Development and Calibration of an Item Response Model that Incorporates Response Time

Tianyou Wang and Bradley A. Hanson

ACT, Inc.

Send correspondence to:
Tianyou Wang
ACT, Inc
P.O. Box 168
Iowa City, IA 52243
wang@act.org
(319) 337-1641

Paper presented to the annual meeting of the American Educational Research Association in Seattle, April, 2001

Abstract

This paper proposed an item response model that incorporates response time. A parameter estimation procedure using the EM algorithm was developed. The procedure was programmed and evaluated with both real and simulated test data. The results suggest the estimation procedure works well in estimating model parameters. By utilizing response time data, estimation of person ability parameters can be improved. Potential applications of this model are discussed. Directions for further study are suggested.

Development and Calibration of an Item Response Model that Incorporates Response Time

Tianyou Wang and Bradley A. Hanson, ACT, Inc.

In psychophysics and cognitive psychology, response time has long been a research topic of interest. Psychologists who study response time are mainly interested in inferring the organization of the mental process from the distribution of response time to different components of the mental process (e.g., Luce, 1986). A general result from this research is the well-known speed-accuracy trade-off function (SATF) which describes how a subject's accuracy to certain task changes as the response time changes (see Luce, 1986. P81; Roskam, 1997).

In educational measurement, however, the response time data has traditionally been ignored in practice probably due to the fact that it was difficult to collect response time data at the individual item level with paper-pencil testing. Computer-based testing makes response time data at the item level readily available to measurement professionals. For this reason, there has been an increased interest in response time in recent years. So far, existing research has focused on modeling the distribution of response time and its relationship with other variables such as item difficulty, examinee ability and test taking strategies (e.g., Thissen, 1983; Schnipke & Scrams, 1997; Segall, 1987; Parshall, Mittelholtz, & Miller, 1994; Scrams & Schnipke, 1997; Swygert, 1998). Research has indicated a strong but complicated across-examinee relationship between response time and response accuracy (See Schnipke & Scrams, 1998 for an overview of the literature). Generally, the relationship largely depends on context and content of the test. All this research has treated response time as a dependent variable or outcome variable and response accuracy and response time were not included in a single model except in a few cases. Those few exceptions (e.g., Verhelst, Verstralen, & Jansen, 1997; Roskam, 1997) are exclusively applied to speeded test because of the nature of their model specifiction. For example, in Roskam's (1997) model, the probability of a correct response to an item is specified as:

$$P(U=1\mid t)=\frac{\theta t}{\theta t+\varepsilon}=\frac{\exp(\xi+\tau-\sigma)}{1+\exp(\xi+\tau-\sigma)} \quad , \tag{1}$$

where θ is the person ability, ε is item difficulty, and *t* is response time, and ξ, σ and τ are the logarithms of θ, ε, and *t*. It can be seen that as *t* goes to infinite, $P(U=1\mid t)$ will approach one no matter how hard the item is. Therefore this type of model can only applied to speeded tests because a basic assumption of speeded tests is that when time is unlimited, the answers are always correct. In most educational assessment settings, tests are designed to be power tests, which means that even given unlimited time, not every student will get a near perfect score. Further more, even for power tests, there is usually a time limit for the test administration. Even though that time limit is typically adequate for examinees in the middle and upper ability range, the time limit still has an effect on examinee performance. A model is needed that incorporates response time and can be applied to power tests.

The objectives of this paper are to (a) propose an item response model that incorporates response time, which can be applied to power tests, (b) formulate an item parameter estimation procedure for this model, (c) calibrate some real test data with this model and compare the results with calibration under the 3PL model using BILOG, (d) use simulation techniques to evaluate the item parameter estimation procedure, and (e) discuss its potential applications to measurement settings, particularly to computerized adaptive testing (CAT).

A Item Response Model that incorporates Response Time

This paper develops a model that incorporates the response time into the usual 3 parameter logistic (3PL) model which can be applied to power tests. With this model, the probability of correct response to item j by examinee i can be given as:

$$P\left(x_{ij}=1\,|\,\theta_i,\rho_i,a_j,b_j,c_j,d_j,t_{ij}\right)=c_j+\frac{1-c_j}{1+e^{-1.7a_j\left[\theta_i-\left(\rho_i d_j/t_{ij}\right)-b_j\right]}} \quad , \qquad (2)$$

Where a, b, c, and θ are usual IRT parameters, *d* is an item slowness parameter, *t* is the response time by this examinee on this particular item, ρ is an examinee slowness parameter, and θ is the similar person ability parameter as in the regular 3PL model. These two slowness parameters determine the rate of increase in correct answer probability as a result of increase in response time. This model treats response time as only a conditional variable and does not model how the examinee decides to spend a certain amount of time on a particular item. As the time increases, the term in the exponent decreases with a marginally decreasing rate. Because of the minus sign, this will effectively increase the overall term within the parenthesis and consequently the probability of correct response. For lack of a better term, we will call this model the four-parameter logistic response time (4PLRT) model because each item will have four parameters.

Figure 1 provides some examples how the probability of correct response increases as response time increases for some hypothetical items and examinees under this model. First of all, these curves bear the common characteristics of the speed-accuracy trade-off function (more precisely, the conditional accuracy function or sometimes called the micro speed-accuracy tradeoff function in Luce, 1986, p. 245). Comparing the curves with all but one parameters being the same can reveals how these parameters change the relationship between response time and correct response probability. It can be seen that the larger the *d* and ρ parameters, the slower the probability converges to its asymptote. For that reason, these two parameters should be called the item slowness parameter and the person slowness parameter. The top three curves all converge to the same asymptote because they all have the same $a(\theta - b)$ term, whereas the bottom curve has a different asymptote because it is for a different θ value.

Figure 1 also show the correct response probabilities do not converge to 1, but to some values less than 1. Unlike the models developed for speeded tests (Verhelst, Verstralen, & Jansen, 1997; Roskam, 1997), as response time goes to infinity, the overall term in the exponent does not increase to infinity, but converges to $a(\theta - b)$. This is achieved by putting a negative term of the inverse of response time in the exponent of the logistic function rather than a positive term of a increasing function of response time as did in the Verhelst, Verstralen, & Jansen (1997) and the Roskam (1997) models. That means spending unlimited time does not guarantee a correct answer. In this way, this model can be applied to the power tests.

The conventional 3PL model can be viewed as the limiting case of this model; that is, the case there is no time limit for answering each item. Because in realistic testing situations there is always a certain time limit, this model provides a more realistic description of the item response mechanism than the 3PL model.

Like the 3PL model, this model also has some indeterminacy problems. In addition to the indeterminacy of the θ scale as in the 3PL model which can be fixed in the same way as is done for the 3PL model, there is another indeterminacy due to the product of ρ and *d* in the term $(\rho d/t)$. One way to fix the problem is to set the scale for one of the parameters ρ or *d* to have a fixed mean and standard deviation (SD), and let the scale of the other parameter be automatically tied to the scale of θ and the unit of *t* (such as second or minute. Throughout this paper, second is used as the unit of time).

This model has many potentially useful applications. One direct application is that it may help estimate the item parameters for power tests with time limits. If the actual response mechanism is close to what is described in this model, taking response time into consideration in the item calibration process might even help estimate the regular item parameters a, b, and c. A second potential application is that response time might be used to help infer examinee ability (θ). This can be used in computerized adaptive testing (CAT) to make the provisional ability estimates converge faster to the true ability level, and thus reduce the number of items administered and testing time. This model can also be used to make inferences about examinees' other characteristics such as if they can solve problems quickly as well as accurately. It can also be used to study the optimal test taking strategy for tests with time limits. Other potential applications of the model are to detect random guessing behavior and to help deal with incomplete tests for the computerized adaptive testing situation. All these potential uses need to be verified with simulated and real test data. The first step needed before such research can be conducted is to develop a calibration program.

It should be noted that the model presented here is only a partial description of the test taking process. A more complete description should include a model that models the distribution of response time. With the distribution of response time, the joint distribution of correct response and response time and the marginal distribution of correct response can be derived. For the present, we focus on this partial model with response time being treated as a conditional variable. Our main goal for this paper is to develop a calibration procedure for this partial model.

Parameter Estimation for the 4PLRT Model

A parameter estimation procedure is developed using the EM algorithm as described in Woodruff and Hanson (1996). Response time is treated as a fixed rather than a random variable, like an independent variable in a regression model. Therefore, response time and item responses are treated differently, with only the item responses considered as observed realization of random variables in the observed and complete data likelihood. The EM algorithm finds 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. The difference between the EM procedure for this model and that for the 3PL model is the two dimensional nature of the person parameter space. The E-step will involve double integrals that take more computing time. Another major source of increase in computing time is the aggregate statistics for items (usually noted as n, r) are not available in the M-step with this model. The full description of the estimation procedure is seen in the appendix. The procedure was programmed in the C++ language and evaluated with both real and simulated test data.

Calibration and Evaluation of the 4PLRT Model with Real Test Data

A set of 20 ACT Mathematics items was administered to a group of 1161 examinees via computers. The response data and response time data were input into the calibration program. The response data were also input into BILOG discarding the response time data.

The resulting parameter estimates were tabulated in Table 1. The *a, b* and *c* parameter estimates from the two models were plotted against each other in Figure 2. Both Table 1 and Figure 2 show the two models produced very similar *a, b* and *c* estimates. The correlations between these pairs of estimates were found to be 0.940, 0.974 and 0.986 for *a, b* and *c*, respectively. The fact that the *c* parameter has stronger similarity than the *a* and *b* parameters make sense because the *c* parameter should be not affected by difference in the two models.

The *d* parameter estimates from the 4PLRT model vary considerably across items. The relationship between this parameter and the other parameters was investigated by computing their correlations. The correlations between the *d* parameter and *a, b* and *c* are found to be 0.484, 0.339 and -.2166, respectively. These correlations suggest that more discriminating and more difficult items takes more time to converge to their asymptotic correct answer probabilities. These results should be replicated with more real test data in future studies.

Evaluation of the Estimation Procedure with Simulations

Method

The model and the estimation procedure developed in this paper are evaluated using simulated data. To simulate the response time, Thissen's (1983) model for response time was used. The model is described as

$$\begin{aligned}&\log(t_{ij}) = \upsilon + s_i + u_j - gz_{ij} + \varepsilon_{ij}\\&\varepsilon_{ij} \sim N(0,\sigma^2)\\&where, z_{ij} = a_j(\theta_i - b_j)\end{aligned} \qquad (3)$$

v is the overall mean, s is a person slowness parameter, u is item slowness parameter, g is the log-linear relationship between response time and examinee ability. The generated response time data are used to generate the item response data using Equation 2 along with the examinee true ability parameters that are generated from a standard normal distribution. For this initial stage, we only examined a special case of the model; that is, we fixed the person slowness parameter in Equation 2 as constant and only let the item slowness parameter vary (from a uniform distribution from 0 to 10). The true a, b, and c parameters were taken from another set of ACT Mathematics item parameters.

Two different test lengths (20 items and 60 items) and three sample sizes (1000, 2000, and 4000) were used in this simulation. The summary statistics of the true item parameters for these two sets of items are in Table 2. The simulated data were generated and calibrated 100 times for each of the six conditions. The bias, standard error (SE) and root mean square error (RMSE) were computed for each item parameter across 100 replications. The means and standard deviation (SD) of these error indices were computed across items. (For bias, the means of the absolute values were computed). The correlations between the true and estimated item parameters were also computed for each item parameter for each replication. The means and SD of these correlations across replications were computed.

To study the effect of ignoring response time when response time does have an effect on examinee performance, we also calibrated a sample response data with a regular 3PL model using BILOG. By comparing these a, b, and c parameters to the true parameters and the estimated parameters with the response time model, we can study the effect of omitting response time in the calibration.

Preliminary Analysis of the Simulation Model

In order to examine how the different components in Equation 3 affect the calibration of the model, we experimented with some different variations of Thissen's model. The issue of using different variations first arose when we discovered that when the full model in Equation 3 is used to generated the response time, the d parameters in Equation 1 can not be properly estimated.

The true and estimated parameters under different condition were contained in Table 3. It is interesting to see that the way the response time was generated had a major effect on the parameter estimates, particularly for the item slowness parameter, d. When Thissen's model was implemented with all the terms inside, the d parameters all shrink to near zero and the a and b parameter estimates were also negatively affected. When the z term was dropped from Equation 2, then all the parameter estimates were quite accurate. We hypothesize that it is because with the z term in Equation 2, response time is negatively (because of the minus sign) correlated with the term $a(\theta - b)$ in Equation 2 (whether this is a realistic situation needs more empirical verification, current literature does not strongly support this correlation), which causes some identifiability problems. By deleting the z term in Equation 3, the response time is basically not correlated with examinee ability and item difficulty. This result seems to indicate that the model parameters can be accurately estimated only when the response time is not correlated with $a(\theta - b)$, although the final conclusion needs to be verified with more data.

Results from the Simulation Study

Table 4 contains the aggregate error indices for each of the item parameters. Note the magnitude of these values should be interpreted in accord with the scale of the parameters. For example, a smaller value for the c parameter than for the d parameter does not necessarily mean the c parameter is better estimated than the d parameter. Overall, these error indices seem to suggest that the parameters are reasonably well estimated. Increasing the sample size consistently results in smaller errors. Increases in the test length also consistently reduce SE and RMSE, but not always bias.

Table 5 contains the average correlation values between the estimated and true item parameters averaged across replications. Note that the values of correlation reflect SE more than the bias. Again, the correlations indicate these item parameters were well estimated, particularly for the b and the d parameters. The effects of test length and sample size are consistent with those seen in the error indices.

Figure 3 plots the estimated and true person parameters for a sample of 2000 simulees taking the 60-item test estimated under both the 3PL (BILOG) and the 4PLRT model. The correlation between the 3PL and the true θs is .884, whereas the correlation between the 4PLRT θ estimates and the true θs is .937. This increase in correlation by using the 4PLRT model rather than the 3PL model suggests that if in fact the 4PLRT model is true (i.e., response time plays a role), incorporating response time in estimating θ will result in more precision.

Figure 3 also shows that the person slowness parameters ρ s are not as well estimated as θs. The correlation between the estimated ρ s and true ρ s is .649. The reason why this parameter is not well estiamted should be further investigated.

Conclusion and Discussions

The 4PLRT model proposed in this paper utilizes an important source of data in educational measurement made available by computerized testing; namely the response time.

The advantage of this model relative to models presented in the existing literature is that it can be applied to power tests.

This paper developed a parameter estimation procedure for the 4PLRT model and evaluated it with both real and simulated test data. The results showed that the estimation procedure works well and produced reasonably accurate parameter estimates. With the real test data, the a, b and c parameter estimates from the 4PLRT model are very similar to those from the 3PL model calibrated with BILOG. The item slowness parameters seem to have strong correlation with the discrimination and difficulty parameters. Results based on simulated data show that if the response time affects the correct answer probability, ignoring response time data will have an adverse effect in estimating examinee ability.

In summary, these results give clear indication that this model provides promising capabilities in utilizing the response time data. As discussed previously, this model can be used to enhance measurement quality and handle many complicated issues otherwise difficult to handle, particularly in computerized adaptive testing settings. With the ever-increasing popularity of the computer-based testing, it is almost certain that there will be more and more need to utilize response time data to improve measurement quality. Future studies should apply this model to additional real test data and investigate how much can be gained in incorporating response time in the model.

In addition to helping estimate the person ability parameters, utilizing response time data with the 4PLRT model also renders it possible to make inferences about another important dimension of person characteristics: namely the slowness in problem solving. It may be debatable whether and how this information about persons should be used in various educational settings may be an entirely different set of questions, but it is good to know that this information can be obtained.

Reference

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

Parshall C. G., Mittelholtz, D. & Miller, T. R. (1994, April*). Response time: An investigation into determinants of item-level timing.* In C. G. Parshall (Chair), Issues in the development of a computer adaptive placement test. Symposium conducted at the meeting of the National Council on Measurement in Education, New Orleans.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.), In W. J. van der Linden and 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.

Segall, D. O. (1987). *Relation between estimated ability and test time on the CAT-ASVAB.* Unpublished manuscript.

Swygert, K. A. (1998). *An examination of item response times on the GRE-CAT.* Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

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.

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.

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 describes an application of the EM algorithm as described in Woodruff and Hanson (1996) to compute parameter estimates of the 4PLRT model for dichotomous items.

Model

The probability of a correct response to item j for a randomly sampled examinee with ability and speed parameters θ_i and ρ_i is given by the 4PLRT model as

$$P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j[\theta_i - (\rho_i d_j / t_{ij}) - b_j]}} \,, \tag{A1}$$

where $\boldsymbol{\delta}_j = (a_j, b_j, c_j, d_j)$ are item parameters for item j, and t_{ij} is the amount of time examinee i takes to respond to item j. The values θ_i and ρ_i are realizations of latent random variables that determine the probability of examinee i answering item j correctly. The latent variable associated with θ_i will be called the ability latent variable, and the latent variable associated with ρ_i will be called the speed latent varible.

In this paper 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 is π_{kl}. With this assumption the joint distribution of the latent variables has a multinomial distribution with probabilities $\pi_{kl}, k = 1, \ldots, K, l = 1, \ldots, L$ (the set of all π_{kl} is denoted $\boldsymbol{\pi}$). The notational convention used in this paper is that q_k, $k = 1, 2, \ldots, K$ and $u_l, l = 1, \ldots, L$ are the possible values of the two latent variables, whereas θ_i and ρ_l are unspecified values of the latent variables for examinee i which can equal any of the q_k and u_l.

Data

The model treats response times as fixed rather than random, like an independent variable in a regression model. Therefore, response times and item responses are treated differently, with only item responses considered observed realizations of random variables in the observed and complete data likelihoods.

Observed Data. The observed data are the responses of a sample of N examinees to J dichotomous items. The item responses are contained in a $N \times J$ matrix $\mathbf{Y}$, where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^t$, $\mathbf{y}_i$ 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.

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 $[(\mathbf{y}_1, \theta_1, \rho_1), (\mathbf{y}_2, \theta_2, \rho_2), \ldots, (\mathbf{y}_N, \theta_N, \rho_N)]$.

Response Times. The time examinee i took to respond to item j is t_{ij}. The times that examinee i took to respond to all the items are $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots t_{iJ})$. The $N \times J$ matrix containing the response times for all examinees to all items is $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N)^t$.

EM 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. To simplify the computations maximum likelihood estimates will be found for the conditional observed likelihood of the item responses given the response times. Parameter estimates will be found that maximize the following observed data likelihood (conditional on response times):

$$L(\mathbf{Y} \mid \mathbf{T}, \Delta, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} \prod_{j=1}^{J} P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j)^{y_{ij}} [1 - P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j)]^{1-y_{ij}} \right) \quad \text{(A2)}$$

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

The corresponding likelihood for the complete data conditional on response times is:

$$L(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{T}, \Delta, \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)^{y_{ij}} [1 - P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)]^{1-y_{ij}} f(\theta_i, \rho_i \mid \boldsymbol{\pi}), \quad \text{(A3)}$$

where $f(\theta_i, \rho_i \mid \boldsymbol{\pi}) = \pi_{kl}$ if $\theta_i = q_k$ and $\rho_i = u_l$. The log-likelihood corresponding to Equation A3 is

$$\begin{aligned}&\log[L(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\rho} \mid \mathbf{T}, \Delta, \boldsymbol{\pi})] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} \Big(y_{ij} \log[P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)] + (1 - y_{ij}) \log[1 - P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)] + \log[f(\theta_i, \rho_i \mid \boldsymbol{\pi})] \Big)\end{aligned}$$

$$= \sum_{j=1}^{J} \sum_{i=1}^{N} \{y_{ij} \log[P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)] + (1 - y_{ij}) \log[1 - P(\theta_i, \rho_i \mid t_{ij}, \boldsymbol{\delta}_j)]\}$$

$$+ \sum_{i=1}^{N} \log[f(\theta_i, \rho_i \mid \boldsymbol{\pi})] \,. \qquad \text{(A4)}$$

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 A4 over the conditional distribution of the missing data ($\boldsymbol{\theta}$, $\boldsymbol{\rho}$) given the observed data (**Y**), fixed values of the response times (**T**), and fixed values of the parameters $\Delta^{(s)}$ and $\boldsymbol{\pi}^{(s)}$ obtained in the M step of iteration $s - 1$ (starting values for the parameters are used for $\Delta^{(0)}$ and $\boldsymbol{\pi}^{(0)}$). The expected complete data log-likelihood is given by (Woodruff and Hanson, 1996):

$$\phi(\Delta) + \psi(\boldsymbol{\pi}) \qquad \text{(A5)}$$

where

$$\phi(\Delta) = \sum_{j=1}^{J} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} \{y_{ij} \log[P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j)] +$$

$$(1 - y_{ij}) \log[1 - P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j)]\} \, f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)}) \qquad \text{(A6)}$$

and

$$\psi(\boldsymbol{\pi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} \log[f(\theta_i, \rho_i \mid \boldsymbol{\pi})] \, f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)}) \,. \qquad \text{(A7)}$$

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 responses $\mathbf{y}_i$, observed response times $\mathbf{t}_i$ and parameter values of $\Delta^{(0)}$ and $\boldsymbol{\pi}^{(0)}$ is (Woodruff and Hanson, 1996):

$$f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)}) = \frac{f(\mathbf{y}_i \mid q_k, u_l, \mathbf{t}_i, \Delta^{(s)}) \pi_{kl}^{(s)}}{\sum_{k'=1}^{K} \sum_{l'=1}^{L} f(\mathbf{y}_i \mid q_k, u_l, \mathbf{t}_i, \Delta^{(s)}) \pi_{k'l'}^{(s)}}$$

$$= \frac{\pi_{kl}^{(s)} \prod_{j=1}^{J} P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j^{(s)})^{y_{ij}} [1 - P(q_k, u_l \mid t_{ij}, \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}}{\sum_{k'=1}^{K} \sum_{l'=1}^{L} \pi_{k'l'}^{(s)} \prod_{j=1}^{J} P(q_{k'}, u_{l'} \mid t_{ij}, \boldsymbol{\delta}_j^{(s)})^{y_{ij}} [1 - P(q_{k'}, u_{l'} \mid t_{ij}, \boldsymbol{\delta}_j^{(s)})]^{1-y_{ij}}} \,, \qquad \text{(A8)}$$

The E step consists of computing the conditional probabilities in Equation A8 which are used to compute the derivatives of $\phi(\Delta)$ and $\psi(\boldsymbol{\pi})$ in the M step.

M Step

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

$$\pi_{kl}^{(s+1)} = \frac{n_{kl}^{(s)}}{N}\,. \tag{A9}$$

where

$$n_{kl}^{(s)} = \sum_{i=1}^{N} f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)})\,, \tag{A10}$$

and $f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)})$ is given by Equation A9.

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:

$$\begin{aligned}
\frac{\partial\phi(\Delta)}{\partial a_j} &= 0 \\
\frac{\partial\phi(\Delta)}{\partial b_j} &= 0 \\
\frac{\partial\phi(\Delta)}{\partial c_j} &= 0 \\
\frac{\partial\phi(\Delta)}{\partial d_j} &= 0\,.
\end{aligned} \tag{A11}$$

using $f(q_k, u_l \mid \mathbf{y}_i, \mathbf{t}_i, \Delta^{(s)}, \boldsymbol{\pi}^{(s)})$ computed in the E step at iteration s.

Table 1. Item Parameter Estimates for 20 ACT Math Items Calibrated using the 4PLRT and the 3PL Models.

	4PLRT Model				3PL Model (BILOG)		
Item	a	b	c	d	a	b	c
1	1.0585	-0.7920	0.1724	1.9075	1.0036	-0.8069	0.1690
2	1.1757	-0.6201	0.2689	0.7682	1.0650	-0.6823	0.2512
3	1.0189	-0.6134	0.1694	1.4534	0.9497	-0.6262	0.1670
4	1.3573	-0.3716	0.1462	0.2889	1.2234	-0.4153	0.1402
5	1.0150	-0.0644	0.2576	0.0690	0.9407	-0.0740	0.2582
6	0.6996	-0.8598	0.1720	0.1321	0.6576	-0.9220	0.1707
7	0.8727	-0.0798	0.2039	0.3678	0.8279	-0.0498	0.2167
8	1.2477	-0.8789	0.0890	9.9204	0.9844	-0.4775	0.0840
9	1.7428	-0.6523	0.1126	9.9083	1.5805	-0.3800	0.1270
10	1.3443	-0.4211	0.1322	9.8975	1.2155	-0.2329	0.1133
11	0.8154	-0.1078	0.1665	3.8295	0.7860	-0.0166	0.1727
12	1.3032	0.3069	0.3262	9.8315	1.3716	0.6902	0.3535
13	1.4892	1.0953	0.3532	9.2318	1.2027	1.3073	0.3394
14	1.4821	0.7874	0.2354	6.7653	1.5569	1.0731	0.2452
15	1.2934	0.3104	0.2965	9.7058	1.2431	0.5828	0.3037
16	0.6530	0.2333	0.0931	9.9003	0.6453	0.6130	0.1109
17	1.8339	0.1447	0.0548	9.9611	1.4352	0.5050	0.0581
18	0.7671	0.2343	0.0840	9.5824	0.8156	0.5783	0.1184
19	1.7604	0.4114	0.1168	9.8686	1.7798	0.7100	0.1336
20	0.8772	0.8415	0.2301	1.9538	0.9110	0.9621	0.2482

Table 2. Descriptive Statistics for the True Item Parameters Used in the Simulation Study.

20-Item Test	a	b	c	d
Mean	1.0208	0.3844	0.1580	5.3603
Median	0.9430	0.3730	0.1660	5.8435
Standard Deviation	0.2746	1.0139	0.0537	3.0369
Kurtosis	-0.2001	-0.4648	-0.9133	-1.0680
Skewness	0.6507	-0.2669	-0.4176	-0.3811
Minimum	0.6280	-1.4930	0.0580	0.1400
Maximum	1.6510	2.1990	0.2300	9.4960
60-Item Test				
Mean	1.0352	0.3243	0.1485	4.6711
Median	0.9965	0.2725	0.1535	4.6415
Standard Deviation	0.2583	0.9317	0.0494	2.9019
Kurtosis	-0.7320	-0.5126	-0.8017	-1.1416
Skewness	0.3929	-0.2652	-0.0559	0.1330
Minimum	0.6280	-1.8160	0.0580	0.1400
Maximum	1.6510	2.1990	0.2510	9.7980

Table 3. The True and Estimated Item Parameters for Different Response Time Data and Model.

True item parameters					With z term				Without z term				Regular 3PL model		
Item	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c
01	0.717	-1.493	0.159	7.055	0.737	-0.709	0.168	0.073	0.823	-1.225	0.167	5.450	0.803	-0.493	0.156
02	0.826	-1.364	0.173	5.795	0.677	-0.773	0.176	0.049	0.720	-1.305	0.154	4.857	0.707	-0.571	0.158
03	0.929	-1.175	0.156	3.019	0.783	-0.916	0.184	0.053	0.832	-1.090	0.151	2.358	0.835	-0.773	0.156
04	0.707	-0.156	0.211	0.140	0.737	0.001	0.277	0.081	0.853	0.188	0.337	0.085	0.839	0.178	0.331
05	1.040	-0.087	0.227	8.145	0.808	0.521	0.188	0.091	1.107	-0.047	0.215	6.957	1.019	0.879	0.210
06	1.651	-0.223	0.065	0.454	1.935	-0.106	0.110	0.020	1.846	-0.238	0.101	1.025	1.795	-0.099	0.098
07	1.148	0.047	0.183	8.626	1.073	0.884	0.193	0.069	1.111	0.251	0.177	7.435	1.072	1.237	0.184
08	0.920	-0.116	0.223	3.735	0.763	0.124	0.174	0.091	0.897	-0.272	0.196	4.959	0.836	0.300	0.181
09	0.628	-0.045	0.212	8.714	0.564	1.083	0.234	0.192	0.737	0.037	0.241	8.716	0.663	1.414	0.227
10	1.084	0.498	0.184	9.496	1.011	1.508	0.203	0.326	1.422	0.380	0.206	9.871	0.959	1.994	0.197
11	0.957	0.289	0.135	5.249	0.952	0.706	0.120	0.096	1.059	0.078	0.113	6.595	0.921	0.937	0.098
12	0.911	1.356	0.097	0.535	0.911	1.490	0.103	0.109	1.005	1.457	0.112	0.282	0.938	1.521	0.107
13	1.135	1.096	0.230	4.687	0.627	1.733	0.195	0.353	1.071	1.052	0.240	6.125	0.768	2.055	0.218
14	1.411	0.457	0.058	6.227	1.259	1.050	0.066	0.046	1.505	0.462	0.075	6.538	1.297	1.343	0.067
15	1.373	1.186	0.094	2.638	1.481	1.373	0.089	0.072	1.536	1.314	0.086	1.437	1.417	1.563	0.084
16	0.779	1.107	0.189	8.298	0.479	1.573	0.128	4.252	0.495	1.073	0.130	9.795	0.492	2.472	0.146
17	0.741	1.026	0.121	5.892	0.858	1.554	0.164	0.258	0.820	1.199	0.152	5.815	0.817	1.951	0.156
18	1.333	1.370	0.192	9.110	1.146	1.931	0.149	0.986	1.451	0.907	0.156	9.689	0.844	2.648	0.143
19	1.250	1.715	0.107	6.951	1.212	2.343	0.121	0.374	1.462	1.961	0.114	4.085	1.174	2.633	0.112
20	0.875	2.199	0.144	2.439	1.242	2.283	0.178	0.804	1.062	2.071	0.172	3.916	0.986	2.680	0.170

Table 4. The Across-Item Means and SDs (in parentheses) for Absolute Bias , SE, and RMSE of the Item Parameter Estimates

N	a			b			c			d		
	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
20-Item Test												
1,000	0.1256	0.3295	0.3556	0.0290	0.1702	0.1747	0.0115	0.0285	0.0324	0.6877	1.3588	1.5671
	(0.1062)	(0.2983)	(0.3132)	(0.0233)	(0.0716)	(0.07)	(0.0144)	(0.0108)	(0.0144)	(0.6298)	(0.7095)	(0.8696)
2,000	0.0629	0.1827	0.1957	0.0273	0.1246	0.1301	0.0073	0.0235	0.0257	0.7100	1.0465	1.2945
	(0.0601)	(0.116)	(0.1267)	(0.0217)	(0.048)	(0.0457)	(0.0101)	(0.0104)	(0.0123)	(0.5303)	(0.6231)	(0.7675)
4,000	0.0322	0.1215	0.1269	0.0226	0.0909	0.0952	0.0058	0.0189	0.0205	0.6291	0.7498	1.0080
	(0.0261)	(0.0584)	(0.0614)	(0.0161)	(0.0322)	(0.0314)	(0.0073)	(0.0106)	(0.0117)	(0.4785)	(0.4452)	(0.6049)
60-Item Test												
1,000	0.0816	0.1982	0.2169	0.0411	0.1395	0.1485	0.0112	0.0264	0.0301	0.4525	1.0644	1.1904
	(0.0593)	(0.1199)	(0.1296)	(0.0255)	(0.0406)	(0.037)	(0.0118)	(0.0094)	(0.012)	(0.4317)	(0.5348)	(0.6257)
2,000	0.0365	0.1306	0.1366	0.0394	0.1040	0.1133	0.0077	0.0222	0.0242	0.4088	0.7668	0.8915
	(0.0246)	(0.0648)	(0.0673)	(0.0195)	(0.0298)	(0.028)	(0.0087)	(0.0098)	(0.0116)	(0.3374)	(0.4164)	(0.4968)
4,000	0.0205	0.0902	0.0932	0.0309	0.0766	0.0839	0.0054	0.0171	0.0184	0.4284	0.5391	0.7053
	(0.0142)	(0.0392)	(0.04)	(0.0143)	(0.0241)	(0.0238)	(0.0071)	(0.0092)	(0.0109)	(0.3044)	(0.3032)	(0.401)

Table 5. The Across-Replication Means and SDs (in parentheses) for Correlations Between Estimated and True Item Parameters.

N	a	b	c	d
20-Item Test				
1,000	0.7218	0.9862	0.8101	0.8467
	(0.1453)	(0.0086)	(0.0699)	(0.0752)
2,000	0.8530	0.9923	0.8743	0.9012
	(0.0695)	(0.0043)	(0.0492)	(0.0521)
4,000	0.9170	0.9959	0.9151	0.9508
	(0.0365)	(0.0017)	(0.038)	(0.025)
60-Item Test				
1,000	0.7988	0.9883	0.8224	0.9027
	(0.0564)	(0.0027)	(0.0369)	(0.0297)
2,000	0.8878	0.9937	0.8766	0.9478
	(0.0319)	(0.0014)	(0.0286)	(0.0144)
4,000	0.9395	0.9965	0.9184	0.9733
	(0.0145)	(0.0008)	(0.0242)	(0.0067)

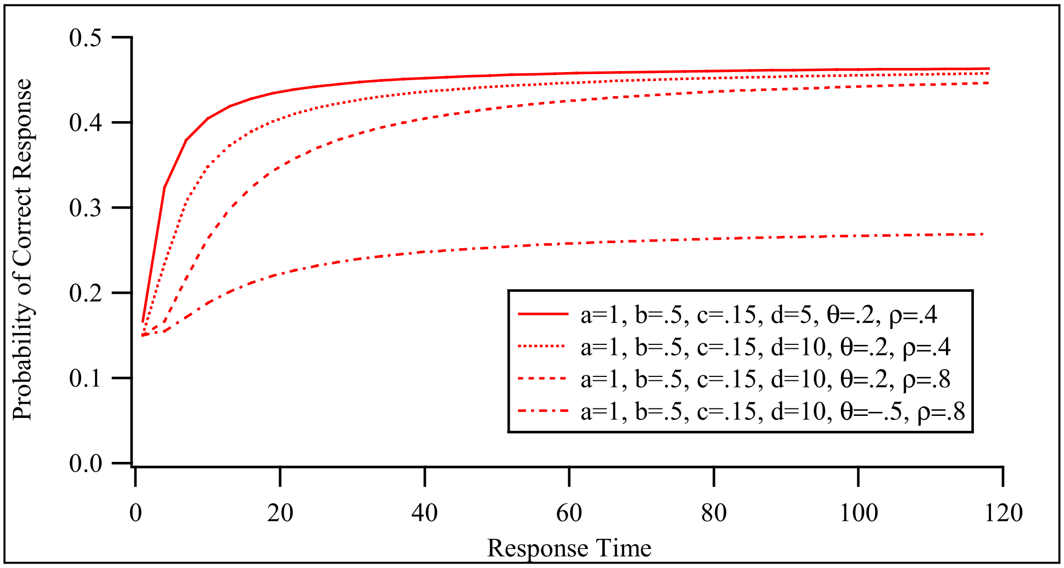

Figure 1. Relationship between Response Time and Probability of Correct Response under the 4PLRT Model.

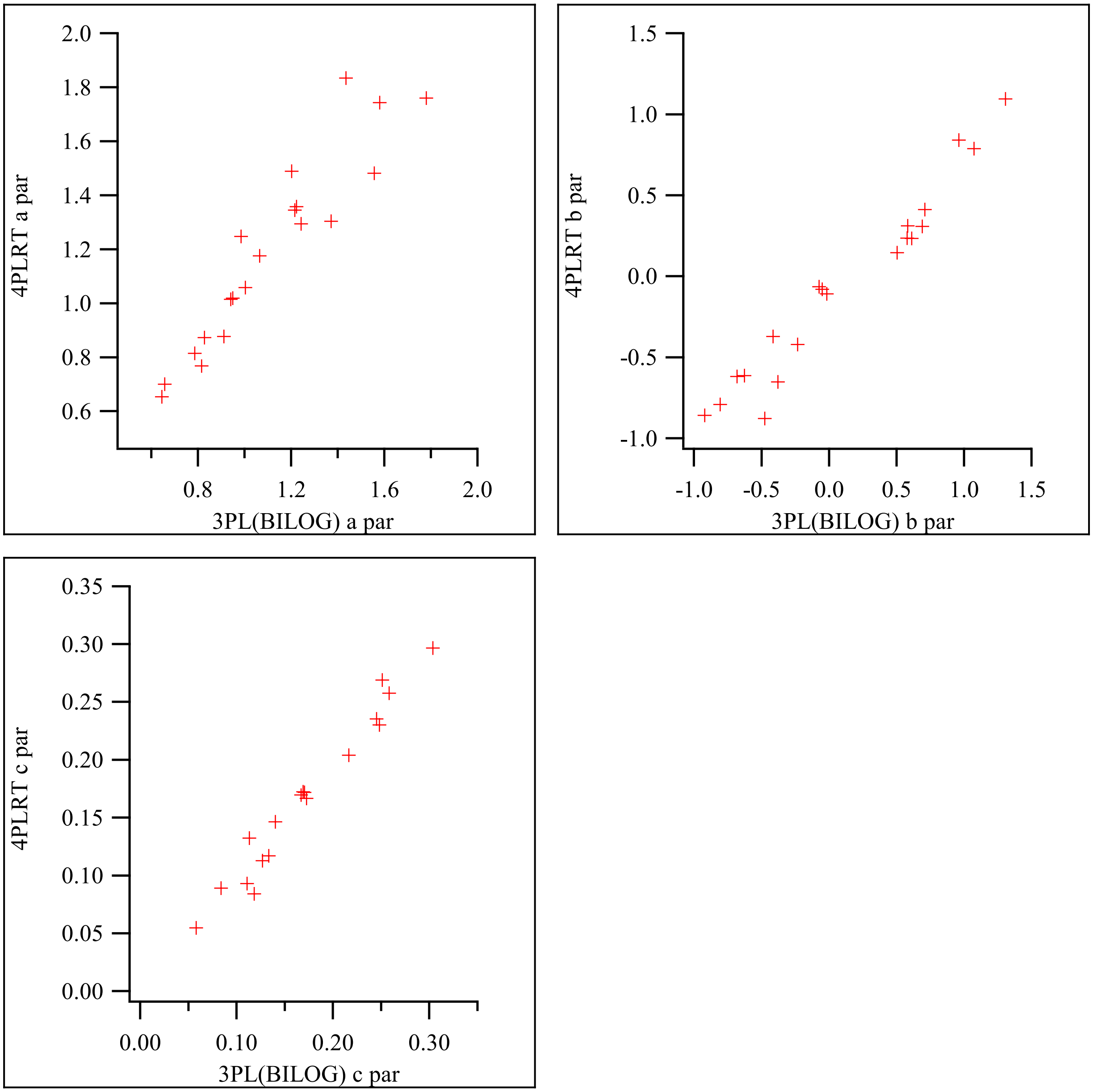

Figure 2. Item parameter estiamtes from the 4PLRT model versus the BILOG estimates for the Math items.

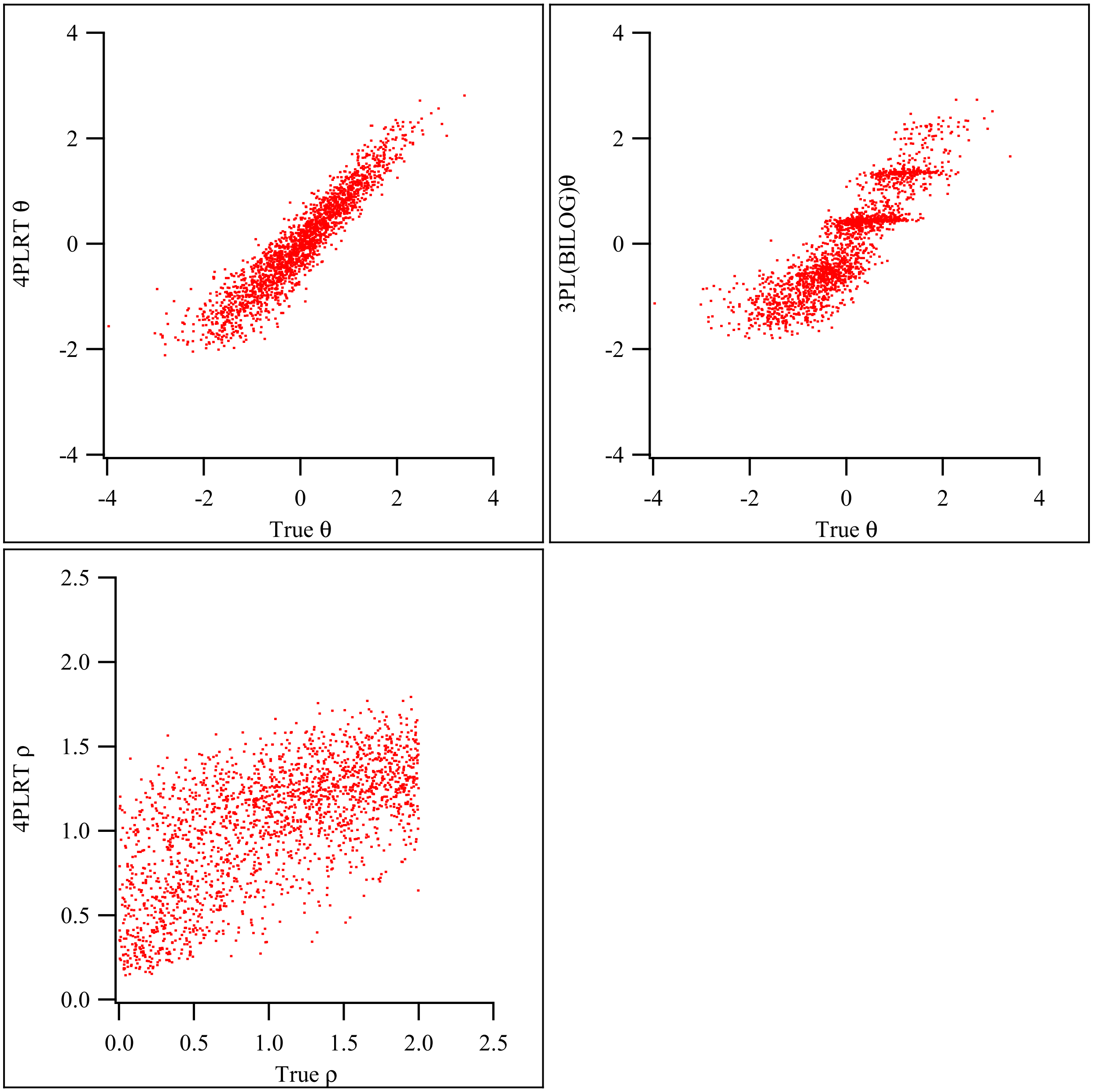

Figure 3. Plots of estimated and true person parameters.