An Evaluation of the IRT Models Through Monte Carlo Simulation

Tyler Baur, Dylan Lukes

Faculty Sponsor: Sherwin Toribio, Department of Mathematics

ABSTRACT

Item response theory (IRT) models are commonly used in the realm of educational and psychological testing. The three commonly used IRT models for responses that are graded as either correct or incorrect are the One-parameter, Two-parameter, and Three parameter IRT models. These models are used to calibrate standardized tests and to assess human latent traits such as student's overall intelligence, vocabulary prowess, or mathematical ability. The effectiveness of these tests is regularly debated by students, especially those who are required to take these examinations, as their future may depend on how they are assessed by these IRT models. One good example is the Graduate Record Exam (GRE), which are taken by students and used by universities to assist them in the selection of applicants to their graduate program. The main objective of this research was to study the validity of each IRT model in assessing students' ability. In particular, the effect of the sample size and exam length to the accuracy of the estimates of students' ability and item characteristics were studied. The results were based on an extensive Monte Carlo simulation study done using the statistical software R. Through these simulation studies, we hope to discover the IRT model that is most effective in calibrating standardized tests and in assessing student's raw abilities, and the acceptable exam length and sample size to obtain accurate parameter estimates. Preliminary results indicate that the Three-parameter model is problematic as the estimates of the item characteristics (difficulty, discrimination, and guessing parameters) obtained using this model are biased. Our results in evaluating IRT models confirmed some of our conjectures of the preliminary phase of our investigation. The Three-parameter model proved problematic, as correlations between actual and estimated parameters were deficient. However, the One-parameter IRT model and Two-parameter IRT model performed well as correlations between actual and estimated parameters for difficulty and ability parameters were very well and similar from model to model.

INTRODUCTION

Item response theory (IRT) models are mathematical models that assess latent human characteristics and quantify underlying traits. IRT models are recurrently used in psychological and educational testing in which researchers endeavor to measure underlying abilities of examinees such as intelligence, mathematical prowess, or scholastic aptitudes. These characteristics cannot be quantified directly as one would measure height or eye color. Instruments in the form of exams or questionnaires are commonly used to assess desired latent variables. IRT models are used to calibrate questions and determine particular item characteristics such as difficulty and discrimination. The effectiveness of these tests is regularly debated by students, especially those who are required to take these examinations, as their future may depend on how they are assessed by these IRT models. One good example is the Graduate Record Exam (GRE), which are taken by students and used by universities in the selection of applicants to their graduate program.

The IRT models that we studied are for dichotomous responses – an examinee can either get the item correct or incorrect. These models also assume that examinee's ability is uni-dimensional and that the probability of answering an item correctly is an increasing function of their ability. Thus, individuals with higher ability have higher chance of correctly answering the given question. Although there are other external factors that may contribute to exam results, such as anxiety, motivation, or ability to work quickly, one assumes these to be extraneous.

IRT models have several advantages over the Classical Test Theory models. The property of invariance is inherent within IRT models as item and ability parameters are invariant. That is, ability estimates obtained from different sets of items will be the same, and item parameter estimates obtained from different examinees will be the same. In other words, parameters that characterize an item do not depend on the ability of the examinees used to

calibrate the items, and the parameters that characterize an examinee do not depend on the items included in the exam. Hence, one is able to compare the abilities of different students who took different exams. The three commonly used IRT models for dichotomous response items are, the One-parameter or Rasch model, the Two-parameter IRT model, and the Three-parameter IRT model. The One-parameter model assumes that each item on the test is characterized by a single difficulty parameter, b_j. Items that are more difficult will have higher b_j values than easy items. The Two-parameter IRT model uses both a difficulty parameter (b_j) and a discrimination parameter (a_j) to characterize an item. The discrimination parameter represents how well an item on an exam distinguishes between students of low and high abilities. The greater the capability of an item to do so, the greater will be the value of the discrimination parameter. Finally, the Three-parameter IRT model incorporates a guessing parameter (c_j) into the model. The guessing parameter is the probability that an examinee will be able to get the correct response to an item by pure chance. Each of these models expresses the likelihood of a student with a particular ability (θ_i) to correctly answer a given item in the test. These IRT models are listed below:

Table 1. IRT Models

One-Parameter	$Pr(Y_{ij} = 1) = F(\theta_i - b_j)$
Two-Parameter	$Pr(Y_{ij} = 1) = F(a_j \theta_i - b_j)$
Three-Parameter	$Pr(Y_{ij} = 1) = c_i + (1- c_i) F(a_i \theta_i - b_j)$

The variable Y_{ij} represents the response of examinee *i* to item *j, w*here $Y_{ij} = 1$ when the *i*th examinee gave the correct response to the *j*th item and $Y_{ij} = 0$ otherwise. In all three models, the inverse of the function F is called the link function. Two commonly used link functions in IRT are the standard normal cdf and the standard logistic link. For our research, the standard logistic link was used. Some additional assumptions adhered to within our research included that only one latent or ability trait was to be measured, items are answered independently (i.e., examinees do not increase their chance of getting an item correctly based on previous items), and examinees work on the test independently from one another (i.e., no cheating is allowed).

OBJECTIVES

The main objective of our research is to study the effectiveness of these IRT models in estimating the latent ability of examinees and the accuracy in estimating the item parameters. In particular, we desired to study the accuracy of the parameter estimation process for the One-parameter, Two-parameter, and Three-parameter IRT models. We want to study how the sample size (number of examinees) and the length of the exam affects the accuracy of the parameter estimates, and determine the proper sample size and number of items required to accurately estimate the students' ability for each model. In addition, we also plan to do a comparison of the strengths and weaknesses of these 3 IRT models in terms of how accurately they estimate the model parameters. For example, we want to answer question like, "Can the Rasch One-parameter model correctly estimate students' abilities if two parameter data is used instead of one parameter data?" Finally, this research also intends to study how well the new "ltm" package for the statistical software R is able to implement the estimation procedure to obtain the estimates for the different IRT models.

METHOD

The estimation procedure needed to obtain the parameter estimates in IRT models is quite complex and is extremely difficult to perform manually. In our research, we used the open source statistical package R to perform all calculations and simulations (program codes can be found in the appendix).

To begin, a data matrix was created by writing an algorithm to create an *i* by *j* data matrix composed of Y_{ij} variables representing the response of the *i*th examinee on the *j*th item. As stated above, these responses were dichotomous in nature that follows a Bernoulli random variable in which the probability of answering correctly, p_{ij}, was modeled by a function of the various ability and item parameters. These parameters were generated from a probability distributions that reflects the kind of values they take in theory. The distribution used for each parameter can be seen in Table 2.

Table 2. IRT Parameter Distributions

Parameter	**Distribution**

θ_i	N(0,1)
b_j	N(0,1)
a_j	Unif(.2,2)
c_i	Unif(.2,.25,.33,.5)

Both the ability parameter (θ_i) and the difficulty parameter (b_j) were generated from a standard normal distribution, the discrimination parameter (a_j) was generated from a uniform distribution with a range from .2 to 2, and finally, the guessing parameter (c_j) was generated from a discrete uniform distribution with possible values of .2, .25, .33 and .5 to model the probability of guessing the correct answer from a multiple-choice item. A total of $n + 3k$ parameter values were generated, n values for ability and k values for each of the 3 item parameters. Depending on which IRT model was being evaluated, the appropriate parameter values were selected to be included in the model. Using these generated parameter values, the $n \times k$ probability matrix of p_{ij}'s was computed using the appropriate IRT model. Finally, the simulated $n \times k$ data matrix was generated by sampling from a Bernoulli distribution using success probabilities taken from the $n \times k$ probability matrix of p_{ij}'s.

Once the actual ability, difficulty, discrimination, and guessing parameters were generated, and the creation of the simulated data set was completed, the latent trait model (ltm) R package, written by Dimitris Rizopoulos (Rizopoulos, 2006), was used to estimate the model parameters. The parameter estimates obtained from this package were compared with the actual generated values to assess their accuracy. This procedure was repeated 100 times for different sample sizes and exam lengths. For each of the One-parameter, Two-parameter, and Three-parameter models, five different sample sizes (n=100, 250, 500, 1000, and 5000) and five exam lengths (k = 5, 10, 15, 20, and 30) were used. Ideally, Monte Carlo simulations should have at least 5,000 iterations, but due to time constraint and lack of computing power, only 100 iterations were completed for each of the 75 different combinations of n, k, and model.

Once both the actual parameter estimates and estimated parameter estimates for each respective model were obtained from the algorithms using R, a meticulous comparison was needed. Correlations between the actual parameters values and estimated parameter values were computed for each of the 25 different scenarios for each of the three models. These correlations compared the mean parameter estimates obtained through the 100 iterations using Monte Carlo Simulation to the actual parameter values used in generating the simulated data. Different program for each model was made to obtain the desired correlations for each of the 25 scenarios. For example, correlations between the actual difficulty parameter value and the estimated difficulty parameter for the two parameter model were obtained for each of the n = 100, 250, 500, 1000 and 5000, and the k = 5, 10, 15, 20, and 30 combinations. This was repeated for each parameter and each model. Also, to check how well the ltm package in estimating the model parameters, we found it useful to check for any unusual estimates obtained within each model scenario. Any estimate that was at least 3 units from the actual value was what we defined as an unusual estimate. Finally, we created a program on R to estimate parameters for the Rasch model using two parameter data created as described in the above process.

RESULTS/DISCUSSION

In the evaluation of the accuracy of the estimation process for the difficulty parameter (b_j), we concluded that overall the two parameter IRT model obtained good estimates when two parameter data were used. Most correlations were greater than .99 in value. However, there were outlying correlations when n = 100 examinees were used, as correlations were found to be sub par. In all likelihood, this can be attributed to the small sample size. In Figures 1, one can see a general rising trend in correlations between the actual difficulty parameter and the difficulty parameter estimates. Correlations noticeably increase as the number of examinees (n) increase. On the other hand, in most cases, the number of items (k) within a given exam had limited impact upon the values of our correlations. In conclusion, when the Two-parameter IRT model was used to estimate the difficulty parameters for a two parameter data set, the estimation of the difficulty parameter was far more contingent upon the number of examinees than that of the number of questions upon a proctored exam. Figure 1 gives a succinct visual summary of these results. In this figure, the horizontal axis represents the number of items in the exam, the vertical axis represents the correlations between actual and estimated item difficulties, and the separate lines are the different results for different sample sizes.

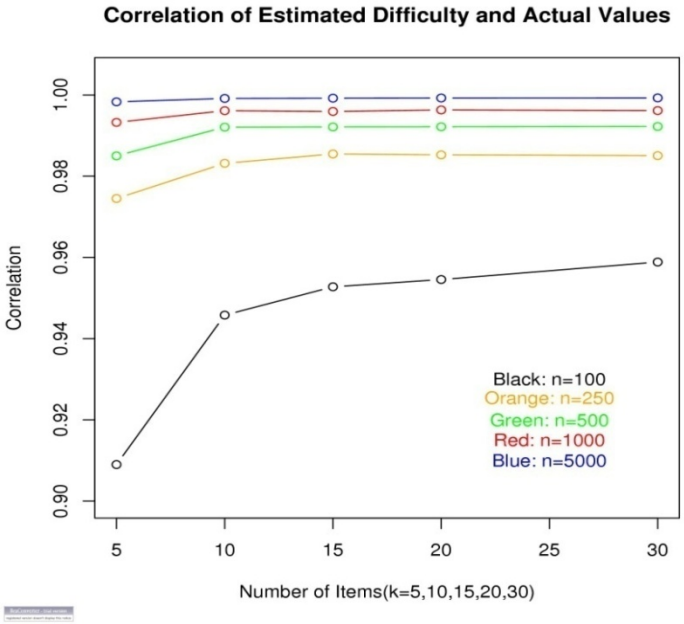

Figure 1. Correlations of Estimated Difficulty and Actual Values

Next, in our evaluation of the accuracy of the estimation process for the discrimination parameter (a_j), we concluded that overall the Two-parameter IRT model obtained comparably lower correlations than those achieved for the difficulty parameter (b_j) given a two parameter data set. Once again, we observed that as the number of examines rose the correlations of the actual discrimination parameter and the estimated discrimination parameter improved. In smaller number of examinees, the number of items (*k*) more greatly influenced correlations for the better. However as the number of examinees (*n*) increased, the impact of the number of items upon correlations was dampened. Figure 2.0 below graphically displays the correlation results for the estimated discrimination parameter and actual values given identical labels for both the horizontal and vertical axis as in Figure 1.

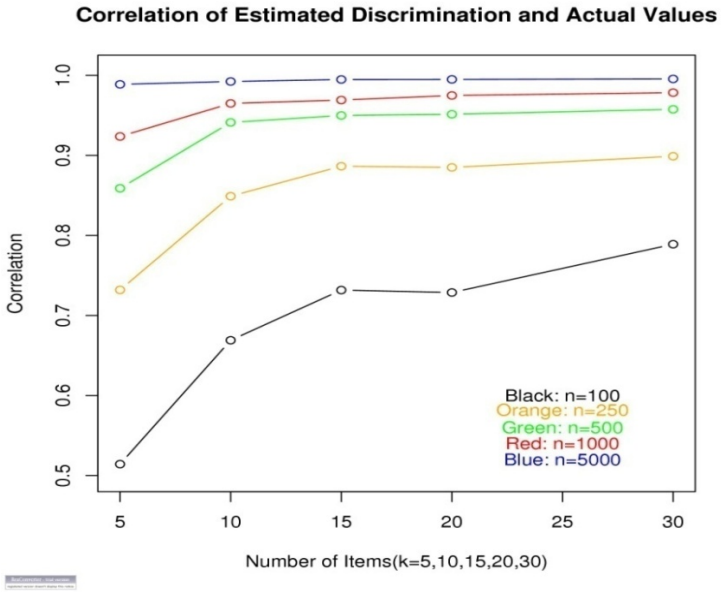

Figure 2. Correlations of Estimated Discrimination and Actual Values

Finally, in our evaluation of the accuracy of the estimation process for the ability parameter (θ_i), we concluded that overall when the Two-parameter IRT model was used, correlations between estimated and actual ability parameters were noticeably worse than both the difficulty and discrimination parameter correlations. Correlations for n =500, 1000, and 5000 are nearly duplicitous in nature. In addition, most correlations have a value less than .90. It is interesting to note that for all levels of the number of examinees, the number of items (*k*) impacts correlations significantly. Therefore, we observe a deviation from the above trends for both the difficulty and discrimination parameter correlations. However, the adage of the law of large numbers still holds seeing that correlations of estimated ability and actual values continue to increase as the number of examinees increases. Figure 3 below graphically displays the correlation results for the estimated ability and actual values.

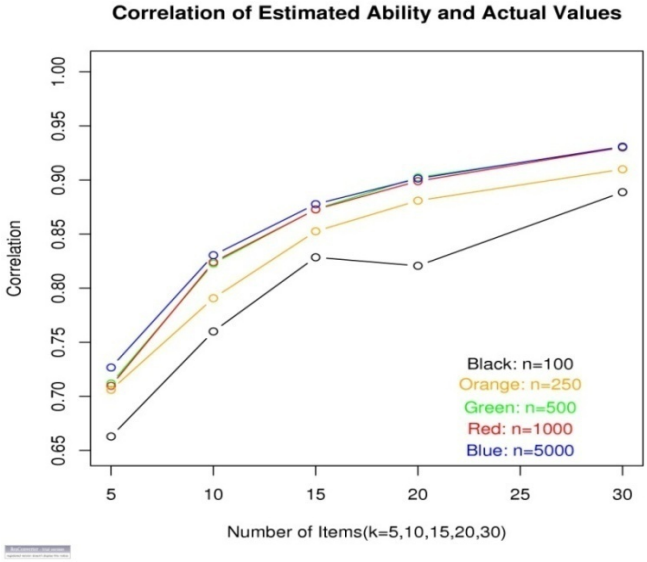

Figure 3. Correlation of Estimated Ability and Actual Values

In order to save space, graphs and figures for the results of the Three-Parameter and One-Parameter estimation process were omitted. For the Three-parameter estimation process, correlations for both the estimated difficulty and actual values and the estimated discrimination and actual values were very low. Correlations for the estimated difficulty and actual values and the estimated discrimination and actual values were heavily dependent upon the number of examinees. Once again, a general increasing pattern was observed, as the number of examinees (n) increased correlations improved dramatically. Simply stated, initial low correlations within the Three-Parameter model are attributable to small sample sizes. As for the correlations between the estimated guessing parameters and actual values, the results were extremely low, with many negative correlations. These are very poor and undesirable results. This result indicates the inadequacy of the Three-parameter model to accurately estimate the guessing parameter (c_j). However, it is interesting to note that even with the dire results with regard to difficulty, discrimination, and guessing parameter correlations, the correlations of estimated ability and actual values have a similar shape as previously recorded for the Two-parameter model. However, correlations as an aggregate are not as lofty in comparison to the Two-parameter correlation of actual values and estimated ability. Within the One-parameter estimation process, correlations of estimated difficulty and actual difficulty values were very high. A large majority of correlations were greater than .99 in value. Once again, the case of n=100 examinees proved to be an outlier, as correlations were significantly lower than all other comparable number of examinee levels. In addition, correlations continued to improve as the number of examinee increased; however not by as significant of a degree as other models. This is mainly attributable to the fact that correlations for the difficulty parameter had little room for improvement seeing that many of them were already quite superior even with low number of examinees. Finally, based on the trend of correlations of estimated ability and actual values for the One-Parameter Rasch model, one receives the impression that ability appears to be independent of sample size. Therefore, as the number of examinees increases, the correlation remains nearly identical as those obtained for previously lower sample sizes. This is quite thought provoking seeing that in all other models, ability improved with increases in both examinees and items.

CONCLUSION

In conclusion, the Two-parameter IRT model estimations produced good estimates. High correlations for difficulty, discrimination, and ability parameters corroborate this claim. With careful consideration, we suggest that sample size of n = 500 and k =15 should be used to obtain best parameter estimates. The Three-parameter IRT model is problematic and nearly inappropriate in its estimation process. We advise not to use this model if one's focus is dominantly upon estimating item characteristics. However, ability estimates are decent. The One-Parameter Rasch model performed quite well with respect to producing high correlations among estimated ability parameters and actual values, and estimated difficulty parameters and actual values. For this particular model we advise using at least k = 15 items. Exams with 15 items yielded reliably good estimates.

ACKNOWLEDGEMENTS

We want to extend a candid and heartfelt thank you to the University of Wisconsin-La Crosse for providing us with the financial support and making it possible to endeavor our research. Last but certainly not least, we want to express our deep felt gratitude and thankfulness to our faculty advisor, Dr. Sherwin Toribio, whom labored diligently and patiently with us in this process.

REFERENCES

Baker F. B. and Kim S-H. (2004), *Item Response Theory: Parameter Estimation Technique, 2nd Ed.* New York: Marcel Dekker.

Hambleton, Ronald K., H. Swaminathan, and H. Jane Rogers. *Fundamentals of Item Response Theory* London: Sage Publications, Inc., 1991.

Hambleton, Ronald K., and Hariharan Swaminathan. *Item Response Theory: Principles and Application*s. Boston: Kluwer Nijhoff Publishing, 1985.

McIntire, Sandra A., and Leslie A. Miller. *Foundations of Psychological Testing: A Practical Approach*. 2nd ed. London: Sage Publications, Inc., 2007.

R Development Core Team (2006). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rizopoulos, Dimitris (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, Volume 17, Issue 5.

APPENDIX

The following lines are an abbreviated version of the code used to create the data matrix used in research:

```
gen_data_irt=function(n,k,model)
{
# Sample command: gen_data_irt(1000,50,3)
theta=rnorm(n,0,1)                                  # generates n ability scores
b=rnorm(k,0,1)                                      # generates k discrimination parameters
a=runif(k,min=.2, max=2)                            # generates k difficulty parameters
c=sample(c(.2,.25,.3333,.5),replace=TRUE,size=k)    # generates k random
c_mat=matrix(rep(c,n),nrow=n,byrow=T)               # guessing parameters
th_mat=matrix(rep(theta,k),ncol=k)
a_theta=t(t(theta))%*%a
b_mat=matrix(rep(b,n),nrow=n,byrow=T)
a_th_b=a_theta-b_mat                                # matrix of (a*theta-b)
pr=exp(a_th_b)/(1+exp(a_th_b))                      # matrix of logit(a*theta-b)
#Two Parameter Model
if (model==2){
data=matrix(rbinom(length(pr),prob=pr,size=1),nrow=nrow(pr))
}
list(data=data,a=a,b=b,theta=theta,c=c)
}
```

This next code was used to extract item and ability parameters for one set of Two-Parameter data:

```
ltm_est_final=function(n,k)
{
```

```
temp=gen_data_irt(n,k,2)                                    #Generates Data
ability=ltm(formula = temp$data~z1,IRT.param=TRUE)          #Performs the Estimation
c=coefficients(ability)                  #Gives the Estimated coefficients(difficulty and discrimination)
b=c[,1]                                  #Estimated Difficulty Parameter (Not the correct one)
a=c[,2]                                  #Estimated Discrimination Parameter
b=a*b                                    #Estimated Difficulty(corrected)
data=temp$data
f=factor.scores(ability)
s=f$score.dat[,1:k]
s=as.matrix(s)
colnames(s)=NULL
z=f$score.dat[k+3]                       #extracted Latent Ability Scores( In order of permutations)
z=t(t(z))                                #creates a matrix and gives z dimensions
# This  nested for loop creates a vector that is in the
# same order as our original theta vector.
z_est=1:n
for (l in 1:n)
{
     for (j in 1:(length(z[,1])))
     {
     if (identical(s[j,],data[l,])==TRUE){z_est[l]=z[j]; break}
     }
}
list(data=data,a_0=temp$a,b_0=temp$b,theta=temp$theta,c=c,a=a,b=b,f=f, z=z, z_est=z_est)
}
This final code performed the Monte Carlo Simulations:
ltm_MC=function(n,k,t)
{
mat=matrix(rep(1:9,t),c(t,9))
i=1
     while(i<=t)
     {
     est=ltm_est_final(n,k)
     b=sum((abs(est$b-est$b_0)>3)*1)
     cor_b=cor(est$b_0,est$b)
     b_max=max(abs(est$b))
     a=sum((abs(est$a-est$a_0)>3)*1)
     cor_a=cor(est$a_0,est$a)
     a_max=max(abs(est$a))
     theta=sum((abs(est$z_est-est$theta)>3)*1)
     cor_th=cor(est$theta,est$z_est)
     th_max=max(abs(est$z_est))
     i=i+1
     }
list(mat=mat)
}
```