Estimation of Item Parameters as a Method for Supporting Learner/Examinee Assessment

Mitra Mirzarezaee [1,2], Kambiz Badie[1], Mehdi Dehghan[1,3], Mahmood Kharrat[1]

1-Iran Telecommunication Research Center (ITRC), Tehran, Iran
2- Dept. of Computer Eng., Islamic Azad University-Science and Research Branch
3- Dept. of Computer Eng., AmirKabir University of Technology, Tehran, Iran

Abstract

Item Response Theory (IRT) is a model for expressing the association between an individual's response to an item and the underlying latent variable (ability) being measured by the instrument. Item Characteristic Curves (ICCs) are one of the basic blocks of an Item Response Theory, and their parameters (difficulty, discrimination and guessing) must be estimated accurately. The estimated parameters will subsequently be used to form ICCs of an exam upon which other latter judgments about examinees' abilities will be made. Regarding the importance of assessment in learning process and reaching accurate estimations about learners' abilities, this paper is focused on a comparative approach for finding the best technique of estimating these parameters. The criterion for such an optimization is the chi-square goodness of fit. Results show that Genetic Algorithms obtain the best estimations among two other applied techniques.

Keywords: *Item Response Theory, Item Characteristic Curve, Intelligent Tutoring System, Assessment, Genetic Algorithm, Simulated Annealing, Gradient Descent.*

1. Introduction

Intelligent Tutoring Systems (ITS) have to provide individualized tutoring and instructions. They have a multidisciplinary approach, and have been benefited by many fields of research such as psychology, education, artificial intelligence, networking and so on, to build finally an electronic system of personalized teaching and learning. During a course, an Intelligent Tutoring System has to observe learners' abilities and improvements to decide on the next steps of tutoring. Therefore they are to be concerned with experiments of educational assessment.

In educational assessment, we observe what students say, do, or make in a few particular circumstances and attempt to infer what they know, can do, or have accomplished more generally. Some links in the chain of inference depend on statistical models and probability-based reasoning, and it is with these links that terms such as validity, reliability, and comparability are typically associated—"psychometric principles," as it were. Familiar formulas and procedures from test theory provide working definitions and practical tools for addressing more broadly applicable qualities of the chains of argument from observations to inferences about students, as they are applied to familiar methods of gathering and using assessment data [1].

The increasing need for psychometrically-sound measures calls for better analytical tools beyond what traditional measurement theory (or classical test theory, CTT) methods can provide [3]. Item Response Theory (IRT) has a number of advantages over CTT methods to assess learning

outcomes, including (a) detailed descriptions of the performance of individual test items; (b) indices of item- and scale-level precision that are free to vary across the full range of possible scores; (c) assessments of item- and test-level bias with respect to demographic subgroups of respondents; (d) measures of each examinee's response-profile quality and consistency; and (e) computer-adaptive test (CAT) administration, which can dramatically reduce testing time without sacrificing measurement precision[2].

Because of these advantages, IRT is being applied in many research areas to develop new measures or improve existing measures, to investigate group differences in item and scale functioning, to equate scales, and to develop computerized adaptive tests [4].

While the basic concepts of IRT are straight forward, the underlying mathematics is somehow advanced compared to that of CTT, and it is difficult to examine some of these concepts without performing a large number of calculations to obtain usable information which is sometimes a hard and also a time-consuming task.

The goal of this paper is to test some different applicable techniques of optimization to improve IRT functions. Here the focus is on "item characteristic curves".

Organization of this paper is as follows: in section two and three, IRT and its Item Characteristic Curves are introduced. In section four the problem of estimating ICC parameters are explained and it is shown how optimization can be performed using different optimization techniques. Finally the results are compared and the best optimization method is proposed.

2. Item Response Theory

In the past decade, applications of item response theory (IRT) in measurement have been increased considerably, because of its utility in item and scale analysis, scale scoring, and adaptive testing. IRT is a model-based measurement in which trait level estimates depend on both persons' responses and the properties of the questions that were administered [3].

Item response theory (IRT) methods seek to model the way in which latent psychological constructs manifest themselves in terms of observable item responses; this information is useful when developing and evaluating tests, as well as estimating examinees' scores on the latent characteristics in question [2].

3. Item Characteristic Curve

Item characteristic curve is the basic building block of IRT, on which all other constructs depend. It has three technical properties to describe it. *Difficulty* describes where the item functions along the ability scale. It is perhaps a location index. *Discrimination* describes how well an item can differentiate between examinees having abilities below and above the item location. This property essentially reflects the steepness of the curve in its middle section. The steeper the curve, the better the item can discriminate. The third property is *guessing* factor that shows the probability by which an examinee can get items correctly by chance. Thus the probability of correct response includes a small component that is due to guessing [5].

Using these three descriptors, one can describe the general form of the item characteristic curve. These descriptors are also used to discuss the technical properties of an item. It should be noted that these properties say nothing whether the item really measures some facets of the underlying ability or not; that is a question of validity.

There are three basic mathematical models for item characteristic curve to show the relation of the probability of correct response to ability. Each model employs one or more parameters whose numerical values define a particular item characteristic curve. These models are *logistic models*, *Rash* or *one parameter models* and *three parameter model*s [5]. Such mathematical models are needed if one is to develop a measurement theory that can be rigorously defined and is amenable to

further growth. In addition, these models and their parameters provide a vehicle for communicating information about an item's technical properties. Equation for the three-parameter item characteristic curve model is as follows [5]:

$$P(\theta) = c + (1-c)\frac{1}{1+e^{-a(\theta-b)}} \qquad (1)$$

where

b is the difficulty parameter,
a is the discrimination parameter,
c is the guessing parameter,
θ is the ability level.

$P(\theta)$ shows the probability of correct response at ability level θ. The value of *c* does not vary as a function of the ability level. Thus the lowest and the highest ability examinees have the same probability of getting the item correct by guessing. In practice the range of ability levels vary from -3 to +3, the discrimination is between -2.8 and 2.8, and values above 0.35 are not considered acceptable for parameter *c* [5].

4. Estimating Parameters of a Curve

Since the actual values of the parameters of the items in a test are unknown, one of the tasks performed when a test is analyzed under IRT, is to estimate these parameters. The obtained item parameters estimates then provide necessary information for the technical properties of test items. It is wroth mentioning that under Item Response Theory, item parameters are independent of the distribution of examinees over the ability scale. From a practical point of view, it means that the parameters of the total item characteristic curve can be estimated from any segment of the curve.

In a typical test, a sample of *M* examinees responds to the *N* items in the test. The ability scores of these examinees will be distributed over a range of ability levels on the ability scale. These examinees will be divided into *J* groups along the scale so that all the examinees within a given group can have the same ability level θ_j, and there will be m_j examinees within group *j*, where *j=1, 2, 3… J*. Within a particular ability level of θ_j, the observed proportion of correct response is $p(\theta_j) = \frac{r_j}{m_j}$, which is an estimate of the probability of correct response at that ability level [5].

4.1. Application of Maximum likelihood Estimation

In order to find the item characteristic curve that best fits the observed proportion of correct response; we should first select a model for the curve to be fitted. The procedure used to fit the curve is based upon *Maximum likelihood Estimation* [5].

MLE endeavors to find the most "likely" values of distribution parameters for a set of data by maximizing the value of what is called the "likelihood function"[6]. Under this approach, initial values for the item parameters, are established a-prior. Then using these estimates, the value of $p(\theta_j)$ is computed at curve model. The agreement of the observed value of $p(\theta_j)$ and the computed value is determined across all ability groups. Then adjustments to the estimated item parameters are found that results in better agreement. The process of adjusting is continued until the adjustments get so small that little improvement is possible. At this point, the estimation procedure is terminated.

Although the actual MLE procedure is rather complex mathematically and entails very laborious computations that must be performed for every item in a test. In what follows after introducing the chi-square goodness of fit index, we try to find the optimum values of item

parameters using three different techniques of optimization based on minimizing the goodness of fit index.

4.1.1 Chi-Square Goodness of Fit Index of an Item

The agreement of the observed proportion of correct response to those yielded by the fitted curve is measured by the chi-square goodness of fit index. The chi-square test is used to test whether or not a sample of data came from a population with a specific distribution [7]. It is proven that if data is binned, and uncertainties are Gaussian, then χ^2 test is equivalent to MLE [8].

For our purpose chi-square goodness of fit index is defined as follows [5]:

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{[\rho(\theta_j) - P(\theta_j)]^2}{P(\theta_j)Q(\theta_j)} \qquad (2)$$

where

J is the number of ability groups,

θ_j is the ability level of group j,

m_j is the number of examinees having ability θ_j,

$\rho(\theta_j)$ is the observed proportion of correct response for group j,

$P(\theta_j)$ is the computed probability of correct response for group j,

If the value of the obtained index is greater than a criterion value, the item characteristic curve specified by the value of the item parameter estimates is said not to fit the data. The criterion value for the goodness of fit depends on two factors: degree of freedom and percentile points of rejecting the null hypotheses.

4.2 Gradient Descent Technique

After formulating an objective or energy function $E(f)$ and setting the optimality criteria, the simplest way of founding minimum energy or the optimum values for parameters (*a*, *b* and *c*) is to perform gradient descent. Start with an initial configuration; iterates with

$$f^{(t+1)} = f^{(t)} - \mu \nabla E(f^{(t)}) \qquad (3)$$

where

$\mu > 0$ is a step size and $\nabla E(f)$ is the gradient of the energy function.

until the gradient converges to a point f^* for which $\nabla E(f^*) = 0$. Considering $E(f) = \chi^2$ for our case, the gradient to minimize chi-square is composed of the following components

$$\frac{\partial \chi^2}{\partial a} = \frac{\partial \chi^2}{\partial P}\frac{\partial P}{\partial a}; \qquad \frac{\partial \chi^2}{\partial b} = \frac{\partial \chi^2}{\partial P}\frac{\partial P}{\partial b}; \qquad \frac{\partial \chi^2}{\partial c} = \frac{\partial \chi^2}{\partial P}\frac{\partial P}{\partial c} \qquad (4)$$

where

$$\frac{\partial \chi^2}{\partial P} = \sum_{j=1}^{J} \frac{m_j([-2(\rho(\theta_j) - P(\theta_j)][P(\theta_j) - P^2(\theta_j)])}{[P(\theta_j) - P^2(\theta_j)]^2} - \sum_{j=1}^{J} \frac{m_j[1 - 2P(\theta_j)][\rho(\theta_j) - P(\theta_j)]^2}{[P(\theta_j) - P^2(\theta_j)]^2} \qquad (5)$$

and

$$\frac{\partial P}{\partial a} = \sum_{j=1}^{J} \frac{(1-c)(\theta_j - b)e^{-a(\theta_j - b)}}{(1+e^{-a(\theta_j - b)})^2}$$
$$\frac{\partial P}{\partial b} = \sum_{j=1}^{J} \frac{-a(1-c)e^{-a(\theta_j - b)}}{(1+e^{-a(\theta_j - b)})^2} \qquad (6)$$
$$\frac{\partial P}{\partial c} = \sum_{j=1}^{J} (1 - \frac{1}{(1+e^{-a(\theta_j - b)})})$$

At the end of each iteration loop, we update parameter values with the following formulas.

$$a = a - \mu_1 \frac{\partial \chi^2}{\partial a}; \quad b = b - \mu_2 \frac{\partial \chi^2}{\partial b}; \quad c = c - \mu_3 \frac{\partial \chi^2}{\partial c} \qquad (7)$$

A different step size is chosen for updating each parameter, and we would decrease their values as reaching a local minimum to slow down the search.

4.3 Simulated Annealing Technique

Simulated annealing is an optimization method that derived from statistical mechanics. Annealing is the physical process of heating up a solid and then cooling it down slowly until it crystallizes. As the temperature is reduced, the atomic energies decrease. A crystal with regular structure is obtained at the state where the system has minimum energy [10].

The algorithm consists of a sequence of iterations. Each iteration consists of a randomly change in the current solution to create a new solution in the neighborhood of the current. The change in the cost function is computed to decide whether the newly produced solution can be accepted as the current solution or not. If the change is negative, the solution is accepted. Otherwise, it is accepted according to Metropolis's criterion based on boltzman's probability [10].

The following are the characteristics applied for finding the optimum values of item parameters: solutions are composed of three parameters *a*, *b* and *c* initialized in proper ranges. Evaluation is based on the chi-square goodness of fit index and a cooling schedule.

In designing the cooling schedule, four parameters must be specified. These are as follows: an initial temperature 1000, the geometric cooling rule as the temperature update rule by the following formula:

$$T_{i+1} = cT_i \qquad i = 0,1,2,... \qquad (8)$$

where *c* is a temperature factor which is a constant smaller than 1. In our simulation *c* is 0.25, number of iterations to be performed at each temperature step is 20 and stopping criterion for the search is convergence of achieved optimum solutions.

4.4 Genetic Algorithm Optimization Technique

The evolution of a population of individuals is what a genetic algorithm does. GAs operate on a population of potential solutions applying the principle of survival of the fittest to produce (hopefully) better and better approximations to a solution [8].

Chromosomes are the way of coding the solutions composed of genes (characters). For our case the chromosomes are composed of three genes *a*, *b* and *c* parameters respectively. The valid range of each parameter is specified a prior. For the population representation, real coding of values is selected and the population size is set to 20. Having decided the chromosome representation into the decision variable domain, it is possible to assess the performance or fitness of individual members of a population. The chi-square goodness of fit index is chosen as the objective function to characterize an individual performance. The selection procedure is stochastic

universal sampling with discrete recombination crossover of rate 1 and real mutation with the rate of 0.05 as genetic operators. Other technical parameters are as follows: generation gap of 0.8, max number of generations 1200, insertion rate of 0.9.

We used a multiple population approach, the use of which has shown, in most cases, to improve the quality of results obtained using single population GAs [9]. In a multiple population GA, each subpopulation is evolved over generations by a traditional GA and from time to time individuals migrate from one subpopulation to another. The amount of migration and patterns of migration determines how much genetic diversity can occur. The technical parameters of applied multiple-population GA is as follows: number of subpopulations 8, migration rate 0.2, number of genes/migration 20 and number of individuals /subpopulation is 20.

As the fitness of a population may remain static for a number of generations before the superior individual is found. A common practice is to terminate algorithm after a pre-specified number of generations and then test the quality of the best members [9]. If no acceptable solutions are found, the GA may be restarted or a fresh search initiated.

5. Experimental Results and Conclusions

The algorithms were tested within two distinct phases. In the initial phase, a data set of ten different ability levels was used to estimate the parameter values. For each ability level, a group of ten answers generated by students was used, and the probability of correct response for the related ability level was computed. Examinees' test responses (0/1s) were generated as follows: the program reads in a file of calibrated item parameters and generates normally distributed random variables to represent examinees' ability levels. The probability of an examinee obtaining a correct response is calculated and then this probability is compared with a uniform random number to decide the examinee's item response. If the probability is larger than the random number, the examinees are credited a correct response (i.e. an item score of 1), otherwise, a zero. The generated examinees' answers were then used to calculate the observed probability of correct response, $\rho(\theta_j)$.

In this phase, algorithms were tested with more than 1000 different data sets. Here is the result of running these three proposed methods on sample input data sets:

Table 1 – chi-square for the first data set

Method	Data sets				
	1	2	3	4	5
GD	11.757	15.1298	7.9996	22.090	8.6896
SA	11.701	0.0002	5.5802	4.4643	7.8414
MPGA	**11.665**	**0.0001**	**4.8655**	**4.4643**	**7.6530**

Table 2 – chi-square for the second data set

Method	Data sets				
	1	2	3	4	5
GD	10.614	24.765	17.031	7.9995	8.6896
SA	**6.6328**	4.4644	**10.4268**	4.0932	**3.6767**
MPGA	**6.6328**	**4.4643**	**10.4268**	**4.0719**	**3.6767**

Table 3 – chi-square for the third data set

Method	Data sets					
	1	2	3	4	5	6
GD	57.38	9.863	9.116	14.87	7.999	13.98
SA	5.185	9.913	9.563	6.530	5.799	10.86
MPGA	**4.622**	**9.8606**	**8.3686**	**6.526**	**5.3372**	**10.53**

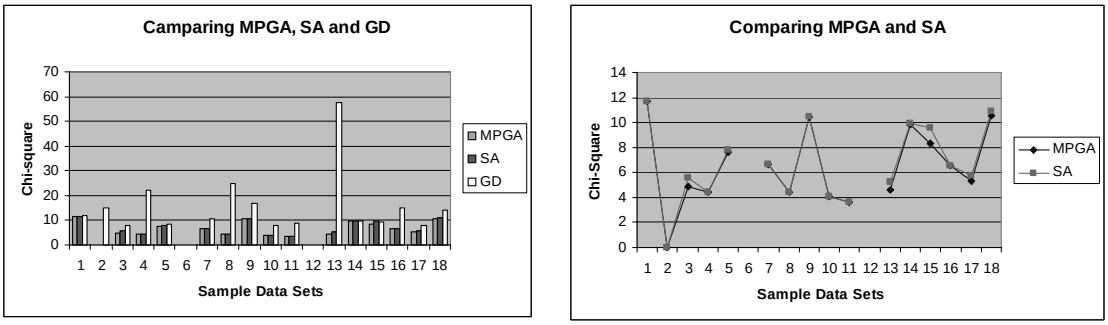

Fig 1- Results of Running MPGA, SA and GD on the Sample Data Sets

According to the tables and diagrams, genetic algorithm and gradient descent show the best and the worst results on the generated data sets respectively, and simulated annealing is placed in the middle. In a gradient descent algorithm, if energy function is convex, convergence is guaranteed by the algorithm. When it is not the case, as it is probably true with our problem, the algorithm gives only a local minimum. Recalculation of $\nabla E(f)$ for new types of ICCs, is another drawback of using Gradient Descent. Therefore for the second phase, we only examined the other remaining methods.

In the second phase, a data set of five different ability levels (weak, average, good, very good, excellent) from a group of 270 students of different primary schools in Tehran whom ability levels are known a-prior, were selected. The collected data (students' answers) was used for estimation of item parameters of 60 multiplication questions. Table 4 shows the results.

Table 4 – Comparison of SA and MPGA

Method	Average Chi-Sqr	Standard Derivation	Best Fit	Worst Fit	Fit-Rate
SA	3.962524	3.450872	0.1319	17.5905	0.822581
MPGA	**2.806771**	**2.030203**	**0.0773**	**9.8126**	**0.951613**

In this table, best (worst) fit is the minimum (maximum) value of the calculated chi-squares within the sample data set which belongs to the best (worst) fitted curves. Fit-Rate shows the percentage of fit within the sample. Data is said to be fitted by the curve if its chi-square value is less than or equal to the number of clusters.

As the results show GA finds a better optimum parameter values than what a simulated annealing does. So, Multiple-Population Genetic Algorithm (MPGA) is proposed as a solution to estimate parameters of an item. It is wroth mentioning that since GA is not proper for online (real-time) operations, in case that it is unavoidable for finding the optimum parameter values, one should pay the trade offs and use a quicker method.

References

[1] Mislevy, R.J., Wilson, M.R., Ercikan, K., Chudowsky, N., Psychometric principles in student assessment, In D. Stufflebeam & T. Kellaghan (Eds.), *International Handbook of Educational Evaluation.*, the Netherlands: Kluwer Academic Press, 2002.

[2] Robert J. Harvey, Allen L. Hammer, Item Response Theory, Virginia Polytechnic Institute & State University and Consulting Psychologists Press Inc.,harvey.psyc.vt.edu/Documents/, 2000.

[3] van der Linden, W., & Hambleton, R. K., Handbook of modern item response theory, Heidelberg: Springer-Verlag, 1997.

[4] Bryce B. Reeve, An Introduction to Modern Measurement Theory, National Cancer Inst., 2002

[5] Frank B. Baker, *The Basics of Item Response Theor*y , ERIC clearinghouse on Assessment and Evaluation, 2001.

[6] Earl Gose and etal, *Pattern Recognition and Image Analysis*, Prentice Hall, USA, 1996.

[7] Hashemi Parast, *Probability and Statistic in Science and Engineering*, khajeh nasir university, fourth edition, 1993 (farsi).

[8] Mitsuo Gen, Runwel Cheng., *Genetic Algorithms and Engineering Design*, John Wiley & Sons Inc., 1997.

[9] A. Chipperfield, P.Fleming, H. Pohlheim, Genetic Algorithm Toolbox for use with Math lab, *IEE Colloqium on Applied Control Techniques Using MATLAB*, 1995.

[10] D.T. Pham, D. Karaboga, *Intelligent Optimization Techniques: genetic algorithms, tabu search, simulated annealing and neural network*, Springer, London, 2000.