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 Stability of Item Parameter Estimates in the Context of a Second Language Competency Test

Doreen Ready

Thesis submitted to the School of Graduate Studies
and Research in partial fulfillment of the requirements for the
M.A. degree in Education

University of Ottawa

ISBN 0-612-04988-4

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Acknowledgements

I would like to thank Dr. Marvin Boss for his encouragement to undertake this project as an M.A. thesis. It has been the best of times and the worst of times but always worthwhile with Dr. Boss's support.

I would also like to thank Dr. Marc Gessaroli and Dr. Bruno Zumbo for their interest and suggestions and when all else failed for their sense of humour. I am also grateful to Beatrice Magyar for her assistance and patience with yet another round of revisions.

Table of Contents

List of Tables

Abstract

The purpose of this study was to examine the feasibility of using IRT models to equate test versions of an English second language (ESL) test through the use of subsets of linking items. The test was designed to measure global comprehension at the intermediate level, of listening and reading texts and general knowledge of vocabulary, grammar and structure as measured by a cloze text. The data were subsets of listening, reading and cloze items used on two or more occasions. The procedure used was to ascertain the stability of the item parameter estimates from correlations and other descriptive measures. The estimates were obtained using three estimation methods (LOGIST, BILOG, NoHarm) and three IRT models. In addition the unidimensionality of the data sets was examined using a method described by Gessaroli and De Champlain (1991). The results suggest that IRT models may not be suitable with test data such as those used in this study. Failure could not be consistently linked to too few candidates, issues of statistical dimensionality, too few items, or the difficulty of the material in which the target subsets were embedded. If an expanded study yields similar results, then the issue is one of validity, whether the definition of ESL global comprehension at the intermediate level is compatible with how an ability is defined under an IRT model.

Chapter I — Introduction

Item response theory models provide advantages to the test developer and test user that cannot be matched by classical test theory. The most important advantage "is that given a set of test items that have been fitted to a latent trait model (that is, item parameters are known), it is possible to estimate an examinee's ability on the same ability scale from any subset of items in the domain of items fitted to the model" (Hambleton & Cook, 1977, p. 90). Thus, ability estimation is independent of the items chosen for the estimation. The obverse of this is that item parameter estimations are invariant across sub-groups of examinees. A further advantage is that latent trait models provide a measure of the precision of the ability estimate at each ability level unlike the single standard error of measurement provided by classical test theory. Thus, provided that the test data can be fit to an IRT model, IRT would appear to be the better option for the test developer and test user.

As is well known, there are generally three IRT models available (see, for example, Hambleton & Cook, 1977) for the estimation of item and ability parameters when the test consists of dichotomously scored binary data that is also unidimensional. All three models provide an estimate of a candidate's ability. The difference lies in the number of item parameters estimated. The 1PL model only provides an estimate of b, a measure of item difficulty. The 2PL model provides both b and a, a measure of the item's discriminating power. The 3PL model provides, in addition to a and b, an estimate of c, the lower asymptote of the Item Characteristic Curve (ICC) which is also known as the pseudo-guessing parameter.

The advantages outlined above also offer potential solutions to two problems that have always been somewhat contentious. These are the development of parallel forms of a test and the equating of ability estimates obtained from different tests measuring the same ability.

The resolution of both of these problems in the context of second language testing is of particular importance in this study. Since Henning (1984) endorsed the use of IRT models for second language testing applications, these models have become increasingly important to practitioners in the field. This is particularly true of test developers faced with the problem of creating parallel versions of second language tests both for purposes of placement into second language programs and for purposes of certification of a desired (or necessary) level of second language ability. The problem has been complicated by the fact that in many instances only two

or three hundred candidates write a particular form of the test. This has lead to a heavy reliance on the one parameter (1PL) or Rasch model which has been accepted for the most part with very few reservations as to the limitations of this model or of IRT models in general. The objective of this research is to determine whether it is appropriate to use IRT models with data obtained from a particular second language test by verifying the stability of the parameter estimates obtained for the same test items on two or more occasions. In the next chapter a review of the literature relevant to this research is presented.

Chapter II — Review of the Literature

Researchers tend to approach the problem of the limitations of IRT models from two different perspectives. First, given no violation of any of the underlying assumptions (both in general and for particular models) what are the practical constraints which may still affect the stability of the estimates? These include: 1) the effect of test length and sample size; 2) the effect of the ability distribution of the population and 3) the effect of item context. Second, given the appropriate practical conditions for stability of the estimates, what are the theoretical constraints which may affect the stability of the estimates? These include: 1) the effect of violation of the general assumptions of the models (unidimensionality and local independence); 2) the effect of violation of some/all the assumptions of particular models; and 3) the effect of the estimation procedure used.

If consensus amongst researchers can be used as an indicator of resolution of the issues, then it is evident that practical concerns have been more easily addressed than violation of theoretical assumptions.

One of the reasons why there is no consensus on some issues is that most of the studies are not easily generalizable. Researchers sometimes work with computer generated data and sometimes with real test data. They use different models with a definite tendency to use either the 1PL or 3PL model. This is partly because of the perceived value of these two models. The 1PL model, because it only estimates one item parameter, can be used with smaller sample sizes. The 3PL model requires large sample sizes, but it has the advantage of estimating a pseudo-guessing parameter, a desirable characteristic when working with multiple choice data where guessing may very well be a problem. In fact, a great deal of the research using this model has been undertaken by researchers employed by commercial organizations doing large scale testing. A third factor which makes generalizability of studies difficult is that researchers use a variety of significance tests in establishing the appropriacy of their conclusions. Under these circumstances it is sometimes only possible to present all the views without reaching any firm conclusions regarding particular issues.

In what follows, the six factors identified above as being relevant to the issue of item and ability estimation stability are addressed separately. Where possible, an attempt is made to either arrive at a consensus or to summarize conflicting views.

Test Length and Number of Subjects Required

This is an issue on which most researchers are now in agreement. It is well established that the number of subjects required increases with the number of parameters being estimated. Hambleton, Swaminathan, and Rogers (1991) suggest approximately 200 subjects for the 1PL model, 500 for the 2PL model, and 1,000 or more for the 3PL model, although as Hulin, Drasgow, and Parsons (1983) have pointed out there are no definitive rules because numbers vary according to the purpose of the study (p. 105). For example, the initial item calibration of long tests (60 items) may not require large numbers of examinees. Hulin, Lissak, and Drasgow (1982) report with 60 items, 200 examinees yielded good correlations between Θ and $\hat{\theta}$ for both the 2PL and 3PL models.

An idea which appears to have gained validity in practice is that when using the 1PL or Rasch model, it is permissible to use fewer than 100 subjects in test development work. It is difficult to know how this idea became so firmly entrenched, especially in the area of second language testing, because even those researchers who are strong proponents of the model are in general agreement about the number of subjects required. In fact there is only one study which dealt specifically with less than 100 examinees (Lord, 1983a). In this study Lord makes a comparison of the 1PL and 2PL models in terms of which model yielded a smaller sampling variance than the number correct score. His conclusion was that the 1PL was slightly superior to the 2PL in terms of the examinee's true score when sample size was less than 100 or 200. In fact, this study is extremely limited since the issue of parameter estimation is not addressed at all.

All other evidence with regard to the number of examinees required with the Rasch model points to the fact that a minimum of 200 is required (Wright, 1977; Forster, 1978; Whitely & Dawis, 1974) for any kind of stable estimation. Gustaffson (1980) says that if using the 1PL model where the assumptions of no guessing and $a = 1$ have been violated that a minimum of 10,000 responses are required for accurate assessment of invariance. For example,

given a 50-item test, it would be necessary to have a minimum of 200 subjects to meet the requirement of 10,000 responses.

Test length is an issue which has been addressed from two perspectives. What is the minimal test length required for a test to give stable parameter estimates and what is the minimal length required to serve as an anchor test for test equating purposes? In this latter context, a subset of items is embedded in two tests where all the other items are different. This subset can be used subsequently to equate the other items to each other.

As with the number of subjects required (Hulin, Drasgow, & Parsons, 1983) the answer as to how many items are enough for stable estimates depends on the purpose. Hambleton and Cook (1983), using simulated data and the 3PL model, concluded that a minimum of 20 items and 200 subjects were required for ability estimates if the researcher's primary concern is with values in the middle region of the ability range. Swaminathan and Gifford (1979) using the 3PL model found that the number required depended upon the parameter being estimated. In general, correlations between a and â improved with increasing test length and sample size ($r_{a\hat{a}}$ = -.02 for a 15 item test (N (number of subjects) = 50); $r_{a\hat{a}}$ = .88 for an 80 item test (N = 1000). They also noted that the shape of the ability distribution, an issue to be addressed in the next section, also affected the precision of the estimates. The requirements for stability of the b estimates were less stringent. A 10-item test and 50 examinees yielded a correlation between b and $\hat{b}$ of .80. Hulin, Lissak, and Drasgow (1982) concluded that if the aim is accuracy of the estimates, large numbers of test items are not required but relatively large numbers of subjects are. But, as an example, they still cite a minimum test length of 30 items with 500 examinees for the 2PL model.

Researchers investigating the effect of test length for anchor tests used for test equating purposes suggest far fewer than 20 items. An anchor test is a subset of items embedded in two different tests which permits equivalent scaling of all the items on the two tests. Hambleton, Swaminathan, and Rogers (1991) give examples of anchor tests less than 10 items long. Wingersky and Lord (1984) using a 3PL model suggest five items although this was in the context of a procedure which estimated all the item parameters on both tests concurrently. It would seem that 10 or fewer items might be enough although evidently the researcher would have to establish the appropriacy with great care.

In general there is agreement amongst researchers on the issue although there are no absolute rules. It is still at the discretion of the individual to determine whether numbers may be varied depending upon use and circumstances, provided that adequate checks of estimate invariance are carried out after each analysis.

Ability Distribution of the Subject Population

Several researchers have noted an effect of the shape of the ability distribution on the stability of parameter estimation. Wingersky and Lord (1984) using the 3PL model and a 45-item second language test noted that the standard errors of the estimates were better with either a rectangular or bimodal distribution then they were with a normal distribution. In fact the authors state that the standard errors produced using a rectangular distribution were nearly as small as those produced using a bell shaped distribution of abilities with a sample size almost four times as large. Evidently one of the advantages of a rectangular distribution is that there are enough responses at the extremes of the ability range for more accurate parameter estimation of extremes of ability and of item difficulty. Swaminathan and Gifford (1979) found that a skewed distribution adversely affected the correlation between a and â for a 10-item test regardless of the sample size. The same was true for a uniform (-1.7 to 1.7) ability distribution for 10-, 15-, and 20- item tests at all sample sizes. Stocking (1988) looked at information functions as a means of finding the optimum ability ranges for estimating item parameters in the 1PL, 2PL, and 3PL models. Although she does not claim that this is the same as investigating the accuracy of the parameter estimates, she does draw some conclusions to aid the test practitioner. In the case of the 1PL model, if the range of item difficulty is thought to be broad, then the ability range of the population should be equally broad. The 2PL model requires an even broader range of abilities in the test population than the 1PL model because examinees above and below b are required to measure a. For the 3PL model only abilities well below b are useful in estimating c. Her conclusion is that better results may be obtained by sampling all levels of ability equally; on this she is in agreement with Wingersky and Lord.

The other concern investigated by several researchers is whether the populations from repeated administrations of a test contribute to parameter instability. Cook, Eignor, and Petersen (1982) investigated temporal stability with a view to the feasibility of creating an item bank for the SAT. In their literature review they cited Divgi (1981) who had noted that the assumption

of stability was probably satisfied well enough for applications such as the equating of intact tests but expressed reservations about the validity of the assumptions when dealing with individual items as would be the case for an item bank. The authors analyzed both subsets of items and several tests from different content areas, SAT verbal and mathematics items and achievement tests in biology, American history, and social studies, using the 3PL model. They chose versions of the tests that differed in length of time between administrations. The statistics used included correlations of the parameter estimates obtained at the two administrations, scatter plots of the two sets of difficulty parameters and the two sets of discrimination parameters, descriptive statistics for each test administration, and the mean of the absolute difference between the item response functions calculated for the two administrations of each test (this provided a descriptive statistic for a comparison of the relative difference for the two administrations for all the different tests). They found the greatest stability in SAT mathematics items and the least in the achievement tests especially history and social studies. This finding is most probably related to the fact that the content domain for tests of subjects like history is so broad that very few examinees would have studied identical material. Their general conclusion was that stability of the estimates was related to the type of test and differences in ability and not at all to the length of the time lapse between administrations.

Cook, Eignor, and Taft (1984) looked at the problem of stability from the view point of the effect of curriculum on the estimates. As they state in their rationale for the study, one of the problems facing the test practitioner is that of defining the population from which the samples are drawn. Their data consisted of test responses to a biology achievement test collected at one spring and two fall administrations which were analyzed using both the 3PL model and classical test methods. The results were very convincing. They concluded that the properties of a curriculum-bound achievement test such as they had used clearly depended upon when in the student's course of study the student elected to take the test.

Kingston and Dorans (1982) working with sections of the GRE found that the instability was particularly apparent with regard to estimates of the c parameter.

The evidence appears to be quite clear that content tests are especially susceptible to instability over time because of the effects either of recency of instruction or of changing

emphasis in curriculum. Whether this might also be true of second language tests has not been investigated.

Thus, it appears that test developers need to carefully consider the characteristics of the sample in terms of the range of proportion of abilities included and the similarity of the ability distribution from one administration to another.

The Effect of Item Context

As item response theory models became more widely used, several studies were undertaken with regard to the stability of parameter estimates for conditions involving item context. Since the evidence was quite convincing, there is now general agreement on the conditions requiring caution on the part of the researcher and no new studies have been undertaken.

One of the earliest studies on contextual effects was that of Whitely and Dawis (1976) using the 1PL (Rasch) model. They looked at the parameter estimates of 15 anchor items embedded in seven tests of varying degrees of difficulty and found that the b value obtained for the majority of the 15 items changed depending upon whether they were embedded in an "easy" or "difficult" test form. They concluded that the difficulty of the test context, although influencing some individual item difficulties, was not sufficient to account for the observed pattern of differences so that some other unidentified factor(s) also influenced the stability of the estimates.

Yen (1980) using the 1PL and 3PL models tried to identify what those factors might be. As in the study of Whitely and Dawis contextual effects were observed in the stability of the b parameters for both the 1PL and 3PL models and, in addition, even stronger effects were noted on the slope estimates of the 3PL model. In fact, the context effects were stronger for the 3PL model estimates in general, but that is not unexpected considering that sample sizes were relatively small (number of examinees in the 300 range). Yen looked at four factors: 1) inclusion of extraneous items; 2) difference in the number of items scaled; 3) systematic differences in the sequence in which the items appeared; and 4) unspecified other factors. The only factors which could be linked to the stability of the estimates were differences in the

number of items scaled (very slight effect on slope estimates) and the effect of sequence (only partially consistent). Evidently the factors at work are not easily identified but no further studies have been published in this area.

The only factor definitively identified as affecting the stability of the parameter estimates is that of the level of difficulty of the test in which the target items are embedded. There may be some effect due to the number of items scaled and the sequence in which the items appear, but the evidence is not strong. Regardless of whether all the factors can be identified, the evidence of contextual effects is quite convincing. It would seem prudent to keep this in mind, especially in the context of using anchor items for test equating and item banking. It reinforces the need for enough anchor items to be able to establish the presence of sufficiently strong correlations between the parameter estimates from different test administrations so that the test developer can feel confident in proceeding to the next step.

Violation of the Assumptions of Unidimensionality and Local Independence

The general understanding of the relationship between unidimensionality and local independence is that given unidimensionality, the assumption of local independence holds (Lord & Novick, 1968; Gustaffson, 1980; Bejar, 1983). Goldstein (1980) dissents from this view in that he feels there is little empirical evidence to support this idea.

There are two main areas of investigation bearing on the assumption of unidimensionality. One is the robustness of the estimates in the face of varying degrees of violation of the assumption and the other is the definition of unidimensionality itself and how to test it.

Reckase (1979) did a pioneering study in investigating the effect on parameter estimations of varying degrees of multidimensionality using both the 1PL and 3PL models. The data consisted of nine tests of varying factorial complexity. Five of the tests were real data obtained from course examinations and the Missouri Statewide Testing Program (all were factor analyzed for use in this study). Four others were computer generated to simulate different factor configurations and comprised a one-factor test, a two-factor test, a nine-factor test in which the factors were correlated and a nine-factor test in which the factors were not correlated. All tests

were 50 items long. Reckase first correlated the parameter estimates and fit statistics with the theoretical and empirical loadings from the factor analyses. Then he compared the item statistics to the strength of the first factor using regression techniques. Third, he looked at the relationships between the results of the 1PL model, the 3PL model, classical item analysis, and factor analysis. To do this he used principal component analysis on the item statistics. In the case of what factor was being measured by the two models when the data consisted of more than one independent factor, he concluded that the 3PL model estimated the dominant factor while the 1PL model estimated the sum of the factors. Where the first factor was large relative to the others, both models measured the first factor. How large should the first factor be in order to obtain stable estimates? Reckase said that both models were similar in that the first factor should account for at least twenty percent of the variance. Below twenty percent reasonable estimates of ability might be expected but item parameters would be unstable. It should be noted that stability in his terms does not refer to the stability of the estimates over test administrations or of stability in comparison with a set of generated (therefore known) values. Rather it is the stability of the parameter estimates in relation to which factor(s) will influence what is estimated by the IRT model.

Drasgow and Parsons (1983) further investigated the effects of multidimensionality. While Reckase had looked at data with an underlying dominant trait and several weaker latent traits that affected clusters of items, they looked at data sets where the underlying trait became incrementally less dominant or "prepotent" as they termed it. Their investigation was carried out using both the 2PL and 3PL models. They used simulated data consisting of 50 items and five levels of dominance of the latent trait over the various subscales comprising the test. They then looked at the correlation between the estimated ability parameters and the subjects' known score on the latent trait. The correlations between the estimated and known values for the first two data sets were .96 and .94. The first data set had been generated to be unidimensional and the second data set had intercorrelations between pairs of subtests that ranged from .68 to .94. For the third data set, the correlation of the estimated and known values was .84 and the intercorrelations between pairs of subtests ranged from .46 to .60. The fourth data set correlation of estimated and known ability values was .74 (intercorrelations between pairs of subtests ranged from .25 to .39). Drasgow and Parsons concluded that at this point it would no longer be appropriate to apply an IRT model. It should be noted that they only examined ability estimates and not item parameter estimates.

Kingston and Dorans (1982) also investigated the effects of non-unidimensionality on the stability of parameter estimation using the 3PL model. The study was done in the context of comparing three test equating methods, two based on classical test methods and one based on IRT. They used the verbal section of the GRE which, through factor analysis, they had shown represented two factors, discrete verbal items and reading comprehension items. In their summary discussion they reported that the discrimination parameter estimates of discrete verbal items were more stable in a homogeneous context (discrete verbal items only) than when obtained in a heterogeneous context (discrete verbal and reading comprehension together). The b estimates were somewhat less sensitive to the effects of heterogeneous item sets, and compared to a and b, the c estimates were least stable under these conditions. Unfortunately since the authors were preoccupied with comparisons among the equating methods, they have not provided detailed correlations of the estimates obtained under the two conditions so it is difficult to judge the importance of these effects in numerical terms in the context of the study.

Thus, although researchers are in agreement that multidimensionality is a possible threat to the stability of parameter estimation and affects some estimates more than others, there is not enough evidence to provide concrete guidelines across different conditions.

In general, their work points to the fact that there can be only one dominant dimension if one of the IRT models is to be successfully applied. This work leads naturally into the idea of essential dimensionality as opposed to the idea of traditional dimensionality where only one trait is present (Stout, 1990). The definition of essential unidimensionality allows for the possible presence of minor dimensions along with the one dominant dimension.

Work has progressed so that a statistical method to assess the dimensionality of test data in accordance with the definition of essential dimensionality is now available. Several studies have been reported (Stout, 1987; Nandakumar, 1987, 1991) in which this statistical method has been used with different data sets (both computer generated and real) with some success. Stout's concept of essential dimensionality is similar to McDonald's (1991) definition of unidimensionality as the weak principle of local independence. The weak principle of local independence is based on the correlational version of the principle of local independence and states that the covariance $(u_i\, u_j\, /\Theta) = 0$. Stout states that the covariance $(u_i\, u_j\, /\Theta) \cong 0$ as the number of items approaches infinity. In practice these two definitions of unidimensionality are

indistinguishable. A promising statistic also now exists to test unidimensionality according to McDonald's definition (Gessaroli & De Champlain, 1991). Thus, recent work has permitted both a slightly less stringent definition of unidimensionality and statistical methods to assess fit to the unidimensional assumption.

Evidently, for each data set the assumption of unidimensionality should be tested and any violation needs to be investigated thoroughly to assess the extent of the effect on item parameter estimation. Indeed, Bejar (1983) would go so far as to say that the issue should be addressed each time a test is administered for his feeling is that in practice dimensionality is situation specific.

Model Fit

The choice of which model is the most appropriate for a given set of data is probably one of the most difficult decisions facing the potential user. According to Traub (1983) it is impossible to satisfactorily model test data obtained from dichotomously scored responses to multiple choice items testing educational achievement. One of the fundamental reasons for this is "that there exists no basis in inductive logic for concluding that a model fits data" (p. 57) which he feels applies to all IRT models.

For many test developers, the choice of model must be to some extent guided by the assumptions implicit in the models. An assumption implicit in both the 1PL and 2PL models is that there is little or no guessing in candidate responses since neither model estimates c. An additional assumption for the 1PL model is that the discriminating power of the items is equal and thus items vary only in terms of difficulty.

From an examination of the literature it is evident that there are three main areas of interest. These are the problems related to the accuracy or appropriacy of the various measures of goodness of fit, the comparison of the fit of two or more models to particular data sets, and the preoccupation with attempting to define the strengths and limitations of the 1PL or Rasch model, the area of greatest controversy.

It has been evident for some time (Traub, 1983; Hambleton & Murray, 1983) that there are problems with statistics developed to measure both overall goodness of fit and item misfit. This has to do with the fact that the power of these test statistics increases with sample size. "Thus a model is almost certain to be rejected for ill-fit if the sample size is sufficiently large" (Traub, p. 57). Traub and Lam (1985) also point out that the other problem associated with overall goodness of fit statistics is that the sampling distributions for samples of realistic size are unknown. Their solution to the problem is to work with nested models and thus be able to compare the statistics obtained fitting different models to the same data. This also provides a rational basis for comparing the fit of one model with that of another. Hambleton and Rovinelli (1973) did a study on detecting misfitting items with increasing sample size. Again, as with statistics of overall fit, the number of misfitting items increased substantially indicating that with sufficient sample size the rejection rate of misfitting items becomes an artifact of the method.

If this is not a promising route of investigation, how can the test practitioner arrive at the decision of which model to choose? Hambleton and Murray (1983, p. 72) summarize the three possible approaches that are available:

a) determine if the test data satisfy the assumptions of the model;
b) determine if the expected advantages derived from the use of the item response model (for example, invariant item and ability estimates) are obtained;
c) determine the closeness of fit between predictions and observable outcomes (for example, test score distributions) utilizing model parameter estimates and the test data.

In fact, the authors conclude that the best approach incorporates analyses that include all three strategies. They also emphasize that test practitioners should treat statistical indices with caution. They stress the value of comparing the residuals derived from fitting two or more models as being a particularly useful tool.

This last point leads naturally into the second area of interest to researchers, the comparison of models using various statistical means to assess their efficacy in providing best fit (Hambleton & Murray, 1983; Way & Reese, 1991; Yen, 1981; Hambleton & Traub, 1971, 1973). In general the studies tend to show that the 3PL model provides better fit at the item

level than the 2PL or 1PL models unless the data are simulated to fit these latter models. In a comparison of the 1PL, 2PL, and 3PL models, Hambleton and Traub (1971) concluded that even when guessing was a factor, if only high-ability examinees are of interest the 2PL model still provides acceptable ability estimates compared to the 3PL model. The 1PL model, on the other hand, remains efficient only in the presence of little or no guessing and a limited range of discrimination parameters. Hambleton and Traub (1973) report further that the 2PL model will provide the greatest improvements over the 1PL model when applied to data from short tests where the variability of the discrimination parameters is considerable. It is interesting to note that in spite of the fact that researchers have noted the superiority of the 2PL model over the 1PL model in their findings, the 1PL model is still the preferred model even when sufficient sample size for good estimation with the 2PL model is not an issue.

Part of the preference for the 1PL model (Divgi, 1986, p. 284) is that it has some special properties that the more general models lack:

1. Estimation of ability is convenient. The maximum likelihood estimate (MLE) of ability for each score can be computed and stored. Then it is only necessary to look up the estimate corresponding to each examinee's score.
2. Conditional Maximum Likelihood (CML) estimates of item parameters are consistent. Unconditional Maximum Likelihood (UCON) estimates are not. CML estimates cannot be obtained with 2P and 3P models.
3. It is possible to compare difficulties of two items, or abilities of two persons, without having to estimate any other parameters, difficulties or abilities (Rasch, 1966). Rasch calls this property "specific objectivity."

These properties and the fact that good estimates can be obtained with smaller sample sizes have lead to a situation where the 1PL model is the most frequently used model for smaller scale testing. Consequently there is an extensive body of literature devoted to the advantages and limitations of the 1PL Rasch model.

The advantages have been stated previously; the limitations are those imposed by the assumptions for using the model. These are that there will be little or no guessing on the part of examinees and that all items have equal discriminating power. Researchers address the

violation of assumptions in two ways. Wright and Stone (1979) suggest that both misfitting persons and misfitting items should be removed from the analysis thus preserving the integrity of the assumptions. Other researchers (Hambleton & Traub, 1971, 1973; Dinero & Haertel, 1977; van de Vijver, 1986) have examined the robustness of the estimates given varying degrees of violation of the assumptions. In this way they hope to provide guidelines to test practitioners in determining the seriousness of any violation of the assumption in their data. Both of these approaches have their critics.

The critics of the approach of removing misfitting items and persons (Goldstein, 1979; Whitely, 1977; Gustaffson, 1980; Lumsden, 1980; Traub, 1983; Divgi, 1986) take the position that this is an artificial solution to the problem. These authors take issue with Wright's (1977) assertion that if data do not fit the model it is because the data are not suited for any kind of measurement. Goldstein (1979) makes the point that by fitting data to the model and not the model to the data, the test developer is allowing the model to dictate the contents of the test. This may in fact lead to a situation where what is being measured will be restricted to only parts of a curriculum and will not meet the required educational objectives. Gustaffson (1980) also agrees that this is a valid criticism and suggests that items should be subdivided into groups with similar discriminating powers and then analyzed accordingly. Whitely (1977) expresses the same concern as Goldstein in that by adhering to the equal discrimination assumption of the model, test developers will inadvertently restrict the content of the test. Goldstein also envisions that when a test is used on a national scale, differences in teaching emphasis, experiences of the examinees, etc., could lead to different results and call into doubt the conclusions reached based on those different outcomes. He also points out the dangers of creating an item bank using the Rasch model, where over time differences in curriculum emphasis, for example, could change the values of the item parameters considerably.

Other authors (Tall; 1981; Wood, 1978) echo these same concerns on both practical considerations and also on theoretical grounds. Thus, it appears there is a consensus on this issue. Misfit to the Rasch model should not be addressed by removing items or examinees unless very few items are involved.

The approach of investigating robustness of estimates given violation of the assumptions yields results from different researchers that are more difficult to reconcile. There is general

agreement on the robustness of estimates with more than minimal guessing. Several authors (Hambleton & Traub, 1971, 1973; Slinde & Linn, 1979; Gustaffson, 1980; van de Vijver, 1986) agree that more than minimal guessing adversely affects the robustness of the estimates. A consensus of the results of robustness studies when violation of the equal discrimination assumption occurs is not so obvious. Hambleton and Traub (1973) state that estimates remain robust providing the range of discrimination is small. Dinero and Haertel (1977), on the other hand, assert that estimates remain robust for varying item discrimination. Divgi (1986) has criticized their study for its use of a single sample of 75 examinees and because he felt that there was something wrong with the computer programs they used. Gustaffson (1980) also adds that in connection with test equating ability estimates are seriously affected when tests differ in mean discrimination. Boettcher Barnes and Wise (1988) found that the Rasch model was remarkably robust to deviations in homogeneity of item discrimination. This study like most of the others already discussed was done using simulated data but, unlike the others, the authors used a comparison of item characteristic curves (ICC) recovered from the parameter estimates as their criterion for measuring robustness. As Hulin, Drasgow, and Parsons (1983) have pointed out, this is somewhat less stringent than comparisons made using known parameters and parameter estimates. van de Vijver (1986), again using simulated data, also found that estimates were robust with varying item discrimination (0 to 2.0) and his study used comparisons of correlations, standard deviations, bias estimates, and root mean square error estimations. Evidently given simulated data sets, estimates do display robustness given varying departures from uniform item discrimination. It is not possible to form a consensus as to how wide a range of discrimination is permissible nor is it possible to know whether the same results would hold with real data sets.

Estimation Procedures

The effect of the estimation procedure on parameter invariance is an extremely complex area and the most thorough review from which most of this discussion is drawn is that of Baker (1987). The two main areas of discussion are: 1) the efficacy of various estimation procedures when estimating item and ability parameters simultaneously, and 2) the accuracy of the estimates obtained for the three item response models.

According to Baker there is an inextricable link between the estimation procedures and the computer programs used to implement them. Thus, the availability of particular computer programs has influenced what researchers have investigated with regard to the estimation procedures themselves. The programs that have been most widely disseminated and for the longest time are LOGIST (Wingersky & Lord, 1973) and BICAL (Wright & Mead, 1976); both are based on joint maximum likelihood estimation (JMLE). Consequently, the characteristics of JMLE have been thoroughly investigated and in studies where the results obtained from programs using different estimation procedures have been examined, they are usually compared with those obtained from LOGIST. Because of its importance in the literature, JMLE will be discussed first, followed by discussions of the other estimation procedures described in Baker (1987), classical test estimation, conditional and incomplete maximum likelihood estimations, Bayesian estimation, and marginal maximum likelihood estimation.

Although JMLE has been the most widely used over the last twenty years, there are problems associated with the procedure. In fact, several authors besides Baker (Hambleton & Cook, 1977; Swaminathan, 1983; Traub & Lam, 1985) have discussed the fact that under conditions of estimating item and ability parameters simultaneously, there is no guarantee that as the number of examinees is increased that estimates of the item parameters will converge to their true values (the same is true of the ability estimates where the number of items increases).

Researchers of JMLE have examined three pertinent areas: bias of the estimates, the accuracy of recovered parameter estimates, and goodness of fit.

The studies investigating bias are not conclusive since they either represent a restricted set of conditions (Lord, 1983b) or do not provide comparisons of the differences between the parameter and its estimator (Wright & Douglas, 1977). It is still true that where bias was specifically examined there is some evidence that bias exists. Swaminathan and Gifford (1983) did a simulation study examining the consistency and bias of the item parameter estimates under a 3PL model. To examine bias they did 20 replications of the 20-item test, 200 subject condition. They found that there was an overestimate of small values of a and accurate estimates of large values of a. In the case of b values, there was a slight underestimate of negative values of the b parameter and close estimates of large positive values of the b parameter. Lord (1983b) studied item bias using simulated data with $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{\theta}$ parameters roughly equal to those

obtained with 2,995 examinees on a 90-item verbal test. In the case of item difficulty, easy and medium difficulty items had a negative bias and difficult items were positively biased. The bias in the estimates of item discrimination was always positive. The bias in the estimate of the c parameter was negative for all items. In general, he noted that if the item parameter estimate had a large standard error, then the bias was approximately .1 of its standard error and very infrequently was greater than .2 of its standard error. Lord concluded that because standard errors are inversely proportional to sample size, when the number of examinees is large, the numerical value of the bias would probably be negligible. Thus, in these studies bias did not seem to present much of a problem, although with such a small number of studies, it is impossible to reach a definitive conclusion.

Results of studies examining the accuracy of recovered parameter estimates tend to provide support for the idea that to some extent accuracy is model dependent. Thissen and Wainer (1982) investigated the asymptotic standard errors of the item parameters for the 1PL, 2PL, and 3PL models under the assumption that the Θ's of the examinees were known. Tables of the minimum asymptotic standard errors were reported for combinations of parameter values under each of the three models. They presented an interesting set of results for the 2PL and 3PL models when $c = 0$. Even though the numerical values of a and b were the same, the information matrices were not. When an item was easy and had low discrimination the standard errors for the 2PL models were roughly .09 of those reported for the 3PL model. They also found that the asymptotic standard errors for item difficulty under the Rasch model were consistently smaller than those obtained for the other two models.

Lord (1975) used simulated data for a 90-item test and 2,995 examinees and the item parameters were matched approximately to those of an existing test. The correlation between the discrimination estimates and their parameters was .92. When c was overestimated, a also tended to be overestimated. Item difficulty was overestimated for large absolute values of b ($b > 3.0$) and slightly underestimated for medium values of b. Overall the correlation between the difficulty estimates and the parameter values was .98. The estimates of c generally underestimated the parameter values. It appears that accurate estimates are much more difficult to obtain using the 3PL model as compared to the 1PL or 2PL models because of estimation of the c parameter.

Other researchers have also reported difficulty in the estimation of the c parameter. Kolen (1981) found for three tests that 92%, 53%, and 39% of the c parameters were not successfully estimated. Lord (1980) suggested that the c parameter cannot be estimated unless a considerable lower tail of the ICC is present in the range of Θ scores employed. McKinley and Reckase (1980) found that c might be poorly estimated even if this criterion were met. Thus one of the ongoing problems which affects the 3PL parameter estimations is accurate estimation of the c parameter.

Studies have also been done to investigate goodness of fit. The most relevant are those of Yen (1981) and McKinley and Mills (1985). Yen used a 30-item test and 1,000 examinees, and obtained estimations using all 3 models. She found that in general a lack of fit occurred when the model used to make the estimations had fewer parameters than the model for which data were generated. The single exception was that of the 2PL model used to estimate the parameters of simulated data generated for the 3PL model. McKinley and Mills used simulated data generated for the 3PL and 2PL models. They obtained similar results to Yen with respect to lack of fit of models with fewer parameters estimating the parameters of data generated to fit a model with more parameters. They noted the same exception to this rule as Yen, the 2PL model fit the data generated to fit the 3PL model. It appears that JMLE, in general, yields acceptable results with 1PL and 2PL models but the 3PL model presents some problems with regard to the accuracy of the estimates because of the problems associated with estimation of the c parameter.

Other estimation procedures besides JMLE do exist and the number of studies which have been done reflects to some degree when the computer program using the procedure was developed. Most studies have been done comparing the results obtained with particular computer programs with the results obtained with LOGIST.

Urry (1976) was primarily responsible for the development of an estimation procedure linked to classical test theory and for the computer program (ANCILLES) used for the calculations. He investigated the item parameter recovery properties of his program. Using 100 items and samples of 2,000 and 3,000 examinees, he found that the root mean square of the difference between the estimates and the item parameters decreased over the stages of the procedure. The correlations when N = 2,000 were $r_{bb} = .996$, $r_{aa} = .915$ and $r_{cc} = .760$. Ree

(1979) compared the item parameter estimates from both ANCILLES and LOGIST. He used three simulated datasets with 2,000 examinees and 80 items with respect to parameter recovery and found that the two programs yielded similar results. Swaminathan and Gifford (1983) using n (number of items) = 10, 15, 20, N = 50, 200, 1,000 found that LOGIST provided better estimates of item parameters with shorter tests and smaller sample sizes than did ANCILLES while with larger samples and test lengths, the results were comparable. Thus, it appears that for long tests and large sample sizes, this procedure may be a viable alternative to JMLE, especially if cost of computer time is a consideration, for ANCILLES can be run for much less cost than LOGIST.

Conditional maximum likelihood (CML) and incomplete maximum likelihood (ICON) are only feasible with the 1PL model since there is no simultaneous estimation of item and ability parameters. Only the item parameters are estimated; the ability parameters are estimated from the sufficient statistic (the total correct raw score).

The problem inherent in the method as reported by Wright and Douglas (1977) is that because of the accumulation of the round-off errors in the calculation of the symmetric functions only tests of 10 to 15 items can be analyzed. In an attempt to obviate this problem they developed an iterative procedure for evaluating the symmetric functions which they called the incomplete maximum likelihood approach. Even this approach, however, merely delayed the onset of the problem to tests of 20 to 30 items.

Another procedure discussed by Baker (1987) is the Bayesian approach. Swaminathan and Gifford (1982, 1985) have developed procedures for simultaneous estimation of parameters with the 1PL and 2PL models. In general their results suggest that with smaller data sets, there may be some advantage in using the Bayesian approach. With large data sets the values obtained are similar to those obtained using JMLE. There are some problems inherent in this approach; it tends to regress the non-central parameter estimates toward the mean and it requires assumptions about the prior distributions of the parameters which may be difficult to obtain in practice. Lord (1984) examined the characteristics of the Bayesian approach from a statistical point of view. He stated that one cause of the regression of the parameter estimates toward the mean is the use of "high" priors such as those obtained from previous test administrations. As a result a more diffuse prior would be preferable. He also stated that the use of priors has some

practical advantages: 1) infinite estimates of Θ do not occur; 2) item discrimination estimates will not become infinite; and 3) the estimates of c will not come out at implausible values even with easy items. He concluded that priors should probably be used for the a and c parameters as regression toward the mean has less consequence than for the b and Θ values.

ASCAL developed by Vale and Gialluca (1985) for the Assessment Systems Corporation uses a pseudo-Bayesian approach. They suggest that the values obtained using ASCAL are as good as those obtained using LOGIST with 3PL generated data and N = 2000. Evidently more work needs to be done using the Bayesian approach before a definitive statement can be made concerning the relative merits of the procedure in comparison with JMLE.

Marginal maximum likelihood (MMLE) employs a two-stage process. The program developed based on MMLE is BILOG (Mislevy & Bock, 1982, 1984, 1986). The first stage of the estimation is the expectation step in which the provisional values of the item parameters are used to compute for each item the "expected" number of examinees at each quadrative point and the number of these correctly responding to an item (Bock & Aitkin, 1981). The second stage is the maximization (M)-step in which the improved estimates of the item parameters are obtained using a conventional maximum likelihood estimation logit analysis such as that used in the item stage of the JMLE procedure. The remaining feature of the MMLE approach is that the normal prior distribution of Θ can be replaced by some empirically defined distribution of the examinees over the Θ scale. Thus, rather than integrating over the normal density, the quadrative procedures are applied to the empirical distribution. Once the item parameters have been estimated they are considered as known and the examinee's Θ levels can be estimated as a separate process. Yen (1987) using data generated under the 3PL model (N = 1000, n = 10, 20, 40) found that while LOGIST was usually faster than BILOG, BILOG almost always produced more accurate estimates of the individual item parameters. A comparison of the item characteristic functions showed that BILOG made better estimates for the 10-item test and BILOG and LOGIST made comparable estimates on the 20- and 40-item tests. Like Bayesian estimation, the MMLE appears to provide more accurate estimation when using a small number of items, but as yet there are not enough studies to reach a definitive conclusion.

It should be pointed out that there are other procedures not discussed by Baker. Swaminathan (1983) who discusses the same estimation procedures as Baker mentions several

of these. These are the harmonic analysis (computer program NoHarm) approach (McDonald, 1967) empirical Bayes procedures (Meredith & Kearns, 1973) and Jacknife techniques (Wainer & Wright, 1980).

In summary, JMLE has been the estimation procedure which has been studied the most because suitable computer programs have been available for some time. It does have some problems inherent in the procedure and this is most evident with the 3PL model where accurate estimates of c are sometimes difficult to obtain. Both Bayesian and the MMLE procedures look promising especially with regard to shorter length tests, but more research needs to be done in this regard.

Summary of the Findings

It is evident from the above discussion that the decision about whether to choose an IRT model requires a careful examination of the factors which may threaten the invariance of parameter estimation.

Insufficient numbers of items or subjects can affect the stability of the estimates adversely. Hambleton, Swaminathan, and Rogers (1991) suggest approximately 200 subjects for the 1PL model, 500 for the 2PL model, and 1,000 or more for the 3PL model. As Hulin, Drasgow, and Parsons (1983) have pointed out, the number of subjects must be treated flexibly, for the numbers required may vary depending upon the purpose of the analyses. Although the idea of using 100 subjects or fewer with the 1PL model has gained some validity in practice, researchers (Wright, 1977; Forster, 1978; Whitely & Dawis, 1974; Gustaffson, 1980) are in general agreement that a minimum of 200 subjects are required in this context.

As is the case for the number of subjects required, the test length required depends to some extent on the purpose of the analyses. Hambleton and Cook (1983) concluded that a minimum of 20 items and 200 subjects were required to estimate ability parameters using the 3PL model if the researcher's primary concern is with values in the middle region of the ability range. Swaminathan and Gifford (1979) found that the number required depended upon the parameter being estimated. Hulin, Lissak, and Drasgow (1982) concluded that if accuracy of

the estimates is important, large numbers of items are not required but relatively large numbers of subjects are.

In the case of anchor tests, subsets of items embedded in two or more tests for the purpose of test equating, the number of items required can be fewer than that required for a test. Hambleton, Swaminathan, and Rogers (1991) cite examples of anchor tests with less than 10 items. Wingersky and Lord (1984) suggest five items in the context of a procedure for estimating all the item parameters concurrently.

The ability distribution of the subject population can affect the stability of the parameter estimates obtained. Wingersky and Lord (1984) noted that the standard errors of the estimates were better when either a rectangular or bimodal distribution was used instead of a normal distribution. Swaminathan and Gifford (1979) found a skewed distribution adversely affected the a parameter estimates for a 10-item test regardless of sample size and obtained similar results for a uniform (-1.7 to 1.7) ability distribution for 10, 15, and 20-item tests at all sample sizes. Evidently better results are obtained using a distribution that samples all levels of ability equally (Stocking, 1988).

Researchers (Cook, Eignor, & Petersen, 1982; Cook, Eignor, & Taft, 1984) investigating temporal stability of the estimates identified several conditions which adversely affected the parameter estimates. They found that stability of the estimates was related to the type of test (achievement tests especially history and social science were affected), to the ability distribution of the examinees, and in the case of a curriculum-bound test to when in the student's course of study the student chose to take the test.

The investigations into the effect of item context are particularly pertinent to the use of anchor tests as a means of item and test equating because context is difficult to control precisely. Whitely and Dawis (1976) found that the b parameter estimates of a subset of items were affected by whether the items were embedded in an easy or a difficult test. They also concluded that the observed effects could not be completely explained by the level of context difficulty. Some other unidentified factor(s) also influenced the stability of the estimates.

Yen (1980) tried to identify what those factors might be. She was only partially successful in that the number of items scaled and the sequence of item presentation appeared to account for some of the observed instability but no other factors could be linked to these effects. No further work has been done in this area so although there is general agreement that context may affect parameter invariance, there is no complete list of what all the factors might be.

The theoretical consideration which affects all three IRT models is the effect of the violation of the assumptions of unidimensionality and local independence. There is general agreement that given unidimensionality, the assumption of local independence holds (Lord & Novick, 1968; Gustaffson, 1980; Bejar, 1983).

Based on the work of Stout (1987) and Nandakumar (1987, 1991) and the work of Gessaroli and De Champlain (1991), there are now two statistics available for testing for unidimensionality. If the data are multidimensional, then it is up to the potential user to determine whether the parameter estimates obtained are stable enough to permit the use of an IRT model.

Reckase (1979) suggests that if the first factor accounts for at least twenty percent of the variance, the parameter estimates obtained will be stable from the point of view of what factors will be measured. He did not address the issue of stability of the estimates as measured by correlations between estimates obtained at different times.

Drasgow and Parsons (1983) found the correlation between ability estimates varied considerably under conditions of multidimensionality. When the correlations between pairs of subscales on a test were between .25 and .39, they concluded that the application of an IRT model was completely untenable. They did not examine the stability of item parameter estimates. Kingston and Dorans (1982) compared the stability of the item parameter estimates under homogeneous and heterogenous (two-factor) conditions. They found that the b estimates were most stable, a estimates less so, and c estimates were least stable under the two-factor condition. On the evidence presented it is not possible to formulate precise guidelines to aid the potential user.

The choice of which model is the most appropriate for a given set of data is probably one of the most difficult decisions facing the potential user. It must to some extent be guided by the assumptions implicit in the model, but there is evidence to suggest that models are somewhat robust to violation of the assumptions.

There are problems (Traub, 1983; Hambleton & Murray, 1983) with the statistics developed to measure overall goodness of fit and item misfit. One has to do with the fact that the power of the test statistics increases with sample size so that a model is almost certain to be rejected if the sample size is large enough. Hambleton and Rovinelli (1973) found a similar phenomenon with item misfit as sample size increased more items were identified as misfitting.

Traub and Lam (1985) have pointed out the other problem associated with fit statistics, the sampling distributions for samples of realistic size are unknown. They suggest working with nested models as a method of dealing with the problem. Hambleton and Murray (1983) suggest incorporating three approaches to evaluate model fit: determine 1) if the data satisfy the assumptions of the model, 2) if the expected advantages of IRT are obtained, and 3) the closeness of fit between predictions and observable outcomes. They also stress the value of comparing residuals derived from fitting two or more models as being particularly useful.

In general, studies (Hambleton & Murray, 1983; Way & Reese, 1991; Yen, 1981; Hambleton & Traub, 1971, 1973) tend to show that the 3PL model provides better fit at the item level than the 2PL or 1PL models unless the data were simulated to fit these latter models. Yen (1981) and McKinley and Mills (1985) found that the overall fit of the 2PL model to data generated to fit the 3PL model was almost as good as the fit of the 3PL model. There is evidence to suggest that the 2PL is better than the 1PL model, but this is not reflected in the preoccupations of researchers.

The model of greatest interest in terms of the number of studies reported is the 1PL or Rasch model. Partly, this is because of the property of specific objectivity (Rasch, 1966) and partly, because the model can be used successfully with fewer subjects ($\geq$ 200) than the numbers required for the other two models.

Violation of the assumptions of the 1PL model have been addressed in two ways. Wright and Stone (1979) suggest removal of both misfitting items and persons. Other researchers (Goldstein, 1979; Whitely, 1977; Gustaffson, 1980; Lumsden, 1980; Traub, 1983; Divgi, 1986) suggest that it may lead to a restriction on the contents of the test.

Other researchers have examined the effects of violation of the assumptions on the robustness of the estimates. Several authors (Hambleton & Traub, 1971, 1973; Slinde & Linn, 1979; Gustaffson, 1980; van de Vijver, 1986) agree that more than minimal guessing adversely affects the robustness of the estimates. In the case of violation of the assumption of equal discrimination, the evidence is conflicting. Dinero and Haertel (1977), Boettcher Barnes and Wise (1988) and van de Vijver (1986) found that the estimates were robust even with large differences in discrimination values. Hambleton and Traub (1973), Gustaffson (1980) and Divgi (1986) felt that robustness of the estimates was only established when the range of discrimination values was narrow.

The investigation into the effect of the estimation procedure on the stability of the estimates has been somewhat limited to those procedures for which programs were available. Thus, because of the long term availability of LOGIST and BICAL, the JMLE has been the procedure which has been most thoroughly investigated. Most studies using the parameter estimates obtained from programs using other estimation procedures have been done comparing those results with those obtained from LOGIST.

Several authors (Baker, 1987; Hambleton & Cook, 1977; Swaminathan, 1983; Traub & Lam, 1985) have discussed the fact that, with JMLE, under conditions of estimating item and ability parameters simultaneously there is no guarantee that as the number of examinees is increased that estimates of the item parameters will converge to their true values (the same is true of the ability estimates where the number of items increases).

Where bias of the estimates using JMLE has been specifically examined (Swaminathan & Gifford, 1983; Lord, 1983), it has been shown to exist. Results of studies examining the accuracy of recovered parameter estimates tend to support the idea that to some extent accuracy is model dependent. Thissen and Wainer (1982) found that for the 2PL and 3PL models, when $c = 0$, the minimum asymptotic standard errors of the 2PL model tended to be lower than those

obtained using the 3PL model. They also found that these values were smaller for the Rasch model than for the other two models. Lord (1975), Kolen (1981), and McKinley and Reckase (1980) found problems in the accurate estimate of the c parameter.

In goodness of fit studies, Yen (1981) and McKinley and Mills (1985) found that in general lack of fit occurred when the model used to estimate the data had fewer parameters than the model for which the data were generated. The single exception to this was when the 2PL model was used to estimate the data generated to fit the 3PL model where the fit was almost as good as the fit of the 3PL model.

Other estimation procedures for which studies have been done are a procedure linked to classical test theory, conditional and incomplete maximum likelihood, Bayesian estimation and marginal maximum likelihood estimation.

With large data sets, the method linked to classical test theory (Urry, 1976) provides results as accurate as those obtained with LOGIST with the added benefit of being more economical to run in terms of computer time (Urry, 1976; Ree, 1979; Swaminathan & Gifford, 1983). Conditional and incomplete maximum likelihood are only feasible with the 1PL model since there is no simultaneous estimation of item and ability parameters. The problem reported by Wright and Douglas (1977) is that because of the accumulation of round-off errors in the calculation of the symmetric functions, only tests from 20 to 30 items can be analyzed.

Both the Bayesian (Swaminathan & Gifford, 1982, 1985) and the marginal maximum likelihood estimation (Yen, 1987) procedures show promise of providing more accurate estimates with small data sets than those obtained with LOGIST (JMLE). Given the issues in the measurement literature, the question of interest is how these issues have been dealt with in second language testing applications.

Applications of IRT in Language Testing Contexts

IRT models became more widely used in second language test development after Henning (1984) outlined the benefits of using the 1PL model for small scale test development work. In fact, until 1989 the 1PL model was the only one reported in studies appearing in the *Language*

Testing Journal and the focus was not on investigating the limitations of the model but rather on using it as a tool. Chen and Henning (1985) used the 1PL model to investigate item bias. They took two subsamples from the 312 students who wrote the English as a Second Language Placement Examination at UCLA. Their subsamples were 77 students whose native language was Chinese and 34 students whose native language was Spanish. Comparisons of the item response patterns were made using both classical and IRT computations of difficulty. The authors do note that the hypothesis of invariance of populations may not be fully met in their data set, but they still see the IRT results as confirmed because they mirror the results obtained using classical test methods. They also examined the number of misfitting items for each of the two groups, but abandoned their efforts because the number of misfitting items was so small. Given the small number of subjects it is not surprising that this approach failed.

Woods and Baker (1985) wrote a guide to item response theory, and as part of their discussion they illustrated how to use the 1PL model for test equating purposes. Their example provided for three anchor or common items on two tests and they advise, since the b parameter estimates from two administrations will be different, that the test developer take an average of the two values for each of the three items against which all the other items can then be calibrated. An investigation of the stability of the estimates is not considered as part of the process in the context of their discussion.

Theunissen (1986) discussed the feasibility of developing an item bank of texts each with its own set of questions which he calls clusters rather than a bank of individual items. His solution is to treat the clusters of items as individual tests for which a test information function can be calculated based on the sum of the individual item information functions of the item cluster. This method would appear to offer a solution as to how to proceed in creating an item bank for a second language test made up of texts and questions, rather than discrete point items. However, there still may be a dimensionality problem.

De Jong and Glas (1987) discuss the use of the 1PL model to validate second language listening comprehension tests. They examined the response patterns of second language and native speakers to a listening comprehension test in terms of the pattern of item misfit of the two groups. They found that the misfitting items which were the same for both groups had all been judged as items requiring an interpretative understanding of the text. This provided theoretical

evidence which supported the statistical indication of misfit and allowed them to conclude that second language proficiency items needed to be those requiring only a literal understanding of the text. The Rasch model appears to have worked well in this context, but it would have been interesting to know whether classical item analysis techniques would have yielded similar results.

From the papers discussed above, it appears that for researchers in second language test development, the focus of study is not on the limitations of IRT models in this or any other context. Rather, the focus is on using IRT models as a means of resolving the problems associated with second language testing.

Purpose of the Study

The purpose of this study is to examine whether IRT models are appropriate in the context of a particular second language test as measured by whether item parameter estimates are stable for 1) short subsets (i.e., a set of questions pertaining to the same text) of items (between 4 and 7 items) and 2) longer subsets of items (> 27 items). The other issue to be addressed is whether a language test which purports to measure different skills (listening, reading) is unidimensional. Hambleton, Dirir, and Des Brisay (1993) suggest that an ESL test measuring different skills can be unidimensional. The data in this study comprise subsets of items from English as a second language (ESL) tests. Each subset has the potential to serve as a set of anchor items for equating all the items on the two tests provided the item parameter estimates are stable enough. The research questions which arise in the context of this study are the following:

1) Are second language tests measuring more than one skill unidimensional?
2) Are item parameter estimates for a subset of items affected by the other subsets of items in the test? It is expected that differences as measured by the level of difficulty in the subsets of items affect the stability of the item parameters of the embedded subsets.
3) Do the number of items in a subset affect the stability of the estimates over occasion? It is expected that the parameter estimates for longer subsets of items should exhibit more stability than the shorter subsets especially for those subsets with more than 27 items.

4) Are item parameter estimates stable over different sample sizes? It is also expected that extreme variability of sample size adversely affects the stability of the estimates regardless of test length.

5) Does estimation program (BILOG, LOGIST, NoHarm) affect stability of estimates over occasion? From the literature it is expected that BILOG will yield more stable estimates than LOGIST. There are no studies available in which the results obtained from NoHarm have been compared with those obtained using the other two procedures. Given the problems associated with JMLE, NoHarm will probably yield better results than LOGIST for the 2PL model.

6) Is stability of item parameter estimates affected by IRT model? It is expected that because of the difficulty of accurately estimating the c parameter especially where N = < 1000 that the estimates obtained using the 3PL model will yield the least stable results. Although examinees are asked not to guess, given the importance of the test it is likely that some examinees will. In this case, the 2PL model should yield more stable results than the 1PL model.

The results should help to determine whether there are limitations to the application of IRT models to second language data. This is especially in the context of institutional tests where numbers of students taking particular versions vary and where the focus is on global comprehension of longer texts and on testing different skills within the same test.

Chapter III — Methodology

In this chapter the methodology for the study is presented. The chapter is divided into four sections: measuring instruments, test data and examinee samples, procedures, and data analysis.

Measuring Instruments

The instruments are versions of a test known as the English Proficiency Test, a test of second language proficiency used to determine the minimum level of second language competence required for fulfilment of the University of Ottawa's bilingual requirements. The tests have been developed for the population entering the university; thus, they are comprised of texts readily understood by educated young adults. Potential sources for reading comprehension texts include introductory textbooks, university documents, and magazine and newspaper articles where the main purpose is to popularize new ideas from science, industry, education, etc. Sources for listening comprehension texts include excerpts from lectures and radio broadcasts where again the main purpose is to try to popularize new ideas for a non-specialist audience.

All questions are about the literal meaning of the text and include recognition of the main idea and supporting ideas, recognition of examples that support or contradict the main idea, location of specific pieces of information, recognition of paraphrases of words/phrases in the text, and the understanding of both implied and explicit relationships between elements in the text. Each test version has the same format and consists of seven texts, each with a set of questions pertaining to the global comprehension of the text. The skills tested are listening comprehension (three texts with a maximum of 18 items), reading comprehension (three texts with a maximum of 18 items), and general knowledge of vocabulary, grammar, and structure as measured by a cloze text (28 to 32 items). This latter text has had words deleted on a rational basis in order to test specific points of second language knowledge.

All questions are in multiple choice format with four alternatives. Candidates are specifically asked not to guess and told to mark alternative E if they feel there is no correct answer. Three versions are prepared each year and, for security reasons, individual texts are

used approximately every eighteen months. The bank of texts from which each version is drawn consists of 24 listening texts, 24 reading texts, and 9 cloze texts. The order of presentation is always the same, first the listening texts, then the reading texts, and finally the cloze text. The language of presentation is always the targeted second language although the instructions are also written in the candidate's presumed first language (French). For listening comprehension, the questions for each text are written in the test booklet and the texts are presented on tape. Time is given to read the questions before the presentation of each text. After hearing the text once, examinees have time to answer the questions before hearing the text a second time. Additional time is then given to check their answers to the questions. Presentation of the listening texts takes approximately 30 minutes. Evidently, to some extent the ability to read in English is also being measured in the listening comprehension section of the test. The reading texts and questions and the cloze text with four alternative word choices are printed in the test booklet. One hour is given for completion of this part of the test. The time allowed is sufficient for the majority of examinees to be able to complete the test.

All texts and the accompanying questions are validated before being used in a test by being field tested with groups of students of known second language ability. If revision is needed, it is done and the texts are field tested again. The process is repeated until the classical item statistics are at a satisfactory level and the mean difficulty used in field testing is such that students at the intermediate level (these are students enrolled in classes known to be at that level) obtain a pass mark of 50%. Three part scores, one for listening comprehension, one for reading comprehension, and one for the cloze test, are each weighted to form approximately one third of the possible total. Although the three part scores are reported, it is the test score total which is used to determine whether the student has passed or failed. The pass mark, 50%, represents an intermediate level of second language ability for the receptive skills tested. The total score is considered to be a fairer means of determining the pass level because it allows for the fact that many students do not have a uniform profile in their skill levels and can compensate in one area for slight deficiencies in another area. All students who obtain a mark of 45 to 49% are permitted to appeal and be retested by writing another version of the part of the test which they failed. Until now the test developers have relied on classical test methods to assess the difficulty level of the different test versions. If an IRT model is appropriate, then it could be used in subsequent test equating and item banking.

Test Data and Examinee Samples

The data consist of responses to subsets of items of varying lengths, short subsets (n = 4 to 7), longer subsets (cloze tests) n = 28 to 32. These target subsets have been used on at least two occasions and are embedded in different versions of the University of Ottawa's English Proficiency Test. Over two administrations there are ten versions of the English Proficiency Test in which eight subsets are embedded (two listening subsets, three reading subsets, and three cloze tests). Although the focus of the study is on the stability of the item parameter estimates of these subsets, the parameter estimates are obtained in the context of the test in which they are embedded. Because of the different times of the year (April, September, December) the test is administered, the number of examinees varies considerably (N = 102 to 618) with the largest number always being in September. The different test versions showing the embedded subsets for the first administration are described in Table 1, for the second administration in Table 2 and for the third administration in Table 3.

Table 1 **Descriptions of the Test Versions Used for the First Administration Showing the Embedded Subsets**

Test Version	Embedded Subsets	Type of Subset	N of Items Subset	N of Items Test	N of Examinees	Overall Test $\bar{X}$ %	Confidence Interval %
April 1991	Computers Snowmobiles	L.C.* Cloze	5 32	66	119	58	± 10.6
September 1990	Sleepcycles Lufts Mother Teresa	L.C. R.C.** Cloze	6 6 30	61	499	67	± 10.8
December 1991	Hazardous Home	R.C.	7	63	177	59	± 11.3
December 1990	Volunteer Work	R.C.	4	63	148	58	± 10.5
September 1989	Acid Rain	Cloze	28	61	328	68	± 10.2

* L.C. = Listening Comprehension
** R.C. = Reading Comprehension

Table 2 **Descriptions of the Test Versions Used for the Second Administration Showing the Embedded Subsets**

Test Version	Embedded Subsets	Type of Subset	N of Items Subset	N of Items Test	N of Examinees	Overall Test $\bar{X}$ %	Confidence Interval %
September 1993	Computers Lufts	L.C.* R.C.**	5 6	63	525	62	± 11.4
April 1993	Hazardous Home Volunteer Work	R.C. R.C.	7 4	63	117	62	± 10.2
December 1992	Sleep Cycles Acid Rain	L.C. Cloze	6 28	63	175	59	± 10.5
September 1992	Snowmobiles	Cloze	32	71	618	68	± 10.6
April 1992	Mother Teresa	Cloze	30	67	102	63	± 9.4

* L.C. = Listening Comprehension
** R.C. = Reading Comprehension

Table 3 **Descriptions of the Test Versions Used for the Third Administration Showing the Embedded Subsets**

Test Version	Embedded Subsets	Type of Subset	N of Items Subset	N of Items Test	N of Examinees	Overall Test $\bar{X}$ %	Confidence Interval %
September 1989	Lufts	R.C.*	6	61	328	68	± 10.2

* R.C. = Reading Comprehension

The number of embedded subsets in any one version of the test varies from one to three. When more than one subset is embedded in a test version, for security reasons, the same pattern was not repeated on the subsequent administration. The overall mean of the test versions varies from 56% to 68%; in general, the higher means are those of the September versions of the test with the exception of September, 1993. In September, 1993, all students regardless of ability wrote the test where previously many of the weaker students enrolled in courses without trying it. Thus the mean is lower than in previous years.

As was reported by Whitely and Dawis (1976), the difficulty of the test material in which subsets are embedded can affect the stability of the item parameter estimates. They reported only the overall difficulty of the test forms. Those tests evaluated as being "difficult" or as "easy" had means that were .10 different from the mean of the target items. Table 4 shows the level of difficulty of each subset, the average difficulty of the rest of the test, the difference between the subset mean and the overall test mean, the average difficulty of the other subsets measuring the same skill and the difference between the subset mean and the mean of the other subsets measuring the same skill. This permits comparisons similar to those made in Whitely and Dawis study.

Table 4 **A Comparison of the Subset Means, the Overall Test Means and the Same Skill Subset Means**

Embedded Subset	Time	Mean Embedded Subset	Overall Test Mean	Difference Between Mean & Overall Test Mean	Mean of Other Subsets Testing Same Skill	Difference Between Subset Mean & Mean of Same Skill Subsets
Computers	1	.64	.61	.03	.64	.00
	2	.63	.60	.03	.62	.01
Sleep Cycles	1	.57	.70	- .13	.75	- .18
	2	.48	.62	- .14	.58	- .10
Hazardous Home	1	.62	.65	- .03	.74	- .12
	2	.67	.58	.09	.62	.05
Volunteer Work	1	.69	.56	.13	.64	.05
	2	.68	.56	.12	.58	.10
Lufts	1	.70	.69	.01	.49	.21
	2	.64	.69	- .05	.72	- .08
	3	.64	.60	.04	.64	.00
Acid Rain	1	.73	.66	.07	—	—
	2	.63	.58	.05	—	—
Mother Teresa	1	.68	.69	.01	—	—
	2	.64	.69	.05	—	—
Snowmobiles	1	.60	.63	.03	—	—
	2	.61	.66	- .05	—	—

From Table 4, subset "Computers" and the three cloze subsets ("Acid Rain", "Mother Teresa" and "Snowmobiles") are very similar in difficulty to the test in which they are embedded (i.e., the differences in means are much less than .10). In the case of "Sleepcycles", the pattern is quite similar at the two administrations, but there is a substantial difference (i.e., much greater than .10) between the mean of the target subset and the other test material. The overall test means and the means of the other subsets testing the same skill are all much easier than the embedded subsets. For "Hazardous Home", the overall test mean is similar to the subset mean but the mean of the other reading subsets is easier (> .10). At the second administration, the overall test mean is more difficult (close to .10) and the mean of the other reading subsets is slightly more difficult than the mean of the embedded subset. For "Volunteer Work", the pattern is similar at the two administrations. Both overall test means are more difficult (both > .10) than the embedded subset means. In the case of the other reading subset means, the mean at the second administration is .10 more difficult while the mean of the first administration is less difficult (.05). For "Lufts", the overall means at all three administrations are close to the means of the embedded subsets. For the means of the other reading subsets, the mean on the first administration is of concern (.21 difference).

In summary, a difference in parameter estimation because of the difficulty of the other material in the test might be expected for "Sleep Cycles", "Hazardous Home", "Volunteer Work" and "Lufts" first administration. The estimates for the cloze tests and subset "Computers" do not appear to be threatened by the difficulty of the material in which they are embedded.

Procedures

The first step was to ascertain the dimensionality of each of the ten versions of the test using the program NoHarm. There are two statistics generated by the program which are used in determining the dimensionality of the data set. The values in the residual matrix are converted to Z scores and the percentage of values greater than 1.96 is calculated. When this value is approximately .05 or smaller, it provides evidence of model fit. Three χ^2 values (Gessaroli & De Champlain, 1991) are also calculated and the one used in this study is χ^2 B, F, and H. The χ^2 value provides evidence of fit when it is not significant. The dimensionality of the data sets was analyzed using each of the three IRT models. Because one of the ways in

which model fit is ascertained is through the stability of the item parameter estimates, estimates using all three models were obtained.

From the literature it is also evident that the stability of the estimates may be affected differentially by the estimating procedure. Therefore estimates were obtained using three different estimating procedures: joint maximum likelihood (LOGIST), marginal maximum likelihood (BILOG), and non-linear factor analysis (NoHarm).

The effect of other factors mentioned in the literature (in addition to dimensionality and estimating procedure) which may affect invariance could only be judged indirectly. These factors were the variability in the number of subjects (105 to 618) and the difficulty of the test material in which the subset was embedded (this particularly threatened the estimates for "Hazardous Home", "Volunteer Work" and the first administration of "Lufts").

Data Analysis

The dimensionality of the data sets was judged using the criteria mentioned earlier (the percent of Z scores > 1.96 and the χ^2 statistic). The classical item statistics were calculated including the difficulty (p value) and the biserial. The stability of the item parameter estimates was determined from the correlation of both b and a parameter estimates over the two (three) administrations. The means and standard deviations were examined for the b and a parameter estimates. As well the means and standard deviations of the error estimates for the a and b parameter estimates for which these statistics are calculated (BILOG and LOGIST) were examined. NoHarm does not provide error estimates of the calculated a and b parameters. Lastly, the a and b parameter estimates from the second administration of each subset were equated with the estimates from the first administration and the absolute differences between the original values and the equated values were calculated.

Chapter IV — Results and Discussion

In this chapter the results and discussion are presented. The issue of dimensionality is discussed first in the context of the ten complete data sets. All of the dimensionality studies were done using NoHarm. For reference, the classical item statistics (p values and biserials) were calculated and are presented in Appendix A. Then the descriptive statistics for the a and b parameters for each of the three methods used in the study and the results of the correlations (as measures of the stability of the estimates) are presented for each subset in turn. The descriptive statistics of the error estimates associated with the a and b parameter estimates for LOGIST and BILOG and the absolute differences between the equated a and b parameter estimates (using Time 1 as the standard) and the original estimates are also presented.

Dimensionality of the Data Sets

The dimensionality of the data sets was first examined using a 1PL model. Because of the fact that the same constraints must be applied to the a parameter with solutions of more than one factor (thus yielding the same results) it was not possible to obtain valid estimates of the 1PL model with more than one factor. None of the data sets was judged to be unidimensional under a 1PL model.

Next the dimensionality of the data sets was examined using a 2PL model, and the results are presented in Table 5.

Table 5 Test of Dimensionality with a 2PL Model

Test	N of Factors	Percent Z > 1.96	χ^2 B, F & H	df	p*
September 1989	1	.06	1890.4	1769	.02**
	2	.02	1177.3	1709	n.s.
	3	.01	985.6	1650	n.s.
September 1990	1	.14	4051.5	1769	.000
	2	.02	1323.5	1709	n.s.**
	3	.02	1146.2	1650	n.s.
September 1992	1	.20	6902.7	2414	.000
	2	.04	2446.5	2344	n.s.**
	3	.02	1680.8	2275	n.s.
September 1993	1	.13	3259.1	1890	.000
	2	.02	1328.6	1828	n.s.**
	3	.01	1139.1	1767	n.s.
December 1990	1	.03	1514.3	1890	n.s.**
	2	.01	1213.0	1828	n.s.
December 1991	1	.02	1450.7	1890	n.s.**
	2	.01	1258.5	1828	n.s.
December 1992	1	.02	1386.7	1890	n.s.**
	2	.02	1232.9	1828	n.s.
April 1991	1	.03	1657.1	2079	n.s.**
	2	.01	1374.6	2014	n.s.
April 1992	1	.02	1512.8	2144	n.s.**
	2	.01	1270.8	2078	n.s.
April 1993	1	.02	1394.2	1890	n.s.**
	2	.01	1232.1	1828	n.s.

* n.s. = $<.05$ ** chosen best fitting model

Both a one and two factor solution were tried for all ten data sets in order to make sure that if the one factor model appeared to give adequate fit, the two factor model did not substantially improve on the fit of the one factor model. In the case of the September data sets where in most instances the one factor solution did not appear to fit, a three factor solution was also tried to see whether it provided any substantial improvements over the two factor solution. Since no substantial changes were observed in either the percent $Z > 1.96$ or χ^2, the more parsimonious model was chosen in all cases.

From Table 5, it appears that the September data sets do not fit a one factor model with the possible exception of September, 1989, which according to the criteria is somewhat borderline but on the grounds of parsimony was judged to be unidimensional. The other data sets, three from December administrations and three from April administrations, appear to be unidimensional. Because of the fact that the data sets judged to be unidimensional are also those with the smaller numbers of candidates (<325), there is a question as to whether these results represent a failure of the method because of the small number of subjects or whether they are truly valid. It is interesting to note, however, that no data set was judged to be unidimensional using a 1PL model so that there is some evidence to suggest that the results are valid and are not simply due to the small number of subjects.

In order to see whether a 3PL model yielded better results in the case of the September data sets, one, two and three factor solutions were tried and the results are presented in Table 6.

Table 6 Test of Dimensionality with a 3PL Model

Test	N of Factors	Percent Z > 1.96	χ^2 B, F & H	df	p*
September 1989	1	.06	1868.1	1769	n.s.**
	2	.02	1281.9	1709	n.s.
	3	.02	1104.0	1650	n.s.
September 1990	1	.14	4040.4	1769	.000
	2	.02	1323.1	1709	n.s.**
	3	.01	1140.8	1650	n.s.
September 1992	1	.21	6874.2	2414	.000
	2	.04	2424.3	2344	n.s.**
	3	.02	1690.9	2275	n.s.
September 1993	1	.13	3238.1	1890	.000
	2	.02	1311.0	1828	n.s.**
	3	.01	1147.5	1767	n.s.

* p = <.05 ** chosen best fitting model

In general, the results are identical to those obtained using the 2PL model. For the September, 1989 data set, the percent of Z >1.96 remains the same but the χ^2 is clearly non-significant for the one factor 3PL model solution.

For the three data sets for which a two factor solution was judged to be appropriate, the factor structure was unrelated to task type (listening, reading, cloze) or to the content of the various subsets (three listening topics, three reading topics). A preliminary examination of particular items which loaded in a similar way failed to show any obvious relationship among the items. For example, these items were not testing a similar skill (e.g., all items dealt with paraphrases) as judged by those who had prepared the items. It would appear that any attempt to uncover what the factor structure might represent would require a separate study.

In summary, from the results of the analyses, seven of the data sets were judged to be unidimensional using a 2PL model and the remaining three were judged to consist of two factors using a 2PL model. In the case of those data sets requiring a two factor solution, no substantial change was observed using a 3PL model. In terms of Research Question 1, it is evident that second language tests measuring more than one skill may be unidimensional depending upon the time used. These results are consistent with the observation of Bejar (1983) who feels that dimensionality is situation specific.

Subset Results

In general the means and standard deviations of the estimated errors are unremarkable in that no consistently small or large values are found within particular subsets. This is because the size of the values is somewhat dependent upon sample size and that frequently varies within subsets across the two administrations. Large values also occur in the case of subset "Lufts", for example, when other items in the rest of the test cannot be estimated successfully regardless of the fact that the number of subjects is fairly large (325). It should be noted, however, that the errors are consistently smaller in the case of BILOG although this may be because of a difference in the scales used for the calculations. In addition there are no error estimates (a or b) for the 3PL model September, 1989 (Time 1 "Acid Rain" and Time 1 "Lufts") estimated by LOGIST. It appears that the program could not make adjustments for certain anomalous items when estimating the c parameters and thus could not meaningfully calculate error estimates. In fact when the program NoHarm was used under a 3PL model using the c parameter estimates generated by BILOG, one of the c parameter estimates had to be held at 0 in order for the program to complete the required calculations. When there are any exceptions noted, they are reported in the discussion of the results for each subset as they occur.

Listening Subsets

The descriptive statistics and the correlations of the a and b parameter estimates for subset "Computers" are presented in Tables 7a, b, c and d.

Table 7a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Listening Subset "Computers" (n=5) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.71	.76	-.40	.61	-.70	.81	-.31	.60	—	—	—	—
	2PL	-.44	.72	-.25	.67	-.52	.72	-.18	.66	-.53	.83	-.19	.69
	3PL	-.18	.88	-.11	.97	-.09	.88	.26	.82	—	—	—	—
a	2PL	.84	.40	.67	.36	.70	.21	.64	.30	.65	.22	.61	.26
	3PL	1.01	.49	.73	.27	.89	.15	.87	.19	—	—	—	—

Table 7b **Correlations Between the a and b Parameter Estimates Obtained for Subset "Computers" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.99**	.99**	.99**	.99**	.99**	.93**	.92*
a	—	—	.81	.73	.63	.88	.64

* .05 ** .01

Table 7c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Subset "Computers" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	1.56	.83	.53	.26	.22	.02	.09	.01
2PL	1.47	1.85	.90	.67	.23	.06	.12	.05
3PL	.59	.25	.51	.41	.28	.07	.18	.04
Error Estimates — a parameters								
1PL	.36	.04	.18	.02	.17	.06	.08	.03
2PL	.45	.23	.18	.04	.24	.04	.16	.03

Table 7d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Subset "Computers"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.09	.06	.21	.12	.10	.04	.13	.05	.08	.05	.32	.14
3PL	.33	.26	.13	.09	—	—	.29	.23	.03	.02	—	—

The means and standard deviations (Table 7a) appear to be similar across estimation procedures and administration times. There is a slight decrease in the value of the mean of the b estimates as the number of parameters being estimated increases which is to be expected. The standard deviations are unremarkable. The mean values of the b parameter estimates are not the same over the two administrations although the means of the p values calculated over two administrations are almost identical (.64, .63). The means and standard deviations for the a parameter estimates are also within a reasonable range. For the standard errors of the estimates (Table 7c), in general, the size of the mean values is smaller for the administration with the largest number of subjects (Time 2, N = 525). The standard deviation of the error estimates of the 2PL model, Time 1, LOGIST b estimate is much larger than any of the others. The values of the error estimates for the a parameters are in general smaller in keeping with the

narrower range of values of the a parameters, and the standard deviations are small except for 3PL, Time 1, LOGIST. The means and standard deviations of the absolute differences (Table 7d) observed between the original a and b parameter estimates at Time 2 and the equated estimates (using Time 1 as the reference) are similar across estimating procedures and unremarkable. The values are considered to be similar when the means are less than .35 and the standard deviations are smaller than the means since this is the pattern observed across subsets with very few noted exceptions.

The correlations (Table 7b) of the b estimates are all above .90 for all three models across estimating procedures. None of the a parameter estimate correlations are statistically significant and the correlations range from .63 to .88. The correlation is higher for LOGIST for both the 2PL and 3PL estimates.

The descriptive statistics and correlations of the a and b parameter estimates for subset "Sleep Cycles" are presented in Tables 8 a, b, c and d.

Table 8a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Listening Subset "Sleep Cycles" (n=6) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.41	.57	.05	.73	-.38	.60	.12	.80	—	—	—	—
	2PL	-.46	.57	-.16	1.04	-.42	.60	-.01	.83	-.46	.62	-.14	.97
	3PL	-.34	.73	.31	1.09	-.09	.88	.26	.82	—	—	—	—
a	2PL	.64	.19	.55	.26	.59	.16	.57	.21	.60	.16	.57	.29
	3PL	.96	.67	1.17	.71	.99	.54	.94	.42	—	—	—	—

Table 8b **Correlations Between the a and b Parameter Estimates Obtained for Subset "Sleep Cycles" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.80	.80	.72	.78	.77	.35	.62
a	—	—	.96**	.93**	.94**	.28	.74

** .01

Table 8c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Subset "Sleep Cycles" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	.49	.22	.78	.39	.10	.01	.18	.02
2PL	.81	.79	7.28	14.00	.13	.04	.24	.10
3PL	.67	.71	3.17	7.19	.19	.10	.28	.16
a parameters								
1PL	.16	.02	.30	.10	.07	.01	.11	.03
2PL	.20	.08	.48	.23	.19	.11	.28	.06

Table 8d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Subset "Sleep Cycles"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.30	.19	.20	.14	.33	.17	.06	.05	.06	.05	.30	.31
3PL	.07	.03	.13	.10	—	—	.21	.05	.11	.06	—	—

The means and standard deviations (Table 8a) are similar across estimation procedures and administration times. Unlike subset "Computers", the mean of the b parameter estimates is easier for the 2PL model rather than for the 1PL model. The standard deviations are similar in value to those observed previously for subset "Computers". Only the values for the 2PL and 3PL models, Time 2, LOGIST are somewhat larger than the other values. Like subset "Computers", the means of the b parameter estimates are not the same over the two administrations but in this case, the means of the p values are also different (.57, .48). The means and standard deviations of the a parameter estimates are unremarkable.

For the standard errors of the estimates (Table 8c), like subset "Computers", the values in general are smaller for the administration time with the larger number of subjects (Time 1,

N = 499). However, there appear to be additional problems in estimating at the time with fewer subjects because both the means and standard deviations for the 2PL and 3PL models at Time 2 are large. The values of the error estimates for the a parameter estimates are similar to those obtained for subset "Computers".

The means and standard deviations of the absolute differences (Table 8d) observed between the original a and b parameter estimates at Time 2 and the equated estimates (using Time 1 as the reference) are also similar to those observed for subset "Computers".

For subset "Sleep Cycles", the observed pattern of correlations (Table 8b) is much different from that of subset "Computers". None of the correlations of the b parameter estimates is statistically significant and all are below .90. For the a estimates, however, for the 2PL model all three estimating procedures show correlations above .90. The highest correlation is that of LOGIST (.96). The 3PL a parameter correlation for LOGIST is smaller than that for BILOG (.28 and .74).

In summary, the results for the two listening subsets are different in terms of the observed pattern of correlations. The b estimates for subset "Computers" are judged to be stable (correlation $\geq$ .90) while those of "Sleep Cycles" are not. Nothing can be seen in the various descriptive statistics for the b parameter estimates that would suggest an explanation of the observed differences. The only observed difference was that in the case of "Sleep Cycles" the mean of the p values over the two administrations was different (.57, .48) while for "Computers", it was the same (.64, .63). In the case of the a estimates, the results are different. The correlations for "Computers" are in the moderate range (.63 to .88) while for "Sleep Cycles" the correlations for the a estimates generated by the 2PL model indicate a high degree of stability for all three estimating procedures. Again there is no indication from the various descriptive indices why this should be so. The values observed for subset "Computers" are similar in magnitude to those observed for subset "Sleep Cycles".

In terms of the possible effect of dimensionality or of the numbers of candidates on the outcomes, it should be noted that both subsets appeared in one test version that was unidimensional and one that was not and in one test version with fewer candidates (119 and 175 "Sleep Cycles") and one with more (500).

Reading Subsets

The descriptive statistics and the correlations of the a and b parameter estimates for subset "Hazardous Home" are presented in Tables 9 a, b, c and d.

Table 9a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Reading Subset "Hazardous Home" (n=7) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.62	.51	-.89	.42	-.58	.52	-.86	.47	—	—	—	—
	2PL	-.52	.46	-.82	.37	-.50	.45	-.76	.41	-.52	.47	-.81	.46
	3PL	-.18	.71	-.42	.53	.00	.54	-.26	.45	—	—	—	—
a	2PL	.86	.23	.66	.23	.64	.19	.79	.16	.63	.21	.77	.17
	3PL	1.33	.61	1.05	.55	.89	.15	1.31	.57	—	—	—	—

Table 9b **Correlations Between the a and b Parameter Estimates Obtained for Subset "Hazardous Home" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.72	.75	.49	.58	.55	.10	.52
a	—	—	.65	.58	.38	.40	.66

Table 9c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Subset "Hazardous Home" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	1.45	.60	1.30	.50	.18	.02	.21	.02
2PL	1.17	.92	1.19	.60	.18	.03	.20	.06
3PL	.53	.8	.65	.62	.24	.04	.23	.06
a parameters								
1PL	.26	.02	.37	.05	.12	.03	.18	.04
2PL	.39	.16	.50	.24	.22	.05	.46*	.45

Table 9d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Subset "Hazardous Home"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.06	.03	.16	.06	.01	.00	.16	.04	.19	.05	.25	.11
3PL	.05	.04	.98	1.04	—	—	.28	.06	.43	.41	—	—

As was the case for the listening subsets, the descriptive statistics (Table 9a) have similar patterns across estimating methods and across models. The means of the b parameter estimates are not the same for the two administrations as was seen previously for the listening subsets. Like subset "Sleep Cycles" the means of the p values for the two administrations are also not the same (.62, .67). The standard deviations for the b parameter estimates are within the range of values observed previously. The means and standard deviations of the a parameter estimates are also similar to those observed previously.

The standard errors of the estimates (Table 9c) for the b parameter estimates are similar in magnitude to those observed previously for administrations with a smaller number of subjects (both administrations for "Hazardous Home" had small numbers of candidates: 117, 177). The

standard errors of the estimates for the a parameters are unremarkable except for 3PL, Time 2 BILOG where the standard deviation is quite large and where it has already been noted that BILOG did not calculate an error estimate for item 6.

For the absolute differences between the original a and b parameter estimates (Time 2) and the equated values (Table 9d), most of the values are similar to those observed previously with the exception of the BILOG values for the 3PL model. For subset "Hazardous Home", none of the correlations (Table 9b) for either the a or b parameter estimates are statistically significant and all are well below .90.

The descriptive statistics and the correlations of the a and b parameter estimates for subset "Volunteer Work" are presented in Tables 10 a, b, c and d.

Table 10a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Reading Subset "Volunteer Work" (n=4) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.96	.56	-.87	.58	-.97	.61	-.87	.61	—	—	—	—
	2PL	-.94	.66	-.77	.50	-.98	.69	-.75	.52	-1.06	.80	-.81	.58
	3PL	-.61	.32	-.27	.33	-.75	.65	-.26	.49	—	—	—	—
a	2PL	.67	.10	.83	.09	.64	.09	.78	.05	.63	.14	.74	.06
	3PL	.78	.03	1.26	.62	.70	.12	1.08	.10	—	—	—	—

Table 10b **Correlations Between the a and b Parameter Estimates Obtained for Subset "Volunteer Work" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.99*	.99*	.95*	.96	.96	.18	.95*
a	—	—	.27	-.22	.53	.77	.54

* .05

Table 10c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Subset "Volunteer Work" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	.96	.36	1.48	.68	.20	.01	.21	.01
2PL	1.16	1.10	.96	.38	.26	.12	.21	.04
3PL	.56	.24	.38	.11	.29	.10	.25	.03
a parameters								
1PL	.25	.03	.36	.05	.13	.01	.19	.04
2PL	.28	.10	.50	.31	.15	.02	.31	.05

Table 10d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Subset "Volunteer Work"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.14	.04	.34	.23	.21	.02	.23	.01	.17	.03	.27	.19
3PL	5.00	6.38	.07	.03	—	—	.52	.55	.38	.02	—	—

The means and standard deviations (Table 10a) are similar in pattern to those observed previously and in terms of magnitude are unremarkable with the exception of the standard deviation of the a parameter estimate for 3PL, Time 2 LOGIST and BILOG indicating higher discrimination at Time 2. The corresponding mean values of the b parameter estimates are different at the two administrations while the means of the p values are not (.69, .68).

The standard errors of the estimates (Table 10c) of the b parameters are similar to those observed previously and are large for LOGIST. For the a parameters, the error estimates are also similar to those observed previously with the exception of 3PL, Time 2 LOGIST where the standard deviation is somewhat larger than for the other values.

For the absolute differences (Table 10d) between the original a and b parameter estimates (Time 2) and the equated values, most of the values are similar to those observed previously with the exception of the LOGIST values for the 3PL model (larger than any other reported values for any subset).

For the correlations (Table 10b), the results for subset "Volunteer Work" are different from those observed for subset "Hazardous Home". For the 1PL and 2PL models all the correlations for the b parameter estimates are above .90. For the 3PL model, the results are very different for the two estimating methods. The correlation for LOGIST is very low (.18) while the correlation for BILOG is .95. This is probably what is reflected in the very large value observed above for the absolute difference of the b parameters for 3PL LOGIST. The correlations of the a parameter estimates for the 2PL model are low. In fact for BILOG the correlation between the two sets of estimates is negative. The estimates for the 3PL model are moderate for both estimating procedures.

The descriptive statistics and the correlations of the a and b parameter estimates for subset "Lufts" are presented in Tables 11 a, b, c and d.

Table 11a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Reading Subset "Lufts" (n=6) Across Three Administrations**

Parameter	Model	LOGIST						BILOG						NoHarm					
		Time 1		Time 2		Time 3		Time 1		Time 2		Time 3		Time 1		Time 2		Time 3	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-1.05	.94	-.77	.49	-.33	.49	-1.00	.86	-.79	.47	-.43	.49	—	—	—	—	—	—
	2PL	-1.16	1.12	-.75	.45	-.35	.56	-1.18	1.14	-.75	.43	-.46	.58	-1.40	1.52	-.76	.44	-.35	.58
	3PL	-.63	1.00	-.25	.63	.08	.67	-1.09	1.28	-.48	.82	-.16	.87	—	—	—	—	—	—
a	2PL	.63	.30	.66	.16	.63	.20	.59	.23	.60	.13	.62	.15	.57	.22	.61	.12	.61	.13
	3PL	.66	.30	.87	.36	.89	.39	.74	.26	.83	.20	.92	.25	—	—	—	—	—	—

Table 11b **Correlations Between the a and b Parameter Estimates Obtained for Subset "Lufts" At Two Administrations**

Parameter	Model	Time 1 - Time 2			Time 1 - Time 3			Time 2 - Time 3		
		LOGIST	BILOG	NoHarm	LOGIST	BILOG	NoHarm	LOGIST	BILOG	NoHarm
b	1PL	.78	.78	—	.84*	.87*	—	.82*	.83*	—
	2PL	.68	.84*	.80	.21	.98**	.98**	.81	.80	.76
	3PL	.74	.86*	—	.89**	.91**	—	.92**	.95**	—
a	2PL	.55	.55	.24	.43	.38	.27	.84*	.79	.63
	3PL	.86*	.92**	—	.71	.77		.69	.51	—

* .05 * .01

Table 11c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Subset "Lufts" At Two Administrations**

Error Estimates — b parameters									
	Time 1			Time 2			Time 3		
Statistic	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
				LOGIST					
$\overline{X}$	1.17	6.08	*	.64	.71	.44	.52	.89	.55
SD	.91	12.60	*	.23	.37	.32	.24	.89	.70
				BILOG					
$\overline{X}$	.14	.26	.30	.11	.13	.19	.09	.11	.16
SD	.02	.23	.16	.01	.03	.04	.00	.04	.05

* No values estimated.

a parameters						
	Time 1		Time 2		Time 3	
Statistic	2PL	3PL	2PL	3PL	2PL	3PL
			LOGIST			
$\overline{X}$	.25	*	.16	.19	.17	.20
SD	.08	*	.01	.07	.02	.07
			BILOG			
$\overline{X}$	.10	.15	.08	.15	.07	.17
SD	.04	.05	.01	.06	.01	.07

Table 11d Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Subset "Lufts"

Administration	Model	Difference b						Difference a					
		LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Time 1 with Time 2	2PL	.40	.20	.38	.17	.43	.09	.14	.07	.07	.06	.95	.69
	3PL	.15	.10	.18	.10	—	—	.25	.06	.11	.06	—	—
Time 1 with Time 3	2PL	.21	.18	.22	.15	.37	.08	.08	.04	.09	.05	.28	.50
	3PL	.17	.14	.10	.04	—	—	.27	.11	.20	.02	—	—
Time 2 with Time 3	2PL	.08	.10	.05	.05	1.21	.04	.04	.03	.08	.01	.28	.14
	3PL	.04	.04	.10	.06	—	—	.05	.03	.13	.07	—	—

The means and standard deviations (Table 11a) appear to be similar across estimation procedures. The standard deviations of the b parameter estimate means are unremarkable with the exceptions of Time 1 for all three models and estimating procedures where the values are larger than for the other two times. The mean values of the b parameter estimates are different over the three administrations (Time 3 being hardest) while for the means of the p values, Times 2 and 3 are identical (.63) and Time 1 is easier (.70).

For the standard errors of the estimates (Table 11c) of the b parameters, in general, the values are similar except in the case of Time 1 where the means and standard deviations are much larger than the other values for the 1PL and 2PL models. For the 3PL model LOGIST, no values were calculated because of the problem noted previously. The error estimates for the a parameters are unremarkable.

For the absolute differences (Table 11a) between the original a and b parameter estimates and the equated estimates for the b and a parameters, the smallest differences are observed at Time 2 with Time 3. The largest differences in the equated b parameter estimates observed are at Time 1 with Time 2, 2PL model, all three estimating procedures and for NoHarm Time 1 with Time 3. The largest differences in the equated a parameter estimates as Time 1 with Time 2, NoHarm, and Time 1 with Time 3, NoHarm (SD .50).

The results of the correlations (Table 11b) reflect the patterns of the other two reading subsets. Sometimes the correlations are high enough to indicate stability of the estimates (as in the case of "Volunteer Work") and at other times the correlations indicate instability of the estimates (like "Hazardous Home"). The least successful pairing of administrations in terms of stability of the estimates of the b parameters is Time 1 with Time 2. Only two of the correlations are statistically significant (BILOG, 2PL and 3PL) but the values are less than .90. For the a parameter estimates, there are also two statistically significant correlations, LOGIST and BILOG for the 3PL model.

The most successful pairing in terms of the number of significant correlations for the b parameter estimates is Time 1 with Time 3. Six of the seven correlations are statistically significant (three $>.90$). For the 1PL model the results for LOGIST and BILOG are identical. For the 2PL model only the estimates for BILOG and NoHarm are stable (both .98). For the 3PL model, the correlations are similar for the two estimating procedures but the value is $>.90$ for BILOG. For the a parameter estimates, all the correlation values can be characterized as being low (for the 2PL model) and moderate (for the 3PL model).

For the pairing of Time 2 and Time 3 for the b parameter estimates, four of the six correlations are statistically significant but only the values obtained for the 3PL model (LOGIST and BILOG) are $>.90$. One of the values for the a parameter estimates is statistically significant (LOGIST, 2PL). The rest are in the moderate range.

In summary, as was the case for the listening subsets, the results are mixed for the reading subsets. The a and b estimates for "Hazardous Home" showed instability. The b parameter estimates for "Volunteer Work" were highly stable for the 1PL and 2PL models across estimating procedures but not for the 3PL model (only BILOG yielded stable estimates). None of the a parameter estimates was invariant. For subset "Lufts", the results across three administrations were similarly uneven. Generally Time 1 and Time 3 gave the most stable results and this is in spite of the fact that there were problems in estimating the 3PL model at Time 1 (September 1989 as reported above).

As was stated earlier, there is nothing remarkable in the descriptive statistics that could account for the observed variability. For example, the standard deviations for the a parameter

estimates for BILOG are similar for the 2PL and 3PL models at Time 1 with Time 2 and yet the correlations are quite different (.55 and .92). It would appear that the descriptive statistics have very little explanatory power in terms of the observed pattern of correlations. For subset "Lufts", the most successful pairing (Time 1 with Time 3) is of test versions judged to be unidimensional (Time 1) and two dimensional (Time 3). The number of candidates for the two administrations was 328 and 499. For "Hazardous Home" and "Volunteer Work" all the data sets were judged to be unidimensional and the number of candidates was small (177 and 117 for the former and 148 and 117 for the latter). It is interesting to note that the two subsets were embedded in the same data set at Time 2.

Cloze Texts

In contrast to the listening and reading subsets where there was some success in obtaining stable estimates (correlation $\geq$ .90), there were no estimates that met the criterion for the three cloze texts. The results of the descriptive statistics and the correlations of the a and b parameter estimates for cloze text "Acid Rain" are presented in Tables 12 a, b, c and d.

Table 12a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Cloze Text "Acid Rain" (n=28) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-1.10	.74	-.67	.76	-1.13	.79	-.63	.81	—	—	—	—
	2PL	-.99	.71	-.77	.96	-.94	.69	-.58	.73	-.99	.76	.70	.86
	3PL	-.74	.79	-.60	1.02	-.54	.70	-.21	.69	—	—	—	—
a	2PL	.90	.35	.80	.44	.85	.30	.78	.32	.87	.35	.76	.34
	3PL	1.14	.57	.96	.52	1.15	.50	1.00	.41	—	—	—	—

Table 12b **Correlations Between the a and b Parameter Estimates Obtained for Cloze Text "Acid Rain" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.82**	.82**	.73**	.83**	.80**	.71**	.83*
a	—	—	.65**	.68**	.74**	.71**	.74**

** .01

Table 12c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Cloze Text "Acid Rain" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	1.21	.71	1.5	.98	.15	.02	.19	.02
2PL	1.21	1.52	3.50*	8.21	.16	.09	.21	.11
3PL	**	**	3.60	10.17	.21	.09	.25	.11
a parameters								
1PL	.24	.04	.35*	.11	.12	.03	.15	.07
2PL	**	**	.32	.10	.23	.11	.24	.12

** No values estimated * Item 6 no value estimated

Table 12d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Cloze Text "Acid Rain"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.19	.17	.03	.03	.11	.05	.09	.06	.03	.02	.28	.10
3PL	.07	.06	.11	.08	—	—	.18	.04	.15	.09	—	—

The means and standard deviations (Table 12a) are similar in magnitude and pattern to those observed previously. The magnitude of the mean values for the b parameter estimates are

different from Time 1 to Time 2, and in this case the mean of the p values is also different (.73, .63).

The standard errors (Table 12c) of the b parameter estimates for LOGIST reflect what has been observed previously. The magnitude of the values especially for the 2PL and 3PL models is larger for Time 2 where the number of subjects is smaller (175). The absolute differences (Table 12d) between the original estimates of the a and b parameters (Time 2) and the equated values (using Time 1 as the reference) are similar in magnitude to those observed previously.

The correlations (Table 12b) for the b parameter estimates are all statistically significant and range from .71 to .83. For the 1PL model the correlations for LOGIST and BILOG are identical. For the 2PL model the correlation of the BILOG estimates is slightly better than those for the other two estimating procedures. The pattern is the same for the 3PL model. The correlation for the estimates obtained from BILOG is better than that for the estimates obtained from LOGIST (.71 and .83). The correlations for the a parameter estimates are also all statistically significant. For the 2PL model the correlation for NoHarm is slightly better than that of BILOG which in turn is slightly better than that for the estimates from LOGIST. For the 3PL model again the correlation for BILOG is slightly better than that for LOGIST.

The results of the descriptive statistics and correlations of the a and b parameter estimates for cloze text "Mother Teresa" are presented in Tables 13 a, b, c and d.

Table 13a **Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Cloze Text "Mother Teresa" (n=30) Across Two Administrations**

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.91	.90	-.66	.84	-.83	.90	-.56	.81	—	—	—	—
	2PL	-.94	.98	-.73	.98	-.73	.82	-.55	.79	-.79	.87	-.72	1.04
	3PL	-.84	1.10	-.45	1.09	-.33	.85	-.16	.75	—	—	—	—
a	2PL	.69	.32	.87	.44	.69	.28	.83	.28	.68	.27	.77	.31
	3PL	.84	.53	1.16	.59	.88	.37	1.12	.39	—	—	—	—

Table 13b **Correlations Between the a and b Parameter Estimates Obtained for Cloze Text "Mother Teresa" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.77**	.77**	.74**	.80**	.75**	.68**	.74**
a	—	—	.27	.35	.36*	.25	.27

* .05 ** .01

Table 13c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Cloze Text "Mother Teresa" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	.87	.59	1.17	.89	.11	.02	.20	.02
2PL	2.64	6.36	4.12	9.37	.15	.07	.24	.10
3PL	2.22	5.38	2.72	6.52	.21	.10	.27	.09
a parameters								
1PL	.19	.05	.64	.15	.09	.03	.21	.10
2PL	.21	.10	.55	.22	.15	.07	.35	.16

Table 13d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Cloze Text "Mother Teresa"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.30	.24	.04	.03	.15	.06	.21	.12	.19	.03	.14	.12
3PL	.10	.07	.03	.02	—	—	.32	.06	.24	.02	—	—

The means and standard deviations (Table 13a) are similar in magnitude and pattern to those observed for "Acid Rain" except contrary to the pattern of slightly increasing difficulty with the number of parameters estimated, LOGIST 2PL for both administration are slightly easier than the 1PL means. The magnitude of the mean values for the b parameter estimates are different from Time 1 to Time 2 and the means of the p values are slightly different (.68, .64).

The standard errors (Table 13c) of the b parameter estimates for LOGIST for both administrations are large (except for 1PL, Time 1 which is similar to values observed previously). This is unexpected at Time 1 because of the large number of subjects (499). The absolute differences (Table 13d) between the original estimates of the a and b parameters (Time

2) and the equated values (using Time 1 as the reference) are similar to those observed for "Acid Rain" with the exceptions of the b estimates 2PL LOGIST and the a estimates 3PL LOGIST.

The correlations (Table 13b) of the b parameter estimates are all statistically significant and range from .68 to .80 (a little lower than the range observed for "Acid Rain"). For the estimating procedures the pattern of correlations is similar to that of "Acid Rain" (identical for the 1PL model, better for BILOG for both the 2PL and 3PL models). The correlations of the a parameter estimates are much lower than those observed for "Acid Rain". Only the 2PL NoHarm correlation is statistically significant but it is still in the same range of magnitude as the other values.

As has been observed previously for cloze text "Acid Rain" and the listening and reading subsets, the descriptive statistics are unremarkable in terms of explanatory power. Although the correlations for the a parameter estimates are different, for example, the descriptive statistics are quite similar.

The descriptive statistics and correlations of the a and b parameter estimates for cloze text "Snowmobiles" are presented in Tables 14 a, b, c and d.

Table 14a Descriptive Statistics of the a and b Parameter Estimates for All Three Estimating Procedures for Cloze Text "Snowmobiles" (n=32) Across Two Administrations

Parameter	Model	LOGIST				BILOG				NoHarm			
		Time 1		Time 2		Time 1		Time 2		Time 1		Time 2	
		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
b	1PL	-.50	.74	-.51	.48	-.38	.77	-.38	.47	—	—	—	—
	2PL	-.56	.91	-.49	.52	-.38	.73	-.36	.52	-.49	1.00	-.38	.54
	3PL	-.27	.96	-.26	.64	-.03	.72	.07	.46	—	—	—	—
a	2PL	.73	.31	.65	.18	.76	.25	.67	.20	.76	.34	.69	.21
	3PL	.97	.53	.86	.40	1.02	.40	1.00	.37	—	—	—	—

Table 14b **Correlations Between the a and b Parameter Estimates Obtained for Cloze Text "Snowmobiles" At Two Administrations**

	1PL		2PL			3PL	
Parameter	LOGIST	BILOG	LOGIST	BILOG	NoHarm	LOGIST	BILOG
b	.64**	.65**	.77**	.72**	.78**	.65**	.66**
a	—	—	.33	.42*	.37*	.27	.36*

* .05 ** .01

Table 14c **Descriptive Statistics of the Standard Errors of the a and b Parameter Estimates for LOGIST and BILOG for Cloze Text "Snowmobiles" At Two Administrations**

Error Estimates — b parameters								
Model	LOGIST				BILOG			
	Time 1		Time 2		Time 1		Time 2	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1PL	1.42	.98	.57	.22	.22	.02	.10	.00
2PL	3.88	8.17	.65	.98	.23	.13	.10	.05
3PL	2.82	6.14	.51	.86	.27	.12	.15	.08
a parameters								
1PL	.36	.08	.13	.02	.17	.05	.07	.01
2PL	.48	.19	.16	.05	.32	.19	.16	.06

Table 14d **Descriptives Statistics of the Mean Absolute Differences Between the Original a and b Parameter Estimates (Time 2) and the Equated a and b Parameter Estimates for LOGIST, BILOG and NoHarm for Cloze Text "Snowmobiles"**

Model	Difference b						Difference a					
	LOGIST		BILOG		NoHarm		LOGIST		BILOG		NoHarm	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
2PL	.28	.21	.13	.10	.32	.10	.12	.07	.06	.04	.31	.36
3PL	.12	.11	.02	.02	—	—	.12	.11	.03	.02	—	—

The means and standard deviations (Table 14a) are similar in magnitude and pattern to those of cloze test "Mother Teresa". In this case, the mean values for the b parameter estimates

are almost identical at Time 1 and Time 2 which is also true of the means of the p values (.60, .61).

The standard errors (Table 14c) of the a and b parameter estimates are within the same range as has been observed previously with the exception of the b parameter estimates Time 1 LOGIST which was also the time with fewer subjects (119). The absolute differences (Table 14d) between the original estimates of the a and b parameters (Time 2) and the equated values (using Time 1 as the reference) are similar to those observed for the other two cloze tests with the exception of 2PL LOGIST.

The correlations (Table 14b) of the b parameter estimates are all statistically significant and the range is slightly lower than that observed for "Mother Teresa" (.64 to .78). For the estimating procedures across models the pattern is somewhat different from those observed for the other two cloze texts. For the 1PL the results are similar for the two estimating procedures. The 2PL results are somewhat different (NoHarm and LOGIST are .78 and the BILOG estimate is lower .72). Unlike the other two cloze texts where there was a modest difference in favour of BILOG for the 3PL model, the correlations are the same for the two estimating procedures. The results of the correlations of the a parameter estimates are slightly better than those observed for "Mother Teresa" but still in the modest range.

As stated above, the descriptive statistics are unremarkable and cannot be used to corroborate the observed pattern of correlations for "Snowmobiles". Similar anomalies appear in the descriptive statistics of the error estimates for the b parameters of LOGIST (elevated means and standard deviations 2PL and 3PL models Time 1, large standard deviations at Time 2) but again since the correlations of the estimates for LOGIST are equal to or better than those of BILOG it is impossible to know how to interpret these observations.

Summary and Discussion of the Findings in Terms of the Pertinent Research Questions

In summary, the results for the cloze tests are similar and unlike the results obtained for the listening and reading subsets in that there are no correlations of the b parameters high enough ($\geq .90$) to be able to state that there is evidence for stability of the estimates. This is in spite of the fact that it was hypothesized that because of the larger numbers of items, it was

more likely that stable estimates would be obtained with the cloze tests (Research Question 3). It happens that all three cloze texts were embedded in test versions that for one administration had reasonably large numbers of subjects (328, 499, 618) and for the second administration considerably smaller numbers of subjects (175, 102, 199) so it is impossible to determine whether the results would have been more stable with larger numbers of subjects (Research Question 4). In a study by Hale, Stansfield, Rock, Hicks, Butler and Oller Jr. (1988), the authors concluded that stability of the item parameter estimates of a cloze test similar to these tests would not be a problem based on a large sample size (11,290 examinees).

The question of why item parameter estimates of cloze tests are not stable with small numbers of subjects may be related to the problem of defining precisely what the cloze test is measuring. Originally the cloze test was regarded (Oller, 1979) primarily as a reading test but subsequent research has shown that this may not always be the case (Alderson, 1984) and that different cloze texts do not necessarily measure precisely the same things (Klein-Braley, 1984). Alderson (1984) found that the cloze results in his study correlated better with grammar and vocabulary subscores than with the reading subscore from a test of general proficiency in ESL. Hale *et al.* (1988) come to a similar conclusion, for they found in a multiple regression analysis that their cloze test results were best accounted for by all the subscores (listening, reading, structures grammar, vocabulary, written expression) of the TOEFL and not by reading alone. Thus, it appears that cloze tests may tap a broader base of ESL knowledge and cannot be considered to be primarily a test of reading comprehension. In this case it is not suprising that with small numbers of subjects the parameter estimates are not stable especially at an intermediate level of ESL proficiency where there is no guarantee that the same proportion of candidates will share a particular item of knowledge since their ESL learning experiences, maternal language and background knowledge will have differed widely.

From the evidence for the reading subsets and the listening subsets, it does not appear that either the dimensionality of the data sets or the number of subjects plays a crucial role in predicting the stability of the item parameter estimates since stable estimates were obtained under all variations of these conditions (Research Question 4).

One of the hypotheses (Research Question 2) made concerning the stability of the estimates was whether the difficulty of the material in which the subsets were embedded might

affect the outcome. Specifically using the criterion of Whitely and Dawis, it was stated that a threat existed for subsets "Sleep Cycles", "Hazardous Home", "Volunteer Work" and "Lufts", Time 1. It would appear that this hypothesis is not bourne out by the results. While it is true that stable estimates were not obtained for "Sleep Cycles" and "Hazardous Home", the results were stable for the other two, so some other explanation is necessary to account for the results obtained with "Sleep Cycles" and "Hazardous Home". Since the a estimates for "Sleep Cycles" were stable it is even a less likely explanation in this case.

It may be possible to provide an explanation of the inconsistency of the listening and reading results from second language research. It is evident from a review of the literature that although a certain amount of work has been done in trying to understand the processes which underlie listening and reading in a second language, research conclusions are still somewhat tentative.

In order to understand how second language learners go about answering questions on a listening exercise, Hawkins (1985) and Buck (1990) have used think-aloud protocols. What their work suggests is that individual listening items measure a variety of skills which change according to each respondent (e.g., for one listener the item may be primarily a test of vocabulary while for another, it is a test of inferencing skills). In addition to the fact that it is difficult to know precisely what items measure, there is also the problem of background knowledge and how it interacts with second language (L2) proficiency in enabling L2 learners to arrive at the correct answer.

The processes underlying L2 reading have been explored somewhat more thoroughly because of the research done in first language (L1) reading. A great deal has been done by Carrell (1984, 1985, 1991) in assessing the effect of L1 reading skills and their transfer to L2 reading. For advanced L2 learners, there is a strong link between L1 and L2 reading skills but this is not the case for less advanced L2 learners (Clarke, 1980; Grabe, 1986) where lack of linguistic knowledge interferes with the reading process regardless of the level of L1 reading skills.

It has also been acknowledged (Carrell, 1983) that background knowledge can play an important role in enabling L2 readers to successfully understand the text. This phenomenon is

particularly important when the subject population is not strong as is the case with those writing the English Proficiency Test.

Given that questions may be testing a variety of skills which are not easily definable, familiarity with the topic may be especially important at lower levels of L2 ability. Thus, it is not surprising that enough variability occurs from one group to another to create instability in the item parameter estimates. This was even the case for subset "Lufts" where the number of subjects on which the estimates were based was fairly large (minimum 328).

Comparison of the Stability of the Estimates Across Estimating Procedures

From the literature, it is desirable to obtain correlations of .90 or better for the b parameter estimates over two administrations if the goal is to equate the items of one test with those of another. According to Hambleton, Swaminathan, and Rogers (1991) stable estimates should be obtainable even with anchor tests of fewer than 10 items. Stable estimates for the b parameter estimates were not obtained for subsets "Sleepcyles", "Hazardous Home", "Lufts" (Time 1 with Time 2), "Acid Rain", "Mother Teresa" and "Snowmobiles". The pattern of correlations of the b and a parameters for those subsets which did produce stable estimates are presented in Table 15.

Table 15 **Comparison of the Stability of the b Parameter Correlations for Subsets Meeting the Criterion (≥.90) Across Estimating Methods (Including the a Parameter Correlations)**

Subset	Model	LOGIST		BILOG		NoHarm	
		b	a	b	a	b	a
Computers	1	.99	—	.99	—	—	—
	2	.99	.81	.99	.73	.99	.63
	3	.93	.28	.92	.74	—	—
Volunteer Work	1	.99	—	.99	—	—	—
	2	.95	.27	.96	.22	.96	.53
	3	.18	.77	.95	.54	—	—
Lufts	1	*	—	*	—	—	—
Time 1-Time 3	2	.21	.43	.98	.38	.98	.27
	3	.89	.71	.91	.77	—	
Lufts	1	*	—	*	—	—	—
Time 2-Time 3	2	*	—	*	—	*	—
	3	.92	.69	.95	.51	—	—

* <.90

For subset "Computers" the results are similar across estimating procedures. The correlations are identical for the 1PL and 2PL models and slightly lower for the 3PL model For "Volunteer Work", the results are similar across estimating procedures for the 1PL and 2PL models but not for the 3PL model (only the BILOG estimate is >.90). The correlations are slightly better for the 1PL model than for the 2PL and 3PL models. For subset "Lufts" the results are somewhat mixed. For "Lufts" (Time 1 with Time 3) the 1PL model did not produce stable estimates. The 2PL estimates were stable for BILOG and NoHarm and the 3PL estimates were stable for BILOG and close enough for LOGIST to be considered to have met the criterion (3% variance difference between the two estimates). For "Lufts" (Time 2 with Time 3) only the 3PL model produced stable estimates. The correlation of the BILOG estimate is slightly better than that of the LOGIST estimate.

To summarize, where stable estimates were obtained (total = 9 occasions where the correlations ≥.90), the correlations were identical or virtually so on seven occasions. On the remaining two occasions (one using the 2PL model and one the 3PL model), BILOG produced stable estimates and LOGIST did not. For the 2PL model NoHarm was as successful as BILOG in producing stable estimates.

For the a parameter estimates obtained with the stable b parameter estimates, the results are less clear cut. For the 2PL model, the correlations of the a parameter estimates are somewhat better for LOGIST than they are for BILOG (although none of the correlations indicate any real stability). The NoHarm correlations are the smallest on two occasions and the largest on one. In the case of the 3PL model, the pattern of correlations is not consistent (in two cases, the correlations for BILOG are better and in two cases, the correlations for LOGIST are better).

In conclusion, although it appears that stable estimates for the b parameters are slightly more likely using BILOG rather than LOGIST, neither program produced estimates of the a parameters which could be described as similarly stable ($\geq .90$). For the 2PL model where the correlations of the b parameter estimates were the same, the correlations of the a parameters were slightly better using LOGIST. For the 3PL model, the correlations were sometimes better using LOGIST and at other times better using BILOG. In general the results for the 2PL model were similar using BILOG and NoHarm (Research Question 5).

In terms of model fit (Research Question 6), it was predicted that the 3PL model would probably be more appropriate than the other two models. In fact on those occasions where stable estimates were obtained, the 3PL model estimates were judged to be stable on all four occasions. The 2PL model results were judged to be stable on three of the four occasions and the 1PL model results were judged to be stable on two of the four occasions. If dimensionality is crucial then, these latter results are difficult to reconcile with the results of the unidimensionality study where the data were never judged to be unidimensional using a 1PL model. From a theoretical viewpoint the 1PL model would not be considered appropriate although stable estimates were obtained in two cases. Unidimensionality cannot account for the discrepancy observed with regard to the slightly better success rate of the 3PL model over the 2PL model since the dimensionality results were similar for the two models.

It would appear that the most consistent explanation for the observed difference between the 2PL and 3PL models is the presence of the pseudo guessing parameter which in the case of this data provides a slightly more flexible framework for calculation of the b parameter estimates. In a similar manner, the fact that the 2PL model was somewhat better than the 1PL

model is probably related to the fact that the a parameter estimates are not fixed at 1 (Hambleton & Traub, 1973).

Strengths and Limitations

One of the main strengths of this study is that it was conducted using real test data. From the studies discussed earlier it appears that very little work has been done which focuses on the stability of the estimates obtained with data from a second language test measuring global comprehension. The particular test data used do, however, limit any conclusions that may be drawn from the results. They apply to tests using a similar format (texts with global comprehension questions, cloze texts which do not focus on reading comprehension) and at a level of difficulty aimed at identifying students at the intermediate level of second language ability.

It was also possible to examine the effect of dimensionality on the stability of the estimates although the results were not conclusive. It appears that the issue of dimensionality as a statistical concept does not necessarily preclude the use of IRT models with second language data. This is consistent with the findings of Choi (1992) who examined the reading sections of two ESL tests. The dimensionality of what is being measured at the level of the individual item may be of more importance when the number of subjects is small (>100 but <600), in that even 30 or 40 candidates with particular background knowledge would be enough to change target item parameter estimates from one administration to another.

One of the main limitations is that this was a naturalistic study. This limited the number of listening and reading subsets that could be examined and since the conditions were not identical for each subset, made the process of comparison difficult. This is particularly true in the case of subset "Lufts" where the results were inconsistent and it was impossible to satisfactorily explain why this was so. Because of the small number of texts that were examined, it is also difficult to know whether IRT models could have been applied at a more successful rate than was the case here.

The issue of the stability of the estimates obtained from the cloze tests is also not resolvable in the context of this study. There was some evidence to suggest that more stable

estimates might be obtained with larger numbers of subjects. However, it might also be related to some property inherent in the texts used in this study or to the issue mentioned above, the influence of a group of candidates having a slightly different language profile from those on previous administrations.

Suggestions for Future Research

In order to expand on what has been learned here, it would be useful to repeat this experiment using an intact test on three occasions (two of which are September administrations). In that way it should be possible to gain a better idea of the consistency with which stable estimates are obtained across texts, to explore further what occurs at two administrations where the test is likely to have more than one dimension and to see whether the stability of the estimates obtained from a cloze text is related to the number of subjects or to some other factor(s).

With a sufficient number of candidates, it might be useful to undertake a study in which item parameter estimates were obtained using multidimensional IRT models to see whether the estimates might be improved.

If the conclusions reached are substantially the same as those reached in this study, then, providing that the populations can be assumed to be similar, the issue that emerges is that of validity. The suggestion from the literature is that at intermediate levels of ability items may measure different microskills depending on the individual respondent. If this is the case, then more work needs to be done in this area so that test developers can be more confident about what it is they are testing and whether the conclusions they draw from their tests are valid. It may be that this is not a problem, that given sufficient numbers of appropriate items, the language learner's skills can be measured successfully because with sufficient repetition enough microskills are demonstrated to constitute a valid measure of ability. But there is reason to proceed with caution until more is known.

Chapter V — Summary

In many ways this study suggests more questions than it resolves. From the evidence, it appears that IRT models are not consistent in providing stable item parameter estimates with a second language test measuring global comprehension at the intermediate level. It is not clear why the method fails although failure cannot be consistently linked to too few candidates, issues of statistical dimensionality, too few items, or the difficulty of the material in which the target subsets were embedded.

It may be linked in some way to the crucial role played by background knowledge in comprehension of the text at the intermediate ESL level. If this is so it raises the issue of the validity of the results obtained when so much remains unknown about what is being tested.

It appears that marginal maximum likelihood estimation (BILOG) is slightly more consistent in producing stable estimates than joint maximum likelihood (LOGIST) although for the 2PL model, NoHarm (non-linear factor analysis) was as successful as BILOG. None of the corresponding a parameter estimates were judged to be stable (although some higher correlational values were observed in cases where the correlation of the b parameter estimates failed to reach the criterion ($\geq$.90)). Therefore, although it would be feasible to equate estimates for the 2PL and 3PL models, not much reliance can be placed on the accuracy of the equated a parameter values obtained.

The purpose of this study was to examine the feasibility of using IRT models to equate test versions by means of subsets of linking items. Given the evidence accumulated so far, it does not appear to be a viable method with these data sets. Failure may be related to the fact that this is a naturalistic study, but if an expanded study yields similar results, then the issue becomes one of validity, whether the definition of ESL global comprehension at the intermediate level is compatible with how an ability is defined under an IRT model. The results of this study also underline the fact that testers need to be more cautious in how they interpret test scores since a great deal still remains to be done before the processes of acquiring and demonstrating second language proficiency are more fully understood.

References

Alderson, J.C. (1984). The cloze procedure and proficiency in English as a foreign language. In *Issues in Language Testing Research*, J.W. Oller Jr. (ed.). Rowley, MA: Newbury House Publishers Inc.

Baker, F.B. (1987). Methodology review: item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, 11(2), 111-141.

Bejar, I.I. (1983). Introduction to item response models and their assumptions. In *Applications of Item Response Theory*, R.K. Hambleton (ed.). Vancouver, B.C.: Educational Research Institute of British Columbia.

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

Buck, G. (1990). The testing of second language listening comprehension. Unpublished doctoral dissertation. Department of Linguistics and Modern English Language, University of Lancaster.

Boettcher Barnes, L.L., & Wise, S.L. (1988). Corrections for guessing in the Rasch model: a simulation. Paper presented at the annual meeting of the AERA, New Orleans.

Carrell, P.L. (1983). Some issues in studying the role of schemata or background knowledge in second language comprehension. *Reading in a Foreign Language*, 1(2), 81-92.

Carrell, P.L. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, 18, 441-469.

Carrell, P.L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly*, 19, 727-752.

Carrell, P.L. (1991). Second language reading: reading ability or language ability. *Applied Linguistics*, 12, 159-179.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155-163.

Choi, J.-C. (1992). An application of item response theory to language testing. *Theoretical Studies in Second Language Acquisition, Volume 2*. New York: Peter Lang Publishing Inc.

Clarke, M.A. (1980). The short-circuit hypothesis for ESL reading — or when language competence interferes with reading performance. *Modern Language Journal*, 64, 203-209.

Cook, L.L., Eignor, D.R., & Petersen, N.S. (1982). A study of the temporal stability of IRT item parameter estimates. Paper presented at the annual meeting of AERA, New York (ERIC 219415).

Cook, L.L., Eignor, D.R., & Taft, H.L. (1984). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Paper presented at the annual meeting of AERA, New Orleans.

De Jong, J.H.A.L., & Glas, C.A.W. (1987). Validation of listening comprehension tests using item response theory. *Language Testing*, 4, 170-192.

Dinero, T.E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1, 581-592.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23(4), 283-298.

Divgi, D.R. (1981). Potential pit falls in applications of IRT. Paper presented at the annual meeting of NCME, Los Angeles.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.

Forster, F. (1978). Everything you wanted to know about the Rasch model (but were afraid to ask). *Portland Public Schools Occasional Papers in Measurement*, No. 17, (ERIC 189099).

Gessaroli, M.E., & De Champlain, A.F. (1991). Assessing test dimensionality using an approximate x^2 statistic. Paper presented at AERA, Chicago, IL.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 33, 234-246.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5(2), 211-220.

Grabe, W. (1986). The transition from theory to practice in teaching reading. In *Teaching Second Language Reading for Academic Purposes*, D. Eskey and F. Dubin (eds.) Reading, MA: Addison-Wesley, 25-47.

Gustaffson, J.E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33, 205-233.

Hale, G.A., Stansfield, G.W., Rock, D.A., Hicks., M.M., Butler, F.A., and Oller Jr., J.W. (1988). *Multiple-choice Cloze Items and The Test of English as a Foreign Language.* (RR88-2). Princeton, NJ: Educational Testing Service.

Hambleton, R.K., & Cook, L.L. (1983). The robustness of item response models and effects of test length and sample size on the precision of ability estimates. In *New Horizons in Testing*, D. Weiss (ed.). New York: Academic Press, 31-49.

Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*. 14, 75-96.

Hambleton, R.K., Dirir, M., & Des Brisay, M. (1993). New measurement models and methods for constructing language tests. *Carleton Papers in Applied Language Studies*, X, Carleton University Press.

Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In *Applications of Item Response Theory*, R.K. Hambleton (ed.). Vancouver, B.C.: Educational Research Institute of British Columbia.

Hambleton, R.K., & Rovinelli, R.J. (1973). A Fortran IV program for generating examinee response data from logistic test models. *Behavioral Science*, 17, 73-74.

Hambleton, R.K., & Traub, R.E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, 26, 195-211.

Hambleton, R.K., & Traub, R.E. (1971). Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 24, 273-281.

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

Hawkins, B. (1985). Is an "appropriate response" always so appropriate? In *Input in Second Language Acquisition*, S.M. Gass and C.G. Madden (eds.).

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing*, 1, 123-133.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory Application to Psychological Measurement*. Homewood, Ill.: Dow Jones-Irwin.

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.

Kingston, N.M., & Dorans, N.J. (1982). The feasibility of using item response theory as a psychometric model for the GRE test. (GRE Board Professional Report GREB No. 79-12P, ETS RR 82-12.) Princeton, N.J.: Educational Testing Service.

Klein-Braley, C. (1984). A cloze is a cloze is a question. In *Issues in Language Testing Research*, J.W. Oller Jr. (ed.). Rowley, MA: Newbury House Publishers Inc.

Kolen, J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 18, 1-11.

Lord, F.M. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (RR-84-30-ONR). Princeton, N.J.: Educational Testing Service.

Lord, F.M. (1983a). Small N justifies the Rasch model. In *New Horizons in Testing*, D. Weiss (ed.). New York: Academic Press, 51-61.

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

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, N.J.: Erlbaum.

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., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lumsden, J. (1980). Discussion: Session 7. In *Proceedings of the 1979 Computerized Adaptive Testing Conference*, D.J. Weiss (ed.). Minneapolis: University of Minnesota, Department of Psychology, 345-347.

McDonald, R.P. (1991). Testing for approximate dimensionality. Paper presented at the International symposium on Modern Theories in Measurement: Problems and Issues, Montebello, Quebec.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monograph*.

McKinley, R.L., & Mills, C.N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9, 49-57.

McKinley, R.L., & Reckase, M.D. (1980). *A comparison of the ANCILLES and LOGIST parameter estimation procedure for the three-parameter logistic model using goodness of fit as a criterion* (research report 80-2). Columbia, MO: University of Missouri, Tailored Testing Laboratory.

Meredith, W., & Kearns, J. (1973). Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. *Psychometrika*, 38, 533-554.

Mislevy, R.J., & Bock, R.D. (1986). *PC-BILOG: Item Analysis and Test Scoring with Binary Logistic Models* (computer program). Mooresville, IN: Scientific Software Inc.

Mislevy, R.J., & Bock, R.D. (1984). *BILOG: Item Analysis and Test Scoring: Logistic Model*. Mooresville, IN: Scientific Software.

Mislevy, R.J., & Bock, R.D. (1982). *BILOG: Maximum Likelihood Item Analysis and Test Scoring with Logistic Models for Binary Items*. Chicago: International Testing Services.

Nandakumar, R. (1991). Assessing the dimensionality of a set of items. Comparison of different approaches. Paper presented at the meeting of the AERA, Chicago, Ill.

Nandakumar, R. (1987). Refinement of Stout's procedure for assessing latent trait dimensionality. Unpublished doctoral dissertation. University of Illinois, Urbana-Champaign.

Oller Jr., J.W. (1979). *Language Tests at School*. London: Longman Group Limited.

Rasch, G. (1966). An individualistic approach to item analysis. In *Readings in Mathematical Social Science*, P.F. Lazarsfeld and N.W. Henry (eds.). Chicago: Science Research Associates, 89-107.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 4(3), 207-230.

Ree, M.J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement*, 3, 371-385.

Slinde, J.A., & Linn, R.L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement*, 16, 159-165.

Stocking, M.L. (1988). *Specifying Optimum Examinees for Item Parameter Estimation in Item Response Theory* (ETS RR 88-57-ONR). Princeton, N.J.: Educational Testing Service.

Stout, W.F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-325.

Stout, W.F. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.

Swaminathan, H. (1983). Parameter estimation in item response models. In *Applications of Item Response Theory*, R.K. Hambleton (ed.). Vancouver, B.C.: Educational Research Institute of British Columbia.

Swaminathan, H., & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349-364.

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

Swaminathan, H., & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175-191.

Swaminathan, H., & Gifford, J.A. (1979). Estimation of parameters in the three-parameter latent-trait model. *Laboratory of Psychometric and Evaluation Research (Report No. 90)*. Amherst, Mass.: University of Massachusetts, School of Education.

Tall, G. (1981). The possible dangers of applying the Rasch model to school examinations and standardized tests. In *Issues in Evaluation and Accountability*, C. Lacey & D. Lawton (eds.). London: Methuen, 189-203.

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

Theunissen, T.J.J.M. (1986). Text banking and test design. *Language Testing*, 3, 1-8.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In *Applications of Item Response Theory*, R.K. Hambleton (ed.). Vancouver, B.C.: Educational Research Institute of British Columbia.

Traub, R.E., & Lam, Y.R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology*, 36, 19-48.

Urry, V.W. (1976). Ancillary estimators for the item parameters of mental tests. In *Computers and Testing: Steps Towards the Inevitable Conquest* (PS-7C-1), W.A. Gorham (chair). Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, 14-18.

Vale, C.D., & Gialluca, K.A. (1985). ASCAL: a microcomputer program for estimating logistic IRT item parameters. St. Paul, MN: Assessment Systems Corporation.

van de Vijver, F.J.R. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement*, 10(1), 45-57.

Wainer, H., & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, 45, 373-391.

Way, W.D., & Reese, C.M. (1991). *An investigation of the use of simplified IRT models for scaling and equating the TOEFL test* (TR-90-29). Princeton, N.J.: Educational Testing Service.

Whitely, S.E. (1977). Models, meanings and misunderstandings: some issues in applying Rasch's theory. *Journal of Educational Measurement*, 14, 227-236.

Whitely, S.E., & Dawis, R.V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, 36, 329-337.

Whitely, S.E., & Dawis, R.V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 11, 163-178.

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., & 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, N.J.: Educational Testing Service.

Wood, R. (1978). Fitting the Rasch model: a heady tale. *British Journal of Mathematical and Statistical Psychology*, 21, 27-32.

Woods, A., & Baker, R. (1985). Item response theory. *Language Testing*, 2, 117-140.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

Wright, B.D., & Douglas, G.A. (1977). Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement*, 37, 573-586.

Wright, B.D., & Mead, R.J. (1976). BICAL: calibrating items with the Rasch model. *Research Memorandum*, 23, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17, 297-311.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 4, 245-262.

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

Appendix A

Table 1 **Classical Item Statistics for the Listening Subsets Given at Two Administrations**

Subset	Item	Time 1		Time 2	
		Difficulty	Biserial	Difficulty	Biserial
Computers	1	.42	.34	.39	.44
	2	.55	.56	.60	.39
	3	.66	.66	.61	.61
	4	.77	.68	.78	.70
	5	.78	.59	.76	.77
Mean		.64(A)*	.57	.63(S)	.58
Alpha		.573		.522	
Sleep Cycles	1	.60	.52	.58	.29
	2	.79	.72	.65	.59
	3	.33	.51	.27	.54
	4	.55	.63	.44	.59
	5	.43	.60	.34	.69
	6	.73	.32	.61	.24
Mean		.57(S)	.55	.48 (D)	.49
Alpha		.538		.522	

* Time of year of test administration: A = April; S = September; D = December.

Table 2 **Classical Item Statistics for Reading Subset "Lufts" Over three Administrations**

Item	Time 1 (S)		Time 2 (S)		Time 3 (S)	
	Difficulty	Biserial	Difficulty	Biserial	Difficulty	Biserial
1	.60	.47	.64	.53	.58	.58
2	.45	.56	.44	.49	.45	.53
3	.78	.72	.65	.63	.67	.52
4	.67	.52	.62	.64	.61	.69
5	.93	.70	.80	.69	.84	.70
6	.79	.22	.69	.40	.67	.30
Mean	.70	.53	.64	.56	.64	.55
Alpha	.560		.590		.554	

Table 3 **Classical Item Statistics for the Reading Subsets Given at Two Administrations**

Subset	Item	Time 1		Time 2	
		Difficulty	Biserial	Difficulty	Biserial
Hazardous Home	1	.62	.50	.62	.75
	2	.75	.73	.80	.69
	3	.68	.60	.69	.65
	4	.63	.45	.76	.52
	5	.53	.46	.54	.68
	6	.65	.65	.68	.61
	7	.45	.40	.63	.52
	Mean	.62(D)	.54	.67(A)	.63
	Alpha	.644		.364	
Volunteer Work	1	.55	.63	.51	.62
	2	.73	.58	.74	.64
	3	.80	.43	.79	.58
	4	.67	.60	.68	.66
	Mean	.69(D)	.56	.68(A)	.63
	Alpha	.494		.496	

Table 4 Classical Item Statistics for Cloze Test "Acid Rain" at Time 1 and Time 2

Item	Time 1 (S)		Time 2 (A)	
	Difficulty	Biserial	Difficulty	Biserial
1	.72	.38	.69	.28
2	.87	.44	.81	.33
3	.88	.65	.73	.82
4	.59	.65	.50	.70
5	.80	.71	.73	.59
6	.52	.79	.28	.76
7	.80	.60	.78	.54
8	.86	.80	.67	.61
9	.88	.61	.76	.67
10	.82	.80	.86	.66
11	.68	.78	.61	.65
12	.85	.80	.80	.81
13	.70	.85	.61	.79
14	.35	.57	.33	.46
15	.56	.52	.51	.51
16	.53	.39	.57	.21
17	.89	.53	.80	.54
18	.64	.81	.48	.83
19	.82	.39	.64	.53
20	.63	.57	.54	.61
21	.80	.80	.65	.65
22	.65	.88	.48	.73
23	.73	.84	.59	.64
24	.77	.37	.68	.44
25	.64	.84	.39	.64
26	.61	.59	.66	.41
27	.87	.48	.79	.48
28	.86	.75	.72	.68
Mean	.73	.65	.63	.59
Alpha	.891		.876	

Table 5 **Classical Item Statistics for Cloze Test "Mother Teresa" at Time 1 and Time 2**

Item	Time 1 (S)		Time 2 (A)	
	Difficulty	Biserial	Difficulty	Biserial
1	.33	.70	.26	.69
2	.43	.63	.35	.60
3	.48	.55	.47	.67
4	.78	.78	.67	.71
5	.73	.67	.58	.76
6	.77	.32	.73	.20
7	.61	.55	.62	.53
8	.52	.17	.50	.31
9	.89	.53	.83	.60
10	.86	.52	.77	.38
*11	.31	.31	.39	.46
*12	.40	.47	.28	.68
*13	.56	.26	.74	.45
14	.85	.82	.83	.58
15	.87	.86	.71	.83
16	.67	.60	.57	.70
17	.79	.85	.69	.83
18	.84	.76	.76	.81
19	.81	.80	.73	.60
20	.72	.42	.72	.51
21	.95	.81	.91	.83
22	.86	.56	.84	.46
23	.72	.87	.63	.59
*24	.13	.50	.46	.71
25	.74	.76	.60	.60
26	.82	.82	.77	.57
27	.41	.49	.38	.42
28	.87	.82	.83	.74
29	.88	.49	.83	.49
30	.89	.42	.87	.78
Mean	.68	.60	.64	.60
Alpha	.857		.888	

* These items had one or more distractors changed

Table 6 Classical Item Statistics for Cloze Test "Snowmobiles" at Time 1 and Time 2

Item	Time 1 (A)		Time 2 (S)	
	Difficulty	Biserial	Difficulty	Biserial
1	.42	.63	.48	.70
2	.79	.63	.80	.68
3	.64	.66	.64	.67
4	.59	.45	.57	.34
5	.31	.65	.28	.55
6	.86	.61	.88	.62
7	.73	.24	.72	.25
8	.36	.85	.36	.79
9	.50	.36	.59	.52
10	.87	.32	.83	.37
11	.50	.67	.54	.75
12	.79	.49	.81	.70
13	.72	.54	.70	.54
14	.46	.29	.37	.29
15	.64	.60	.70	.68
16	.47	.46	.44	.47
17	.61	.70	.64	.65
18	.64	.32	.64	.39
19	.35	.76	.45	.75
20	.54	.41	.66	.46
21	.46	.68	.54	.66
22	.49	.61	.55	.71
23	.67	.70	.72	.76
24	.57	.71	.41	.54
25	.58	.68	.56	.74
26	.47	.75	.49	.75
27	.67	.62	.63	.67
28	.70	.57	.70	.75
29	.60	.73	.64	.75
30	.68	.66	.72	.80
31	.78	.77	.77	.78
32	.65	.59	.65	.67
Mean	.60	.58	.61	.62
Alpha	.894		.891	