Speedup Analysis for Parallel Implementation of Model of Response Accuracy and Response Time in Computerized Testing

Tianyou Wang[1] and Jun Ni[2,3]

[1] Center for Advanced Study in Measurement and Assessment, College of Education
[2] Information Technology Services
[3] Department of Computer Science, College of Liberal Arts
University of Iowa, Iowa City, IA 52242, USA
{tianyou-wang,jun-ni}@uiowa.edu

Abstract. Recently, computerized testing has revolutionized the testing industry by offering an array of benefits for both the examinees and the test users. It potentially enhances the quality of education at all levels. Computerized testing also makes available new types of data such as item-response time and human response at the individual item level. However, the present models of these responses demand much computing time. This paper presents an analysis for a parallel implementation of the models by estimating parallel speedup and efficiency. The model provides an insight into the success of advanced computerized testing study using high performance computing (HPC) technology.

Keywords: social implications of IT, educational and psychological testing, computerized testing, HPC, parallel computing, education measurement.

1 Introduction

In the last two decades, computerized testing has gradually evolved from a research and experimental stage to applications. Large-scale standardized testing programs such as the GRE, TOEFL and GMAT have routinely adopted a computerized testing mode. Computerized testing has revolutionized the testing industry by offering highly flexible testing schedules, instant scoring, and high efficiency. Computerized testing has also been used in assisting classroom learning by offering interactive instruction and instant feedback, which can potentially increase the quality of education at all levels.

Computerized testing generates new dimensions of data that are not available on the traditional paper-and-pencil testing, such as the data of item response time at the individual item level. Such availability of new testing data opens a new door to solve the obstacles to psychometric theory and practice. For example, traditional models of the Item Response Theory (IRT) model, as the basic psychometric tools for computerized testing, focused on modeling response accuracy. In those models, each examinee typically has only one person parameter; i.e., the human personal ability parameter. Recently-proposed models can not only model the accuracy of response

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 34–41, 2007.

but also account for response time. A general feature of such models is that each examinee has the person parameters (e.g. speed parameter) in addition to the ability parameters. However, due to the extra person parameters, the parameter calibration procedure of an item response model creates a huge demand on the computing power, especially for the calibration of EM algorithm-based item response models. The EM algorithm treats the person parameters as missing data marginalized through an integration of computing the marginal likelihood function. The models that take consideration of person parameters require additional layers of integrations; thus greatly increasing the computation time. For example, in our previous models with two person parameters, even with a small number of items and a relatively small sample size, one calibration takes about several computational hours per PC with 2.8 GHz. The model computation for a large number of items and/or big sample size makes a single calibration unbearably long. In real psychometric applications, simulations are needed to estimate the precision of the calibration procedures and each simulation involves hundreds or even thousands of calibrations, which makes it almost impossible to accomplish computational tasks using conventional computing facilities.

Previous studies indicate that modeling a distribution of response time is quite intriguing and will probably require different sub-models for different testing situations such as different subject materials and testing time constraints. Typically, in an appropriate psychometric model, large-scale simulations are needed to accurately estimate the precision of the calibration procedures. Such simulations usually contain hundreds of calibrations, which make it almost impossible to carry out, based on a single processor computer and its computing capability. This motivates us to consider an alternative approach – to conduct simulations of EM-algorithm-based calibrations using high performance computing (HPC) technology.

The paper presents a speedup analysis for a parallel implementation of a model of response accuracy and response time in computerized testing. The parallel algorithm and its implementation appropriately simulate the joint distributions of response accuracy and response time for different kinds of testing. The analytical model is based on data decomposition and message passing. The model provides insight into large-scale parallel item-response models in computerized testing. The paper is organized in the following sections. Section 1 presents the related work. Section 2 briefly depicts the item response models and the associated numerical algorithms. Section 3 addresses the parallel implementation, domain decomposition, and performance prediction which is followed by a conclusion.

2 Related Work

One of the benefits of computerized testing is that it makes response-time data available at the level of individual items. This new availability, with different dimensions 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 utilize the response-time data. Schnipke and Scrams [3] provided a comprehensive review on this topic and presented the state-of-the-art research on response time data. Most of the previous models on response time consider the response time as a

dependent variable. These models did not consider the relevance between response accuracy and response time. The few exceptions are by [2,5,7]. The models developed by Verhelst et al and by Roskam were similar in that the probability of correct response approaches 1 as response time goes to infinity. Thus, these models were applicable only to speed measurement tests (i.e., the main purpose is to measurement the examinee's speed) where an unlimited response time is guaranteed to achieve a correct response. Alternatively, Thissen's model [5] applies to power tests (i.e., the main purpose is to measure ability rather than speed). His model used a logistic model for the marginal distribution of response accuracy, and a lognormal model for the marginal distribution of response time. However, this model has a limitation due to the assumption that response accuracy and response time are independent; thus their joint distribution can be expressed as the product of the marginal distributions of response accuracy and response time, although the two marginal distributions share some common parameters. This assumption does not generally hold in realistic testing situations. Recently, Wang and Hanson proposed a four-parameter logistic response-time (4PLRT) model [9] in which the response accuracy and time are modeled correlatively. Besides, the model considers the response time as an independent variable, which impacts the final probability of a correct response. In the model formulation, as response time goes to infinity, the probability of a correct response is consistent with the one described by a regular three parameter logistic (3PL) model. The model is applicable to power tests where an unlimited response time does not guarantee a correct response. The model makes an assumption that the item response time is independent of person parameters. This assumption was given to avoid modeling the distribution of response time and consequently simplifying the calibration procedure. This assumption, however, poses a severe limitation to the applicability of their model. In order to eliminate this limitation, a model of the joint distribution of response accuracy and time is needed. Bloxom [1] presented two different methods to describe such distribution. One way is to model the conditional distribution of response accuracy at a given response time and then to multiply the distribution with the marginal distribution of response time. The other way is to model the conditional distribution of response time with a given response accuracy and then to multiply the distribution with the marginal distribution of response accuracy. In both approaches, there is a variety of methods for modeling the conditional and marginal distributions. Most of these probability distributions have not been thoroughly explored in literature. Wang explored a simple model with the first approach, using a one-parameter Weibull distribution for the marginal distribution of response time [8,10], and achieved limited success in terms of model-data fit and calibration precision. One great obstacle with this line of research is the high demand of computing speed. With only the two person parameters model [8,10], a single calibration for a test of 20 items and 1000 examinees can take more than two hours on a single processor PC. Wang's new model adopted the 4PLRT model as the conditional model, except that it omitted one person parameter in order to avoid three person parameters in the joint distribution because of the infeasibility of computational demand. One possible extension of Wang's new model would be use the 4PLRT model and allow the model for the joint

distribution to have three person parameters. With that extension, people can use HPC technology to increase desired computational power.

2.1 Item Response Models

Below is a description of the first model [8,10] which extends 4PLRT model [9] as the conditional model of response accuracy given response time, without the omission of the person slowness parameter, and using the one-parameter Weibull model for the marginal distribution of 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 joint distribution model, the joint distribution of y_{ij} and t_{ij} is expressed as:

$$f(y_{ij},t_{ij}\mid\theta_i,\rho_{1i},\rho_{2i},\delta_j)=f(y_{ij}\mid t_{ij},\theta_i,\rho_{1i},\delta_j)f(t_{ij}\mid\theta_i,\rho_{2i},\delta_j), \tag{1}$$

The detailed model can be found in [10].

2.2 Model 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:

$$L(\mathbf{Y},\mathbf{T}\mid\mathbf{D},\mathbf{p})=\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\pi_{kl}\prod_{j=1}^{J}f(y_{ij},t_{ij}\mid q_k,u_l,v_m,\mathbf{d}_j)\right), \tag{2}$$

where $\mathbf{D}$ is the set of item parameters for all items $(\mathbf{d}_j, j=1,\ldots,J)$. The corresponding likelihood for the complete data is:

$$L(\mathbf{Y},\mathbf{T},\mathbf{q},\mathbf{r}\mid\mathbf{D},\mathbf{p})=\prod_{i=1}^{N}\left(\prod_{j=1}^{J}f(y_{ij},t_{ij}\mid\theta_i,\rho_{1i},\rho_{2i},,\mathbf{d}_j)\right)f(\theta_i,\rho_{1i},\rho_{2i},\mid\mathbf{p}), \tag{3}$$

where $f(\theta_i,\rho_i\mid\mathbf{p})=\pi_{kl}$ 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 $f(\theta_i,\rho_i\mid t_{ij},\mathbf{p})=f(\theta_i,\rho_i\mid\mathbf{p})$. This assumption was necessary in [9] because in that paper, t_{ij} was treated as a c 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 response time distribution is specified. The details about the log-likelihood can be expressed in [8,10].

$$log[L(\mathbf{Y},\mathbf{q},\mathbf{r}\mid\mathbf{T},\mathbf{D},\mathbf{p})]=\sum_{j=1}^{J}\sum_{i=1}^{N}log[f(y_{ij},t_{ij}\mid\theta_i,\rho_{1i},\rho_{2i},,\mathbf{d}_j)]+\sum_{i=1}^{N}log[f(\theta_i,\rho_{1i},\rho_{2i},\mid\mathbf{p})]. \tag{4}$$

The computations to be performed in the E and M steps of the EM algorithm are described in the next two sections. Since the E-step dominates most of the computation time, while M-step's computation time is relatively little; therefore, the parallel model, is only focused on the E-step domain decomposition. **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 $(\mathbf{q},\mathbf{r})$, given the observed data $(\mathbf{Y},\mathbf{T})$, and the fixed values of the parameters $\mathbf{D}^{(s)}$ and $\mathbf{p}^{(s)}$ obtained in the M step of iteration $s-1$ (with some type of starting values for the parameters are used for $\mathbf{D}^{(0)}$ and $\mathbf{p}^{(0)}$). The expected complete data log-likelihood is given by (Woodruff and Hanson, 1996):

$$\phi(\mathbf{D}) + \psi(\mathbf{p}) \tag{5}$$

where $\phi(\mathbf{D})$ is the log-likelihood which accounts for item parameter. It dominates the overall computation. It can be expressed as

$$\phi(\mathbf{D}) = \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} log[f(y_{ij}, t_{ij} \mid q_k, u_l, v_m, \mathbf{d}_j)] f(q_k, u_l, v_m \mid \mathbf{y_i}, \mathbf{t_i}, \mathbf{D}^{(s)}, \mathbf{p}^{(s)}) \tag{6}$$

The second part, in Equation (11), $\psi(\mathbf{p})$ is the log-likelihood which accounts for the person parameter. It requires less computation time.

$$\psi(\mathbf{p}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} log\pi_{kl} f(q_k, u_l, v_m \mid \mathbf{y_i}, \mathbf{t_i}, \mathbf{D}^{(s)}, \mathbf{p}^{(s)}) \tag{7}$$

Once the two parts are calculated, one can calculate the joint-distribution function of $f(y_{ij}, t_{ij} \mid q_k, u_l, v_m, \mathbf{d}_j)$.

3 Parallel Implementation

3.1 Domain Decomposition and Message Communication

In the proposed item-response model with EM algorithms, the major computation of parameter estimates is based on the maximization the observed data likelihood, given by Equation (5). The first part is $\phi(\mathbf{D})$. Based on the previous experience, it takes most of the computation time. The second part is $\psi(\mathbf{p})$. It also takes time, but not as much as the first term in Equation (5). There are two approaches we can use.

After task decomposition, we conducted the second-layer data decomposition. On total summation, the domain is decomposed into several sub-domains on which each sub-summation will be performed on a single processor, for example,

$$\phi(\mathbf{D}) = \sum_{np=1}^{NP}\{\sum_{j=1}^{J_{np}}\sum_{i=1}^{N_{np}}\sum_{k=1}^{K_{np}}\sum_{l=1}^{L_{np}} log[f(y_{ij}, t_{ij} \mid q_k, u_l, \mathbf{d}_j)] f(q_k, u_l \mid \mathbf{y_i}, \mathbf{t_i}, \mathbf{D^{(s)}}, \mathbf{p^{(s)}})\} \tag{8}$$

$$(J = \sum_{np=1} J_{np}; N = \sum_{np=1} N_{np}; K = \sum_{np=1} K_{np}; L = \sum_{np=1} L_{np})$$

We may also try to decompose partial summation in different manors such as

$$\phi(\mathbf{D}) = \sum_{j=1}^{J}\{\sum_{np=1}^{NP}\sum_{i=1}^{N_{np}}\sum_{k=1}^{K_{np}}\sum_{l=1}^{L_{np}} log[f(y_{ij}, t_{ij} \mid q_k, u_l, \mathbf{d}_j)]f(q_k, u_l \mid \mathbf{y_i}, \mathbf{t_i}, \mathbf{D^{(s)}}, \mathbf{p^{(s)}})\} \quad (9)$$

We are studying the effects of different data-decomposition in terms of computational time, speedup, and efficiency.

Similarly, we can parallel the computation for the second part. In this domain decomposition algorithm, the load balance between the computations for (D) and $\Psi(\mathbf{p})$ on different HPC clusters should be taken into consideration. However, it should provide significant improvement in the efficiency of computations. In addition, since the computations between Φ(D) and $\Psi(\mathbf{p})$ are loosely coupled, this computation can be deployed on distributed environments with less concern of network latency. It can be a successful application for a computational grid.

The parallel implementation can be processed using C/C++ with Message Passing Interface (MPI), a parallel library to access the communications between computational processors. Since there is less communication between processors, we can experiment block/non-block communication, and definitely gather/scatter functions to ensure data communication between the master node and the computational nodes. The sub-summation can be operated by computational node and total summation result can be integrated together by master node.

3.2 Analysis of HPC Benchmark Performance and Preliminary Research

The parallel speedup can be estimated by the following simple expression

$$\begin{aligned} T_{sq} &= T_{sq}^{\Phi} + T_{sq}^{\Psi} = T_{input,sq}^{\Phi} + T_{sum(j,i,k,l),sq}^{\Phi} + T_{input,sq}^{\Psi} + T_{sum(i,k,l),sq}^{\Psi} \\ T_{pa} &= T_{pa}^{\Phi} + T_{pa}^{\Psi} = T_{input,pa}^{\Phi} + T_{sum(j,i,k,l),ps}^{\Phi} + T_{com,ps}^{\Phi} \end{aligned} \quad (10)$$

Where T is the time used to accomplish a task. The subscripts *sq* and *pa* stand for sequential and parallel times, and the subscripts *sq* and *pa* stand for sequential and parallel computing; the subscripts *input*, *sum* and *com* refer to stands for input data, summation (major computation), and data or message communication. The parallel speedup can then be calculated as

$$s_{np} = \frac{T_{sq}^{\Phi} + T_{sq}^{\Psi}}{T_{pa}^{\Phi} + T_{pa}^{\Psi}} \approx \frac{T_{sq}^{\Phi} + T_{sq}^{\Psi}}{T_{pa}^{\Phi}} = \frac{T_{input,sq}^{\Phi} + T_{sum(j,i,k,l),sq}^{\Phi} + T_{input,sq}^{\Psi} + T_{sum(i,k,l),sq}^{\Psi}}{T_{input,pa}^{\Phi} + T_{sum(j,i,k,l),ps}^{\Phi} + T_{com,ps}^{\Phi}} \quad (11)$$

We can set $T_{pa}^{\Psi} = \gamma T_{pa}^{\Phi}$, where γ is the ratio of computation time used for item parameter likelihood distribution(s) and the one used for the person parameter distribution. In general, since $\gamma << 1$, and if we distribute the computation of T_{pa}^{Ψ} on another HPC system, the speedup can be simplified as

$$s_{np} \approx \frac{T_{sum(j,i,k,l),sq}^{\Phi} + T_{sum(i,k,l),sq}^{\Psi}}{T_{sum(j,i,k,l),ps}^{\Phi} + T_{com,ps}^{\Phi}} = \frac{T_{sum(j,i,k,l),sq}^{\Phi} + T_{sum(j,i,k,l),sq}^{\Phi} / N}{T_{sum(j,i,k,l),sq}^{\Phi} / n_p + T_{com,ps}^{\Phi}} \quad (12)$$

where n_p is the total number of processors of the first HPC Cluster. The estimation of speedup can be expressed as

$$s_{np} \approx \frac{(1+1/N)}{1+\omega n_p} n_p \qquad \omega = T^{\Phi}_{com,ps} / T^{\Phi}_{sum(j,i,k,l),sq} \tag{13}$$

where the ω is the ratio of time required for data communication and the time used to calculate the total summation in sequential computing. In general, N is in the range of 1000-5000 and $1/N$ is relatively small. The parallel speedup totally depends on the value of n_p and ω. Based on the previous experience, the ratio ω is also very small. The parallel speedup should be in linear increase. Fig. 1(a) plots the speedup vs. the number of processors used when ω takes different values. The smaller the ω value is, the better the performance using parallel computing. Since N ranges from 1000 to 5000, its influence on the parallel performance can be eliminated. The ω strongly influences the parallel performance. Fig. 1(b) plots the corresponding parallel efficiency vs. the number of processors employed. For a low value of ω, the performance is better. In the EM-based algorithm, the communication is relatively low, and data communication does not dominate the computation and computation is very intensive. This is idealized case for parallel computing. That is because the major computation in each simulation of calibration is based on discrete integration. However, the correlation of the distribution function may require extra compensation time.

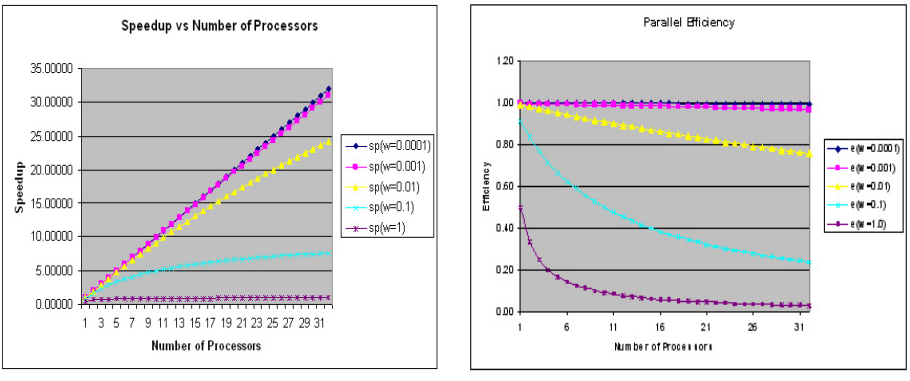

Fig. 1. (a) Estimation of speedup vs. number of processors with N=1000, and =0.0001, 0.001, 0.01, 0.1, 1.0, respectively; (b) Estimation of parallel efficiency vs. number of processors with N=1000, and =0.0001, 0.001, 0.01, 0.1, 1.0, respectively

4 Conclusion

This paper first addresses an interdisciplinary paradigm that integrates computer science and psychometric research that can be applied to computerized testing. The paper presents a parallel analysis of HPC-based the response time data available in computerized testing. Such a model not only improves the measurement precision of

traditional academic ability traits, but also provides valuable information about the examinees' other personality traits such as their problem-solving speed and decision-making strategies. Thus, this line of research and practice can potentially broaden the educational value of computerized testing. One critical phase of developing such item response models is to develop the calibration procedure and evaluate its precision using HPC resources. With the additional person parameters needed to model response time, the simulation of EM algorithm-based calibration procedure can be feasible. The implementation of parallel-algorithm-based simulation in computerized testing studies serves as a pioneer effort in to promote HPC-powered psychometric modeling and computerized testing.

References

1. Bloxom, B.: Considerations in Psychometric Modeling of Response time. Psychometrika. 50. (1985) 383-397
2. Roskam, E. E.: Models for Speed and Time-limit Tests. In W. J. van der Linden & R. K. Hambleton (Eds.). Handbook of Modern Item Response Theory New York: Springer. (1997) 187-208
3. Schnipke, D. L., Scrams, D. J.: Exploring Issues of Examinee Behavior: Insights Gained from Response-time Analyses. In the ETS colloquium on "Computer Based Testing: Building the Foundation for Future Assessment." September 25-26. Philadelphia, PA. (1998).
4. Schnipke, D. L., Scrams, D. J.: Modeling Item Response Times with a Two-state Mixture Model: A New Method of Measuring Speediness. Journal of Educational Measurement, 34, (1997) 213-232
5. Thissen, D.: Timed Testing: An Approach Using Item Response Theory. In New horizons in Testing: Latent Trait test Theory and Computerized Adaptive Testing. D. J. Weiss (Ed.). New York: Academic Press. (1983) 179-203
6. Van Breukelen, G. J. P.: Concentration, Speed, and Precision in Mental Tests. Unpublished Doctoral Dissertation, University of Nijmegen, Netherlands. (1989)
7. Verhelst, N. D., Verstralen, H. H. F. M., Jansen, M. G. H.: A logistic Model for Time-limit Tests. In W. J. van der Linden and R. K. Hambleton (Eds.). In Handbook of Modern Item Response Theory. New York: Springer (1997) 169-185
8. Wang, T. A Model for the Joint Distribution of Item Response and Response Time using a One-Parameter Weibull Distribution. CASMA Research Report, No. 20. Iowa City, IA. CASMA. (2006) (http://www.education.uiowa.edu/casma/documents/20td4plrt.pdf)
9. Wang, T. Hanson, B. A.: Development and Calibration of an Item Response Model that Incorporates Response time. Applied Psychological Measurement. 29 (2005) 323-339
10. Wang, T. Zhang, J.: Optimal Partitioning of Testing Time: Theoretical Properties and Practical Implications. Psychometrika. 71 (2006) 105-120
11. Woodruff, D. J., Hanson, B. A.: Estimation of Item Response Models Using the EM Algorithm for Finite Mixture. ACT Research Report 96-6 Iowa City, IA: ACT, Inc. (1996)