National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

The University of Ottawa

A Comparison of Item Parameter Estimates and ICCs Produced with TESTGRAF and BILOG Under Different Test Lengths and Sample Sizes

By

Liane N. Patsula

A Thesis

Submitted to the Faculty of Graduate Studies

in partial fulfilment of the requirements for the degree

of Masters of Arts in Education

Faculty of Education

Concentration in Measurement and Evaluation

Ottawa, Ontario

June, 1995

ISBN 0-612-04929-9

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

There are many procedures used to estimate IRT parameters; however, among the most popular techniques are those used in the LOGIST and BILOG computer programs. LOGIST requires large numbers of examinees and items (in the order of 1000 or more examinees and 40 or more items) for stable 3PL model parameter estimates. BILOG is a more recent estimation program and, in general, requires smaller numbers of examinees and items than LOGIST for stable 3PL model parameter estimates. It also has been found that, regardless of sample size and test length, BILOG estimates tend to be uniformly more or at least as accurate as LOGIST estimates. For this reason, BILOG is now used as the standard to which new estimation programs are compared.

However, regardless of the smaller sample size and test length needed in BILOG to accurately estimate 3PL item parameters, it has been proven to be quite complex, computer intensive, and still require what some practitioners consider large sample sizes. In instances when there are small sample sizes and test lengths (such as in the classroom), BILOG yields parameter estimates with large biases and large standard errors of estimates. Therefore, there was a need for the development of small-dataset approaches to item parameter estimates. In response to such a need, Ramsay developed the program TESTGRAF which uses nonparametric IRT techniques.

Ramsay claims that the estimation procedures used in TESTGRAF are 500 times faster than using some of the common parametric approaches found in LOGIST and BILOG, that there is no loss of efficiency, and that only as few as a hundred examinees and twenty test questions are

needed to estimate ICCs. However, one must keep in mind that Ramsay states that the produced item parameter estimates are fairly crude, and thus they should not be seen as substitutes for a more serious analysis of the data using the logistic model by programs such as LOGIST or BILOG. Nevertheless, with this limitation in mind, it seems important to examine how results obtained from TESTGRAF compare with results obtained from BILOG.

The purpose of this study was to examine the effects of varying sample size (N = 100, 250, 500, and 1000) and test length (20- and 40-item tests) on the accuracy and consistency of 3PL model item parameter estimates and ICCs obtained from TESTGRAF and BILOG.

Overall, TESTGRAF seemed to perform better or just as well as BILOG. Where large bias effect sizes existed, in all but one case, TESTGRAF was more accurate than BILOG. TESTGRAF was slightly less accurate than BILOG in estimating the $P(\theta)$'s at high ability levels. Where large efficiency effect sizes existed, in all but two cases, TESTGRAF was more consistent than BILOG. TESTGRAF was slightly less consistent than BILOG in estimating the *a* parameter with a sample size of 1000 and in estimating the *c* parameter at all sample sizes.

To date it appears that no researcher has examined the performance of TESTGRAF or has compared it to any other leading program in the field. Hence, this comparison between TESTGRAF and BILOG contributes to our knowledge of both programs and their usefulness in various practical situations. This may lead to a wider use of IRT methods, through the use of TESTGRAF, among educators who develop short tests and/or who are faced with small sample sizes.

Acknowledgements

I would like to acknowledge with appreciation the contributions of all those who helped make the completion of this thesis possible. I would especially like to thank my supervisor Dr. Marc Gessaroli for all of his time, effort, and practical assistance in guiding me through the thesis process and for all his feedback in preparing this final draft. I would also like to thank the members of my thesis committee Dr. Marvin Boss and Dr. Bruno Zumbo for their support and constructive comments along the way. Finally, I would like to recognize and thank my parents for their continued interest and never ending encouragment over the past three years.

Table of Contents

List of Tables

List of Figures

CHAPTER I

Introduction

As is evident in the measurement literature of the past fifteen years, the use of Item Response Theory (IRT) by test developers and educators to analyse test data has become increasingly prominent. This can be attributed to the many stated advantages of IRT models. Specifically, when the fit between model and test data of interest is satisfactory, IRT models are said to provide invariant item and ability parameters (Lord, 1952). "This [invariance] property implies that the parameters that characterize an item do not depend on the ability distribution of examinees [sample-free item parameters] and the parameter that characterizes an examinee does not depend on the set of test items [test-free ability parameters]" (Hambleton, Swaminathan & Rogers, 1991, p. 18). IRT models allow one to predict an examinee's item performance based solely on the examinee's ability and not on the item the examinee is answering, nor on the group the examinee is in.

The three most popular IRT models in common use are the three-parameter logistic (3PL) model (Birnbaum, 1968), the two-parameter logistic (2PL) model (Lord, 1952), and the one-parameter logistic (1PL) model (Rasch, 1960). These models are appropriate for dichotomous item response data. The 3PL model is the most general model and is defined mathematically as:

$$P_i(\theta_j) = c_i + (1 - c_i)\left[1 + e^{-1.7a_i(\theta_j - b_i)}\right]^{-1}.$$

In this model, $P_i(\theta_j)$ is the item characteristic curve (ICC) which defines the probability that examinee j with ability θ will respond correctly to item i. Parameters b_i, a_i, and c_i are the difficulty, discrimination, and pseudo-guessing parameters, respectively, associated with item i. The 2PL and 1PL models are restricted cases of the 3PL model. All three models provide an estimate of an examinee's ability, but as suggested by their names, differ in the number of item parameters they estimate. Inherent in the number of item parameters each model estimates, are the assumptions made with each model. With the 3PL model it is assumed that even an examinee with no knowledge has a non-zero probability of getting the item correct (*i.e.*, by guessing) and, therefore, all three parameters are estimated. With the 2PL model it is assumed that there is little or no guessing (c=0) and, therefore, only the b and a parameters are estimated, item difficulty and item discrimination. With the 1PL model it is assumed that there is little or no guessing and that all items have equal discrimination (a's are equal) and, therefore, only the b parameter, item difficulty, is estimated. Because of the mathematical complexity of estimation procedures, in most practical applications the parameters from any of the models must be estimated by computer programs.

There are many procedures to estimate IRT parameters (for examples see Hambleton *et al.*, 1991, pp. 48-51), however, among the most popular estimation techniques are those found in the LOGIST (Wingersky, Barton, & Lord, 1982; Wingersky & Lord, 1973) and BILOG (Mislevy & Bock, 1984, 1986) computer programs. Both of these programs can be used to estimate IRT parameters for all three logistic models. LOGIST uses the joint maximum likelihood estimation procedure (Birnbaum, 1968) to estimate item and examinee parameters. According to

Wingersky (1983), large numbers of examinees and items (in the order of 1000 or more examinees and 40 or more items) should be used to obtain stable 3PL model parameter estimates in LOGIST. BILOG is a more recent estimation program which uses marginal maximum likelihood (Bock & Aitkin, 1981) and Bayesian estimation procedures (Mislevy, 1986). In general, BILOG requires smaller numbers of examinees and items than LOGIST to obtain stable 3PL model parameter estimates (Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987). Furthermore, these authors found that regardless of test length and sample size, BILOG estimates are almost uniformly more or at least as accurate as LOGIST estimates.

However, regardless of the smaller necessary test length and sample size needed in BILOG to accurately estimate 3PL item parameters, it has been proven to be quite complex, computer intensive, and still requires what some practitioners consider a large sample size (Baker, 1987; Ramsay, 1991). BILOG is suitable for those with access to large datasets, such as commercial testing organizations, but it poses a problem for routine test analysis on a smaller scale. In instances when there are small sample sizes and test lengths, BILOG yields parameter estimates with large biases and large standard errors of estimates (Thissen & Wainer, 1982). For this reason, Baker believed that there was a need for the development of small-dataset approaches to item parameter estimation and that "implementation of such procedures in a 'user-friendly' manner on a microcomputer is an absolute requirement" (p.138). Ramsay attempted to do this in his development of the microcomputer program TESTGRAF (Ramsay, 1989).

TESTGRAF uses a nonparametric IRT estimation technique to estimate $P(\theta)$'s. Ramsay (1991) claims that the estimation procedure used in TESTGRAF is 500 times faster than using

some of the common parametric approaches such as maximum likelihood estimation found in LOGIST and BILOG, that there is no loss of efficiency, and that only as few as 100 examinees and 20 test questions are needed to estimate ICCs. However, one must keep in mind that Ramsay (1989) himself states that the produced item parameter estimates are fairly crude, and thus they should not be seen as substitutes for a more serious analysis of the data using the logistic model by programs such as LOGIST or BILOG. His objective was not to produce robust item parameter estimates, but to replace such numerical summaries by curves (ICCs). Nevertheless, with these limitations in mind, it seems important to examine how results obtained from TESTGRAF compare with results obtained from BILOG.

To date, there does not seem to be any researcher who either has examined the performance of TESTGRAF or has compared it to any other leading program in the field. Researchers could look at the accuracy and consistency of TESTGRAF item parameter and $P(\theta)$ estimates and compare them to results obtained from other estimation procedures. Specifically, factors could be addressed which may affect the accuracy and consistency of TESTGRAF item parameter and $P(\theta)$ estimation such as: 1) a violation of the underlying unidimensionality assumption, 2) model-data fit, 3) sample size, 4) test length, and 5) the ability distribution in the population. Since the utility of TESTGRAF to the practitioner appears to be in the claim that only as few as 100 examinees and 20 test questions are needed to obtain ICCs, only the effects of sample size and test length are addressed in this study.

The purpose of this study was to compare the effects of varying test length and sample size on the 3PL model item parameter and $P(\theta)$ estimates obtained from TESTGRAF and BILOG.

CHAPTER II

Review of the Literature

Before employing IRT models to estimate parameters from test data, there are several factors which researchers must consider. These factors include: 1) the underlying assumption of unidimensionality; 2) model-data fit; 3) sample size; 4) test length; 5) the ability distribution in the population; and 6) the estimation procedure to be used. All of these factors affect the accuracy and consistency of IRT parameter and ICC estimates. In this study, simulated 3PL data were used which controlled for some of these effects. Namely, the test data were simulated to be unidimensional and were simulated based on the 3PL model, and the ability distribution of the population was simulated to have a standard normal distribution. Therefore, the three effects of concern in this study were the effect of sample size, test length, and the effect of the estimation procedure used to obtain accurate and consistent 3PL item parameter estimates and ICCs. A review of the literature of only the effects of interest in this study is provided. The order in which these factors are discussed in the review of literature is sample size, test length, and estimation procedure.

Sample Size

In this section, studies in which the necessary sample size was investigated to obtain accurate IRT parameter estimates and ICCs are reviewed. First, studies by Swaminathan and Gifford (1983), Wingersky and Lord (1984), and Skaggs and Stevenson (1989) in which the effect of sample size on the accuracy of 3PL model estimates was examined are reviewed. Secondly, a study by Hulin, Lissak, and Drasgow (1982) in which the effect of examinee sample size on the recovery of ICCs was examined is reviewed.

Swaminathan and Gifford (1983), Wingersky and Lord (1984), and Skaggs and Stevenson (1989) investigated the effect of sample size on the accuracy of the joint maximum likelihood 3PL parameter estimates obtained from LOGIST with test length held constant and a normal ability distribution of the examinee population. Swaminathan and Gifford (1983) used sample sizes of 50, 200, and 1000 and test lengths of 10-, 15-, 20-, and 80-items. For fixed test lengths, they found that the number of examinees had a slight effect in improving the accuracy of estimation of the *b* and *c* parameters and θ and a large effect in improving the accuracy of the estimation of the *a* parameter especially with 10- and 15-item tests. The largest correlations between true and estimated item and ability parameters for each test length occurred in the largest samples; smaller samples yielded generally poor results. They conclude that the accuracy of item parameter estimation in the 3PL model varies depending on which parameter is being estimated. Wingersky and Lord (1984) used sample sizes of 1500 and 6000 and a test length of 45-items. They found that quadrupling the number of examinees from 1500 to 6000 halved the

standard errors of the estimated item parameters and reduced the largest standard error of the θ estimates sharply, but had little effect on the smaller standard errors of θ estimates. Skaggs and Stevenson (1989) used sample sizes of 500 and 2000 and test lengths of 15- and 35-items to examine the accuracy of item parameters and the accuracy of ICCs. They found that that for both test lengths the average root mean squared error decreased when going from 500 to 2000 examinees. Whereas Wingersky and Lord examined the standard errors of the item parameter estimates obtained with different sample sizes, Skaggs and Stevenson examined the correlations between true and estimated item parameters. They found that $r_{a\hat{a}}$, $r_{b\hat{b}}$,and $r_{\hat{c}c}$ increased when going from a sample of 500 to 2000 examinees. In general, all three studies support the idea that increasing sample size leads to more accurate parameter estimates.

Rather than looking at the accuracy of parameters estimates only, Hulin, Lissak, and Drasgow (1982) also investigated the recovery of 2PL and 3PL ICCs and the correlations between true and estimated parameters from LOGIST. They used simulated data for sample sizes of 200, 500, 1000, and 2000 and test lengths of 15-, 30-, and 60-items. For each test length, for the 3PL ICCs, they found that the root mean squared errors decreased with increased sample size. As well, $r_{\theta\hat{\theta}}$ increased with increased sample size for the 15- and 30- item tests, but for the 60-item test $r_{\theta\hat{\theta}}$ only slightly changed with increased sample size. They also found that for all test lengths, $r_{a\hat{a}}$, and $r_{b\hat{b}}$, increased for all sample sizes.

Based on the above studies, it is difficult to state exactly the number of examinees required to estimate 3PL item parameters. Hambleton, Swaminathan, and Rogers (1991) suggest that 1000 or more examinees are needed to estimate 3PL item parameters. Hulin, Drasgow, and Parsons

(1983) point out that there should be no definitive rules because the number of examinees required to estimate accurately item parameters depends on the purpose of the study and is therefore study specific. There are many other factors which work in conjunction with sample size to affect accurate parameter estimation. Accurate parameter estimation not only depends on the number of examinees and the model chosen, but it also depends on the test length, the ability distribution of the sample, and the estimation procedure used.

Test Length

In this section, many of the studies (Hulin, Lissak, & Drasgow, 1982; Skaggs & Stevenson, 1989; Swaminathan & Gifford, 1983; Wingersky & Lord, 1984) which were examined for the effect of sample size on accurate 3PL IRT parameter estimation, are re-examined to determine the effect of test length when sample size, ability distribution of examinee population, and estimation procedure are held constant on the effect of the accuracy of 3PL parameter estimation and ICCs.

There is a consensus among these researchers that with sample size held constant and a normal ability distribution of the examinee population, the accuracy of the joint maximum likelihood 3PL parameter estimates obtained from LOGIST increased with increased test length. This is to be expected because longer tests tend to be more reliable and test reliability is positively related to test information. The accuracy was measured by observing standard errors of the parameter estimates, correlations between true and estimated parameter values, and root

mean squared errors of ICCs. In general, for the 3PL model with sample size, ability distribution, and estimation procedure used held constant, as test length increases, the accuracy of the parameter estimates and ICCs increase.

Estimation Procedures

In this section, the estimation procedures used with LOGIST, BILOG, and TESTGRAF are reviewed and the estimation procedures used with BILOG, and TESTGRAF to obtain 3PL item and ability parameter estimates are described in detail.

There is an inextricable link between estimation procedures to estimate IRT parameters and the computer programs used to implement them (Baker, 1987). There are many procedures to estimate IRT parameters (*e.g.*, joint maximum likelihood, marginal maximum likelihood, conditional maximum likelihood, joint and marginal Bayesian estimation, Heuristic estimation, and a method based on non-linear factor analysis). The estimation procedure which has been available for the longest and therefore the most widely used and most thoroughly investigated is the joint maximum likelihood (JML; Birnbaum, 1968) estimation procedure found in the LOGIST (Wingersky, Barton, & Lord, 1982; Wingersky & Lord, 1973) computer program.

LOGIST uses a joint maximum likelihood method to estimate simultaneously ability and item parameters for the 1-, 2-, or 3PL model. Although LOGIST has been the most widely used parameter estimation program, there are some problems inherent in the JML estimation procedure used by LOGIST (Hambleton *et al.*, 1991). First, ability estimates do not exist for

examinees with perfect or zero test scores and item parameter estimates do not exist for items that all examinees answered correctly or incorrectly. Examinees and items exhibiting these patterns must be eliminated before estimation may proceed. Secondly, the JML estimation procedure does not yield consistent 2PL and 3PL model estimates of item and ability parameters with small sample sizes and a small number of items (Wingersky, 1983). Finally, in the 3PL model, unless restrictions are placed on the values the item and ability parameters may take, the numerical procedure for finding the estimates can fail. With these problems in mind, many researchers have conducted studies to investigate the accuracy of item and ability estimates produced from LOGIST. In particular, many researchers have focused on JML item and ability estimates of the 3PL model.

In studies where the accuracy of 3PL model estimates produced from LOGIST were examined (Lord 1975, 1983; Swaminathan & Gifford, 1983), there is evidence that bias exists. Using LOGIST, Swaminathan and Gifford (1983) examined the bias of 3PL item parameter estimates with simulated data for a sample size of 200, a test length of 20 items, and for 20 replications. They found that small values of a were overestimated, whereas very large values of a tended to be estimated accurately. With item difficulty, easy items tended to be underestimated while very difficult items tended to be estimated accurately. The c parameter was slightly underestimated for all items. In another study, Lord (1975, 1983), using LOGIST, examined the bias of the 3PL model estimates and the accuracy of recovery of 3PL item parameters with simulated data using a, b, c, and θ parameter estimates roughly equal to those obtained with 2,995 examinees on a 90-item verbal test. Although Swaminathan and Gifford used a small

sample size in their study, Lord found similar results using a larger sample size. In general, he found that JML estimates of ability are biased, which in turn causes the item parameters to be misestimated, even for large sample sizes. In particular, he found that small and medium values of a tended to be overestimated; with item difficulty, easy and medium difficulty items were underestimated and difficult items were overestimated; the c parameter was underestimated for all items. In addition, he found that r_{ad} was .92 and that $r_{b\hat{b}}$ was .99. A correlation for the c parameter was not reported. Lord also noted that if the item parameter estimate had a large standard error, the bias was generally .1 to .2 of its standard error. Lord concludes that because standard errors are inversely proportional to sample size, when the sample size is large, the numerical value of the bias would probably be negligible. Empirical evidence from these studies exemplifies the problems inherent in using LOGIST for parameter estimation.

Alternative parameter estimation procedures are found in BILOG. BILOG is a more recent estimation program (Mislevy & Bock, 1984, 1986) in which marginal maximum likelihood (MML; Bock & Aitkin, 1981) and Bayesian estimation procedures (Bayesian priors can be placed on any item parameters that are difficult to measure; Mislevy, 1986) are used to estimate ability and item parameters for the 1-, 2-, or 3PL model. Before comparing the accuracy of parameter estimates obtained from LOGIST and BILOG, one must compare the parameter estimation procedures (MML and Bayesian) within BILOG.

Yen (1987) and Buhr and Algina (1986) compared the item and ability parameter estimates obtained from different methods that can be implemented with BILOG. Yen (1987) compared the performance of the two procedures with simulated 3PL data for a sample size of 1000 and for

test lengths of 10, 20, and 40 items. She found that the different options produced item and ability parameter estimates that were very similar. Statistics comparing the item parameters and trait estimates never differed more than by .01. Buhr and Algina (1986) compared the item parameter estimates produced by various procedures available in BILOG with simulated data for the 1PL, 2PL, 3PL, and the 3PL with common guessing parameters, for sample sizes of 250, 500, 750, and 1000 examinees, for a normal distribution and for a test length of 39 items. They found that for the most part the various item parameter estimation procedures tended to yield similar results. The major exception to this generalization concerned the Bayesian and maximum likelihood procedures applied to the 3PL model. With 250 examinees, correlations between *a* estimates obtained by the MML and Bayesian estimation procedures averaged about .75. For the *c* parameter estimates the correlation was likewise about .75. For the *b* estimates the correlations averaged about .92. With 500, 750, and 1000 examinees, the *a* estimated correlations increased to between .90 and .95 and the correlations for the *b* and *c* estimates were largely unaffected by changes in the sample sizes. Buhr and Algina highlight this exception, although unfortunately they do not state which procedure was more accurate. Nevertheless, based on the findings from these two studies, in general, it should not matter which estimation procedure one chooses to use in BILOG.

Since BILOG's release in 1983 there have been studies which have compared LOGIST and BILOG (Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987). Qualls and Ansley (1985) compared the performance of the two programs with simulated 3PL data for sample sizes of 200, 500, and 1000 and for test lengths of 10, 20, and 30 items. They found that BILOG

estimates were almost uniformly more accurate than LOGIST estimates. Unfortunately, it was not stated in their paper which options were used in the programs; for BILOG in particular, the options (aside from estimation procedure) chosen could have a substantial effect on the results. Yen (1987) compared the performance of the two programs with simulated 3PL data for a sample size of 1000 and for test lengths of 10, 20, and 40 items. She found that BILOG almost always produced more accurate estimates of item parameters and that in estimating ICCs BILOG was more accurate for a 10-item test. She also found the two programs to be about equally accurate for 30- and 40-item tests. Mislevy and Stocking (1989) compared the performance of the two programs with simulated data for a sample size of 1500 and for test lengths of 15, 30, and 45. In accordance with Qualls and Ansley (1985) and Yen (1987), Mislevy and Stocking (1989) found that with longer tests and larger samples, the programs provide similar item parameter estimates.

In general, BILOG requires smaller numbers of examinees and items than LOGIST for stable 3PL model parameter estimates (Mislevy & Stocking, 1989; Qualls & Ansley, 1985; Yen, 1987). Furthermore, these authors found that regardless of sample size and test length, BILOG estimates are almost uniformly more or just as accurate as LOGIST estimates. For a long time, LOGIST was used as the standard to which to compare other estimation programs; however, because of BILOG's distinct advantages of requiring a smaller sample size and test length to produce more or just as accurate 3PL IRT parameter estimates as LOGIST, BILOG can now be used as the standard to which to compare other estimation programs.

Regardless of the smaller necessary sample size and test length needed in BILOG to accurately estimate 3PL item parameters, it has been proven to be quite complex, computer

intensive, and still require what some practitioners consider a large sample size (Baker, 1987; Ramsay, 1991). BILOG is suitable for those with access to large datasets, such as commercial testing organizations, but it poses a problem for routine test analysis on a smaller scale. In instances when there are small sample sizes and test lengths, BILOG yields parameter estimates with large biases and large standard errors of estimates (Thissen & Wainer, 1982). For this reason Baker believed there was a need for development of small-dataset approaches to item parameter estimation and that "implementation of such procedures in a 'user-friendly' manner on a microcomputer is an absolute requirement" (p.138). Ramsay did just this and developed the microcomputer program TESTGRAF (1989) which uses nonparametric IRT estimation techniques.

The estimation procedures used in BILOG and TESTGRAF are described below.

Estimation Procedures Used with BILOG

In this section, the estimation procedures used with BILOG are described. For item parameter estimation, BILOG implements marginal maximum likelihood (MML) and Bayesian estimation procedures. First, the MML procedure in the context of the 3PL model is described.

Marginal Maximum Likelihood Procedure. Let $P_i(\theta_j)$ be the probability that the jth examinee ($j = 1, \ldots, N$) answers the ith item correctly ($i = 1, \ldots, n$). Let u_{ij} be an indicator variable which takes on the values of 1 for a correct response and 0 for an incorrect response. Let $\underline{\theta}'$

$[\theta_1 \ldots \theta_N]$ be the vector of ability scores and $\underline{a}' = [a_1, \ldots a_n]$, $\underline{b}' = [b_1 \ldots b_n]$, and $\underline{c}' = [c_1 \ldots c_n]$ be vectors of item discrimination, difficulty, and guessing parameters respectively. For the jth examinee the conditional likelihood function for the data is:

$$L\left(\underline{u}_j \middle| \theta_j, \underline{a}, \underline{b}, \underline{c}\right) = \prod_i \left[P_i(\theta_j)\right]^{u_{ij}} \left[1 - P(\theta_j)\right]^{1-u_{ij}},$$

where $\underline{u}' = [u_{1j} \ldots u_{Nj}]$. That is, the likelihood function is conditioned on the jth ability parameter and item parameters for all items.

Each examinee's ability score (θ) is considered to be randomly chosen from a population with ability distribution $f(\theta)$. The marginal likelihood of the data for the jth examinee is:

$$L\left(\underline{u}_j \middle| \underline{a}, \underline{b}, \underline{c}\right) = \int L\left(\underline{u}_j \middle| \theta_j, \underline{a}, \underline{b}, \underline{c}\right) f(\theta)\, d\theta .$$

Essentially, the marginal likelihood is obtained as a weighted average of the conditional likelihoods where the weights are determined by $f(\theta)$. This weighting process removes the dependence on θ, and therefore, in the process of estimating item parameters it is not necessary to estimate the ability parameters for the N examinees. The function maximized in the MML procedure is

$$L\left(\underline{u} \middle| \underline{a}, \underline{b}, \underline{c}\right) = \prod_j L\left(\underline{u}_j \middle| \underline{a}, \underline{b}, \underline{c}\right).$$

To implement the MML procedure it is necessary to make an assumption about the form of $f(\theta)$. In BILOG the default option for $f(\theta)$ is the standard normal distribution. However, the program does permit the user to specify other distributions.

In BILOG, the distribution $f(\theta)$ can be treated in two ways. The first way treats $f(\theta)$ as a distribution to be estimated. Thus the assumed $f(\theta)$ is the basis for starting values in an iterative procedure for estimation $f(\theta)$ and the item parameters. As Mislevy and Bock (1984) note this procedure is similar to the joint maximum likelihood procedure. The second way treats $f(\theta)$ as an assumption about the distribution of ability. The same distribution is used throughout the iterative procedure for estimation the item parameters.

Bayesian Procedures. Bayesian procedures incorporate assumptions about the distribution of item parameters. These assumed distributions are called prior distributions. The incorporation of prior distributions into the estimation procedure make it unlikely for the estimates to occur in regions that are less probable according to the prior distribution. BILOG will use default prior distributions or it allows the user to specify more diffuse or tighter priors. In addition, the user can choose which parameters to place priors on.

Similar to the $f(\theta)$ distribution in the MML procedure in BILOG, when using the Bayesian procedure in BILOG the user can treat the prior distributions in two ways. The user can specify that the priors remain the same at each iteration or that the parameters of the priors be updated on each iteration.

In addition, to choosing the estimation procedure to use in BILOG, BILOG has three options for the treatment of omitted items. Omitted items can be treated as incorrect, not presented, or fractionally correct.

Estimation Procedure Used with TESTGRAF

In this section, the estimation procedure used with TESTGRAF is described. Let $P_{im}(\theta)$ denote the function relating the probability of choosing the option m ($m = 1, \ldots, M$) for item i ($i = 1, \ldots, n$) to the ability level θ. Let j be an index for examinee ($j = 1 \ldots N$). Let y_{imj} be an indicator variable which takes on the values 1 if examinee j actually chooses option m for item i and 0 if examinee j does not choose option m for item i.

TESTGRAF uses nonparametric IRT kernel smoothing techniques as an alternate to maximum likelihood estimation to estimate item and ability parameters. In the case where the examinees' data are dichotomously coded, TESTGRAF estimates the ICC in the following way:

1) A total score X_j, for each examinee j, is computed by calculating the percentage of items answered correctly.
2) The examinees and their exam responses are sorted on the basis of their total scores X_j.
3) The jth examinee is assigned the jth quantile of the standard normal distribution, z_j. where z_j is the value such that the area under the standard normal density function to the left of z_j is equal to $j/(N+1)$.
4) For the mth option for the ith item the indicator values y_{imj} are calculated.

5) The relationship $P_{im}(\theta)$ is estimated by smoothing the relationship between the 0-1 indicator variable values y_{imj} and the standard normal quantiles z_j. Smoothing is in effect a type of local averaging, in which for any ability level θ the probability of choice $P_{im}(\theta)$ at that level is a weighted average of the values of y_{imj} for examinees with ability levels close to θ. The smoothing techniques used is a two-stage kernel smoothing operation. The first stage carries out a very fast smooth for these N pairs of values to estimate $P_{im}(\theta)$ at a small number of argument values, θ_q, ($q = 1, \ldots, Q$). The second stage smoothes these Q estimated values using an adaptive smoothing operation that takes into account the paucity of data at the extreme ability levels.

Ramsay (1991) claims that the estimation procedure used in TESTGRAF is 500 times faster than using some of the common parametric approaches such as JML or MML estimation, that there is no loss of efficiency, and that only as few as a hundred examinees and twenty test questions are needed to estimate ICCs. However, one must keep in mind that Ramsay states that the item parameter estimates are fairly crude and should not be seen as substitutes for a more serious analysis of the data using the logistic model by programs such as LOGIST and BILOG.

Purpose of Study

As stated above, the purpose of this study was to compare the effects of varying test length and sample size on the 3PL model item parameter and $P(\theta)$ estimates obtained by TESTGRAF and BILOG. Three specific research questions were formulated:

1. What is the effect of **test length** on the accuracy and consistency of TESTGRAF and BILOG in estimating:
    a) 3PL item parameters (*a*, *b*, and *c*)?
    b) $P(\theta)$'s at different ability levels?
2. What is the effect of **sample size** on the accuracy and consistency of TESTGRAF and BILOG in estimating:
    a) 3PL item parameters (*a*, *b*, and *c*)?
    b) $P(\theta)$'s at different ability levels?
3. What is the effect of different combinations of **test length and sample size** on the accuracy and consistency of TESTGRAF and BILOG in estimating:
    a) 3PL item parameters (*a*, *b*, and *c*)?
    b) $P(\theta)$'s at different ability levels?

CHAPTER III

Method

In this chapter, the methodology for the study is presented. The method is divided into four sections: test conditions, computer programs, procedure, and data analysis.

Test Conditions

Data corresponding to eight different test conditions were simulated using the 3PL model. Each test condition was defined by some combination of two factors: (1) test length and (2) sample size. Two test lengths were used; n=20 and 40. Twenty items is the minimum number of items claimed by Ramsay for TESTGRAF to estimate $P(\theta)$'s accurately. The 40-item test was comprised of two replications of the 20-item test (*i.e.*, the item parameters for items 21 through 40 were the same item parameters as those for items 1 through 20). As well, four sample sizes were used; N=100, 250, 500, and 1000. The lower sample size of 100 examinees was chosen because Ramsay (1991) claims that only as few as twenty test questions and 100 examinees are needed to accurately estimate $P(\theta)$'s. The upper sample size of 1000 examinees was chosen because it is known that BILOG estimates item parameters accurately at this level (Qualls & Ansley, 1985; Mislevy & Stocking, 1989; Yen, 1987). A crossing of the two levels of test length

with the four levels of sample size resulted in eight distinct test conditions, as shown in Table 1. For each test condition, 100 replications were performed.

Table 1

Summary of Test Conditions

Test Length	Sample Size	Replications
20	100	100
	250	100
	500	100
	1000	100
40	100	100
	250	100
	500	100
	1000	100
		800

Computer Programs

The computer program used to generate unidimensional dichotomous 3PL data was the FORTRAN M2PL Data Generation Program (Ackerman, 1985; modified by Gessaroli, 1994). The two computer programs used to estimate item parameters were TESTGRAF (Ramsay, 1993) and PC-BILOG 3.04 (Mislevy & Bock, 1986). In using TESTGRAF, all default options were used and the number of answer choices was set to four since this reflects what is commonly found in multiple choice tests. In using BILOG, all default options were also used with the following exceptions. As in TESTGRAF, the number of answer choices was set to four. The CASE parameter which determines how data are handled during the estimation process was set to one: CASE=1 is the fastest option and can be used when all examinees have taken all items.

Procedure

In this section, the steps which were used to simulate the data and measure the accuracy and consistency of the item parameter and $P(\theta)$ estimates obtained from TESTGRAF and BILOG are described in detail.

Step 1

The purpose of Step 1 was to select population item parameters. Twenty items were chosen from a 60-item American College Testing (ACT) Math test based on the item parameter

estimates obtained from LOGIST on the 60-item ACT test with 10,000 examinees. Specifically, the 20 items were chosen from the 60-item test based on the following criteria: $.4<a<1.2$, $-1.5<b<1.5$, and $c<.2$. These item parameter estimates were considered as the true item parameters in the subsequent steps. These true item parameters are displayed in Table 2. Because the focus was on choosing items with a variety of *a* and *b* values, a number of items with *c* values equal to zero resulted. Only items that were highly discriminating ($.95<a<1.2$) had non-zero *c* values.

Table 2

True Item Parameters

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	1.020	-0.786	0.000	11	0.601	-0.146	0.000
2	.589	-0.728	0.000	12	1.198	0.111	0.000
3	.911	-0.751	0.000	13	0.459	1.188	0.000
4	.801	-0.700	0.000	14	1.081	0.528	0.075
5	.511	-0.830	0.000	15	0.617	0.216	0.000
6	1.168	-0.080	0.011	16	0.697	1.168	0.195
7	0.597	0.284	0.000	17	1.058	0.821	0.069
8	0.821	0.226	0.000	18	0.590	1.139	0.020
9	0.659	0.268	0.000	19	0.544	0.527	0.000
10	0.967	-0.230	0.038	20	1.139	0.894	0.029

Step 2

The purpose of Step 2 was to generate a sample of unidimensional 3PL data based on the true item parameters. The 40-item test was comprised of two sets of the 20-item test. The underlying distribution of examinee trait levels was assumed to be standard normal (N(0,1)).

Step 3

The *a*, *b*, and *c* item parameters for each item were estimated using TESTGRAF and BILOG.

Step 4

The probabilities of correctly answering an item, $P(\theta)$, (θ= -3.0, -2.9, . . . , 2.9, 3.0), for the 3PL model were calculated for each item, using both the item parameter estimates obtained from TESTGRAF and BILOG in Step 3 and the true item parameters.

Step 5

The difference between the true and estimated item parameters for TESTGRAF and BILOG were calculated for each item using the *a*, *b*, and *c* estimates obtained in Step 3. Similarly, the difference between the true and estimated $P(\theta)$'s at each θ, (θ=-3.0, -2.9, . . . , 2.9, 3.0), for TESTGRAF and BILOG were calculated for each item using the $P(\theta)$ estimates obtained in Step 4.

Step 6

The purpose of Step 6 was to measure the accuracy and consistency of the estimated item parameters and $P(\theta)$'s. Statistical measures of accuracy and consistency are bias and efficiency, respectively. Bias is the average difference of the parameter estimates from the true parameters. An estimator is said to be unbiased if the mean of the sample is equal to the population (true) characteristic to be estimated, in which case bias would be equal to zero. Efficiency is measured by the root mean squared difference (RMSD) between the true and estimated parameters. An estimator is said to be efficient if the RMSD is zero. Given that two estimators show little or no bias, it is reasonable to prefer the estimator with the smaller RMSD. Because the measures of bias and efficiency were treated somewhat differently for the estimated item parameters and $P(\theta)$'s, the steps are presented separately.

Step 6a – Measures of bias and efficiency of estimated item parameters. The average difference and RMSD were calculated across all of the items. The measure of bias was calculated by averaging the difference between the parameter estimates and the true parameters across all of the items. Similarly, the measure of efficiency was calculated by calculating the RMSD between the parameter estimates and the true parameters across all of the items. The result was one measure of bias and one measure of efficiency for each *a*, *b*, and *c* estimate for each of TESTGRAF and BILOG.

Step 6b – Measures of bias and efficiency of estimated $P(\theta)$'s. Before calculating measures of bias and efficiency of the $P(\theta)$'s, the ability (θ) distribution was equally divided, based on proportion of examinees, into three ability levels: low, average, and high. Since the ability

distribution of the examinee population was assumed to be normal, the low, average, and high levels corresponded to θ=-3.0 to -0.44, θ=-0.44 to 0.44, and θ =0.44 to 3.0, respectively. The average difference and RMSD of the $P(\theta)$'s within each ability level were then calculated. The result was one measure of bias and one measure of efficiency for each $P(\theta_{low})$, $P(\theta_{ave})$, and $P(\theta_{high})$ estimate for each of TESTGRAF and BILOG.

Step 7

Steps 2 through 6 were repeated 100 times for the first test condition. The results of Steps 2 through 7 are: 1) 100 measures of bias and 100 measures of efficiency for each *a*, *b*, and *c* estimate for each of TESTGRAF and BILOG and 2) 100 measures of bias and 100 measures of efficiency for each $P(\theta_{low})$, $P(\theta_{ave})$, and $P(\theta_{high})$ estimate for each of TESTGRAF and BILOG.

Step 8

Steps 2 through 7 were repeated for each of the other test conditions.

Data Analysis

The data analysis was conducted in two parts according to the research questions. First, the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating each of the *a*, *b*, and *c* parameters were examined. Second, these effects in estimating $P(\theta)$'s for different ability levels were examined.

Bias and Efficiency of Item Parameter Estimates

The bias and efficiency of TESTGRAF and BILOG in estimating 3PL item parameters were analysed separately. Furthermore, in examining the bias and efficiency of TESTGRAF and BILOG in estimating 3PL item parameters, each parameter was considered separately. That is, one ANOVA was used to examine the bias of TESTGRAF and BILOG in estimating each of the *a*, *b*, and *c* item parameters and one ANOVA was used to examine the efficiency of TESTGRAF and BILOG in estimating each of the *a*, *b*, and *c* item parameters. In total, six 2x4x2 ANOVAs with repeated measures on the last factor were used to obtain measures of effect size for the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating 3PL item parameters. For each ANOVA, the first factor corresponded to the two test lengths (20- and 40-item tests), the second factor corresponded to the four sample sizes (N=100, 250, 500, 1000), and the last factor corresponded to each of the estimation procedures used in TESTGRAF and BILOG.

Bias and Efficiency of $P(\theta)$'s

Similar to above, the bias and efficiency of TESTGRAF and BILOG in estimating $P(\theta)$'s were analysed separately. In total, two 2x4x3x2 ANOVAs with repeated measures on the last two factors were used to obtain measures of effect size for the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating $P(\theta)$'s. For each MANOVA, the first, second, and last factors corresponded to the two test

lengths, the four sample sizes, and the two estimation procedures, respectively. The additional factor corresponded to the three ability levels (low, average, and high).

Traditionally, to examine the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating item parameters and $P(\theta)$'s, one would conduct MANOVAs or ANOVAs and follow-up with the appropriate post-hoc tests if there were any significant effects. However, in the case where there are large sample sizes, one would possibly find many significant effects due to the large amount of power. That is, one would find even small "practically" insignificant differences between the estimation procedures to be "statistically" significant. This is of little interest to practitioners who want to know if there is, in general, a "big" difference between the two procedures in estimating item parameters and $P(\theta)$'s.

One way to circumvent this problem is to use measures of effect size (ES) as an alternative to significance tests. In this study, due to the large number of replications of each test condition, and therefore large power, measures of ES (Cohen, 1992) were used to examine the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating item parameters and $P(\theta)$'s. The ES index used was f^2 [$f^2=R^2(1-R^2)$]. This ES index is defined for squared multiple correlations (R^2), where $R^2=SS_{effect}/(SS_{effect}+SS_{error})$. f^2 is a measure of the amount of sums of squares explained by the effect of interest relative to the amount of sums of squares in the model not explained (SS_{effect}/SS_{error}). Only large ESs were flagged as interesting because, in general, practitioners would like to know whether there is a "big" difference between the bias and efficiency of TESTGRAF and BILOG in estimating item

parameters and $P(\theta)$'s. A large ES corresponds to a value of f^2 greater than .35 (Cohen, 1992). For the purpose of this study, large ESs (>.35) were considered to be important.

CHAPTER IV

Results

The results are presented in two parts according to the research questions. First, the results of the bias and efficiency of TESTGRAF and BILOG in estimating the item parameters are presented. Second, the results of the bias and efficiency of TESTGRAF and BILOG in estimating $P(\theta)$'s at different ability levels are described. Subsequently, the results are summarized according to the research questions.

Bias and Efficiency of Item Parameter Estimates

In this section, the bias and efficiency of TESTGRAF and BILOG in estimating the *a*, *b*, and *c* parameters are considered, respectively. In each part, only ESs due to procedure and all interactions involving procedure are considered – the main effect of estimation procedure (P); the interactions of test length by estimation procedure (LxP); sample size by estimation procedure (SxP); and test length by sample size by estimation procedure (LxSxP). The main effects of test length (L) and sample size (S) and the interaction of test length by sample size (LxS) were of no interest in this study because they did not allow for a comparison of TESTGRAF and BILOG because they did not include the effect of procedure.

The way in which the results are presented in each part is as follows. First, bias ESs for the main effect of P or the interactions of LxP, SxP, and LxSxP in estimating the item parameter are interpreted. Second, efficiency ESs for these effects are interpreted.

Discrimination (*a*)

In this section, bias and efficiency ESs due to procedure and all interactions involving procedure in estimating the *a* parameter are interpreted. Descriptive statistics and ESs for the bias and efficiency of TESTGRAF and BILOG in estimating the *a* parameter are presented in Tables 3 and 4, respectively.

As shown in Table 3, on average, TESTGRAF was less biased ($\overline{X}_{bias a_{TG}}$ =.026 and $\overline{X}_{bias a_B}$ = –.222) and slightly more efficient ($\overline{X}_{eff a_{TG}}$ =.165 and $\overline{X}_{eff a_B}$ =.177) than BILOG in estimating the *a* parameter – TESTGRAF slightly overestimated and BILOG underestimated the *a* parameter.

Table 3

Descriptive Statistics: Bias and Efficiency of *a* Estimates Obtained from TESTGRAF and BILOG

	TESTGRAF			BILOG		
	M	SD	N	M	SD	N
Bias *a*	.026	.046	800	-.222	.080	800
Eff *a*	.165	.045	800	.177	.066	800

Table 4

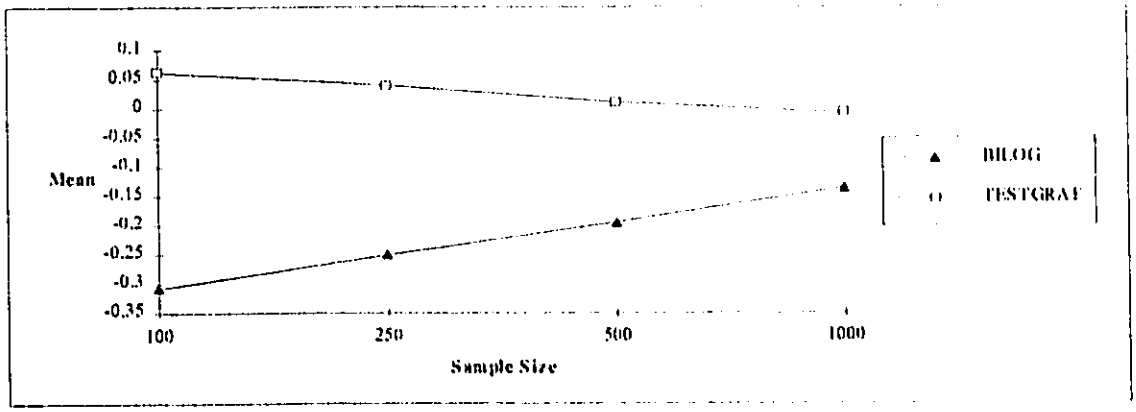
Effect Sizes: Bias and Efficiency of *a* Estimates Obtained from TESTGRAF and BILOG

Effect	ES (Bias)	ES (Eff)
P	71.411 *	.140
LxP	.304	.002
SxP	9.534 *	.371 *
LxSxP	.022	.001

Note. P=Estimation Procedure, L=Test Length, S=Sample Size, and * large effect (> .35).

Bias. As shown in Table 4, there were large bias ESs for the procedure (P) main effect and the sample size by procedure (SxP) interaction in estimating the *a* parameter ($ES_{biasa(P)}$=71.411 and $ES_{biasa(SxP)}$=9.534). The main effect of P can be interpreted by looking at the SxP interaction. The large ES for the SxP interaction suggests that sample size affected the bias of the *a* parameter differently for TESTGRAF and BILOG. By examining Figure 1, it is apparent that: i) the difference between the bias of the two procedures in estimating the *a* parameter decreased as sample size increased, ii) TESTGRAF was less biased than BILOG in estimating the *a* parameter at all sample sizes, and iii) TESTGRAF slightly overestimated while BILOG underestimated the *a* parameter at all sample sizes.

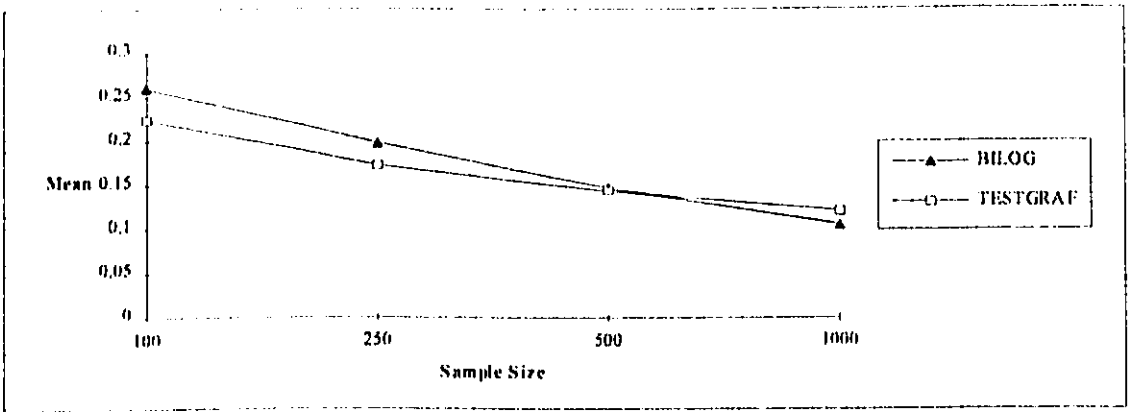
Figure 1. Interaction of Sample Size by Procedure on the Bias of the *a* Estimates.



There were no large bias ESs for the test length by procedure (LxP) and test length by sample size by procedure (LxSxP) interactions in estimating the *a* parameter. This suggests that test length did not affect the bias of the *a* estimates differently for TESTGRAF and BILOG and that there was no large difference between the SxP interactions for the 20- and 40-item tests in the bias of the *a* estimates.

Efficiency. As shown in Table 4, there was a large efficiency ES for the sample size by procedure (SxP) interaction in estimating the *a* parameter ($ES_{effa(SxP)}$=.371). This suggests that sample size affected the efficiency of the *a* parameter differently for TESTGRAF and BILOG. By examining Figure 2, it is apparent that: i) the difference between the efficiency of the two procedures in estimating the *a* parameter decreased slightly as sample size increased, ii) TESTGRAF was more efficient than BILOG in estimating the *a* parameter at sample sizes of 100 and 250, and iii) BILOG was just as efficient or slightly more efficient than TESTGRAF in estimating the *a* parameter at sample sizes of 500 and 1000.

Figure 2. Interaction of Sample Size by Procedure on the Efficiency of the *a* Estimates.



There were no large efficiency ESs for the procedure main effect (P) or for the test length by procedure (LxP) and test length by sample size by procedure (LxSxP) interactions in estimating the *a* parameter. This suggests that on average TESTGRAF and BILOG did not differ largely in efficiency of the *a* estimates, that test length did not affect the efficiency of the *a* estimates differently for TESTGRAF and BILOG, and that there was no large difference between the SxP interactions for the 20- and 40-item tests in the efficiency of the *a* estimates.

Overall, TESTGRAF and BILOG differed largely only in the bias and the efficiency of the *a* estimates for different sample sizes. TESTGRAF was less biased than BILOG at all sample sizes and was more efficient or just as efficient as BILOG at small sample sizes (N=100, 250, and 500). The difference in bias and efficiency between the two procedures became less pronounced as sample size increased. Finally, test length did not affect the bias or efficiency of the *a* estimates differently for TESTGRAF and BILOG.

Difficulty (*b*)

In this section, bias and efficiency ESs due to procedure and all interactions involving procedure in estimating the *b* parameter are interpreted. Descriptive statistics and ESs for the bias and efficiency of TESTGRAF and BILOG in estimating the *b* parameter are presented in Tables 5a and 5b and 6, respectively.

As shown in Table 5a, on average, TESTGRAF was less biased ($\overline{X}_{bias b_{TG}} = -.238$ and $\overline{X}_{bias b_{B}} = -.337$) but slightly less efficient ($\overline{X}_{eff b_{TG}} = .290$ and $\overline{X}_{eff b_{B}} = .226$) than BILOG. The inefficiency of TESTGRAF in estimating the *b* parameter is particularly evident in its standard deviation of the efficiency of .295 compared to the standard deviation of the efficiency of BILOG in estimating the *b* parameter of .061. By examining Table 5b, it is apparent that with a sample of 100 examinees and a 20-item and 40-item test, the standard deviations of the efficiency of the *b* estimate for TESTGRAF are .136 and .755 respectively. In comparison to the other standard deviations of efficiency in Table 5b, these are large standard deviations and account for the large standard deviation of the efficiency of the *b* estimate for the entire sample. TESTGRAF is not very consistent in estimating the *b* parameter with small samples, especially when combined with longer tests.

Table 5a

Descriptive Statistics: Bias and Efficiency of *b* Estimates Obtained from TESTGRAF and BILOG

	TESTGRAF			BILOG		
	M	SD	N	M	SD	N
Bias *b*	-.238	.125	800	-.337	.114	800
Eff *b*	.290	.295	800	.226	.061	800

Bias. As shown in Table 6, there was only a large bias ES for the procedure (P) main effect in estimating the *b* parameter ($ES_{biasb(P)}$=1.232). This suggests that there was a large difference between TESTGRAF and BILOG in the bias of the *b* estimates. By examining Table 5, it is apparent that although both procedures underestimated the *b* parameter, BILOG did so more than TESTGRAF ($\bar{X}_{biasb_{TG}} = -.238$ and $\bar{X}_{biasb_{B}} = -.337$). TESTGRAF was less biased than BILOG.

There were no large bias ESs for the test length by procedure (LxP), sample size by procedure (SxP), and test length by sample size by procedure interactions in estimating the *b* parameter. This suggests that neither test length nor sample size largely affected the bias of the *b* estimates differently for TESTGRAF and BILOG.

Table 5b

Descriptive Statistics: Efficiency of TESTGRAF and BILOG in Estimating the *b* Parameter at Different Test Lengths and Sample Sizes

	TESTGRAF			BILOG		
	M (Eff)	SD (Eff)	N	M (Eff)	SD (Eff)	N
20 items						
100 examinees	.382	.136	100	.302	.058	100
250	.256	.048	100	.237	.042	100
500	.218	.038	100	.203	.030	100
1000	.191	.026	100	.170	.024	100
40 items						
100 examinees	.556	.755	100	.302	.041	100
250	.284	.047	100	.230	.028	100
500	.231	.027	100	.193	.022	100
1000	.204	.019	100	.167	.016	100
For entire sample	.290	.295	800	.226	.061	800

Table 6

Effect Sizes: Bias and Efficiency of *b* Estimates Obtained from TESTGRAF and BILOG

Effect	ES (Bias)	ES (Eff)
P	1.232 *	.057
LxP	.134	.013
SxP	.015	.048
LxSxP	.001	.015

Note. P=Estimation Procedure, L=Test Length, S=Sample Size, * large effect (> .35).

Efficiency. As shown in Table 6, there were no large efficiency ESs for the procedure (P) main effect and the test length by procedure (LxP), sample size by procedure (SxP), and test length by sample size by procedure (LxSxP) interactions in estimating the *b* parameter. This suggests that neither test length nor sample size largely affected the efficiency of the *b* estimates differently for TESTGRAF and BILOG. As already noted, although there were no large efficiency ESs in estimating the *b* parameter, an interesting finding when one looks at the efficiency of TESTGRAF is that the variability of the efficiency varies greatly at N=100 (see Table 5b). The corresponding efficiency of BILOG does not vary greatly and is much more reasonable.

Overall, TESTGRAF and BILOG differed largely only in the bias of the *b* estimates. TESTGRAF was less biased than BILOG. There was no large difference between TESTGRAF and BILOG in the bias or efficiency of the *b* estimates for different test lengths or sample sizes.

Guessing (*c*)

In this section, bias and efficiency ESs due to procedure and all interactions involving procedure in estimating the *c* parameter are interpreted. Descriptive statistics and ESs for the bias and efficiency of TESTGRAF and BILOG in estimating the *c* parameter are presented in Tables 7 and 8, respectively. It is important to recall that these findings are based on data which sere simulated based on item parameter estimates obtained from LOGIST on a 60-item ACT test with 10, 000 examinees, of which many items had *c* estimates equal to zero.

As shown in Table 7, on average, TESTGRAF was less biased ($\overline{X}_{bias_{TG}} = -.092$ and $\overline{X}_{bias_B} = -.146$) but less consistent ($\overline{X}_{effc_{TG}} = .095$ and $\overline{X}_{effc_B} = .066$) than BILOG in estimating the *c* parameter.

Table 7

Descriptive Statistics: Bias and Efficiency of *c* Estimates Obtained from TESTGRAF and BILOG

	TESTGRAF			BILOG		
	M	SD	N	M	SD	N
Bias *c*	-.092	.021	800	-.146	.027	800
Eff *c*	.095	.020	800	.066	.007	800

Table 8

Effect Sizes: Bias and Efficiency of *c* Estimates Obtained from TESTGRAF and BILOG

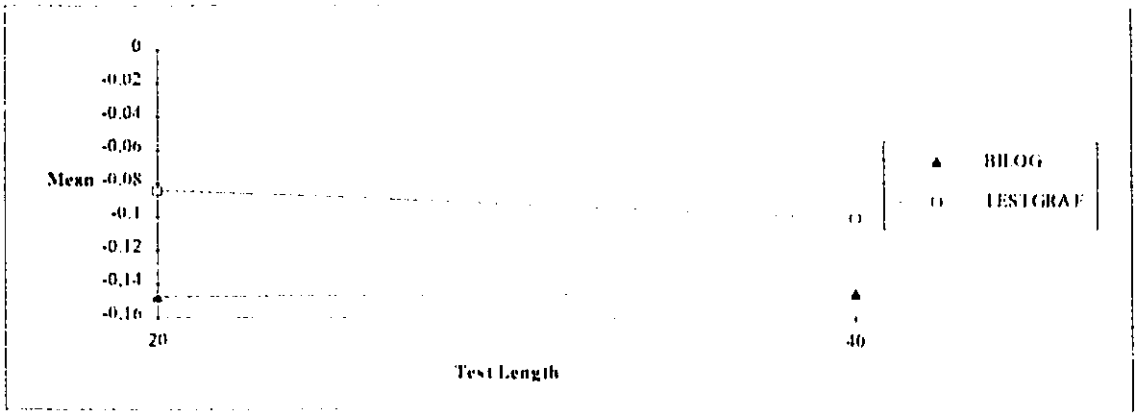
Effect	ES (Bias)	ES (Eff)
P	37.655 *	6.824 *
LxP	1.204 *	.053
SxP	2.293 *	1.294 *
LxSxP	.029	.003

Note. P=Estimation Procedure, L=Test Length, S=Sample Size, * large effect (> .35).

Bias. As is shown in Table 8, there were large bias ESs for the procedure (P) main effect and the test length by procedure (LxP) and sample size by procedure (SxP) interactions ($ES_{biasc(P)}$=37.655, $ES_{biasc(LxP)}$=1.204, and $ES_{biasc(SxP)}$=2.293) in estimating the *c* parameter. The main effect of procedure can be interpreted by looking at the LxP and SxP interactions.

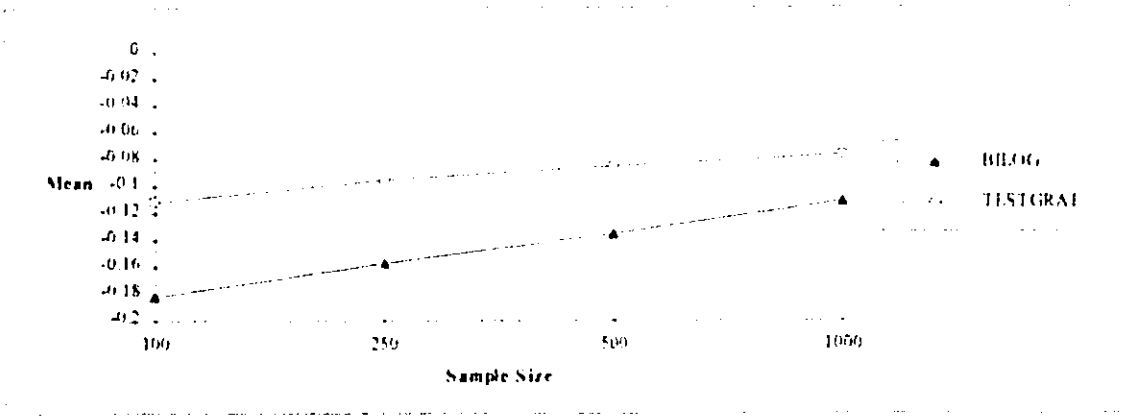
The large bias ES for the LxP interaction suggests that test length affected the bias of the *c* estimates differently for TESTGRAF and BILOG. By examining Figure 3, it is apparent that: i) the difference between the bias of the two procedures in estimating the *c* parameter decreased as test length increased, ii) TESTGRAF was less biased than BILOG in estimating the *c* parameter at both test lengths, and iii) both TESTGRAF and BILOG underestimated the *c* parameter at both test lengths. In essence, TESTGRAF was less biased than BILOG in estimating the *c* parameter, and the difference in bias was more pronounced with the shorter test.

Figure 3. Interaction of Test Length by Procedure on the Bias of *c* Estimates.



The large bias ES for the SxP interaction suggests that sample size affected the bias of the *c* estimates differently for TESTGRAF and BILOG. By examining Figure 4, it is apparent that: i) the difference between the bias between the two procedures in estimating the *c* parameter decreased as sample size increased, ii) TESTGRAF was less biased than BILOG in estimating the *c* parameter at all sample sizes, and iii) both TESTGRAF and BILOG underestimated the *c* parameter at all sample sizes. Again, TESTGRAF was less biased than BILOG in estimating the *c* parameter, and the difference in bias was more pronounced with the smaller sample sizes.

Figure 4. Interaction of Sample Size by Procedure on the Bias of the *c* Estimates.

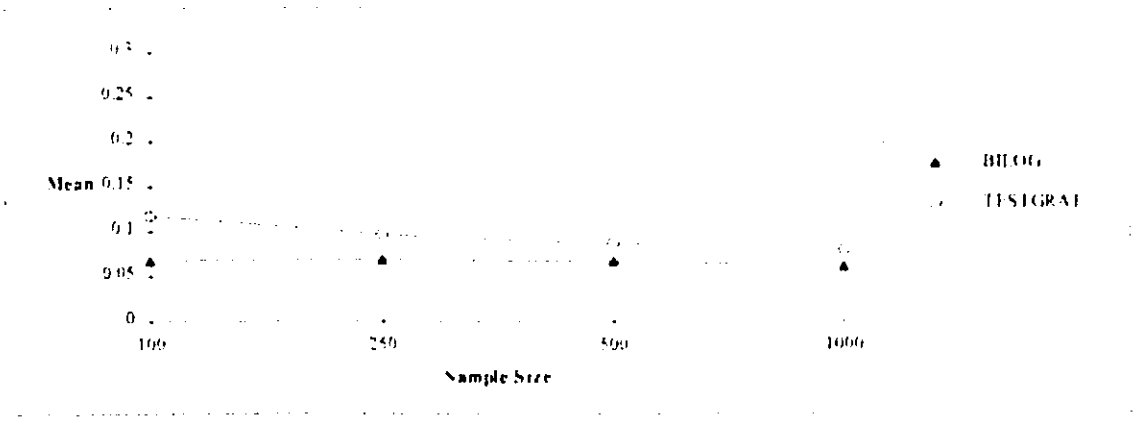


There was no large bias ES for the test length by sample size by procedure (LxSxP) interaction in estimating the *c* parameter. This suggests that the SxP interaction in estimating the *c* parameter did not differ across test lengths.

Efficiency. By examining Table 8, it is apparent that there was a large efficiency ES for the sample size by procedure (SxP) interaction. This suggests that sample size affected the

efficiency of the *c* parameter differently for TESTGRAF and BILOG. By examining Figure 5, it is apparent that: i) the difference between the efficiency of the two procedures in estimating the *c* parameter decreased as sample size increased and ii) TESTGRAF was less efficient than BILOG in estimating the *c* parameter at all sample sizes. In essence, TESTGRAF was less efficient than BILOG in estimating the *c* parameter at all sample sizes and the difference in efficiency between the two procedures decreased as sample size increased.

Figure 5. Interaction of Sample Size by Procedure on the Efficiency of the *c* Estimates.



There was no large efficiency ES for the test length by sample size by procedure (LxSxP) interaction. This suggests that the SxP interaction in estimating the *c* parameter did not differ across test lengths.

Overall, TESTGRAF and BILOG differed largely on the bias of the *c* estimates for the different test lengths and for the different sample sizes. TESTGRAF was less biased than BILOG for both test lengths and for all sample sizes. The differences between TESTGRAF and

BILOG decreased as sample size and test length increased. TESTGRAF and BILOG also differed largely on the efficiency of the *c* estimates for different sample sizes. TESTGRAF was less efficient than BILOG at all sample sizes.

Bias and Efficiency of $P(\theta)$'s

In this section, measures of ES for the main effects and interactions of test length and sample size on the bias and efficiency of TESTGRAF and BILOG in estimating the $P(\theta)$'s at different ability levels are examined. Similar to above, only ESs due to procedure (P) and all interactions involving P are considered – the main effect of estimation procedure (P); the interactions of test length by procedure (LxP); sample size by estimation procedure (SxP); ability level by estimation procedure (AxP); test length by sample size by and estimation procedure (LxSxP); test length by ability level by estimation procedure (LxAxP); sample size by ability level by and estimation procedure (SxAxP); and test length by sample size by ability level by estimation procedure (LxSxAxP). Similar to above, the main effects and interactions which did not allow for a comparison of TESTGRAF and BILOG because they did not include the effect of P were of no interest in this study – the main effects of test length (L), sample size (S), and ability level (A) and the interactions of test length by sample size (LxS); test length by ability level (LxA); sample size by ability level (SxA); and test length by sample size by ability level (LxSxA).

The way in which the results are presented is similar to above. First, bias ESs due to procedure and all interactions involving procedure in estimating the $P(\theta)$'s are interpreted. Second, efficiency ESs for these effects are interpreted.

Descriptive statistics and ESs for the bias and efficiency of TESTGRAF and BILOG in estimating the $P(\theta)$'s are presented in Tables 9 and 10, respectively.

As shown in Table 9, on average, TESTGRAF was less biased ($\bar{X}_{bias P(\theta)_{TG}} = -.014$ and $\bar{X}_{bias P(\theta)_{B}} = -.033$) and slightly more efficient ($\bar{X}_{eff P(\theta)_{TG}} = .047$ and $\bar{X}_{eff P(\theta)_{B}} = .048$) than BILOG in estimating the $P(\theta)$'s.

Table 9

Descriptive Statistics: Bias and Efficiency of $P(\theta)$'s Obtained from TESTGRAF and BILOG

	TESTGRAF			BILOG		
	M	SD	N	M	SD	N
Bias $P(\theta)$	-.014	.019	800	-.033	.019	800
Eff $P(\theta)$	.047	.014	800	.048	.021	800

Bias. As is shown in Table 10, there were large ESs for the procedure (P) main effect, the ability level by procedure (AxP), and sample size by ability level by procedure (SxAxP)

interaction in estimating $P(\theta)$'s ($ES_{bias P(\theta)(P)}$=.965, $ES_{bias P(\theta)(SxAxP)}$=0.681 and $ES_{bias P(\theta)(AxP)}$=6.143). The main effect of P can be interpreted by examining the AxP and the SxAxP interactions.

The large bias ES for the AxP interaction suggests ability level affects the bias of the $P(\theta)$'s differently for TESTGRAF and BILOG. By examining Figure 6, it is apparent that: i) in the low ability range, both TESTGRAF and BILOG underestimated the ICC, but BILOG slightly more so than TESTGRAF; ii) in the average ability range, both TESTGRAF and BILOG slightly overestimated the ICC, but BILOG slightly more so than TESTGRAF; and iii) in the high ability range, TESTGRAF overestimated more than BILOG underestimated the ICC.

Table 10

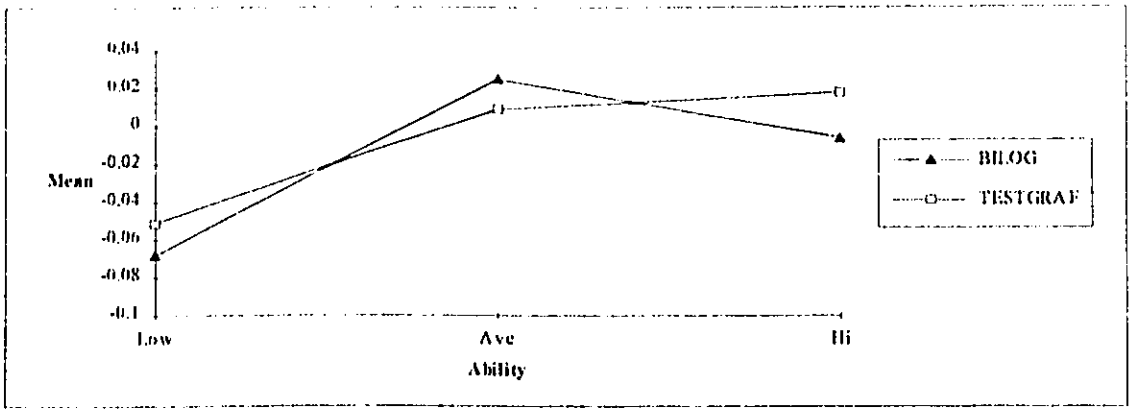
Effect Sizes: Bias and Efficiency of $P(\theta)$'s

Obtained from TESTGRAF and BILOG

Effect	ES (Bias)	ES (Eff)
P	.965 *	.078
LxP	.019	.081
SxP	.236	.107
AxP	6.143 *	.242
LxSxP	.006	.004
LxAxP	.238	.054
SxAxP	.681 *	.289
LxSxAxP	.013	.014

Note. P=Estimation Procedure,
A=Ability Level, L=Test Length,
S=Sample Size, * large effect (> 0.35).

Figure 6. Interaction of Ability by Procedure on the Bias of the $P(\theta)$ Estimates.



The large bias ES for the SxAxP interaction suggests that there was a large difference between the AxP interactions for bias for the different sample sizes for TESTGRAF and BILOG in estimating the $P(\theta)$'s. By examining Figures 7a to 7d, it is apparent that the difference lies at the high ability level. As sample size increased the difference between the two procedures decreased.

Figure 7a. Interaction of Ability by Procedure (N=100) on the Bias of the $P(\theta)$ Estimates.

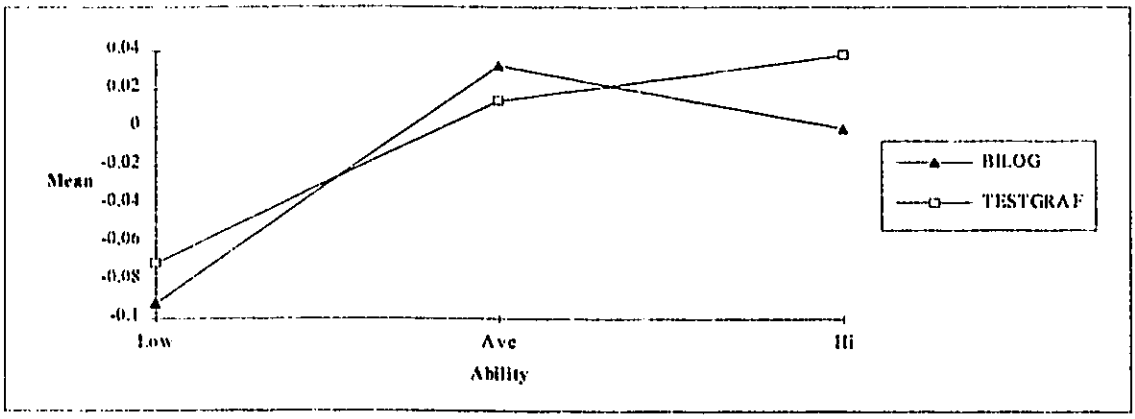


Figure 7b. Interaction of Ability by Procedure (N=250) on the Bias of the $P(\theta)$ Estimates.

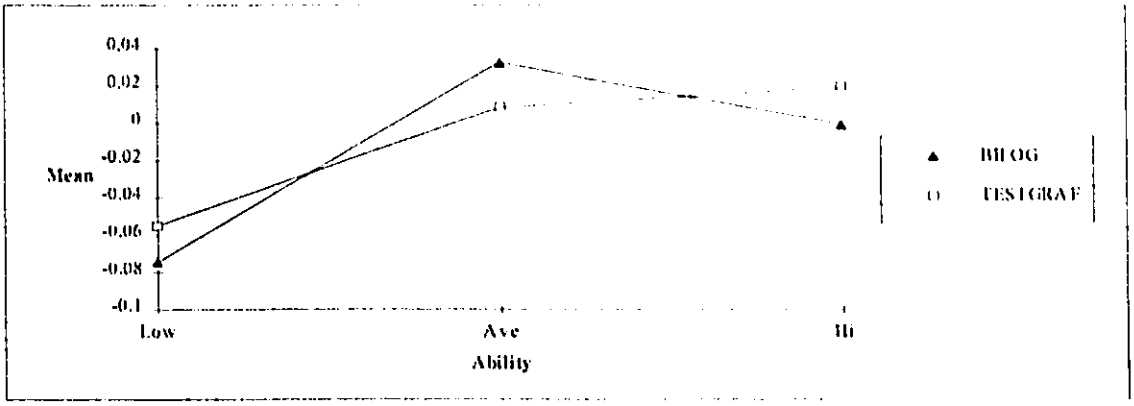


Figure 7c. Interaction of Ability by Procedure (N=500) on the Bias of the $P(\theta)$ Estimates.

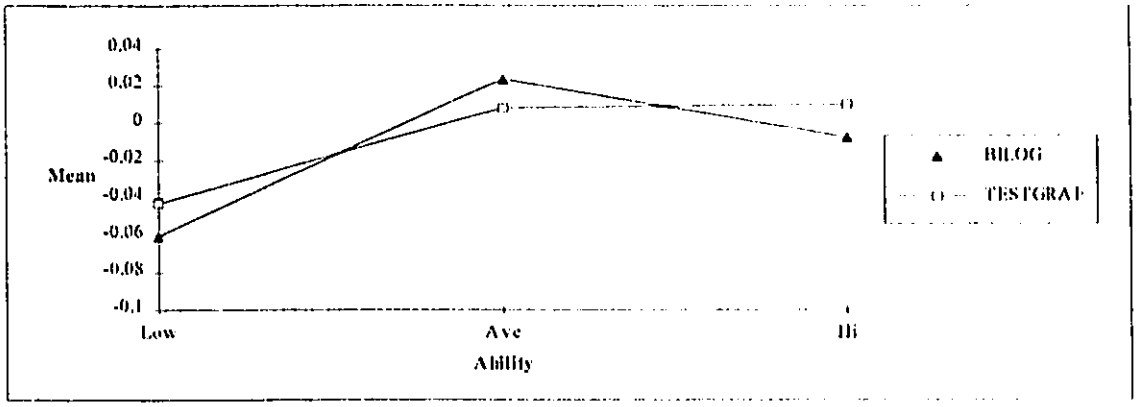
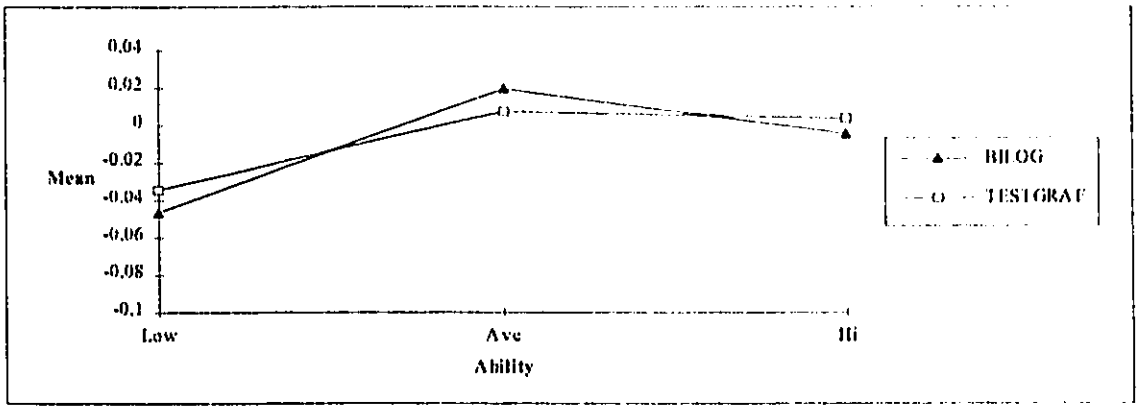


Figure 7d. Interaction of Ability by Procedure (N=1000) on the Bias of the $P(\theta)$ Estimates.



There were no large bias ESs for the test length by procedure (LxP), the sample size by procedure (SxP), the test length by sample size by procedure (LxSxP), the test length by ability level by procedure (LxAxP), and the test length by sample size by ability level by procedure (LxSxAxP) interactions in estimating the $P(\theta)$'s. This suggests that: i) test length did not largely affect the bias of the $P(\theta)$ estimates differently for TESTGRAF and BILOG, ii) sample size did not largely affect the bias of the $P(\theta)$ estimates differently for TESTGRAF and BILOG, iii) there was no difference between the SxP interactions for the 20-item and 40-item tests in the bias of the $P(\theta)$ estimates, iv) there was no large difference between the AxP interactions for the 20-item and 40-item tests in the bias of the $P(\theta)$ estimates, and v) there was no large difference between the SxAxP interactions for the 20- and 40-item tests in the bias of the $P(\theta)$ estimates.

Efficiency. By examining Table 10, it is apparent that the none of the effects had large efficiency ESs. This suggests that test length, sample size, and ability level did not largely affect the efficiency of the $P(\theta)$ estimates differently for TESTGRAF and BILOG.

Summary

In this section, the results are summarized according to the research questions. In accordance with the research questions, accuracy and consistency are used in place of their statistical measures, bias and efficiency.

1 a) There was a large effect of **test length** on the accuracy of TESTGRAF and BILOG in estimating the *c* parameter. TESTGRAF was more accurate than BILOG in estimating

the *c* parameter at both test lengths. There was no large difference between the accuracy or consistency of TESTGRAF and BILOG in estimating the *a* or *b* parameters or in the consistency of TESTGRAF and BILOG in estimating the *c* parameter for different test lengths.

b) There were no large effects of **test length** on the accuracy or consistency of TESTGRAF and BILOG in estimating the $P(\theta)$'s for different ability levels.

2 a) There was a large effect of **sample size** on the accuracy and consistency of TESTGRAF and BILOG in estimating the *a* and *c* parameters. TESTGRAF was more accurate than BILOG in estimating the *a* and *c* parameters and both were more accurate as sample size increased. TESTGRAF was more consistent than BILOG at sample sizes of 100 and 250 in estimating the *a* parameter and BILOG was more consistent than TESTGRAF in estimating the *c* parameter at all sample sizes. There was no large difference between the accuracy or consistency of TESTGRAF and BILOG in estimating the *b* parameter for different sample sizes.

b) There was a large effect of **sample size** on the accuracy of TESTGRAF and BILOG in estimating the $P(\theta)$'s at different ability levels. As sample size increased, the difference between the two procedures decreased. At all sample sizes, TESTGRAF was more accurate than BILOG at the low and average ability levels. There were no large differences between the accuracy or consistency of TESTGRAF and BILOG in estimating the $P(\theta)$'s for low and average ability levels.

3 a) There were no large effects of the interaction of **test length** ***and*** **sample size** on the accuracy or consistency of TESTGRAF and BILOG in estimating the item parameters.

b) There were no large effects of the interaction of **test length** ***and*** **sample size** on the accuracy or consistency of TESTGRAF and BILOG in estimating the $P(\theta)$'s at different ability levels.

Recall that all findings with regard to the c estimates are based on data which were simulated based on item parameters with many c's equal to zero.

CHAPTER V

Discussion

In this section, the results obtained from this study are compared to results obtained from other studies found in the literature. First, results of the effects of sample size and test length on the accuracy and consistency of TESTGRAF and BILOG in estimating item parameters are compared to results obtained from other similar studies. Secondly, the results of these effects in estimating $P(\theta)$'s are compared to results obtained from other studies found in the literature.

Item Parameters

Consistent with Hulin, Lissak, and Drasgow (1982), Skaggs and Stevenson (1989), Swaminathan and Gifford (1983), and Wingersky and Lord (1984), the findings in this study indicate that the accuracy and consistency of both TESTGRAF and BILOG item parameter estimates increased with increased sample sizes and test lengths. In particular, consistent with Swaminathan and Gifford (1983), the findings indicate that increased sample size and increased test length, both, independently, had slight effects in improving the accuracy of the estimation of the *b* and *c* parameters and a large effect in improving the accuracy of the estimation of the *a* parameter. This is an important finding, since the estimation of the *a* parameter is very important in test design and development. The accuracy of the estimation of the *c* parameter is slightly

improved with TESTGRAF as compared to BILOG. However, even with TESTGRAF, it is still underestimated, although, not as much as it is with BILOG. The consistency of the estimation of the *c* parameter is not improved with TESTGRAF as compared to BILOG. Over varying sample sizes, BILOG estimated the *c* parameter more consistently than TESTGRAF.

P(θ)'s

Consistent with Hulin *et. al.*(1982), the findings in this study indicate that the accuracy and consistency of both TESTGRAF and BILOG *P*(θ) estimates increased with increased sample size and test length. However, the problem still remains with the estimation of the *P*(θ) in the lower ability level with small sample sizes. Over varying sample sizes and test lengths, TESTGRAF estimated the 3PL *P*(θ)'s more accurately than BILOG in the low and average ability ranges, and BILOG estimated the *P*(θ)'s more accurately or just as accurately as TESTGRAF in the high ability range. Therefore, when there is interest in estimating *P*(θ)'s based on examinees with high abilities and samples less than 500, it would be best to use BILOG.

CHAPTER VI

Summary and Conclusion

To date no studies in the literature were found where the performance of TESTGRAF was examined or compared it to any other leading program in the field. The discrepancies found between TESTGRAF and BILOG contribute to our knowledge of both programs and their usefulness in various practical situations. Such understanding can lead to a wider use of IRT methods, through the use of TESTGRAF, among educators who develop short tests and who are faced with small sample sizes. In general, the findings from this study indicate that TESTGRAF yields more accurate item parameter and $P(\theta)$ estimates in the low and average ability ranges than does BILOG. Only when there is interest in obtaining: i) consistent a estimates with a sample size of 1000, ii) obtaining consistent c estimates with any sample size, or iii) 3PL $P(0)$ estimates based on examinees with high abilities and samples less than 500; would it be better to use BILOG. Otherwise, it would be best to use TESTGRAF regardless of the item parameter estimate of interest or ability level of candidates.

Three limitations of this study are that simulated data were used, many of the items which were used to simulate the data had c parameters equal to zero, and the default options of both programs were used. Future research could compare the two programs using real test data or simulated data based on item parameters with non-zero c values. However, before further analysing simulated data or analysing real data, it would be interesting to analyse simulated data

and manipulate the various options in both programs. For example, one may set prior distributions on the item parameter estimates in BILOG.

References

Ackerman, T. (1985). M2PL Data Generation Program [Computer Program].

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, 11, 111-141.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Buhr, D. C., & Algina, J. (1986). A comparison of item parameter estimates and of ability parameter estimates obtained by different methods implemented by BILOG. (ERIC Document Reproduction Service No. ED 264 267).

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Hambleton, R, Swaminathan, H., & Rogers, J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications.

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, 249-260.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 40, 205-217.

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (RB-75-33). Princeton, N. J.: Educational Testing Service.

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. Psychometrika, 48, 425-435.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Item analysis and test scoring with binary logistic models [Computer Program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1986). PC-BILOG: Maximum likelihood item analysis and test scoring with logistic models. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Qualls, A. L., & Ansley, T. N. (1985, April). A comparison of item and ability parameter estimates derived from LOGIST and BILOG. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Ramsay, J. O. (1989). TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test Data [Computer Program]. Montreal: McGill University.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika, 56, 611-630.

Ramsay, J. O. (1993). TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test Data [Computer Program]. Montreal: McGill University.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research. [Expanded edition University of Chicago Press, 1980.]

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-Bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model. Applied Psychological Measurement, 13, 391-402.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), New Horizons in Testing (pp. 13-30). Toronto: Academic Press.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.). Applications of Item Response Theory (pp. 45-56). British Columbia: Educational Research Institute of British Columbia.

Wingersky, M. S., & Lord, F. M. (1973). A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses (RM-73-2). Princeton NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Computational Psychometrics, 52, 275-291.