iojes
www.iojes.net
International Online Journal of Educational Sciences
ISSN: 1309-2707

Empirical Investigation of the Stability of IRT Item-Parameters Estimation

Mutasem Akour[1] and Hassan AL-Omari[2]

[1]The Hashemite University, College of Education, Jordan; [2]Jadara University, Faculty of Education, Jordan.

ARTICLE INFO

Article History:
Received 31.08.2012
Received in revised form 30.04.2013
Accepted 01.05.2013
Available online 05.08.2013

ABSTRACT

This study examined the effect of various sample sizes (200, 500, 1000, 5000, 10000, and 20000) and test lengths (15, 30, and 60) on the accuracy of item response theory item- parameters estimation using real test data. Estimates of item parameters were obtained by fitting the three-parameter logistic model. The main findings of this study confirmed those findings in previous studies which used simulated data in that longer tests resulted in more accurate estimates of all item parameters across different sample sizes and across different ability levels, especially at ability levels lower than zero. Item difficulty parameter appeared to be the most sensitive to fluctuations in sample size and test length; whereas, item guessing parameter appeared to be the least sensitive. On the other hand, different samples yielded comparable results in terms of accuracy in estimating the three item parameters. Finally, the minimum requirements for accurate parameters estimation tended to be 500 for sample size and 30 for test length. However, sample sizes as small as 200 can still yield acceptable estimates when combined with test lengths longer than 15.

Keywords:
IRT, item difficulty, item discrimination, item guessing, parameters estimation.

Introduction

Classical test theory (CTT) and item response theory (IRT) are perceived as representing two popular statistical frameworks for addressing measurement problems, such as test development and test-score equating. The major advantage of CTT is that it is based on relatively weak assumptions, which means that these assumptions are easy to meet in real test data (Hambleton & Jones, 1993; Fan, 1998). CTT involves an additive model; observed raw scores are the sum of two components: true scores and error scores, where true scores and error scores are unobservable theoretical constructs (Allen & Yen, 1979). However, person parameters (i.e., true scores) are dependent on the test sample, and item parameters (i.e., item difficulty and item discrimination) are dependent on the examinee sample. These dependencies can limit the utility of these parameters in practical test development (Hambleton & Jones, 1993).

On the other hand, IRT makes stronger assumptions than CTT. In IRT it is assumed that the responses to items on a test can be accounted for by latent traits that are fewer in number than the test items. In fact, most applications of this theory assume that a single latent trait (i.e., examinee ability) accounts for the responses to items on a test (Croker & Algina, 1986). IRT models relate item scores to examinees ability levels and item parameters using nonlinear functions (Yen & Fitzpatrick, 2006). Proponents of Item Response Theory claim that IRT has several advantages over classical test theory (Lord, 1980; Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991). One of the advantages is that examinee ability estimates are independent of the particular sample of items selected for the test (item-free ability estimates). Thus, person

[1] Corresponding author's address: The Hashemite University, P.O. Box 150459, Zarqa 13115, Jordan.
Telephone: (962) 5 390 3333 Ext. 4888
e-mail: mutasem@hu.edu.jo; mutasem_akour@yahoo.com

parameter estimates should remain stable across other samples of test items, and therefore, examinees can be compared even though they might not have taken identical sets of test questions. In addition, item parameters are group independent (person-free item parameter estimate). Thus, item parameters obtained from one group of examinees should remain stable across other groups of examinees.

Although IRT offers many benefits to test developers, it has some limitations. The length of a test and the number of examinees needed for proper estimation of the item parameters are difficult to be determined (Hambleton, 1989). Bigger samples and longer tests are needed to provide accurate estimates, especially when both item and ability parameters are being estimated and when more complex IRT models are being fitted to the data (Hambleton, 1989; Hulin, Lissak, & Drasgow, 1982). For this reason, most applications of IRT are found in large-scale testing situations (Sireci, 1992).

However, many tests are composed of relatively small number of items that are administered to relatively small samples of examinees (e.g., teacher-made tests). Therefore, research was done on IRT with small samples and short tests to help in setting up guidelines for test lengths and sample sizes needed to obtain stable and accurate parameter estimates.

For the three-parameter logistic model (3-PL), Lord (1968) suggested that samples of sizes greater than 1,000 examinees and tests lengths longer than 50 items are needed for adequate estimation of the item discrimination parameter. Moreover, Ree and Jensen (1980) simulated tests of 80 items and sample sizes of 250, 500, 1000, and 2000 using the 3-PL model. This study concluded that sample sizes larger than 500 is needed to yield more accurate estimates for the difficulty and discrimination parameters. However, the guessing parameter showed to be insensitive to increment in sample size. This study stated that stable and accurate estimates of the discrimination and difficulty parameters require large sample sizes over a broad range of ability. The estimation of the guessing parameter requires large number of subjects at very low ability levels.

Using the two-parameter logistic (2-PL) and 3-PL models, Hulin et al. (1982) studied the effect of sample size and test length on the accuracy of item and ability estimates. Sample sizes of 200, 500, 1000, and 2000 were used and test lengths were set at 15, 30, and 60 items. The results indicated that minimum sample sizes and test lengths depend upon the response model and the purposes of an investigation. For the 2-PL model, Hulin et al. (1982) recommended minimum test lengths of 30 and sample sizes of 500. When estimation of item and ability parameters is done simultaneously in the 3-PL model, test lengths of 60 and sample sizes of 1000 are required for highly accurate estimates. However, tradeoffs between sample size and test length were apparent in the results they obtained. For example, results based on 60-item tests and 1000 examinees resembled those based on 30-item tests and 2000 examinees.

Swaminathan and Gifford (1983) investigated the effect of sample size, test length, and the ability distribution on the accuracy of item and ability estimation. They simulated their data using the 3-PL model and sample sizes of 50, 200, and 1000 and test lengths of 10, 15, and 20. The results of that study indicated that sample sizes of 1000 and test lengths of 20 items produced very good estimates of the difficulty and guessing parameters, and fairly good estimates of the item discrimination parameter.

Previous studies indicated that sample sizes larger than 1000 and test lengths longer than 50 are needed for highly accurate estimates of the three item parameters when the 3-PL model is fitted to the data. However, none of these studies was done using real test data; simulated data were used instead. One problem with this approach is that it is not known whether the characteristics of real test data are reflected in the simulated data. When simulating data, ability values are usually sampled from a normal distribution with mean zero and unit standard deviation. However, when using real data it is not guaranteed that ability distributions will be normal. If ability distributions were skewed, this may produce less accurate item parameter estimates than those obtained from distributions that are normal (Seong, 1990; Stone, 1992). Thus, real test data may produce different results as compared to simulated data in terms of the accuracy of estimating the item parameters. In addition, using only simulated data may limit the generalizability of the results to real data. Therefore, this study is trying to reinvestigate the stability of the item parameters estimation using real test data rather than using simulated data.

Moreover, the present study is using the relative efficiency index, which is related to the reliability of a test, in judging the accuracy of the item parameters estimation. This index can be used to indicate which one of two tests of the same latent trait is a better measure of the latent trait at each point on the scale (Crocker & Algina, 2003). And since this latent trait is estimated given the item parameters and the model (Yen & Fitzpatrick, 2006), then this index can be used to judge the accuracy of the item parameters estimation.

In particular, the purpose of the present study is to examine the effect of various sample sizes (200, 500, 1000, 5000, 10000, and 20000) and test lengths (15, 30, and 60) on the accuracy of item parameters estimation. Estimates of the item parameters are obtained by fitting the three-parameter logistic model to real test data. It is hoped that this study would provide a good comparison between the results that would be obtained using real testing data and those obtained using simulated data.

Method

The primary goal of this study was to investigate the stability of IRT item parameters estimation for various sample sizes and test lengths. To accomplish this goal, it is necessary to compare item parameter estimates with "true" item parameter values. Since the true values can not be known a priori, such an investigation is carried out, typically, using simulated data. Fortunately, the estimation procedures can be investigated with real test data that is administered to a large group of examinees. Calibrating the items with the entire population of examinees yields true item parameters (Sawminathan, et. al., 2003). Estimated item parameters can be obtained by calibrating the items with different conditions of test length and sample size.

IRT Models

A thorough description of IRT models (1-PL, 2-PL, and 3-PL) is available in the literature (Lord & Novick, 1968; Lord, 1980; Hambleton, 1989; Hambleton et al., 1991), and so they are not fully described here. The equations for the 1-PL, 2-PL, and 3-PL models, respectively, are presented below:

$$P_i(\theta) = \frac{1}{1 + \exp\left\{-D\bar{a}(\theta - b_i)\right\}} \tag{1}$$

$$P_i(\theta) = \frac{1}{1 + \exp\left\{-Da_i(\theta - b_i)\right\}} \tag{2}$$

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left\{-Da_i(\theta - b_i)\right\}} \tag{3}$$

Where $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ answers item i correctly; a_i is the item discrimination parameter, and D is a scaling factor set equal to 1.7; b_i is the item difficulty parameter, and c_i is the pseudo-chance level, or simply the guessing parameter.

The IRT model used in this study is the three-parameter logistic model, because it can accommodate guessing. Guessing must be considered a possibility on multiple choice items, and hence, the 3-PL model has an advantage over both of the other models in this case (Crocker & Algina, 1986).

Data

The data for this study came from a national 8th grade mathematics exam that is composed of 60 multiple-choice items. This test was administered by the Ministry of Education in Jordan to, approximately, 40,000 examinees.

The design of this study contains two factors: sample size and test length. In accordance with Hulin et al. (1982), three levels of test length were used: 15, 30, and 60. As well, sample sizes of 200, 500, 1000, were used in addition to sample sizes of 5000, 10000, and 20000. A crossing of the three levels of test length with the six levels of sample size resulted in 18 (3x6) distinct conditions, as shown in Table 1.

Table 1. Summary of test conditions

	Sample size					
Test Length	200	500	1000	5000	10000	20000
15	*	*	*	*	*	*
30	*	*	*	*	*	*
60	*	*	*	*	*	*

* Represents each condition of sample size and test length.

The data for each condition was obtained as follows:

- To obtain shorter test lengths (i.e., 15 and 30 items), systematic sampling was used. To form the 15-item test, the first item was selected randomly from among the first four items and every fourth item was selected thereafter. Thus, items 1, 5, 9, .. 57 constituted the 15-item test. For the 30-item test, odd-numbered items were selected.
  This procedure was used because the 60 items in the original test was not ordered according to difficulty or according to any other criteria.
- To obtain smaller sample sizes (less than 40,000 examinees), stratified random sampling was used. That is, for the population of examinees, the scale of ability (θ) was divided into six distinct intervals, from -3 to +3 with an increment of 1. Then, equal numbers of examinees were randomly selected from each interval and aggregated to constitute the required sample size.

Procedure and Analysis

This section describes in detail the steps that were used to investigate the accuracy of the item parameters estimation.

Step1: The purpose of this step was to determine the population item parameters. Item parameters estimated from the 60-item test that was administered to the entire group of 40,000 examinees represented the "true" item parameters to which the sample item parameters would be compared. BILOG (Mislevy & Bock, 1984) was used for the analyses and for computing the "true" values of the item parameters a, b, and c.

Step 2: In order to estimate the sample item parameters, the three-parameter logistic model was fitted to the data of each condition of test length and sample size. Similar to step 1, the computer program BILOG was used for the estimation of item parameters.

Step 3: The accuracy with which item parameters are estimated can be detected by computing the discrepancy between the estimate and the true value. Han, Kolen, and Pohlmann (1997) referred to the differences between estimated and true parameters as an estimation loss. The current study borrowed the Root mean square loss (RMSL) index and used it to judge the accuracy of estimation of each item parameter. One measure of RMSL for each a, b, and c estimate under each condition was obtained by taking the square root of the average of the squared differences between the true and the estimated parameters across all items using the following equation

$$\boldsymbol{RMSL} = \sqrt{\frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{O_i} - \boldsymbol{T_i})^2}{\boldsymbol{n}}} \tag{4}$$

Where O$_i$ is the estimated item parameter (i.e., the estimated a, b, or c), T$_i$ is the corresponding "true" item parameter, and n is the number of items on the test. An inverse relation exists between the stability of the estimated item parameters and the magnitude of this index; the closer to zero the value of this index, the greater the stability of the estimated item parameters. In addition, values of RMSL less than 0.6 were considered as small according to Han, Kolen, and Pohlmann (1997).

Step 4: The purpose of this step was to compute the test information function at different ability levels using the 'true" and the estimated item parameters. Hambleton and Swaminathan (1985) considered the test information function as the IRT analog of test score reliability. The information function for a test, $I(\theta)$, is given by:

$$I(\theta)=\sum_{i=1}^{n}\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \quad \text{or} \quad I(\theta)=\sum_{i=1}^{n}I_i(\theta) \quad (i=1,2,\ldots.,n) \tag{5}$$

Where $I_i(\theta)$ is the item information function, $P_i(\theta)$ and $Q_i(\theta)$ (Q_i = 1- P_i) denote the probability that an examinee with ability θ answers the item correctly and incorrectly, respectively; $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, and n is the number of items.

To compare the test information function that is computed using the estimated item parameters [$I_O(\theta)$] with the test information function that is computed using the "true" item parameters [$I_T(\theta)$], the index of relative efficiency ($RE(\theta)$) is computed, where

$$RE(\theta)=\frac{I_O(\theta)}{I_T(\theta)} \tag{6}$$

The greater this index, the greater the precision of measurement produced by the test with the estimated item parameters as compared to the test with the "true" item parameters (Hambleton, Sawminathan, & Rogers, 1991); that is, the greater the stability of the estimated item parameters. The index of relative efficiency was computed at different ability levels for each condition of test length and sample size.

Results

To answer the research questions, RMSL values were computed for each item parameter (a, b, and c) under different conditions of test length and sample size as shown in table 2. These RMSL values were plotted separately for each item parameter as shown in figures 1, 2, and 3.

For each item parameter (a, b, and c), table 2 and figures 1, 2, and 3 show that increasing sample size resulted in decreasing RMSL values across different test lengths. For the item discrimination parameter and the item difficulty parameter and for test length of 15, the biggest reduction in RMSL values resulted as sample size increased from 200 to 500. However, increasing sample size beyond 1000 had less effect on reducing RMSL values of estimation for both item parameters. RMSL values for the item guessing parameter tended to be more stable across different sample sizes.

Table 2. RMSL of each item parameter estimate under different conditions of test length and sample size.

Item Parameter	Test Length	Sample size					
		200	500	1000	5000	10000	20000
a	15	.28908	.09024	.07489	.04494	.04412	.04259
	30	.13430	.08867	.04290	.00888	.00757	.00389
	60	.11556	.10790	.06375	.01046	.00185	.00069
b	15	1.31426	.27006	.05422	.03233	.02638	.01678
	30	.31074	.13999	.07127	.06967	.06384	.04576
	60	.26776	.10480	.04761	.02779	.00766	.00314
c	15	.02097	.00991	.00512	.00301	.00294	.00288
	30	.03060	.01825	.01289	.00280	.00101	.00044
	60	.02829	.01593	.01060	.00215	.00064	.00011

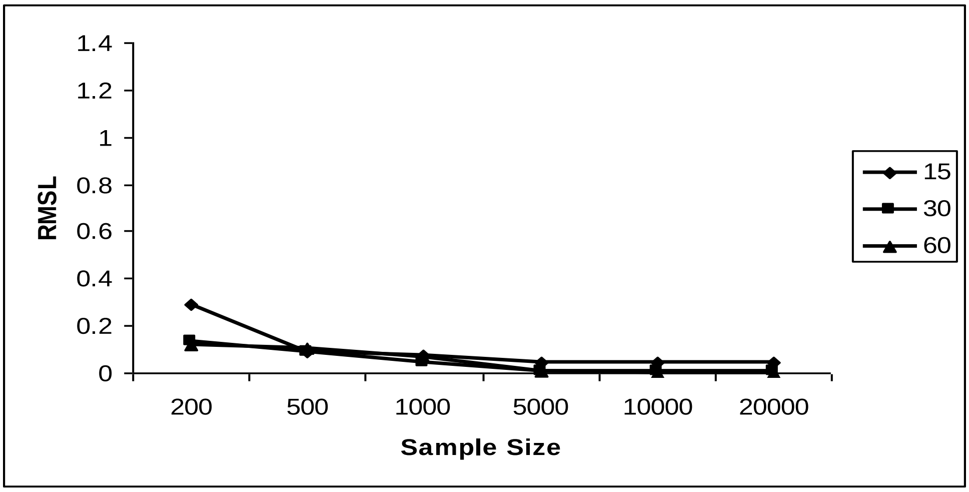

Figure 1. RMSL of item discrimination parameter under different conditions of test length and sample size.

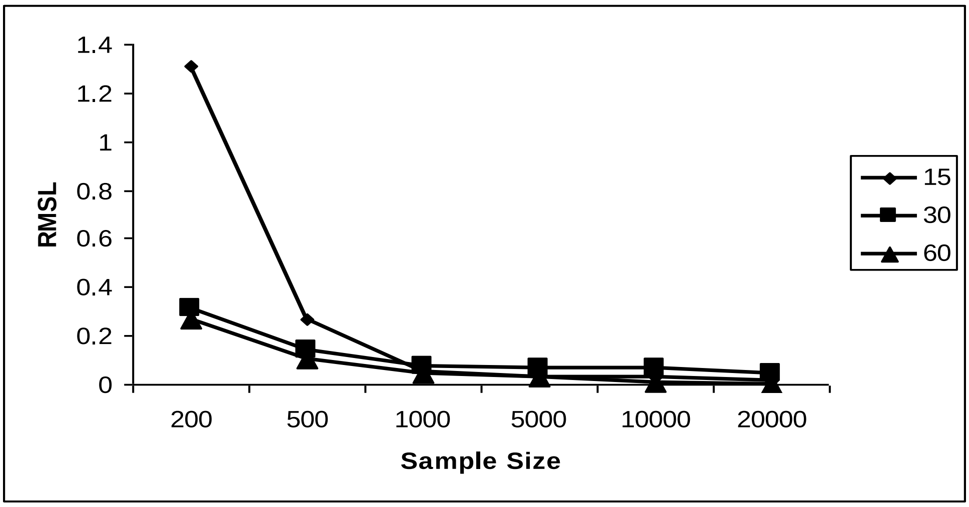

Figure 2 RMSL of item difficulty parameter under different conditions of test length and sample size.

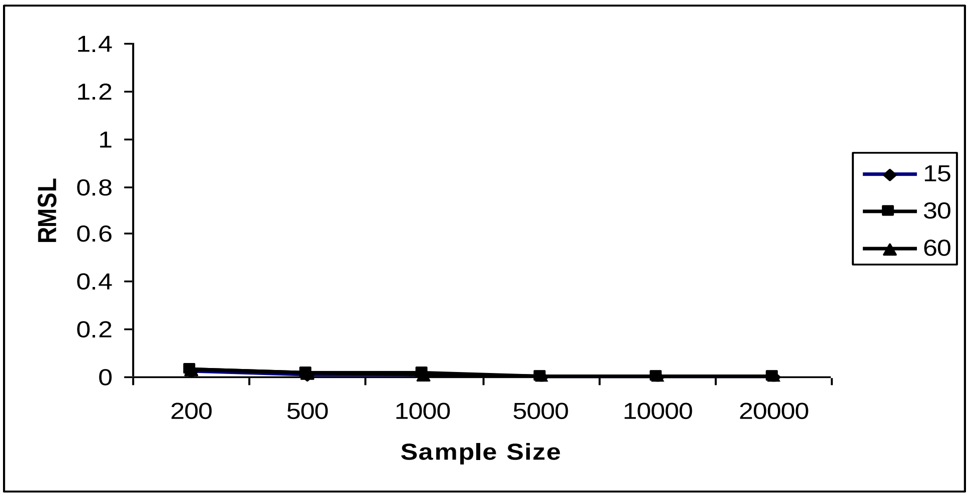

Figure 3. RMSL of item guessing parameter under different conditions of test length and sample size.

For the item discrimination parameter (a), table 2 and figure 1 show that for all conditions of sample size and test length, RMSL values were less than 0.3 which is considered to be small based on Han, Kolen, and Pohlmann (1997). This is an indicative of accurate estimates of this parameter across different conditions of sample size and test length. Moreover, across different sample sizes RMSL values decreased as test length increased. The exception is for sample sizes of 500, 1000, and 5000 as test length increased from 30 to 60. Generally, different test lengths behaved almost similarly for sample sizes larger than 1000.

For the item difficulty parameter (b), table 2 and figure 2 show that for all conditions of sample size and test length RMSL values were less than 0.3, with the exception of sample size of 200 and test length of 15. This indicates that accurate estimates of the difficulty parameter were obtained under different conditions of sample size and test length, except for sample size of 200 and test length of 15. For sample sizes of 200 and 500, RMSL values of estimation decreased as test length increased. However, for sample sizes of 1000 and larger RMSL values increased as test length increased from 15 to 30 and then decreased as test length increased from 30 to 60. In general, different test lengths resulted in, approximately, comparable results for sample sizes of 1000 or larger.

For the guessing parameter (c), RMSL values for all conditions of sample size and test length were less than 0.03. This indicates that the estimation of this parameter under different conditions of sample size and test length was very accurate as compared to the other item parameters. For sample sizes of 5000, 10000, and 20000 table 2 and figure 3 show that RMSL values decreased as test length increased. Whereas, for sample sizes of 200, 500, and 1000 RMSL values of estimation increased as test length increased from 15 to 30 and then decreased as test length increased from 30 to 60. In general, different test lengths yielded almost similar RMSL values across different sample sizes.

For each condition of sample size and test length, the information function was computed at different ability levels, from $\theta = -3$ to $\theta = +3$ with an increment of 1. Information functions were then used to compute the indices of relative efficiency as another measure of accuracy of parameter estimation. These indices were plotted in figures 4 through 7 under different conditions of sample size and test length and at different ability levels.

Figure 4 shows that longer tests tended to be more efficient at different ability levels across different sample sizes. However, figure 4 can be divided into three parts: sample sizes 200 and 500, sample sizes 1000 and 5000, and sample sizes 10,000 and 20,000. For each one of these three parts, comparable indices of relative efficiency, and thus comparable estimates of item parameters, were obtained for each test length across different ability levels. For example, the first part of this figure, which represents sample sizes 200 and 500, yielded comparable indices of relative efficiency for each test length across different ability levels. Similar findings can be obtained from the other two parts of figure 4. That is, each test length yielded comparable results across sample sizes of 1,000 and 5,000, and yielded comparable results across sample sizes of 10,000 and 20,000.

For the upper part of the ability scale (ability values of 1, 2, and 3), figures 4 through 7 show that different conditions of sample size and test length resulted in comparable indices of relative efficiency. However, for lower ability levels (ability values of 0, -1, -2, and -3) bigger samples tended to be more efficient across different test lengths.

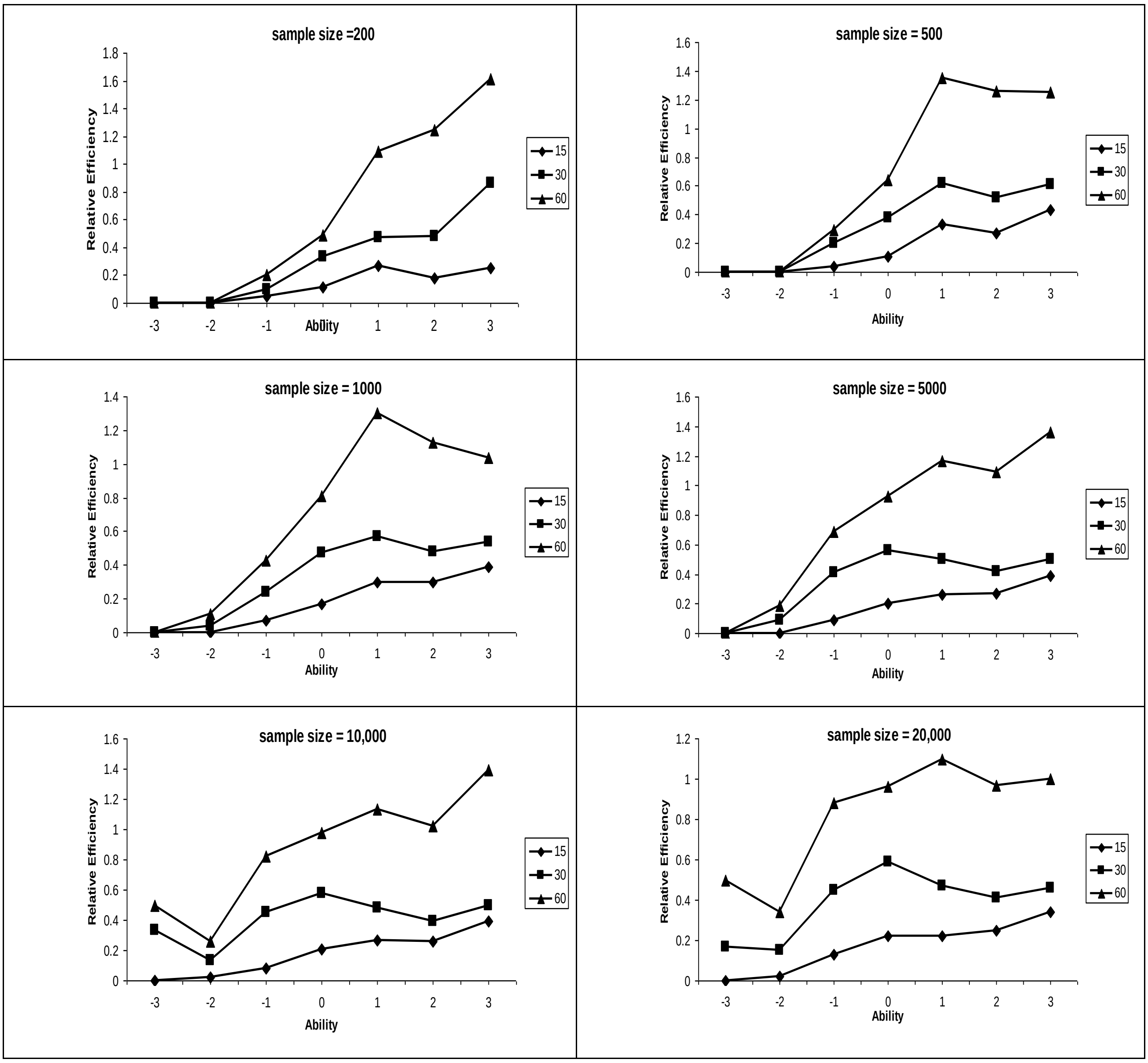

Figure 4. Relative Efficiency for each test length across different sample sizes.

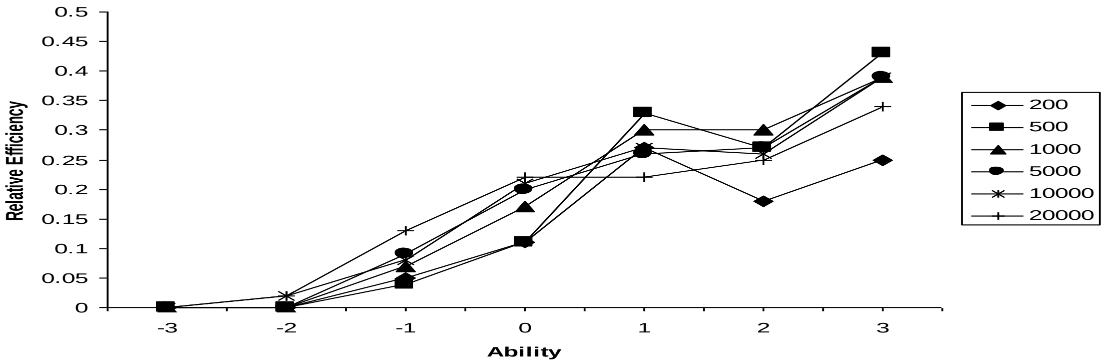

Figure 5. Relative Efficiency at different ability levels across different sample sizes for test length of 15

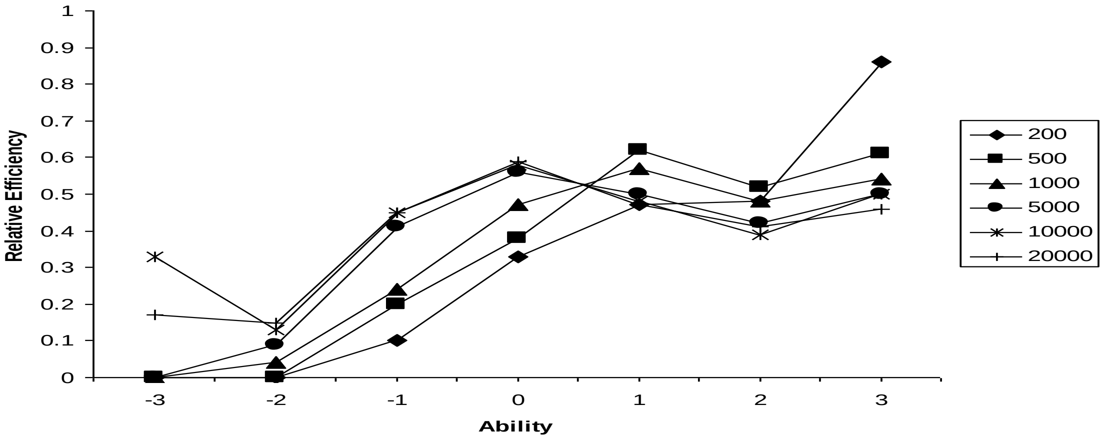

Figure 6. Relative Efficiency at different ability levels across different sample sizes for test length of 30

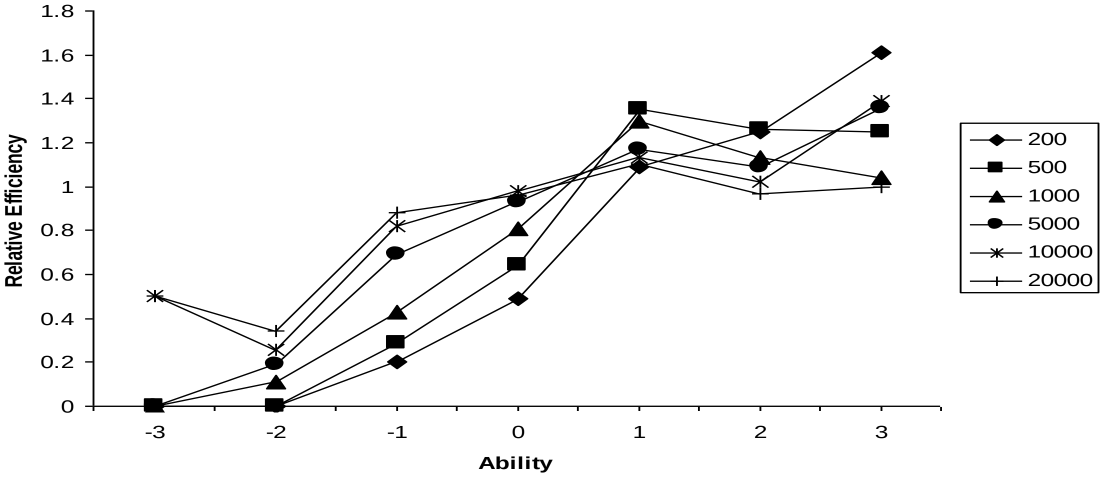

Figure 7. Relative Efficiency at different ability levels across different sample sizes for test length of 60.

Dıscussıon

This study investigated the accuracy of estimating IRT item parameters using two criterions: RMSL values and relative efficiency indexes. The main findings of this study showed that increasing sample size resulted in more accurate item parameters' estimates. The effect of sample size on increasing the accuracy of item parameter estimates were more apparent for both the item discrimination and the item difficulty parameters as compared to the item guessing parameter.

Different conditions of sample size and test length resulted in acceptable accuracy of the estimated item parameters. The exception was the condition of sample size 200 and test length 15 when estimating the item difficulty parameter. To be more precise, sample size can be categorized into three categories: sample sizes less than 1000, sample sizes between 1000 and 10,000, and sample sizes bigger than 10,000. Within each category, different sample sizes would be similar in terms of obtaining accurate estimates of the item parameters. Longer tests within each category is favorable to obtain more accurate estimates especially at the lower part of the ability scale.

The Item difficulty parameter appeared to be the most sensitive to fluctuations in sample sizes. Figure 2 showed big discrepancies in the accuracy of item difficulty parameter estimates when comparing RMSL values of sample sizes less than 1000 with test length of 15 and those of other conditions of sample size and test length. For more accurate estimates of the item difficulty parameter, longer tests are needed with small samples. This agrees with the findings of Lord (1968) and Hulin et al., (1982). However, different sample sizes yielded comparable results. This is inconsistent with the findings of many studies (Lord, 1968; Ree and Jensen, 1980; Hulin et al., 1982; Swaminathan and Gifford, 1983).

The item discrimination parameter behaved similar to the item difficulty parameter in that different sample sizes resulted in comparable accurate estimates of this parameter. This is also inconsistent with the findings of Lord (1968), Ree and Jensen (1980), and Hulien et al. (1982) in that sample sizes bigger than 1000 are needed for accurate estimates of the item discrimination parameter. However, sample sizes as small as 200 resulted in more accurate estimates of the item discrimination parameter as compared to the item difficulty parameter which disagrees with the findings of Swaminathan and Gifford (1983).

The guessing parameter appeared to be the least sensitive to sample size. Different sample sizes resulted in comparable estimates of the guessing parameter. This is consistent with the findings of Ree and Jensen (1980), whereas, it is inconsistent with the findings of other studies (Lord, 1968; Hulin et al., 1982; Swaminathan and Gifford, 1983). . RMSL values of the guessing parameter stayed below 0.03 whereas they reached the level of 0.3 for the discrimination parameter and 1.3 for the difficulty parameter.

Longer tests appeared to result in more accurate estimates of all item parameters and across different ability levels which agrees with the findings of Lord (1968). In fact the three test lengths can be grouped into two groups. The first group contains tests of lengths 30 and 60, since they behaved almost similarly across different sample sizes. On the other hand, the second group contains the test length 15 which behaved differently as compared to the other test lengths.

For different sample sizes, it is recommended to use longer tests to obtain more accurate estimates of all item parameters across different ability values, especially for the lower part of the ability scale. This is inconsistent with the findings of Hulin et al. (1982) in that sample size compensate for test length in obtaining accurate estimates of the item parameters.

In conclusion, the findings of this research which was obtained using real test data agree with the findings of other studies that used simulated data (Ree and Jensen (1980); Hulin et al. (1982); Swaminathan and Gifford (1983); and Lord (1968)) in that longer tests are needed for proper and accurate estimation of the three item parameters. However, it disagrees with the previous studies in that bigger samples are also needed. In this study, different sample sizes yielded comparable estimates of the item parameters. This indicates that small samples can still be used in administering tests and yet not greatly affecting the precision of the estimated item parameters. Sample size of 500 and test length of 30 can be considered as the minimum requirements for accurate item parameters estimation. However, sample size as small as 200 can still yield acceptable item parameters estimates when combined with tests of length 30 or longer.

In addition, this study showed that there is some sort of thresholds in sample size where the accuracy in the estimation of the item parameters changes evidently. This requires more research to be done on the effect of sample size on the accuracy of item parameters estimation through looking at the sample size continuum of being composed of multiple segments, where each segment represents a coherent entity. Finally, this study disagrees with previous studies (Hulin et al. (1982)) in that tradeoffs between sample size and test length are apparent. When it comes to the accuracy in estimating item parameters across different ability levels, especially at lower ability levels, these tradeoffs are not apparent. Still, longer tests are needed for more accurate item parameters estimates across different sample sizes and across different ability levels.

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NY: CBS college publishing.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and psychological measurement, 58*(3), 357-381.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). Washington, DC: American Council on Education and Macmillan.

Hambleton, R. K., Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K.,& Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*.

Newbury Park, CA: Sage.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equpercentile equating. *Applied Measurement in Education, 10*(2), 105-121.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

Lord, F. M. (1968). An Analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mislevy, R. j., & Bock, R. D. (1990). PC- BILOG- item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference*. Minneapolis: University of Minnesota.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311.

Sireci, S. g. (1991). *"Sample-Independent" Item Parameters? An investigation of the stability of IRT item parameters estimated from small data sets*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16.

Swaminathan, H., & Gifford, J. A. (1983). Esimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing*, (pp. 9-30). New York: Academic Press.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rizavi, S. (2003). *Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates* (Research Report LSAC-CTR-98-06). Newtown, PA: Law School Admission Council.

Yen, W. M., & Fitzpatrick, A. R. (2006). . In R. L. Brennan (Ed.), *Educational measurement* (4rd ed., pp. 111-153. American Council on Education and Praeger publishers.