	0.179	0.038	0.172
	c			d		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.016	0.029	0.033	-0.609	0.324	0.690
2	-0.056	0.040	0.069	-0.376	0.242	0.447
3	-0.018	0.025	0.031	-0.491	0.369	0.614
4	-0.051	0.034	0.061	0.089	0.200	0.219
5	-0.073	0.052	0.090	0.296	0.192	0.353
6	0.024	0.030	0.038	0.129	0.166	0.210
7	-0.049	0.037	0.061	0.188	0.301	0.355
8	0.131	0.036	0.136	-3.144	0.926	3.278
9	0.094	0.029	0.098	-2.870	0.618	2.936
10	0.060	0.021	0.063	-2.200	0.574	2.274
11	-0.023	0.027	0.035	-0.215	0.584	0.622
12	-0.012	0.022	0.025	-0.809	1.093	1.360
13	-0.005	0.015	0.016	-0.902	2.143	2.325
14	-0.008	0.016	0.018	0.511	0.850	0.991
15	-0.007	0.019	0.020	-1.030	1.041	1.464
16	0.015	0.018	0.023	-1.105	0.882	1.414
17	0.012	0.010	0.015	-1.840	0.610	1.938
18	0.010	0.017	0.019	-0.949	0.961	1.350
19	0.002	0.012	0.012	-1.431	1.135	1.826
20	-0.048	0.034	0.058	1.353	1.377	1.931
Mean	0.036	0.026	0.046	1.027	0.729	1.330
SD	0.034	0.011	0.033	0.888	0.493	0.917

Table 4: Item Parameter Estimates for the 20-Item Math Test under the Extended 4PLRT Model for Sample Size 4000.

	a			b		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.130	0.067	0.146	-0.044	0.041	0.060
2	-0.207	0.085	0.224	-0.119	0.099	0.155
3	-0.160	0.085	0.181	-0.041	0.085	0.094
4	-0.249	0.074	0.260	-0.149	0.034	0.153
5	-0.111	0.096	0.147	-0.134	0.122	0.181
6	-0.010	0.049	0.050	0.006	0.044	0.044
7	-0.082	0.070	0.108	-0.110	0.114	0.159
8	-0.132	0.092	0.161	0.596	0.140	0.612
9	-0.370	0.140	0.396	0.549	0.120	0.562
10	-0.121	0.088	0.150	0.371	0.069	0.377
11	-0.056	0.052	0.076	-0.047	0.026	0.054
12	-0.117	0.151	0.191	0.014	0.084	0.085
13	-0.081	0.223	0.238	-0.041	0.051	0.065
14	-0.112	0.142	0.181	-0.069	0.031	0.076
15	-0.116	0.148	0.188	0.023	0.083	0.086
16	-0.002	0.052	0.052	0.132	0.037	0.137
17	-0.369	0.105	0.384	0.205	0.037	0.208
18	-0.015	0.048	0.051	0.129	0.032	0.133
19	-0.196	0.175	0.262	0.051	0.100	0.112
20	-0.114	0.085	0.142	-0.275	0.048	0.279
Mean	0.137	0.101	0.179	0.155	0.070	0.182
SD	0.101	0.047	0.097	0.169	0.036	0.161
	c			d		
Item	Bias	SEE	RMSE	Bias	SEE	RMSE
1	-0.019	0.024	0.030	-0.646	0.265	0.698
2	-0.054	0.043	0.069	-0.396	0.190	0.439
3	-0.020	0.034	0.039	-0.530	0.265	0.593
4	-0.055	0.020	0.059	0.036	0.162	0.166
5	-0.050	0.044	0.067	0.211	0.166	0.268
6	0.023	0.023	0.033	0.078	0.130	0.152
7	-0.043	0.040	0.059	0.072	0.229	0.240
8	0.112	0.034	0.117	-2.704	0.524	2.754
9	0.099	0.026	0.102	-3.002	0.616	3.065
10	0.058	0.017	0.061	-2.294	0.462	2.340
11	-0.019	0.018	0.026	-0.232	0.437	0.494
12	-0.007	0.018	0.019	-0.462	0.876	0.990
13	-0.004	0.010	0.011	-0.313	1.765	1.793
14	-0.006	0.012	0.013	0.462	0.740	0.872
15	-0.002	0.018	0.018	-0.562	0.832	1.004
16	0.011	0.013	0.017	-0.881	0.707	1.130
17	0.011	0.008	0.014	-2.081	0.503	2.141
18	0.011	0.012	0.016	-0.995	0.591	1.157
19	0.005	0.010	0.011	-0.804	0.953	1.247
20	-0.042	0.028	0.051	1.251	1.060	1.640
Mean	0.032	0.023	0.042	0.901	0.574	1.159
SD	0.031	0.011	0.031	0.903	0.400	0.872

Table 5: Mean and SD of Correlations between Estimated and True Item Parameters Across Replications.

	a	b	c	d
		N=1000		
Mean	0.8162	0.8877	0.7148	0.9267
SD	0.0878	0.0306	0.1074	0.0443
		N=2000		
Mean	0.8878	0.8977	0.7598	0.9519
SD	0.0426	0.0295	0.0709	0.0218
		N=4000		
Mean	0.9168	0.9145	0.8024	0.9607
SD	0.0313	0.0214	0.0606	0.0154

References

Bloxom, B. (1985). Considerations in Psychometric modeling of response time. *Psychometrika, 50,* 383-397.

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

Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model (2nd ed.) [Computer program].* Mooresville, IN: Scientific Software.

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

Schnipke, D. L., & Scrams, D. J. (1998). *Exploring issues of examinee behavior: insights gained from response-time analyses.* Paper presented at the ETS colloquium on "Computer Based Testing: Building the Foundation for Future Assessment". September 25-26, 1998, Philadelphia, PA.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34,* 213-232.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Van Breukelen, G. J. P. (1989). *Concentration, Speed, and Precision in Mental Tests.* Unpublished Doctoral Dissertation, University of Nijmegen, The Netherlands.

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

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

Wang, T. & Zhang, J. (2006). Optimal partitioning of testing time: theoretical properties and practical implications. *Psychometrika, 71,* 105-120.

Woodruff, D. J., & Hanson, B. A. (1996). Estimation of item response models using the EM algorithm for finite mixture. *ACT Research Report 96-6.* Iowa City, IA: ACT, Inc.

Appendix

This appendix gives the derivatives for Equation 17. First, $\phi(\boldsymbol{\Delta})$ can be further decomposed into these two components:

$$\phi(\boldsymbol{\Delta}) = \zeta(\boldsymbol{\Delta}) + \eta(\boldsymbol{\Delta}), \tag{19}$$

where

$$\zeta(\boldsymbol{\Delta}) = \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L} log[f(y_{ij}|t_{ij}, q_k, u_l, \boldsymbol{\delta}_j)]f(q_k, u_l|\boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}}), \tag{20}$$

and

$$\eta(\boldsymbol{\Delta}) = \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L} log[f(t_{ij}|q_k, u_l, \boldsymbol{\delta}_j)]f(q_k, u_l|\boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}}). \tag{21}$$

Applying Equation 2 and let $P(t_{ij}, \theta_i, \boldsymbol{\delta}_j)$ be simply denoted as P_{ij}, We have

$$\frac{\partial\zeta(\boldsymbol{\Delta})}{\partial\delta_j} = \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}[\frac{(y_{ij}-P_{ij})}{P_{ij}(1-P_{ij})}\frac{\partial P_{ij}}{\partial\delta_j}]f(q_k, u_l|\boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}}). \tag{22}$$

Let

$$w = e^{-1.7a_j[\theta_i-(d_j/t_{ij})-b_j]}, \tag{23}$$

Using Equation 3, $\frac{\partial P_{ij}}{\partial\delta_j}$ can be further expressed for each item parameter as:

$$\begin{aligned}
\frac{\partial P_{ij}}{\partial a_j} &= 1.7(1-c_j)(q_k - \frac{d_j}{t_{ij}} - b_j)\frac{w}{(1+w)^2} \\
\frac{\partial P_{ij}}{\partial b_j} &= -1.7a_j(1-c_j)\frac{w}{(1+w)^2} \\
\frac{\partial P_{ij}}{\partial c_j} &= \frac{w}{1+w} \\
\frac{\partial P_{ij}}{\partial d_j} &= -1.7a_j(1-c_j)\frac{w}{t_{ij}(1+w)^2},
\end{aligned} \tag{24}$$

The second order derivatives can be derived in a similar fashion, but is too lengthy to be presented here.

Likewise, we have

$$\frac{\partial\eta(\boldsymbol{\Delta})}{\partial\delta_j} = \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{\partial log[f(t_{ij}|q_k, u_l, \boldsymbol{\delta}_j)]}{\partial\delta_j}f(q_k, u_l|\boldsymbol{y_i}, \boldsymbol{t_i}, \boldsymbol{\Delta^{(s)}}, \boldsymbol{\pi^{(s)}}). \tag{25}$$

Applying Equation 7, $\frac{\partial log[f(t_{ij}|q_k, u_l, \boldsymbol{\delta}_j)]}{\partial\delta_j}$ can be further expressed as

$$\frac{\partial log[f(t_{ij}|q_k, u_l, \boldsymbol{\delta}_j)]}{\partial b_j} = u_l(q_k - b_j)t_{ij}^2 - \frac{2}{q_k - b_j}. \tag{26}$$

The derivatives with respect to other item parameters are all zero. The second order derivatives with respect with b_j is

$$\frac{\partial^2 log[f(t_{ij}|q_k, u_l, \boldsymbol{\delta}_j)]}{\partial b_j^2} = \frac{-2}{(q_k - b_j)^2} - u_l t_{ij}^2. \tag{27}$$

The second derivatives with respect to other item parameters and the cross partial derivatives all vanish to zero.