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.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

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.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Estimation of Parameters in Item Response Models of Psychological Testing

Weiming Li

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the requirements
for the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

May 1992

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

ISBN 0-315-80955-8

Canada

Abstract

Estimation of Parameters in Item Response Models of Psychological Testing

Weiming Li

In the decade of the 1970s, Item Response Theory (IRT) became the dominant topic for study by psychometricians. Three parameter models, (3-PL model), have received considerable attention because of their applicability to a variety of testing situation where the one- and two-parameter item response models may not be completely adequate.

There are three main approaches to parameter estimation in IRT. Joint maximum likelihood (MLE, Wingersky, 1983), Marginal maximum likelihood (Mislevy and Bock, 1981) and Bayesian approaches (BE, Swaminathan and Gifford, 1986). After assessing all these methods, the author found that some problems of these methods do not appear to have been completely solved.

In this thesis EDE, Experimental Design Estimate, is advocated. Orthogonal Designs can be constructed by Hadamard matrices & some other methods. Uniform Designs are constructed using number theory (Fang, 1980). With Orthogonal $L_{25}(5^6)$ and Uniform designs $U_{25}(25^{20})$ optimal estimators

of item parameters, can be obtained rapidly and easily in the sense of minimizing residuals. Using the goodness of fit as a criteria it is shown through one practical case study, that the residuals of EDE are almost the same as in MLE. But the EDE provides a considerable saving in computer time. From the view point of practice, especially for new test theory which involves more complex models, EDE has a potential applicability in the future.

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to Dr. Y. P. Chaubey for his guidance, encouragement and support in the preparation of this thesis.

CONTENTS

Chapter 1. Introduction

1.1. Standard Test Theory

The basic objective in psychological testing is quantification of certain aptitudes or abilities. An ideal test would be the one which would result in the same score for a person every time it is adminitrated. This score would be called "true" score of the person. But, in practice, to measure the true score of a person, is not an easy task. This necessitates proper design of tests and scoring methods, their evaluation and interpretation. The Classical Test Theory (CTT) and Item Response Theory (IRT) are the responses of the researchers in formulating some of the problems related with psychological testing.

1.1.a. Classical Test theory

In Classical Test Theory, a test score can be viewed as the sum of two components, a "true" score and a random "error" term. Two similar ("parallel") tests are considered to reflect the same true score, but disagree about an examinee's observed score because of the error component. Ideally decisions would be based on true scores; however in practice they must be based

on observed scores. "Reliability", the degree to which the unobservable true scores account for the variance in observed scores, gauges the accuracy with which a test ranks a group of examinees.

The model of Classical Testing Theory is:

$$x = t + e \tag{1}$$

Two unobservable constructs are introduced: true score t and error score e. They are linearly related. The true score for an examinee can be defined as his or her expected test score over repeated administrations of the test (or parallel forms). An error score can be defined as the difference between true score and observed score.

Classical Testing Theory postulates : error scores are random with a mean of zero and uncorrelated with error scores on a parallel test and with true scores:

1. $E(e)=0$;
2. $Corr(t,e)=0$;
3. $Corr(e_1, e_2)=0$, where e_1 and e_2 are error scores on two administrations of a test.

These assumptions can be met easily by most test data sets, and, therefore the models can and have been applied to a wide variety of test development and test score analysis problems. These assumptions also can be restated by :

for $< A, G, \pi, \{X_{ga}, a \in A, g \in G\} >$

$$\sigma^2(X_{ga}) = Var(X_{ga}) < +\infty \tag{2}$$

where A is the set of persons (population of subjects); G is the class of tests under study; π is a probability distribution on A; X_{ga} is a family of random variables. It is assumed that if $a \neq a'$, then for arbitrary g, g' in G (distinct or not), the random variables X_{ga} and $X_{g'a'}$ are independent (Nowakowska, 1983). Today there are countless number of achievement, aptitude, and personality tests that have been constructed with these models.

There might be some shortcomings for using CTT in practical situations: 1) The values of commonly used item statistics in test development, such as item difficulty (proportion of right answers) and item discrimination (correlation between item score and the total score), depend on the particular examinee samples in which they are obtained. The result is that these item statistics are useful only in item selection when constructing tests for examinee population that are very similar to the sample of examinees in which the item statistics were obtained. For example, item discrimination indices tend to be high when estimated from an examinee sample which is heterogeneous in ability than from an examinee sample which is homogeneous in ability, because of the well-known effect of group heterogeneity on correlation coefficients (Lord and Novick, 1968).

2) Increased test score validity can be obtained when the test difficulty

is matched to the approximate ability level of each examinee (Lord, 1980 ; Weiss, 1983). However when several forms of test that vary substantially in difficulty are used, the task of comparing examinees becomes a difficult problem.

3) One of the fundamental concepts, test reliability, in CTT is defined in terms of parallel forms. This concept of parallel measures is difficult to achieve in practice (Hambleton and van der Linden, 1982). Researchers must be content with either lower-bound estimates of reliability or reliability estimates with unknown biases.

4) CTT provides no basis for determining how an examinee might perform when confronted with a test problem. Such information is necessary for test designers.

5) CTT presumes the variance of errors of measurement is the same for all examinees. It is not uncommon to observe that the performance of high-ability on several parallel forms of a test might be expected to be more consistent than the performance of medium-ability examinees.

Therefore psychometricians have been concerned with the development of more appropriate theories of measurements. Presently, perhaps the most popular set of constructs, models, and assumptions for inferring traits is Latent Trait Theory. Considerable attention is being directed currently toward the field of Latent Trait Theory or ITEM RESPONSE THEORY as Lord

(1980) prefers to call this theory.

1.1.b. Item Response Theory

In Item Response Theory, an individual score is made up of the total sum of "item" scores which can be accounted for to a substantial degree by certain parameters depending on various traits, called latent traits. Thus, in IRT, scores on n items are considered to be distributed according to some probability law characterized by these parameters

There are three primary advantages of item response theory models:

1) Assuming the existence of a large pool of items all measuring the same trait, the estimate of an examinee's ability is independent of the particular sample of test items that are administered to the examinee;

2) Assuming the existence of a large population of examinees, the descriptors of a test item, e.g. item difficulty and discrimination indices, are independent of the particular sample of examinees draw for the purpose of calibrating the item;

3) A statistic indicating the precision with which each examinee's ability is estimated is provided. Some shortcomings of CTT can be overcomed by IRT.

There are many mathematical models that have been used in the analysis of educational and psychological test data set in IRT. Each model consists of: 1) an equation linking (observable) examinee item performance and a latent (unobservable) ability and 2) several assumptions.

For dichotomous data, there are Latent Linear (Lazarsfield and Henry, 1968), Perfect Scale (Guttman, 1911), Latent Distance (Lazarsfield and Henry, 1968), One-, Two-, Three- Parameter Normal Ogive (Lord, 1952) and One-, Two-, Three- Parameter Logistic (Birnbaum, 1957, 1958a, 1958b, 1968; Lord, 1980; Rasch, 1960; Wright and Stone, 1979), Four- parameter Logistic (Barton and Lord, 1981) models.

For 3-PL (three parameter logistic) model, the equation linking θ and item parameters a, b, c is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{3}$$

where i=1, 2, ...n;

$P_i(\theta)$= the probability that an examinee with ability level θ answers item i correctly;

D=1.7, a scaling factor;

b_i= the item difficulty parameter;

a_i= the item discrimination parameter (it should not be confused with the notation of a person in page 2);

c_i is a pseudo-chance level parameter, represents the probability of examinees with low ability correctly answering the item.

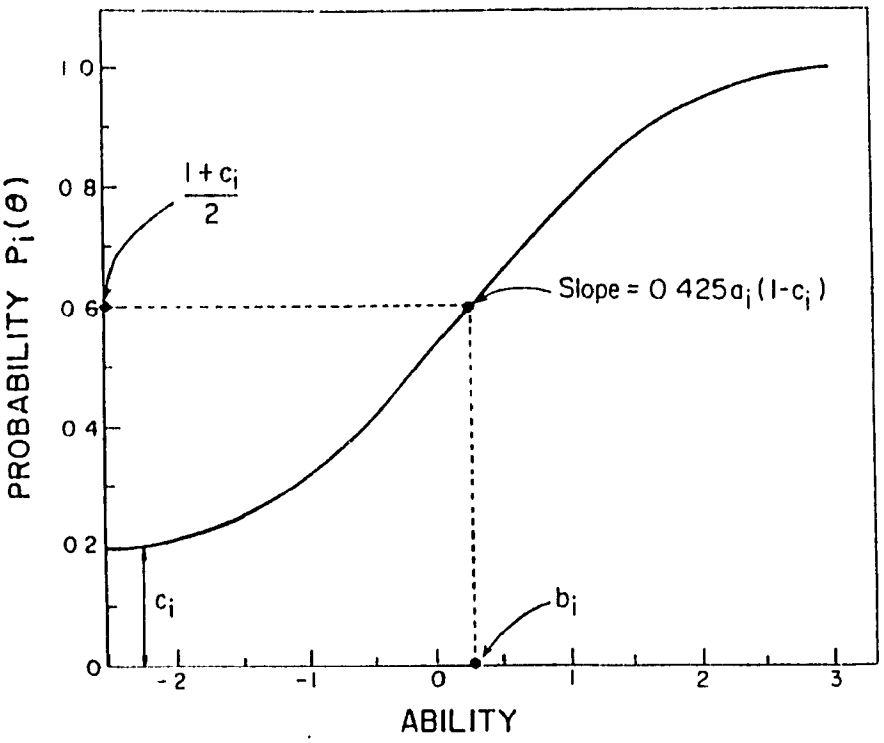

Figure 1: A Typical 3-PL Model Item Characteristic Curve

Figure 1 provided a typical 3-PL item characteristic curve.

When c=0, it reduces to Two-parameter Logistic model:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{4}$$

If c=0 and all b's are equal, it is Rasch Model (Rasch, 1966).

The three-parameter item response model has received considerable attention because of its applicability to a variety of testing situation where the one- and two-parameter item response model may not be completely adequate (Loyd and Hoover, 1980; Slinde and Linn, 1979). In a recently equating study of Test of English as Foreign Language (TOEFL), the result

of the study clearly indicated that the 3-PL model performed better than 2-PL and 1-PL models (Way and Reese, 1990).

The logistic function is chosen as an alternative to normal ogive models, i.e.

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \tag{5}$$

due to its more convenient mathematical properties.

Sometimes we use 1-PL models.

$$P_i(\theta) = c_i + (\gamma_i - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{6}$$

This model differs from 3-PL in that γ_i assumes a value slightly below one.

In Item Response Theory examinees may be compared with each other even when they may have taken quite different subtests of items. A further consequence of the fact that ability can be estimated independently of the choice of items is that equating scores of tests is possible. In addition, the problem of constructing parallel forms of tests is eliminated. It is possible to measure the precision of the ability estimates at each ability level. Thus instead of providing a standard error of measurement that applies to all examinees regardless of test scores, separate estimates of error for each examinee can be produced. Finally, item parameters based on IRT are invariant across different subgroups of examinees. Consequently, they are of

immense value to test developers.

In the decade of 1970s, IRT became the dominant topic for study by psychometrists. But like Einstein's theory of relativity revolutionized physics, it extended rather than supplanted Newton's laws of motion, classical mechanics still works just fine, CTT will continue to be used.

1.2. Statistics and Test Theory

It is only a slight exaggeration to describe the test theory that dominates educational measurement today as application of twentieth century statistics to nineteenth century psychology. The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory (Mislevy, 1989).

Many of the problems in test theory are essentially problems of multivariate analysis in mathematical statistics (Gulliksen, 1961). Much of the recent progress in test theory has been made by treating the study of the relationship between response to a set of test and a hypothesized trait (or traits) of an individual as a problem of statistical inference.

There is an event in the recent history of test theory which has a major impact on the field: the appearance of Lord and Novick's (1968) *Statistical theories of mental test scores* with contributions by A. Birnbaum where a comprehensive and integrated coverage of test theory topics was provided,

in which the fundamental assumptions and essentially stochastic nature of the subject were stressed (Lewis, 1986). For the first time, researchers in the test theory had available the common language of modern statistical inference with which to discuss all aspects of their field. Bock and Wood (1971), in their review of test theory, referred to the book as "a great step forward" (pp. 194).

IRT, with which we are concerned in this thesis, is a practical advance beyond CTT but much more complicated because it is more difficult to work with from the statistical analytical point of view. There are several unsolved statistical issues that need further investigation, e.g. robustness, fit of the model to data etc. Among these a major problem that remains to be solved is that of estimation of parameters in IRT as well as its robustness.

There are three main approaches to parameter estimation in IRT. The quantity typically maximized by each approach is shown below for a test of n items administered to N examinees. $P_i(\theta_j)$ is the probability of success on item i for examinee j at ability level θ_j, $Q_i(\theta_j) = 1 - P_i(\theta_j)$, U_{ij} is the response of examinee j on item i, assumed here to be either 0 or 1, and $g()$ denotes a prior distribution of parameters.

1) Joint maximum likelihood (Wingersky, 1983). It maximizes

$$L(\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}_j) = \prod_{j=1}^{N} \prod_{i=1}^{n} [P_i(\theta_j)]^{u_{ij}} [Q_i(\theta_j)]^{1-u_{ij}}$$

or equivalently

$$logL(\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} logP_i(\theta_j) + (1 - u_{ij}) logQ_i(\theta_j)]$$

2) Marginal maximum likelihood of item parameters (Bock and Aitkin, 1981; Bock and Lieberman, 1970). It maximizes

$$L(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{j=1}^{N} \int_{-\infty}^{\infty} g(\theta_j) L(\theta_j; \mathbf{a}, \mathbf{b}, \mathbf{c}) d\theta_j$$

3) Bayesian approach (Swaminathan and Gifford, 1986). It maximizes

$$f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = L(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) g_1(\theta) g_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$$

or equivalently

$$logf(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = logL(\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) + logg_1(\theta) + g_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$$

There are some problems that do not appear to have been completely solved.

1.3. Outline of This Thesis

In this thesis we are mainly concerned with the estimation of parameters in IRT models, especially the 3-PL model. A Experimental Design Estimate (EDE) is proposed in this thesis after assessing established approaches of estimating parameters in IRT. EDE might become a widely utilized methodology of estimating parameters in latent trait theory. EDE does compare

favorably with MLE because of considerable saving in size of sample and computer cost, ensuring estimators within the parameter space, and having the same degree of goodness of fit.

In Chapter 1, the standard test theory (CTT and IRT) was reviewed very briefly. The hypothesis and different mathematics models were described in Section 1.1. Three main approaches of estimation of parameters were introduced in Section 1.2.

In Chapter 2 the methods of estimating parameters of three-parameter models are emphasized, because of their applicability to testing situations where the one- and two-parameter item response models may not be completely adequate.

Some problems associated with established approaches of estimating parameters do not appear to have been completely solved. The MLE is consistent, efficient and sufficient. But it does not give an $\hat{\theta}$ for a student who responds incorrectly (or correctly) to all items; it does not ensure that estimators remain in the parameter space; sometimes the procedure is divergent. (Section 2.2) Bayesian estimation produces better estimate than MLE by the criteria of "mean squared errors" between estimates and true values. But poor specification of priors may adversely affect the estimates. Sometimes we have to increase the bias of $\hat{\theta}$ in order to reduce the MSE (Section 2.3). The optimization techniques and some miscellaneous approaches are described in Section 2.4 and 2.5 respectively.

In Chapter 3, the procedure of EDE is illustrated in detail. At first the main idea of Three-Stage Designs is described in Section 3.2. Then two different kinds of statistical experimental designs , Orthogonal Design and Uniform Design, are introduced (Sections 3.3. and 3.4.). We mention here how to construct these designs using Hadamard matrices, orthogonal Latin squares and number theory, what their advantages are and where the factors should be put on.

The steps of getting EDE are described in Section 3.4 and a practical case study ($N = 3208, n = 105$) is illustrated in Chapter 4. The test of unidimension of the data set of MET87 is treated in Section 4.1. The ratio of the first two eigen values of the tetrachoric correlation matrix (105×105) shows that it meets minimal criteria (Reckase, 1979). With $L_{25}(5^6)$ and $U_{25}(25^{20})$ "optimal" estimate of item parameters, $\hat{a}, \hat{b}, \hat{c}$, can be reached without any real "experiments" (Section 1.2).

In Section 1.3 we compare these two different kinds of designs according to their mean square error. We found the precision and the efficiency to be almost same. Using the criterias of minimum square error and standardized residuals we conclude that EDE does compare favorably with MLE (Section 1.1).

Chapter 2. Parameter Estimation in IRT

2.1. Introduction

The latent trait models, of Item Response Theory (IRT), have numerous advantages over the classical test models. Perhaps the most important advantage of IRT is that it is possible to estimate an examinee's ability on the same ability scale from any subset of items that have been fitted to the model. This implies that the ability of an examinee can be estimated independently of the particular choice of the number of items and hence represents a major breakthrough in the area of mental measurement.

Because of IRT's advantages it's applications include such academic areas as reading achievement (Rentz and Bashaw, 1977; Woodcock, 1974); psychological variables (Woodcock, 1978); and mathematics, geology, and biology (Soriyan, 1977; Connolly, Nachtman etal, 1974). Lord (1968a, 1977) and Marco (1977) described the application of three parameter logistic model to the analysis of such tests as Verbal Scholastic Aptitude Test (SAT), the Mathematics sections of the Advanced Placement Program (APP), and the College Level Examination Program (CLEP). Yen (1981, 1983) described the application of 3-PL model to the development of the California Tests of Basic

Skills.

Equating of tests based on raw series in CTT is not desirable for reasons of equity, symmetry, and invariance. Equating based on IRT overcomes these problems (Kolen, 1981). IRT appears to be especially useful in test design (or redesign). Utilizing IRT we can select test items to fit target curves, establish item banks, evaluate test score prediction systems, detect item bias, estimate power scores and construct adaptive tests. In educational assessment, IRT makes it possible to establish a stable measurement while allowing assessment instruments to evolve over time (Mislevy, 1989).

Most of the work in IRT has used ability and achievement tests, i.e. in the area of educational and psychological measurements. Some researchers utilized it for attitude surveys, such as Job Description or Supervisory Attitude Survey (Waller,1981; Parsons and Hulin, 1982).

For all these applications of IRT, estimation of parameters in these models is the most important task.

2.2. Maximum Likelihood Estimate

There are currently three main approaches to parameter estimation in item response theory. (Lord, 1986)
1) Joint maximum likelihood, yielding maximum likelihood estimate (MLEs) (Wingersky, 1983).

2) Marginal maximum likelihood (Mislevy and Bock, 1981).

3) Bayesian approaches, in which parameter estimates are usually the mode (or mean) of the posterior distribution of the parameter estimated (Swaminathan and Gifford, 1986).

The probability that an examinee with ability θ obtains a response U_i on item i, where

$$U_i = \begin{cases} 1 & \text{for a correct response} \\ 0 & \text{for an incorrect response} \end{cases}$$

is denoted by $P(U_i|\theta)$. It also can be expressed as

$$P(U_i|\theta) = P(U_i = 1|\theta)^{U_i} P(U_i = 0|\theta)^{1-U_i} = P_i^{U_i} Q_i^{1-U_i} \tag{7}$$

where $Q_i = 1 - P_i$.

If the latent space is complete (in this case, unidimentional), then local independence is obtained, that means the likelihood function is

$$L(\mathbf{u}) = \prod_{j=1}^{N} \prod_{i=1}^{n} P_i^{u_{ij}} Q_i^{1-u_{ij}} \tag{8}$$

When N examinees take a test that has n items, in the 3-PL model, there are N+3n-2 parameters to be estimated because of the "indeterminacy" of scale. Under the transformations

$$\theta^* = \frac{\theta + m}{l}$$

$$b_i^* = \frac{b_i + m}{l}$$

$$a_i^* = l a_i$$

the response function is invariant. Hence we fix the θ's such that their mean is zero and standard deviation is one. It means two constraints have to be imposed.

The logarithm of the likelihood function is

$$\ln L(\mathbf{u}|\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{j=1}^{N} \sum_{i=1}^{n} [u_{ij} \ln P_{ij} + (1 - u_{ij}) \ln Q_{ij}] \qquad (9)$$

where $\mathbf{u}$ is an Nn dimensional vector, θ, and a, b, c are N and n dimensional respectively. Using a multivariate version of the Newton-Raphson procedure we can get the numerical values of the estimators.

The MLE is consistent, sufficient, efficient and asymptotically distributed normally. However:

1) MLEs do not exist for those students who respond incorrectly (or correctly) to all the items.

2) Sometimes there are several maxima in the interval $-\infty < \theta < \infty$. This was first noted by Samejima (1973). It may also happen if the value of the likelihood function at $\theta = -\infty$ or $\theta = \infty$ is larger than the maximum value found in the interval. In the practical applications that Lord (1980, p59) studied, he found that multiple solutions did not occur when the number of items was $\geq$ 20. But in a recent study (Yen, etal, 1991) fourteen multiple-choice achievement tests with from 20 to 50 items were examined, from 0 to 3.1% of them had response vectors with multiple maxima.

3) MLE does not ensure that estimates remain in the parameter space. In some simulations, there are 6 of 35 out of [0, 10] for $\hat{a}$ (Swaminathan and Gifford, 1986).

4) The procedure might be divergent (Swaminathan and Gifford, 1986).

5) Joint MLEs of ability parameters may be biased. This then causes the item parameters to be misestimated. (Lord, 1986)

6) In some cases for 3-PL model (see Equation 3, Chapter 1), e.g. $b = -2$, $s.e. \cong 0.3$, the desired sample size is too large to be of practical use (N=100,000). There is no computer program currently available that will fit the 3-PL model with 100,000 examinees. If the dimensions of existing program were raised to allow data of such magnitudes, some researchers conjecture that it would take the annual revenue of Saudi Arabia to pay such a run (Wainer and Thissen, 1982).

These problems of MLE do not appear to have been completely solved. Bayesian procedures for estimating parameters have been successfully applied in numerous situations.

2.3. Bayesian Approach

Let the joint density of the parameters θ, **a**, **b**, **c** be denoted as $f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c})$. It follows from Bayes' Theorem that the conditional density, $f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}|\mathbf{u})$ is

$$f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}|\mathbf{u}) \propto L(\mathbf{u}|\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) f(\theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) \tag{10}$$

Assume a priori that the parameter vectors θ, **a**, **b**, **c** are independent. Using the hierarchical model, we have

$$\theta_i | \mu_\theta, \sigma_\theta^2 \sim N(\mu_\theta, \sigma_\theta^2) \tag{11}$$

similar iid and normality assumption are made for the parameter b_i:

$$b_i | \mu_b, \sigma_b^2 \sim N(\mu_b, \sigma_b^2) \tag{12}$$

Priors for μ_b and σ_b^2 are specified by assuming that μ_b is uniform and that σ_b^2 has an inverse chi-square distribution with parameters ν_b and λ_b. (Novick and Jackson, 1974, p109)

The prior distribution of a_i can be taken to be the chi square distribution. That is

$$f(a_i|\nu_i,\omega_i)da_i \sim a_i^{\nu_i-1}exp(\frac{-a_i^2}{2\omega_i})da_i \tag{13}$$

The prior for c_i may be taken as the beta distribution with parameters s_i and t_i, assuming that priori $c_1, c_2, ...c_n$ are independent. The Bayesian procedure ensures that the estimates stay in the parameter space. It produces better estimates than the MLE as judged by such criteria as mean squared differences between estimates and true values.

However, Bayesian Estimation (BE):

1) requires specification of the prior regarding an examinee's ability and hence may not be appealing to all.

2) may increase estimation bias, in order to reduce the MSE, minimizing the overall mean square error, of ability parameters. See Figure 3 (Lord, 1986).

3) From the view point of practice, minimizing the MSE is not appropriate (Lord, 1986).

2.4. Optimization Techniques

All the above approaches can be viewed as different types of optimiza-

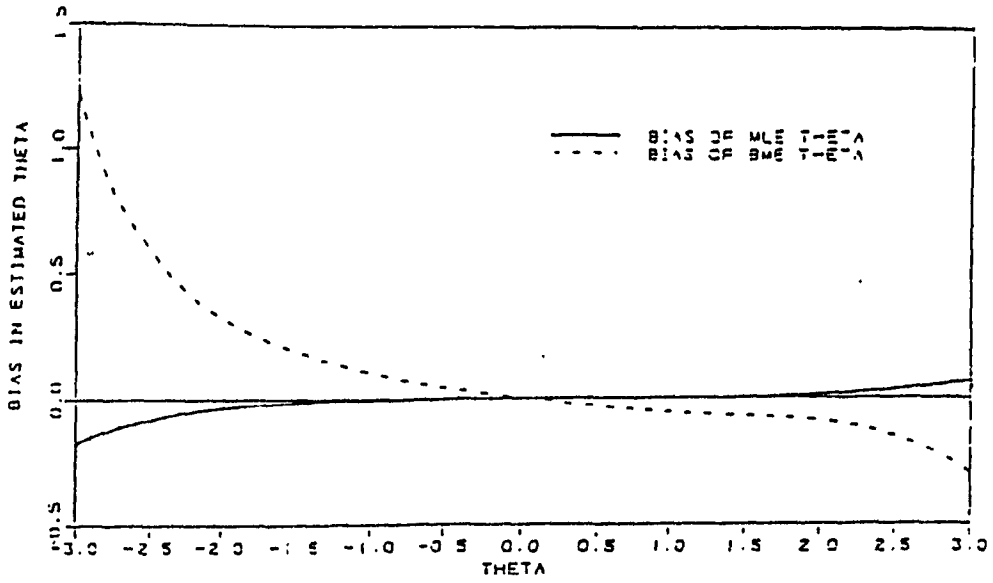

Figure 2: Bias in estimated ability for a 90-item SAT Verbal Test

tion process with different constraints (Suen and Lee, 1989). For 3-PL model, MLE is equivalent to maximize objective function, i.e. maximum likelihood function $L(\mathbf{u}|\theta, \mathbf{a}, \mathbf{b}, \mathbf{c})$, subject to:

a) $\sum_{j=1}^{N} \theta_j = 0$,

b) $\sum_{j=1}^{N} (\theta_j)^2/N = 1$.

For Bayesian estimate, the resulting optimal problem becomes: minimizing

$$f(\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) = L(\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) g_1(\theta) g_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$$

where $g_1(\theta)$ and $g_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$ be the prior distribution of ability and item parameters. Similarly, if a marginal distribution $h(\theta)$ is imposed, it is exactly the marginal likelihood estimate.

There are several constrained nonlinear optimization algorithms available today. GRG-II (Liebman et al. 1986) and MINOS (Murtagh and Saunders, 1987) may be appropriate. But these algorithms can only be used for a very small number of subjects with a short test.

2.5. Miscellaneous Approaches

There are some other kinds of estimating methods, e.g. Approximate Estimation (Urry, 1976, 1977), Kernel Smoothing Approaches (Ramsey, 1991) and Golden Section Search Strategies (Xiao, 1989)

The Approximate Estimates are often useful and provide a considerable saving in computer costs, because of obtaining these estimates may be time consuming and costly in three parameter model.

Under the assumption that (1) the ability is normally distributed with zero mean and unit variance and (2) the model is a two parameter normal ogive, Lord and Novick (1968, pp.377-378) have shown that the biserial correlation between θ and the response U_i to item i, $\rho'_{i\theta}$ is given by:

$$\rho'_{i\theta} = a_i/[1 + a_i^2]^{1/2} \tag{14}$$

where a_i is the discrimination index of item i. Moreover, if γ_i is the normal deviate that cuts off to th e right an area π_i, where π_i is the proportion of examinees who respond correctly to item i, then it is given by:

$$\gamma_i = \rho'_{i0} b_i \tag{15}$$

where b_i is the difficulty of item i.

Unfortunately, ρ'_{i0} cannot be obtained directly. However, it can be shown that the point biserial correlation between the binary scored response to item i and ability θ are related according to

$$\rho_{i0} = \frac{\rho'_{i0}\phi(\gamma_i)}{\pi_i(1-\pi_i)^{1/2}} \tag{16}$$

where $\phi(\gamma_i)$ is the ordinate at γ_i. In order for this to be a reliable estimate, there must be at least 80 items and the KR-20 reliability must be at least 0.90 (Schmidt,1977).

In three parameter model, the item difficulty π'_i is used

$$\pi'_i = c_i + (1-c_i)\pi_i \tag{17}$$

Thus,

$$\rho_{i0} = \rho'_{i0}\phi(\gamma_i)(1-c_i)/(\pi'_i - c_i)(1-\pi'_i)^{1/2} \tag{18}$$

Swaminathan and Gifford (1983) and McKinley and Reckase (1980) have demonstrated that approximations do not compare favorably with the maximum likelihood procedure unless the numbers of examinees and items are very large.

Xiao (1989) suggested Golden Section Search Strategies (GSSS) in computerized adaptive testing (CAT) . She used the golden section ratio t :

$$t = \frac{\sqrt{5}-1}{2} \approx 0.6180339989 \tag{19}$$

to find the optimal solution of ability parameters. She finds GSSS can provide more accurate ability estimates than MLE with the exception of a few very high ability levels, that GSSS is more robust against random guessing than MLE, more effective and cheaper to use.

Chapter 3. Experimental Design Estimate

3.1. Introduction

Birnbaum's (1968) three parameter logistic model, 3-PL logistic, has become a common basis for item response theory modeling, especially within situations where significant guessing behavior is evident. From an analytical point of view, the 3-PL is quite difficult to work with, and as a consequence, some analytical difficulties can also translate into practical problems. Some researchers tried to find improved models.

Pashley (1991) proposed an alternative three-parameter logistic model. He called it *hyperbolic* 3-PL or *hyperbolic* 1-PL. One of the advantages of this model being that it may stabilize related estimation procedure (Lewis, 1990).

His basic idea is using a hyperbola that resembles to the logit transformed 3-PL curve, because the hyperbola exhibits a shape very similar to it. The logic transformat on of 3-PL is

$$\lambda(\theta) = \ln[\frac{P(\theta)}{1 - P(\theta)}] \tag{20}$$

where $P(\theta)$ is defined as in (1).

A general equation for a hyperbola is given by

$$\frac{Z^2}{s^2} - \frac{W^2}{r^2} = 1 \tag{21}$$

where Z and W denote the axis coordinates; and s and r are parameters which define the shape of the curve.

Two more transformations are needed in order for this curve to resemble the logit transformed 3-PL. The first transformation is:

$$Z = Y \cos\alpha - X \sin\alpha; \quad W = Y \sin\alpha + X \cos\alpha. \tag{22}$$

where $\alpha = \tan^{-1}(\frac{s}{r})$; X and Y denote the new coordinates.

The second transformation is :

$$X = \theta - h; \quad Y = \lambda(\theta) - k \tag{23}$$

The result is:

$$\lambda(\theta) = f[\theta - h + \sqrt{(\theta - h)^2 + g}] + k \tag{24}$$

where *f, h and k* are similar to a, b and c respectively in 3-PL Logistic model.

Another new model is based on proportional item response curve (PIRC). In this model, as with the others, it is assumed that the probability of item success is a function of ability level but that the form of that function is the same for all items, except for a constant of proportionality. For example,

if, for a given examinee, the probability of success on an item is half that on second item, the probability of success on the first item is the half the probability of sucess on the second item for all examinees at all levels of ability. The value of one-half used here is only by way of example. Any other fraction could have been used.

There is a *general function* of ability θ, $f(\theta)$. For any particular item, one parameter E_j takes a role of the proportion constant. The estimation is much easier and studies found it is consistent with TOEFL item correlations (Boldt, 1989) and about equally accurate for prediction with 3-PL (Boldt, 1991).

This thesis will not attempt to improve the models, but concentrate how to improve the methods of estimating item parameters in 3-PL models. Because the problem of estimating ability parameters when item parameters are given is reasonably straightforward. It is called *conditional estimation* of θ.

Secondly, estimation of item parameters is the most important task in Adaptive Testing and constructing Item Bank. If item parameters are available a collection of items, tests can be constructed for optimal performance in specific applications such as minimizing classification errors.

With experimental designs the "optimal" estimate of item parameters can be reached in the sense of minimizing residuals without doing any real

"experiments". These estimators are called *Experimental Design Estimate*, EDE, in this thesis. The main idea comes from *Three-stage Designs* which has been used in many countries of the world.

3.2. Parameter Design

In order to gain as much information as possible researchers must plan experiments very carefully in advance. This plan is often refer to as *experimental design*. These variables which can effect the performance of "product", dependent variable y, are called *factors*. Each factor was set at some values which were called *levels*.

Off-line quality control methods are the measurements taken at the product and process design stages to improve product quality. G. Taguchi (1979) has developed a systematic approach to off-line quality control that has been used to a moderate extent in Japan and has attracted attention in a number of other counties, including the United States.

Rather than attempt to find and control some noise factors, Taguchi advocates a three-stage design procedure for off-line quality control : (a) system design; (b) parameter design and (c) tolerance design. In the first stage, a system is designed to fulfill a specific function. The second step, parameter design, attempts to find levels of the controllable factors such that some optimal conditions are reached. In the third stage, it may be necessary to specify narrower tolerances for some of the factors. This final step is con-

sidered only if the reduction in variation achieved at the parameter design stage is insufficient. Thus, parameter design is the key stage for statisticians. We use the same idea to find "optimal" values of item parameters within a parameter space such that the minimum residual can be reached with some experimental designs.

Taguchi himself recommends that orthogonal arrays be used for parameter designs. In the following sections, we will discuss this kind of designs as well as others.

3.3. Orthogonal Designs

An orthogonal design of order n and type $(s_1, s_2, ..., s_l), s_i$ positive integers, is an $n \times n$ matrix $\mathbf{X}$, with entries from $0, \pm x_1, \pm x_2, ...$ satisfying

$$\mathbf{X}\mathbf{X}' = (\sum_{i=1}^{l} s_i x_i^2)\mathbf{I}_n \tag{25}$$

(Geramita and Seberry, 1979). The ideas and methods which we use to construct an orthogonal design are quite varied, and many have been used in the construction of Hadamard matrices.

A matrix whose elements are ± 1 and inner-product of any pair of columns is zero, is called *Hadamard Matrix*. For example following is a Hadamard matrix $\mathbf{H}_4$.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

In fact after omitting the first column, all elements are 1, it is an orthogonal design $L_4(2^3)$, where 4 is the number of trials, 2 is the number of levels and 3 is the number of columns.

Suppose we want construct a orthogonal design which has t levels, t is a prime number, *Orthogonal Latin Square* is used. Generally speaking we can use orthogonal latin square to construct $L_{t^2}(t^m)$ designs. At first we set two fundamental columns as following, we get

$$\begin{pmatrix} 1 & 1 & a_{11} & b_{11} & \cdots & c_{11} \\ 1 & 2 & a_{12} & b_{12} & \cdots & c_{12} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t & a_{1t} & b_{1t} & \cdots & c_{1t} \\ 2 & 1 & a_{21} & b_{21} & \cdots & c_{21} \\ 2 & 2 & a_{22} & b_{22} & \cdots & c_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 2 & t & a_{2t} & b_{2t} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t & 1 & a_{t1} & b_{t1} & \cdots & c_{t1} \\ t & 2 & a_{t2} & b_{t2} & \cdots & c_{t2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t & t & a_{tt} & b_{tt} & \cdots & c_{tt} \end{pmatrix}$$

These square matrices $A, B...$ and C are orthogonal to each other, where $A, B...$ and C are respectively of order t latin squares, for example: A is

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1t} \\ a_{21} & a_{22} & \cdots & a_{2t} \\ \vdots & \vdots & \vdots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tt} \end{pmatrix}$$

The elements of 3rd, 4th...and t-th columns are the elements of $A, B, ...C$ respectively according to raw order. The number of every level in each column equals to t. For each pair of columns the numbers of (i, j), where $i, j = 1, 2, ...t$, all are same. Following is $L_{25}(5^6)$:

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 2 & 2 & 2 & 2 & 2 \\
1 & 3 & 3 & 3 & 3 & 3 \\
1 & 1 & 1 & 1 & 1 & 4 \\
1 & 5 & 5 & 5 & 5 & 5 \\
2 & 1 & 2 & 3 & 1 & 5 \\
2 & 2 & 3 & 1 & 5 & 1 \\
2 & 3 & 1 & 5 & 1 & 2 \\
2 & 1 & 5 & 1 & 2 & 3 \\
2 & 5 & 1 & 2 & 3 & 4 \\
3 & 1 & 3 & 5 & 2 & 4 \\
3 & 2 & 1 & 1 & 3 & 5 \\
3 & 3 & 5 & 2 & 1 & 1 \\
3 & 1 & 1 & 3 & 5 & 2 \\
3 & 5 & 2 & 1 & 1 & 3 \\
1 & 1 & 1 & 2 & 5 & 3 \\
1 & 2 & 5 & 3 & 1 & 4 \\
1 & 3 & 1 & 1 & 2 & 5 \\
1 & 1 & 2 & 5 & 3 & 1 \\
1 & 5 & 3 & 1 & 1 & 2 \\
5 & 1 & 5 & 1 & 3 & 2 \\
5 & 2 & 1 & 5 & 1 & 3 \\
5 & 3 & 2 & 1 & 5 & 4 \\
5 & 1 & 3 & 2 & 1 & 5 \\
5 & 5 & 1 & 3 & 2 & 1
\end{pmatrix}$$

3.4. Uniform Design

Uniform Design is proposed by Fang (1980). The most important feature of it is that much more levels of factors can be arranged in an experiment. It still keeps some advantages of orthogonal design, e.g. experimental points are distributed uniformally over the factor space. So these points have a very good representive property.

Following is $U_{25}(25^{20})$. Comparing with $L_{25}(5^6)$, if we have two factors, Figure 3 shows that U_{25} is more uniformal than L_{25}, because every level of each factor occurs once in U_{25} and, in particular five levels repeat 5 times in L_{25}.

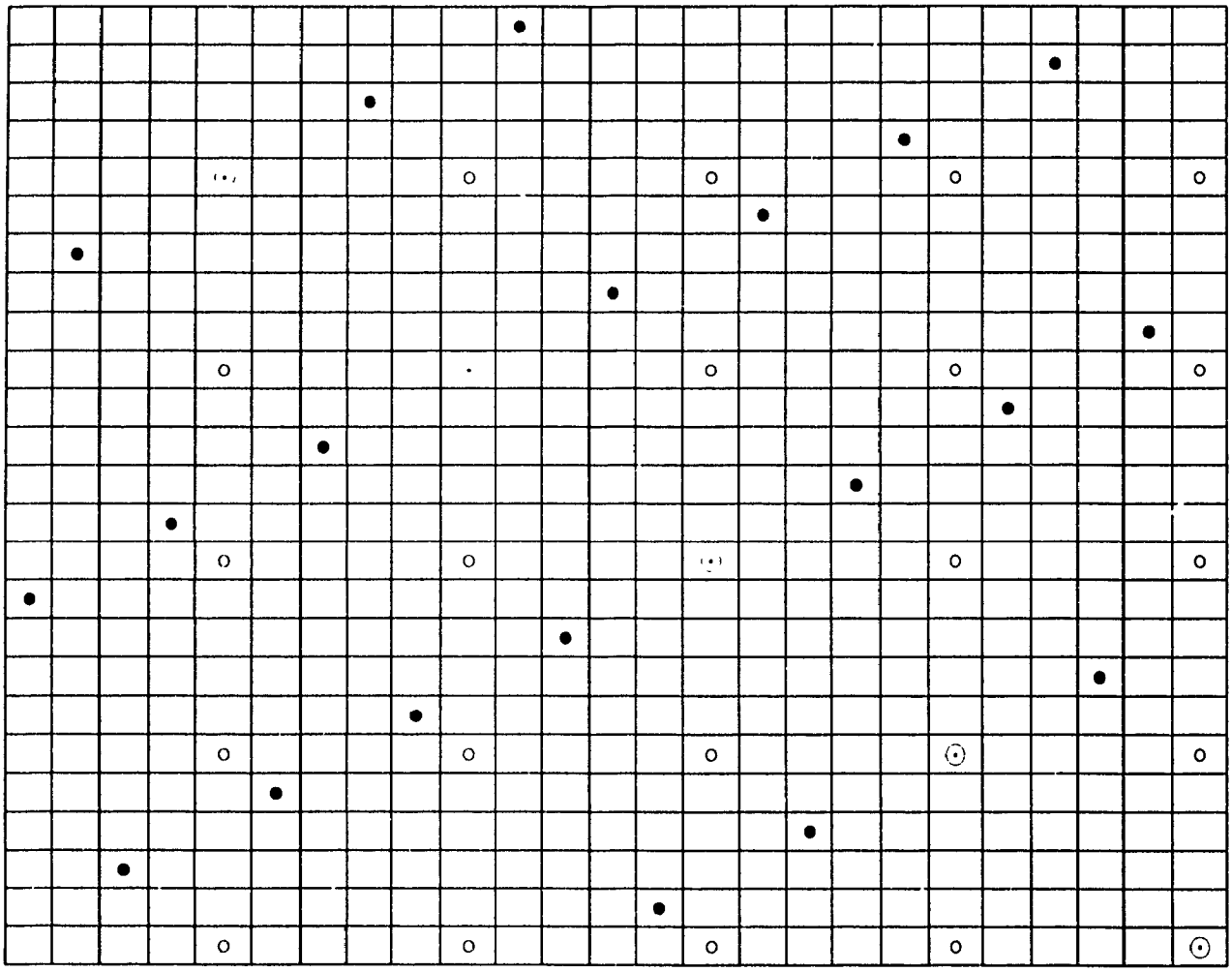

Figure 3. $U_{25}(25^{20})$ and $L_{25}(5^6)$, two factors

In Figure 3, symbol " • " indicates an experimental point in Uniform Design; symbol " ∘ " indicates an experimental point in Orthogonal Design; and " ⊙ " indicates a point in both.

1	2	3	4	6	7	8	9	11	12	13	14	16	17	18	19	21	22	23	24
2	4	6	8	12	14	16	18	22	21	1	3	7	9	11	13	17	19	21	23
3	6	9	12	18	21	24	2	8	11	14	17	23	1	4	7	13	16	19	22
4	8	12	16	24	3	7	11	19	23	2	6	11	18	22	1	9	13	17	21
5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20	5	10	15	20
6	12	18	24	11	17	23	1	16	22	3	9	21	2	8	14	1	7	13	19
7	14	21	3	17	24	6	13	2	9	16	23	12	19	1	8	22	4	11	18
8	16	24	7	23	6	14	22	13	21	1	12	3	11	19	2	18	1	9	17
9	18	2	11	4	13	22	6	21	8	17	1	19	3	12	21	14	23	7	16
10	20	5	15	10	20	5	15	10	20	5	15	10	20	5	15	10	20	5	15
11	22	8	19	16	2	13	24	21	7	18	4	1	12	23	9	6	17	3	14
12	24	11	23	22	9	21	8	7	19	6	18	17	4	16	3	2	14	1	13
13	1	14	2	3	16	4	17	18	6	19	7	8	21	9	22	23	11	24	12
14	3	17	6	9	23	12	1	4	18	7	21	24	13	2	16	19	8	22	11
15	5	20	10	15	5	20	10	15	5	20	10	15	5	20	10	15	5	20	10
16	7	23	14	21	12	3	19	1	17	8	24	6	22	13	4	11	2	18	9
17	9	1	18	2	19	11	3	12	1	21	13	22	14	6	23	7	24	16	8
18	11	4	22	8	1	19	12	23	16	9	2	13	6	24	17	3	21	14	7
19	13	7	1	14	8	2	21	9	3	22	16	4	23	17	11	24	18	12	6
20	15	10	5	20	15	10	5	20	15	10	5	20	15	10	5	20	15	10	5
21	17	13	9	1	22	18	14	6	2	23	19	11	7	3	24	16	12	8	4
22	19	16	13	7	1	1	23	17	11	11	8	2	24	21	18	12	9	6	3
23	21	19	17	13	11	9	7	3	1	21	22	18	16	14	12	8	6	4	2
24	23	22	21	19	18	17	16	11	13	12	11	9	8	7	6	4	3	2	1
25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25

In Figure 3, we put two factors 1st and 9th columns, because unifomality is different from for each pair of $U_{25}(25^{20})$. It will be illustrated in the next

paragraph.

In order to construct $U_{25}(25^{20})$ we set the first column a_i's from 1 to 25. Then we pick up numbers a_i from 1 to 25 which posess the property $(a_i, 25) = 1$. It means the maximum common factor of a_i and 25 is one. Because $25 = 5^2$ and 5 is a prime number, we have $5^2(1 - 1/5)$ columns. The elements $a_{ij} = ja_i \pmod{25}$.

Let n be the number of trials, say, $n = 25$; $a_1, a_2, ..., a_s$ and $b_1, b_2, ..., b_s$ be two groups of positive integers, $a_i \neq a_j, b_i \neq b_j$, for any $i \neq j$, $(a_i, n) = (b_i, n) = 1, i = 1$ to s. $a_1, a_2, ..., a_s$ and $b_1, b_2, ..., b_s$ will generate different s columns. Comparing the following two values.

$$f(a_1, a_2, ..., a_s) = \frac{1}{n} \sum_{k=1}^{n} \prod_{r=1}^{s} \{1 - \frac{1}{\pi} \ln(2 \sin \pi \frac{a_r k}{n+1})\} \tag{26}$$

$$f(b_1, b_2, ..., b_s) = \frac{1}{n} \sum_{k=1}^{n} \prod_{r=1}^{s} \{1 - \frac{1}{\pi} \ln(2 \sin \pi \frac{b_r k}{n+1})\} \tag{27}$$

if $f(a_1, a_2, ...a_s) \leq f(b_1, b_2, ...b_s)$, we conclude the columns generated by the a_i's are more uniformal than these generated by b_i's. In two factories situations of $U_{25}(25^{20})$, the pair (1, 9) is the best.

3.5. Procedure

Experimental Design Estimate (EDE) is a design method for estimating

item parameters. Following are the main steps:

1) chose an experimental design which has some properties you want;

2) take all parameters needed to estimate as "factors", arrange them in proper columns;

3) determine one or more objective function(s) which were used to be the optimization criteria;

4) calculate the values of each "trial" according to theoretical formula;

5) search the optimal levels of "factor" which we will take as initial estimating values of parameters;

6) repeat the iterative procedure until the required precision is reached.

This procedure is very similar to parameter design in off-line quality control except it is unnecessary to collect data after doing real experiments. We will use a set of empirical data to illustrate the application of the method in Chapter 4. But before starting we have to check the unidimensionality for the empirical data.

Chapter 4. Comparison between EDE and MLE

4.1 Unidimensionality

In a general theory of latent traits, it is assumed that a set of k latent traits or abilities underlie examinees performance on a set of test items. The k latent traits define a k dimisional latent space, with each examinee's location in the latent space being determined by the examinee's position on each latent trait.

In many IRT models, unidimensionality is assumed. It is equivalent to the assumption known as *the assumption of local independence*. For multidimensional models, technical developments are limited and applications not possible at this time.

A check on the unidimensionality of MET87 is reported in Table 1. MET87 is the Matriculation English Test of 1987 College Entrance Examination. It is a English proficiency test. There are 105 items. The sample size is 3,208. These values are of the first 10 latent roots of the 105 × 105 tetrachoric correlation matrix. About 23 percent of the total variance was accounted for by the first factor or component, and the ratio of the first to

second eigenvalue was about 8. These statistics do meet Reckase's (1979) minimal criteria for unidimensionality. The data set seems to satisfy the unidimensionality assumption in IRT.

Table 1. First Ten Eigen Values

Eigenvalue Number	Value
1	21.5
2	3.25
3	2.16
4	1.52
5	1.50
6	1.41
7	1.39
8	1.31
9	1.32
10	1.30

4.2. Results

Suppose the factor space is $a \in [0.2]$: $b \in [-2,2]$ and $c \in [0,0.3]$. It also can be extended to any range for which theoretical assumptions are met. Using $U_{25}(25^{20})$, for the 3-PL model, parameters are arranged on the 1st, 9th and 17th columns because this is the most uniformal one after comparing all possible combination, i.e. C_{20}^{3}.

Factor level i corresponds to the value of $t_1 + \frac{i-1}{25-1}(t_2 - t_1)$ if the range of this factor is $[t_1, t_2]$.

The criterion function is:

$$R_i = \sum_{j=1}^{k} [P_i(\hat{\theta}_j)] - F_i(\hat{\theta}_j)]^2 \tag{28}$$

where

$$P_i(\hat{\theta}_j) = c_i - (1 - c_i)\{1 + exp[-Da_i(\hat{\theta}_j - b_i)]\}^{-1}; \tag{29}$$

and $F_i(\hat{\theta}_j)$ is the observed frequency of the correct answers of θ_j groups, estimated by standardized total grade on the ith item; k is the number of homogeneous groups. (in this case study $k = 21$). Denote $\text{MRS}_i = R_i/k$.

In the lth iteration, the values of $(a_i^{(l)}, b_i^{(l)}, c_i^{(l)})$ which correspond to the minimum of R_i among all "trials" are the lth "quasi-optimal" estimates. Taking the $(a_i^{(l)}, b_i^{(l)}, c_i^{(l)})$ as the midpoints of the $(l+1)$th factor space, the length is half of the lth factor space. Table 2 shows the results from the first 8 iteration for Item 1 in MET87.

Table 2. First 8 Iteration for No.1 Item

iteration	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	MRS_1
1	1.33333	-0.16667	0.07500	0.00600
2	1.12500	-0.16667	0.10625	0.00488
3	0.97917	-0.01167	0.06875	0.00282
4	0.87500	-0.08333	0.08137	0.00265
5	0.88691	-0.05208	0.07969	0.00255
6	0.87719	-0.07813	0.08066	0.00253
7	0.87500	-0.08333	0.08099	0.00253
8	0.87516	-0.07913	0.08075	0.00253

After 6 iterations, MRS has already been decrease to 0.00253. We can take this combination of $(\hat{a}_1, \hat{b}_1, \hat{c}_1)$ as "optimal" estimates of item parameters of Item No.1, because its MRS is the minimum.

The stem-and-leaf diagram for the minimum MRS values of 105 items after all iterations is shown as Figure 4. Except for 6 items to which item response might not satisfy the 3-PL logistic model, the MRS values for the rest of 99 items are all less than 0.015. The value of median is Mdn=0.0047 and the mode M_0=0.003.

0.000	0
0.001	2789
0.002	0011215556688889
0.003	000011334441566699999
0.004	0011123567889
0.005	00123333778
0.006	11224556699
0.007	123679
0.008	146
0.009	05
0.010	088
0.011	88
0.012	121
0.013	1
0.014	
0.015	5

Figure 4. Stem & Leaf Diagram of MRS for $U_{25}(25^{20})$

4.3. Comparison between Uniform and Orthogonal Design

Obviously uniform design is not the unique schedule for estimating item parameters. Orthogonal could be better. If we use $L_{25}(5^6)$, the $\hat{a}, \hat{b}, \hat{c}$ are arranged in the first three columns. Following the same steps the values of these estimate can be obtained.

The median of MRS for all 99 items is $Mdn = 0.0045$, and mode $Mo = 0.002$. They are slightly smaller than when using uniform design $U_{25}(25^{20})$. Figure 5 shows its stem-and-leaf diagram.

0.001	23566
0.002	000002233334445555666678999
0.003	01133467888899
0.004	012333356677
0.005	022477
0.006	0112228
0.007	13345889
0.008	03466
0.009	1346
0.010	1
0.011	1117
0.012	79
0.013	4
0.014	06

Figure 5. Stem & Leaf Diagram of MRS for $L_{25}(5^6)$

Furthermore if we do the paired t test on the minimum MRS values of

the total 105 items for the two design methods, the statistic is

$$t = \frac{-0.00061295}{0.007\,10582/\sqrt{101}} = -0.844.$$

We fail to reject the hypothesis that the minimum MRS values are equal.

In order to compare the efficiencies for obtaining the "optimal" estimate, we count the number of iterations for each item in the two design methods. Because $n_0 = 23$, i.e. there are 23 items of which the two designs reach the "optimal" in the same number of iterations; $n_+ = 37$, where $U_{25}(25^{20})$ is more efficient; $n_- = 45$, $L_{25}(5^6)$ is more efficient. We can conclude from the Sign Test that it is not significant.

4.4. Residual Analysis for MLE and EDE

In order to evaluate the results of EDE, the design method, we do a residual analysis. In 1988 researchers at Jianxi Normal University took the same dataset and estimated the item parameters with MLE. (Qian, personal communication, 1989) We eliminated 9 items whose $\hat{b}$ were unreasonably large (e.g. $\hat{b} = 99$). It means there are 9 items whose MLE are outside of the parameter space. Table 3 shows values of these MLEs:

Table 3. MLE of 9 items

No. of item	6	11	13	14	19	26	28	30	87
$\hat{b}$	9.63	60.21		5.21	31.17	353.26	4.34		-11.78
$\hat{a}$			-0.17				-0.20	-0.19	

It is necessary to do a linear transformation for MLE using the "Mean and Sigma" method in order to put all these estimates on the same scale, (Hambleton and Swaminathan, 1985) then we can compare their residuals. The reason is mentioned in Chapter 2, section 2.1. The l is the ratio of s_{MLE} and s_{EDE}, $l = 0.7618/1.1051 = 0.6207$, the $m = \bar{\hat{b}}_{MLE} - l\bar{\hat{b}}_{EDE} = 0.3096$.

We use the following two criteria to compare the two sets of estimates one by one:

1) $\{MRS_i\}$, where $MIRS_i = R_i/k$ for i=1,2,.... 96. That is the MRS of the "optimal" estimates for ith item;

2) $SR_{ij} = [P_i(\theta_j) - F_i(\theta_j)]/\sqrt{MRS_i}$. That is the standardized residual difference.

Let $D = MRS_{EDE} - MRS_{MLE}$. The magnitudes of the D values are shown in Table 4. We use the sing test to compare EDE with MLE in terms of D for its simplicity.

Table 4. Differences of MRS between EDE and MLE

Sign	−			+
$\|D\|$	< 0.01	$[0.01, 0.1]$	> 0.1	< 0.11
Counts	50	22	0	24

Based on the results in Table 4 from the Sign Test,

$$n_+ = 24 \text{ and } n_- = 72$$

the residual difference for EDE is significantly smaller than for MLE. According to the SR_{ij} values, a comparison between the two sets of estimates is shown in Table 8. There is no significant difference.

Table 5. Standardized Residual of EDE & MLE

$\lvert SR_{ij} \rvert$	[0,1]	[1,2)	[2,3)	≥ 3
EDE	1380	555	69	12
MLE	1378	510	90	8

The results of this residual analysis indicate that EDE is almost equivalent to MLE in precision of estimates.

4.5. Conclusion

Experimental Design Estimate (EDE) is proposed in this thesis after assessing established approaches of estimating parameters in Item Response Theory.

EDE might become a widely utilized methodology of estimating parameters in IRT, because it is easy to use and to understand. It ensures that the estimates stay within the parameter space. Its precesion is the same as that of MLE. From Table 5, a 2×3 contigency χ^2 analysis shows that the distributions of residuals of MLE and EDE are almost the same. EDE does compare favorably with MLE because of considerable savings in computer costs.

EDE is more powerful when the number of parameters is large. For example, in the 4-PL model, which is suitable for difficult questions, the number of item parameters is $4n$. Using MLE the amount of working time in computer will increase $1/3$ at least. But using EDE the only difference from 3-PL is putting the $\hat{\gamma}$ on another column, in our case on the $5th$ column of $U_{25}(25^{20})$. The computing time is almost same as before.

In the new test theory students' internal representations of systems, problem-solving strategies, or reconfiguration of knowledge as they learn will be characterized by much more parameters. (Mislevy, 1989) In such complex situation EDE will show its advantages because it does not need more work than before.

EDE can be used in assay tests and in graded response models, even with continuous response data. Essentially it is a non-parameteric method, a distribution-free method. When assumptions are violated it is robust for the estimation.

References

[1] Barton, M.A. and Lord, F.M. An upper asymptote for the three parameter logistic item-response model. *Research Bulletin 81-20.* Princeton, NJ: Educational Testing Service. 1981.

[2] Birnbaum, A. Efficient design and use of tests of mental ability for various decision-making problems. Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Taxas, 1957.

[3] Birnbaum, A. On the estimation of mental ability. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Forc Base, Taxas, 1958(a).

[4] Birnbaum, A. Further considerations of efficiency in tests of mental ability. Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Pandolph Air Force Base, Taxas, 1958(b).

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

[6] Bock, R.D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm, *Psychometrika*, 1981, **46**, 443-460.

[7] Bock, R.D. and Lieberman, M. Fitting a response model for n dichotomously-scored test items, *Psychometrika*, 1970, **35**, 179-197.

[8] Bock, R.D. and Wood, R. Test theory, *Annual Review of Psychology*, 1971, **22**, 193-221.

[9] Boldt, R.F. Latent structure analysis of the Test of English as a Foreign Language, *Language Teaching*, 1989, **6**, 125-142.

[10] Boldt, R.F. Cross-validation of a Proportional Item Response Curve Model, *TOEFL Technical Report*, ETS TR-4, 1991.

[11] Connolly, A.J., Nachtman, W. and Pritchett, E. M. /it Key Math diagnostic arithmetic test. Circle Pines, MN: American Guidance Service, 1974.

[12] Fang, Kaitai. Uniform Design, *Acta Mathematicae Applicatae Sinica*, 1980, **3**, 363-372

[13] Geramita, A.V. and Seberry, J. *Orthogonal Design, quadratic forms and hadamard matrices*, Marcel Dekker INC., New York, 1979.

[14] Gulliksen, H. Measurement of learning of mental abilities. *Psychometrika*, 1961, **26**, 93-107.

[15] Guttman, L. A basis for scaling qualitative data, *American Sociological Review*, 1944, **9**, 139-150.

[16] Hambleton, R.K. and van der Linden, W.J. Advances in item response theoty and applications: An introduction. *Applied Psychological Measurement*, 1982, **6**, 373-378.

[17] Kolen, M. Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 1981, **18**, 1-11.

[18] Lazarsfield, P.F., and Henry, N.W. *Latent structure analysis* . New York: Houghton Mifflin, 1968.

[19] Lewis, C. Test th y and Psychometrika: the past 25 years, *Psychometrika*, 1986, **51**, 11-22

[20] Liebman, J., Lasdon, L., Shrage, L. and Waren, A. *Modeling and optimization with GINO*, Palo Alto, CA: Scientific Press, 1986.

[21] Lord, F.M. A theory of test scores. *Psychometric Monograph*, **7**, 1952.

[22] Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968a, **28**, 989-1020.

[23] Lord, F.M. and Novick, M.R. *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Mass, 1968.

[24] Lord, F.M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, **14**, 117-138.

[25] Lord, F.M. *Applications of item response theory to practical testing problems*, Hillsdale, NJ:Erlbaum, 1980.

[26] Lord, F.M. Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. *Journal of Educational Measurement*, 1986, **23**, 157-162.

[27] Loyd, B.H. and Hoover, H.D. Vertical equating using the Rasch model. *Journal of Educational measurement*, 1981, **18**, 1-11.

[28] Marco, G. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 1977, **14**, 139-160.

[29] McKinley, R.L. and Reckase, M.D. 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*, MD: University of Missouri. 1980.

[30] Mislevy, R.J. and Bock, R.D. *BILOG- Maximum likelihood item analysis and test scoring : LOGISTIC model.* Chicago: International Educational Services, 1981.

[31] Mislevy, R.J. and Bock, R.D. *BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items*, Chicago: International Educational Services, 1982.

[32] Mislevy, R.J. *Foundations of a New Test Theory*, ETS research Report 89-52, 1989.

[33] Murtagh, B.A. and Saunders, M.A. *MINOS 5.1 user's guide.* Report SOL83-20R. Palo Alto, CA: Stanford University, 1987.

[34] Nowakowska, M. *Quantitative Psychology: some chosen problems and new ideas*. Elsevier Science Publishing Company, 1983.

[35] Novick, M.R. and Jackson, P.H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

[36] Parsons, C.K. and Hulin, C.L. An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Psychology*, 1982, **67**, 826-834.

[37] Pashley, P.J. An alternative three-parameter logistic item response model, *Research Report*. ETS RR-91-10, 1991.

[38] Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen:Danish Institute for Educational Research, 1960.

[39] Rasch, G., An individualistic approach to item analysis. In P. Lazarsfeld, and N. V. Henry (Eds.) *Readings in Mathematial social science*. Chicago: Science Research Association, 1966.

[40] Ramsey, J.O. Kernel Smoothing approaches to Nonparametric Item Characteristic Curve Estimation. *Psychometrika*, 1991, **56**, 611-630.

[41] Reckase, M.D. Unifactor latant trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 1979, **4**, 207-230.

[42] Rentz, R.R. and Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, 1977, 161-180.

[43] Samejima, F. A comment on Birnbaum's three parameter logistic model in the laten trait theory. *Psychometrika*, 1973, **38**, 211-223.

[44] Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, **1**, 233-247.

[45] Schmidt, F.M. THe Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement*, 1977, **37**, 613-620.

[46] Slinde, J.A. and Linn, R.L. The Rasch model, objective measurement, equating and robustness. *Applied Psychological Measurement*, 1979, **3**, 437-452.

[47] Soriyan, M.A. Measurement of the goodness-of-fit of Rasch's probabilistic model of item analysis to objective achievement test of the West African Certification Examination. Unpublished doctoral dissertation. University of Pittsburgh. 1977.

[48] Suen, H.K. and Lee, P.S.C. Constraint optimization: a perspective of IRT parameter estimation, a paper presented at *The Fifth International Objective Measurement Workshop*. Berkeley, CA, 1989.

[49] Swaminathan, H and Gifford J.A. Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New Horizons in testing.* New York: Academic Press, 1983.

[50] Swaminathan, H and Gifford J.A. Bayesian estimation in the three-parameter logistic model. *Psychometrika*, **51**, 1986, 589-601.

[51] Taguchi, G. and Wu, Y. *Introduction to off-line quality control* , Central Japan Quality Control Association, 1979.

[52] Urry, V.W. Ancilliary estimators for the item parameters of mental tests, D.C.: Personnel Research and Development Center, U.S. Civil Service, Commission, 1976.

[53] Urry, V.W. Tailored testing: A successful application of latent trait theory, *Journal of Educational Measurement*, 1977, **14**, 181-196.

[54] Wainer H. and Thissen D. Some standard error in item response theory, *Psychometrika*, 1982, **47**, 397-412.

[55] Waller, M.I. A procedure for comparing logistic latent trait models. *Journal of Educational Measurement*, 1981, **18**, 119-125.

[56] Way, W.D and Reese, C.M *An Investigation of the use of Simplified IRT Model for Scaling and Equating the TOEFL Test*, ETS Research Report 90-29, 1990.

[57] Weiss,D.J. (Ed.) *New Horizons in testing*, New York: Academic Press, 1983.

[58] Wingersky, M.s. LOGIST: A program for computing maximum likelihood procedures for logistic models. In R. K. Hambleton (ED.) *Applications of item response theory.* Vancouver: Educational Research Institute of British Columbia, 1983.

[59] Woodcock, R.W. *Woodcock reading master test.* Circle Pine, MN: American Guidance Service, 1971.

[60] Woodcock, R.W. Development and standardization of the *Woodcock-Johnson Psycho-Educational Battery.* Hingham, MA: Teaching Resources Corporation, 1978.

[61] Wright, B.D. and Stone, M.H. *Best test design*, Chicago: MESA, 1979.

[62] Xiao, B.L. Golden section search strategies for computerized adaptive testing, a paper presented at *The Fifth International Objective Measurement Workshop*, Berkeley, CA, 1989.

[63] Yen, W.M. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 1981, **5**, 245-262.

[64] Yen, W.M. Use of the three-parameter model in the development of standardized achievement test. In Hambleton, R.K. (Ed.) *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia, 1983.

[65] Yen, W.M. et al. Nonunique solutions to the likelihood equation for the 3-PL Logistic Model. *Psychometrika*, 1991, **56**, 39-54.