CANADIAN THESES ON MICROFICHE

I.S.B.N.

THESES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche 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 a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche 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 mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

NL-339 (r. 82/08)

Canada

A COMPARISON OF THE USE OF MULTIPLE MATRIX SAMPLING AND EXAMINEE SAMPLING FOR TEST DEVELOPMENT

by Rashmi Garg

Thesis presented to the School of Graduate Studies of the University of Ottawa in partial fulfillment of the requirement for the degree of Doctor of Philosophy.

Ottawa, Canada, 1983

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGMENTS

I am deeply indebted to two special people who have greatly influenced my graduate education. Dr. Marvin Boss and Dr. James Carlson, members of my advisory committee, gave me their continued encouragement and support. The tremendous amount of time they spent molding and revising this manuscript was great help to me. Without their assistance, reaching this point in my graduate studies would have been much more difficult.

My special thanks go to Ashutosh Chowdhury who patiently prepared all the computer programs needed for this research.

I would also like to thank faculty and friends in the Department of Psychology at Laurentian University for sharing with me their experiences, knowledge, and support.

Finally, I would like to express my sincere appreciation to my friends and family who stood by me during very trying times. Particularly my husband, Brij, for his many sacrifices, patience, and encouragement. My mother who gave me a great deal of encouragement and helped me look after the family while I was busy with my graduate studies. Her untimely death has denied us the opportunity of sharing this joy with her.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

INTRODUCTION

The chief purpose of testing is to provide information for decision making related to people, programs and institutions. If the decisions based on the use of test scores are to have any merit the tests must be reliable and valid.

Both the validity and the reliability of a test depend ultimately on the characteristics of its items. Through application of item analysis procedures test constructors are able to obtain quantitative, objective information useful in developing and judging the quality of a test and its items.

Once the item pool has been carefully constructed to measure a certain trait, items should be pretested on a representative sample of adequate size to obtain quantitative information relevant for item analysis. Many different techniques are available for item analysis. The choice among them depends on the nature and purpose of the test. For a norm referenced test measuring a unidimensional trait, the two most common indicators of item quality for characterizing items are item discrimination and item difficulty. One major flaw in both of these indicators is that their estimates are subject to considerable fluctuation unless the sample size is very large.

One principal problem which test constructors face when analyzing items is that of acquiring a representative sample

of adequate size to pretest the items. School officials, industries etc. are reluctant to participate in data collection because of the amount of time which is involved and the small benefit which schools, industries etc. receive. Preliminary tests, besides providing no incentive for testees, are generaly much longer than the final tests. These lengthy tests may produce fatigue and boredom in the individuals, causing them to be careless, which, of course, directly affects the estimation of item indices and test reliability.

This problem in test construction can be overcome by utilizing a multiple matrix sampling technique for data collection in test construction studies. Multiple matrix sampling is a sampling technique in which a pool of items is randomly divided into a number of subtests, and each of these subtests is administered to a different sample of examinees randomly selected from the population.

Multiple matrix sampling is different from traditional examinee sampling in the sense that each individual is required to respond to only a sample of items from the test, whereas in examinee sampling each individual is required to respond to all the items in the test. The main advantage of multiple matrix sampling is the reduction of testing time per examinee.

Research in multiple matrix sampling (MMS) has demonstrated that it is a promising alternative to traditional examinee sampling for estimating group parameters

such as the mean and variance. It has been successfully used for developing test norms and evaluating programs. However no systematic study has been carried out to compare MMS with examinee sampling for estimating item parameters and test characteristics. This study was designed to fill this gap in the test construction literature.

The primary purpose of this study was to explore the use of a multiple matrix sampling technique of data collection for test development. In other words, it was to investigate how the multiple matrix sampling technique of data collection compares with traditional examinee sampling for estimating item parameters such as item difficulty and item discrimination and for developing tests that are reliable and valid. A secondary purpose was to explore the conditions (i.e., varying the subtest size; the examinee sample size; the characteristics of items, etc.) under which the multiple matrix sampling technique is most effective in estimating item parameters.

Significance of the Study

Quantitative information about the characteristics of items and the test has always been valued by measurement specialists. Traditionally this information has been achieved by administering test items to a large sample of examinees. However this approach of collecting quantitative information can be very demanding in terms of examinee time. If the same quantitative information regarding test item and

total test characteristics can be obtained by administering a small sample of items to each examinee, as that obtained by administering the whole test to each examinee, it will be possible for test constructors to pretest items on a large sample of examinees, while at the same time lessening the burden currently placed on schools, industries etc. when they agree to assist in that process.

Organization of the Study

The dissertation is organized as follows: The review of the relevant literature related to item analysis and multiple matrix sampling and the statement of the problem are presented in chapter I. The research design and the methods used to examine the problem are presented in chapter II. The results are presented and discussed in chapter III. Conclusions of the study are stated in chapter IV.

CHAPTER I

REVIEW OF THE LITERATURE

In the development of norm referenced tests one gathers responses to the test items from a sample of examinees that is representative of the population on which the test is to be used. These responses are subjected to item analysis to decide which items should be selected, revised or rejected. Traditionally the examinee sampling method is used to gather the responses to the items. One can also use multiple matrix sampling (MMS) to gather responses to the items. MMS is less demanding in terms of examinee time than examinee sampling. In MMS each examinee is required to respond to only a sample of items as opposed to all the items in the test. The following literature review is divided into two areas of interest: Item analysis and multiple matrix sampling. This is followed by a statement of the problem.

Item Analysis

Item analysis is a general term for a set of methods used to evaluate test items. It is one of the most important aspects of test construction (Kaplan and Saccuzzo, 1982). Tests can be improved through the selection, substitution or revision of items. Item analysis makes it possible to shorten a test and at the same time to increase its validity and reliability (Anastasi, 1976). The reliability and validity of the total test depend entirely on the

characteristics of the items used to build it. Thus the real importance of item analysis arises from the effects of the individual item characteristics upon the characteristics of the entire measuring instrument (Lord and Novick, 1968).

Burt (1921) first applied item analysis with the original Binet scale. By item analysis of the subsequent Binet-Simon scale, Burt secured increases in test reliability from .68 to .91, in validity from .59 to .76. Subsequent investigators examining other tests have shown similar results (Hull, 1928; Neill and Jackson, 1970; Pyrczak, 1973).

A number of different item analysis procedures have been proposed in the measurement literature for studying and characterizing individual items (Allen and Yen, 1979; Lord and Novick, 1968; Ghiselli, Campbell and Zedeck, 1981). However, the two characteristics of items most commonly used for item analysis are item difficulty and item discrimination. In this section item difficulty, item discrimination, research related to item difficulty and research related to item discrimination are discussed.

Item Difficulty:

The most common index of item difficulty (p) for dichotomously scored items is defined as the proportion of correct responses among those who had an opportunity to answer the item. The item difficulty level of an item varies between .00 and 1.00. Zero means that nobody in the population responded to the item correctly. One means everybody responded to the item correctly.

Item difficulty information facilitates the choice of items for the final form. Those items passed by everyone (p=1.0) or failed by everyone (p=0.0) convey no information for discrimination amongst individuals. A primary reason for assessing the difficulty level of items is to permit the construction of tests with specific characteristics for differing purposes. The knowledge of item difficulty is also useful in arranging items in the test. It is helpful to place items that are relatively easy at the beginning of the test. This arrangement gives the individual confidence in approaching the test and also reduces the likelihood of his/her wasting time on items beyond his/her ability (Anastasi, 1976).

Item Discrimination:

An item discrimination index is a measure of how well the item discriminates between persons possessing more of the attribute being measured from persons having only relatively small amounts of this attribute. An item with a very low discrimination compared to other items provides very little information about individual differences in norm referenced tests. For homogeneous tests when item discrimination is evaluated by correlating the item score with the total test score, it also provides information on whether the item measures the same trait or attribute that is measured by the total test. If the correlation between an item score and the total test score is very low, one can say that the item does not tend to measure the same trait or attribute that is measured by the total test. Thus the item discrimination

information facilitates the choice of items in the final tests.

Taking the item difficulty and item discrimination indices into consideration, test constructors develop/design a test that is suitable for the examinees and which all convey the maximum information possible about the differences in the examinees' levels on the trait being measured (Anstey, 1966).

Research Related to Item Difficulty:

There is a disagreement amongst measurement specialists regarding the ideal level of item difficulty for a test. Cronbach and Warrington (1952), Ebel (1972), Ghiselli, Campbell and Zedeck, (1981) and Gulliksen (1945) have argued in favor of items of .5 difficulty. Their argument is based on the fact that in norm referenced tests it is often desirable to have a large test variance. For such tests it is better to have those items that contribute most toward the total test variance than the items that do not contribute much to total test variance. In dichotomously scored items (where items are scored either zero or one), item variance is equal to $p(1-p)$. This variance is maximized when $p=.5$. Since total test variance is related to item variance, it may be maximized when items are of .5 difficulty.

On the other hand, some measurement specialists have argued that test variance is not only affected by item variance, but also by inter-item covariance. Thus when considering a distribution of difficulty, one should also

take into account inter-item covariance. Davis (1951) has stated that if all items in a test are of difficulty .5 and are uncorrelated the test will discriminate well at all ranges of ability. But when the items are correlated, maximum test discrimination would be achieved when the difficulty indices for test items are spread out. When developing a test of a specific trait, since items should be positively intercorrelated, Davis has recommended writing items of varying difficulty. Brogden (1946), Henrysson (1971) and Gutman and Loevingor (Bowers, 1972) have supported Davis.

Furthermore, Anstey (1966) and Scott (1972) have stated that when considering the desired distribution of difficulty level of items, consideration should be given to the total test score distribution desired. The distribution of total test scores depends on the number of items, their difficulties and their intercorrelations. There are differing points of view concerning the ideal shape for a test score distribution. Some authors prefer a normal distribution (Cronbach, 1960), some a rectangular distribution (Ferguson, 1949; Guilford, 1954 ; Humphreys, 1956), and others a distribution that is similar to the true score distribution (Lord and Novick, 1968). Without taking a position on this matter, one may still inquire as to the conditions under which any desired distribution may be obtained.

Anstey (1966) and Lord and Novick (1968) have shown that when the ability distribution is normal and items are

homogeneous (i.e., they are intercorrelated), the affect of the distribution of item difficulties on the distribution of test scores is as follows: If item difficulties are evenly distributed over the ability range an approximately normal distribution of test scores results; if items are concentrated at .5 difficulty an approximately rectangular distribution of test scores occurs; if the distribution of item difficulties is approximately normal, the distribution of test scores is somewhat platykurtic; if the test consists of many more difficult items than easy items, a positively skewed distribution of test scores is evident. Conversely to this, if a test includes many more easy items than difficult items, a negatively skewed distribution results.

Anstey (1966) has further stated that when a test contains relatively heterogeneous items (i.e., zero or very low intercorrelation among items), examinees will tend to be less consistent in the accuracy of their responses and their scores will tend to be more alike than in a homogeneous test. In such tests, generally speaking, when item difficulties are evenly spread over the ability range (assuming the ability distribution is normal), there will be a concentration of scores in the middle and a leptokurtic distribution will result. If item difficulties are concentrated around p=.5, a somewhat normal distribution of test scores would be achieved. However, zero or very low inter-item correlations would imply that items did not measure a common trait, thus making a summative scoring (item scores are summed to achieve the test score) procedure illogical. Three empirical

studies conducted by Anstey (1966) have supported the above results.

On the basis of the above studies and discussion, one can conclude that there is no ideal distribution of item difficulty. The proportion of items to be selected at the different difficulty levels should depend on the type of test score distribution desired, which depends on the purpose of the test and on the degree of homogeneity among items . For example, if one wishes to differentiate examinees at a given level of ability, most of the items in the test should be of a difficulty level which corresponds to that ability level (Lord, 1953; Richardson, 1936a). On the other hand, if a test is not to be limited to any subject population and the examiner wishes to achieve an approximately normal distribution of test scores then item difficulty should be distributed evenly over the ability scale (Richardson, 1951). If a test is required to discriminate at all levels of ability then a rectangular distribution of test scores is desirable.

Research Related to Item Discrimination:

Another important item statistic is the item discrimination index. A desirable characteristic for a test is that it should discriminate over the desired range of ability. Anastasi (1976), Davis (Pyrezak, 1973) and Lord and Novick (1968) have suggested that items that discriminate poorly should be discarded or inspected for possible deficiencies and revised.

Numerous methods of expressing the discriminatory power of an item have been proposed in the measurement literature. As far back as 1935, Long and Sandiford described 23 different methods of expressing item discrimination (Oosterhof, 1976). Guilford (Anstey, 1966) listed 19 methods for calculating item discrimination. However, the method most commomly used by test constructors for assessing item discrimination is the correlation between each item and a criterion or total test score. If the correlation coefficient is high, one would expect correspondence between the trait as measured by the test and the item score. The two most commonly used correlation coefficients with dichotomous items are the point biserial and the biserial correlation coefficients. Both were developed from the Pearson product moment correlation (Ghiselli, Campbell and Zedeck, 1981; Lord and Novick, 1968; Tate, 1955). The point biserial correlation coefficient (point biserial r) is applicable when one of the variables being correlated (the item score) represents a true dichotomy and the other variable (criterion or total test score) is continuous and normally distributed. The biserial correlation coefficient (biserial r) is applicable when one of the variables (the item score) has an underlying continuous normal distribution which has been artificially dichotomized and the other variable is continuous and normally distributed (Lord and Novick, 1968). McNemar (1962) has pointed out that the assumptions for point biserial r are hard to justify when it is suspected that the knowledge required to answer an item is

continuously distributed.

Despite the basic difference in the assumptions for these two coefficients, measurement specialists fail to agree as to which method is most appropriate for estimating item discrimination. For example, Guilford (1965) and Nunnally (1967) have argued in favour of point biserial r on the basis that items selected using this method contribute more toward internal consistency (estimated using the Kuder Richardson formula 20) and test variability than items selected using biserial r. On the other hand, Henryson (1971) and Lord and Novick (1968) have supported biserial r as an index of discrimination. They provided the following argument: point biserial r tends to be higher for medium difficulty (p=.5) items than for very easy or very difficult (p .5) items. In other words, given the way it is calculated, item difficulty sets an upper limit for point biserial r. Biserial r is not affected by the difficulty level of items. If one wants to have an item discrimination index that is independent of item difficulty, biserial r would seem to be a more appropriate index of discrimination than the point biserial r providing that the underlying assumptions are satisfied. Furthermore, the biserial r tends to be more stable from group to group than the point biserial r (Lord and Novick, 1968).

Ashler (1979) has shown that the biserial r underestimates correlation in the presence of guessing. No such studies are reported for point biserial r but it seems likely that the same effect would take place.

Aleamoni and Spencer (1969) and Engelhart (1965) have

reported a high correlation between point biserial r and biserial r indices, obtained on the same data. Bowers (1972) has studied this problem and found that when the distribution of criterion scores is normal both indices lead to selection of the same items if the items with maximum discrimination are selected. However with a skewed distribution of criterion scores the two indices lead to selection of different items. The maximum values of biserial r and point biserial r are not only affected by the difficulty level of the items but also by the distribution of test scores. When the distribution of test scores is symmetrical, maximum biserial r is 1.00, regardless of difficulty whereas point biserial r is maximum for p=.50 and it decreases as difficulty departs from .50. Furthermore, when a distribution of test scores departs from symmetry, i.e., becomes skewed in one or another direction, the maximum biserial r may be greater than one (Lord and Novick, 1968; Richardson, 1936b).

Several investigators (Guilford, 1954; Henrysson, 1963; Zubin, 1934) have indicated that both point biserial r and biserial r between an item and the total test score are spuriously high if the item score is included in the total test score. They have suggested that point biserial r and biserial r should either be corrected or the total test score should be calculated after excluding the item under consideration. Several formulae have been suggested for correcting inflated point biserial r and biserial r. Moreover, Cureton (1966), Wolf (1967) and Berk (1978) have empirically investigated the effect of these corrections on

point biserial r and biserial r, under conditions of varying test length. They found that for tests of 40 or more items, the mean differences between corrected and uncorrected indices was less than .10. The rank order correlation between corrected and uncorrected point biserial r and also between corrected and uncorrected biserial r was .99.

On the basis of the above studies, one can conclude that the choice between biserial r and point biserial r should depend on the nature and purpose of the test. If items for a test are selected on the basis of point biserial r, one would more likely select a majority of items having difficulties around .5, because point biserial r is maximum for p=.5 and it decreases as item difficulty depart from .50. If one wishes to have a test in which item difficulties are spread out or concentrated on either end, it would be more appropriate to use biserial r as it is not affected by item difficulty.

It appears that when developing norm referenced tests one needs to collect statistics of item difficulty and item discrimination. Traditionally these statistics are collected by testing items on a sample of examinees. Cooper and Fiske (1976), Lord and Novick (1968), Neill and Jackson (1970) and Nunnally (1967) have stated that estimates of item parameters are subject to considerable fluctuation unless the sample size is very large. They have indicated that with a small sample size, there are too many opportunities for chance error in item analysis. For this reason, they have suggested that items should be pretested on a large sample of

examinees, preferably 5 to 10 times the number of items. These examinees should be representative of the population in which the final test is to be used. One principal difficulty which test constructors face when analyzing items is that of obtaining a representative sample of adequate size to pretest the items. Many school officials and job supervisors are reluctant to release a large number of individuals for blocks of time for testing. As a result, items are often tried out on a small sample of co-operative individuals. Furthermore, the initial item pool consists of a considerably larger number of items than the number of items one intends to have in the final form of the test. The behavior of examinees, who are faced with more items than they can seriously consider within the time limit, varies considerably. The bolder, less conscientious tend to guess the answers to a large number of items. It is also probably true that even very serious examinees do not finish the test with the same degree of alertness and eagerness that they began it.

One possible solution for the problem discussed above is to give a small number of items to each examinee. If each individual was required to try a very small sample of items rather than all the items in the test, it would be easier to get people to cooperate. As a result a more representative and larger sample of individuals is likely to be obtained. This approach will also make it possible to collect data on a large number of items. This sampling approach is often referred to as multiple matrix sampling, item-examinee sampling, matrix sampling, or incidence sampling.

Multiple Matrix Sampling

The multiple matrix sampling (MMS) model was first outlined by Lord (1962). He developed this model as an alternative to traditionally used examinee sampling in response to the need for developing test norms that are representative and demand less of each examinee.

In a multiple matrix sampling procedure, a set of items is divided into different subtests, and each subtest is administered to a different subgroup of examinees, selected from the population of examinees. In this section the designs used in MMS literature for dividing items into subtests, the mathematical framework in MMS and the research related to MMS are discussed.

Designs used in MMS:

A number of designs have been used in the MMS literature for allocating items to subtests. The most commonly used designs are as follows:

(a) Non-overlapping multiple matrix sampling design (NMMS): These designs represent a class of designs where random samples of items are drawn in such a way that all items are sampled but any given item appears in only one subtest. A set of nonoverlapping matrix samples results when every item is sampled without replacement. This sampling is often referred to as exhaustive NMMS (Sirotnik, 1974).

(b) Multiple matrix sampling with replacement (MMSWR): These represent a class of designs in which a random sample of items is drawn from a pool of items, but these items are

replaced before drawing another sample of items. Thus the items are sampled without replacement within each subtest but with replacement among subtests. In this design it is possible that some items in the set are not included in any subtest whereas other items are included in more than one subtest.

(c) Partially balanced incomplete block design (PBIB) and balanced incomplete block design (BIB): These designs are special cases of MMSWR designs, in which some restrictions are placed on the random sampling of items. In BIB designs, items are sampled among subtests in such a way that each item is repeated an equal number of times among subtests and also each item is paired with each other item an equal number of times. In PBIB designs, each item appears with an equal frequency among subtests, but item pairing does not occur with an equal frequency and for certain designs some item pairs may not occur.

The main difference between NMMS, MMSWR, BIB, and PBIB designs is that in NMMS, each item appears only once among subtests and inter-item data are available only for those items within any given subtest, but not for items occuring in dfferent subtests. In MMSWR designs, it is possible that some items in the set are not included in any subtests, whereas some items are included in more than one subtest. In PBIB designs, inter-item data are available for most of the items depending on the design but not necessarily for all the items in the test. In BIB designs inter-item data are available for all items in the tests.

Mathematical Framework in MMS

The obvious difficulty with MMS is that since all the items are not administered to all of the individuals, one can not estimate parameters in the same way as one would with examinee sampling. Several investigators have focused on the development of formulae for estimating parameters such as test means, variances, standard error of the mean and the standard error of the variances etc. from MMS designs.

Hooke (1956a, 1956b) developed an algebraic function involving symmetric polynomials of the elements in a matrix. These functions are called generalized symmetric means (gsm's) and have the property of being inherited on the average, i.e., the expected value of a gsm in a matrix sample is equal to the gsm in the matrix population. The linear combinations of gsm's called bipolykays can be used to obtain an estimate of the moments of the matrix population. Lord and Novick (1968) and Shoemaker (1973b) have used Hooke's formulation to derive the formulae for estimating the mean, variance and other moments of the matrix population via MMS. An alternative to Hooke's approach was proposed by Sirotnik (1970b, 1974). He derived formulae for estimating population parameters such as the mean and variance using two-way (examinee by item) analysis of variance (ANOVA).

In both Hooke's method of gsm and the ANOVA model, data obtained using NMMS or MMSWR designs are analyzed using individual matrix samples. The point estimates of different moments of the population are obtained on each single matrix.

These estimates are averaged over all matrices to obtain the best estimate (or pooled estimate) of the moments of the population. Since the mean of unbiased estimates is also unbiased, the pooled estimates of the moments are unbiased. When estimating the standard error of the moments, the situation becomes more complicated because the matrices sampled are not independent. Thus neither of the above two methods results in an exact formula for estimating the standard error of the mean, variance and other higher moments (Pandey and Shoemaker, 1975). For other designs such as PBIB and BIB designs, single estimates of the mean and variance of the whole distribution are obtained. Knapp (1973) has suggested that in these designs, there seems to be no way to derive estimates of standard errors of means and variances. Sirotnik and Wellington (1977) and Wellington (1976) have solved the problem of estimating the standard errors of moments. They have extended the use of gsm's defined by Hooke to rectangular arrays. Instead of focusing on the individual matrices in MMS designs, they have used gsm's on an entire configuration of sampled data. They have shown that gsm's computed on the entire configuration of sampled data give unbiased estimates of corresponding gsm's computed for the population. With this general approach, all current matrix sampling designs (e.g., NMMS, PBIB, BIB, etc.) are analyzed in a similar way. Sirotnik and Wellington (1977) have provided formulae in terms of gsm's for deriving estimates of the population mean, variance and their standard errors. They have used the term "incidence sampling" to

refer to the configuration of data points or entries sampled from a matrix population.

Using the incidence sampling approach, Knapp (1979) has developed equations in terms of gsm's to estimate population covariances between two sets of measurements and their standard errors. However, although covariance estimates are unbiased, they have larger standard errors than covariance estimates obtained using those pairs for which paired data are available.

Research Related to MMS:

The literature which currently exists concerning MMS may be divided into the following classes: (a) Studies in which the main purpose was the validation of MMS procedure for estimating the population mean, variance and distribution of test scores. (b) Studies in which the effect of different sampling plans (i.e., effect of variation of sample size, number of subtests, number of items per subtest, etc.) on estimates of population mean and variance was investigated. (c) Studies designed to measure the influences of psychometric characteristics of the test itself and the nature of the distribution of the examinee population on estimation of population parameters. (d) Studies related to context effect. (e) Other studies related to the applicability of MMS in other situations.

(a) The validation studies for means, variances and test score distributions are of two types: a posteriori and a priori. The a posteriori studies were conducted on already

existing data, i.e., data were available for all examinees, whereas for a priori studies, data were collected using MMS.

An initial validation study of the a posteriori type was conducted by Lord (1962). Lord (1962) compared the mean, variance and distribution of test scores obtained on 1000 examinees on a 70 item vocabulary test with estimates of means, variances and the distribution of scores obtained using NMMS designs and examinee sampling. For NMMS designs the 70 items were divided into 10 subtests with 7 items in each subtest and 1000 examinees were divided into 10 random samples with 100 examinees in each sample. For each examinee the item scores on a subtest of 7 items were obtained. For examinee sampling the item scores on the whole test were obtained for ten examinee samples with 100 examinees in each sample. The results showed general superiority of MMS over examinee sampling (in terms of closeness to the population value) for estimating population mean and distribution of the test scores. However for estimating variance MMS tends to be as good as examinee sampling.

Plumlee (1964) administered a 30 item test to a population of 200 clerical applicants and calculated the test mean and standard deviation for the population. She compared this mean and standard deviation with the estimates of means and standard deviations obtained using examinee sampling and MMS (NMMS design). The results showed that the test mean estimated by MMS was closer to the population test mean than the test means estimated by examinee sampling (there were ten examinee samples and one MMS). The test standard deviations

estimated by the examinee sampling were closer to the population test standard deviation than the test standard deviation estimated by MMS. However, Plumlee has stated that the standard deviation estimated by MMS was within the error of measurement that would be expected in giving another form of the test. Cook and Stufflebeam (1967) sampled without replacement (i.e., NMMS design). They compared estimates of population means and variances obtained using examinee samples of four different fractions of total population (10%, 25%, 33%, and 50%) with estimates of population means and variances obtained using MMS of various subtest sizes (10%, 25%, 33% and 50% of total items). Their results lend further support to Lord's (1962) and Plumlee's (1964) conclusions. However, the results with respect to different sizes of the subtest in the MMS showed no strong relationship betweeen the size of the item sample subtest and the efficiency of the estimation of the test score distribution. In all the above studies, when comparing MMS procedures with examinee sampling procedures, the number of observations was kept constant. An observation was defined as a response to one item by an individual.

Owens and Stufflebeam (1969) in an a priori study collected data from examinees' samples of various sizes (.06, .12, and .24 of the population) and compared these results with estimates of the test score distribution based on MMS (NMMS design) of various subtest sizes (.06, .12, and .24 of the items). The results indicated that MMS provides estimates of population means and variances that are as good

as traditionaly-used examinee sampling. Cahen, Romberg and Swirner (1970) in an a priori study used two parallel forms of the same test on the same examinees in a number of schools. Data on one form were obtained using a MMS procedure, whereas data on another form were obtained using the examinee sampling procedure. The correlation of estimated test means and variances of the two forms obtained from 81 schools was calculated. They concluded that the relative position of schools with regard to test mean and variance remained the same regardless of the sampling procedure. In all the above a priori studies the number of observations when comparing examinee sampling with MMS was kept constant.

On the basis of the above review one can conclude that when the number of observations are kept constant MMS gives more precise estimates of the population mean than examinee sampling. Estimates of the population variance are less precise with MMS, but are generally not significantly worse than examinee sampling estimates.

(b) The research for the most efficient item sample size and item sample design represents the second trend in the MMS literature. In a series of post-hoc studies, Shoemaker (1970a; 1970b; 1971b and 1972) systematically manipulated the number of subtests, the number of items per subtest and the number of examinees responding to each subtest to determine the relative merits of several MMS procedures. The item subtests were sampled using NMMS or MMSWR designs. The general finding that showed up consistently in these studies

was that an important factor in the estimation of population mean and variance was the number of observations. As the number of observations increased estimates of population means and variances obtained from different MMS plans also improved. However, for estimating the population mean for a given number of observations, Shoemaker concluded that the most precise estimate (in terms of standard error of estimate) of the population mean involved giving the smallest possible number of items to each examinee and conversely, the least precise estimate involved giving all items to few examinees.

Barcikowski (1974), and Moy and Barcikowski (1974), examined the issue of item subtest sizes as a part of their investigations. They simulated data by computer. The item subtests were sampled using NMMS and MMSWR designs. The general findings of these studies which held the number of observations constant were as follows: When estimating the population mean the estimates improved as the number of items per subtest decreased with resulting increases in the number of examinees tested. The most precise estimates of the population mean (in terms of standared error) involved giving the fewest possible items to most people and conversely, the least precise estimates were obtained by giving the whole test to a lesser number of examinees. When estimating the population variance the most precise estimates of the population variance for tests with item discriminations in the range of .05 to .50, were obtained from MMS plans in which each subtest contained .25 to .50 of the total number

of items.

In most of the research in MMS, exhaustive nonoverlapping designs are used. Relatively little research is available with regard to relative merits of different designs. Lord (1962) and Lord and Novick (1968) have stated that sampling without replacement is better than sampling with replacement. Furthermore, exhaustive sampling is better than nonexhaustive sampling because in estimating the population mean the omission of even one item could have a considerable effect on the standard error of the mean. Cook and Stufflebeam (1967) and Plumlee (1964) have supported Lord's statement. Knapp (1968) has recommended the use of BIB designs for estimating group parameters. He has stated that in BIB designs, inter-item covariances can be obtained for all the items whereas in NMMS designs these data are available only for those items that are within any given matrix sample, not for items in different matrix samples. Since test variance and test reliability (calculated using the Kuder Richardson formula 20) take into account inter-relationships among items, BIB designs would seem to be better designs to use than NMMS designs. Shoemaker (1973a) has pointed out that BIB designs are often very hard to obtain and difficult to implement because the number of subtests required is excessively large. This limitation becomes extremely serious when the number of items exceeds 50. He recommends the use of PBIB design when inter-item data are required. Generally in these designs, inter-item data may not be available for all items, but they are more

practical to use than BIB designs. Shoemaker (1973a) compared estimates of the first four moments obtained using a PBIB design with estimates obtained using a MMSWR design where all items were sampled. He found that PBIB design was better than MMSWR. However, the superiority of PBIB design was more apparent in the estimation of the mean than for higher moments. These findings are consistent with the findings of Scheetz (1976).

On the basis of above studies one can draw the following conclusions: The standard error of the mean (SE $\hat{\mu}$) and standard error of the variance (SE$\hat{\sigma}^2$) decrease generally with an increase in the number of observations; when estimating the population mean for a given number of observations, increasing the number of subtests is preferable to increasing the number of items per subtest. The converse is true for estimating the population variance; for all practical purposes NMMS designs provide almost as good estimates of the population mean and variance as PBIB or BIB designs. The gain in precision which may occur when using PBIB or BIB designs is probably not worth the effort of constructing the large number of subtests required in PBIB or BIB designs. However, PBIB OR BIB would be necessary if data on inter-item correlation are required.

(c) The influence of psychometric characteristics of the test itself and the nature of the distribution of the examinee population on different MMS plans have been investigated by Barcikowski (1974), Moy and Barcikowski (1974) and Shoemaker (1971b,1972).

Shoemaker (1971b) varied the variance of item difficulty indices, the shape of the test score distribution, the number of subtests, the number of items per subtest and the total number of observations. In another study Shoemaker (1972a) also varied the test reliability, expressed as an alpha coefficient, in addition to the above variables. In both of these studies Shoemaker varied one of the variables at a time while holding the other variables constant. The general findings (other than the one expressed earlier) of these two studies were as follows: As the variance of item difficulty indices increased (from .00 to .05), the SE $\hat{\mu}$ increased and the SE $\hat{\sigma}^2$ decreased; The higher level of test reliability (alpha coefficient equal to .95) resulted in larger SE $\hat{\mu}$ than the lower level of reliability (alpha coefficient equal to .80), the converse was true for estimating population variance; the skewed test score distribution resulted in smaller SE $\hat{\mu}$ and SE $\hat{\sigma}^2$ than the normal test score distribution.

Forysth (1976) has shown analytically that in the MMS procedure of estimating the mean, as used by Shoemaker (1971b, 1972a), increases and decreases in the standard error of the mean are related to increases and decreases in examinee variance. He concluded that it is more reasonable to attribute changes in the standard error of the mean estimated via MMS to changes in examinee variance than to changes in other data base characteristics as stated by Shoemaker (1971b, 1972a).

Barcikowski (1972, 1974) and Moy and Barcikowski (1974) investigated the effects of item difficulty and item

discrimination of test items on the selection of a MMS plan. The NMMS and MMSWR designs were used for dividing items into subtests. They selected several hypothetical tests in which the range of item difficulties and/or the range of item discriminations were/was varied. For each test, data were generated on the computer. For a given number of observation the results obtained from different MMS plans were compared among themselves and also with examinee sampling. The results show that the range of item difficulty indices had no effect on which MMS plans gave the most precise estimate (in terms of standard error) of population means and variances. However the range of item discrimination indices affected the estimation of population means and variances obtained through different MMS plans. For tests with a high level of item discrimination (where item-total biserial correlations were .45 or higher) the best estimates of population means and variances were obtained (while holding the number of observations constant) with the MMS plans which gave the fewest number of items per subtest to a maximum number of examinees. In this case MMS provided more precise estimates of population means and variances than examinee sampling. For tests in which item discriminations were generally below .50, the same findings for the estimation of the population means were obtained, but the most precise estimates of the population variances resulted from MMS plans which contained .25 to .50 of the items in each subtest. In these cases MMS provided estimates of population variances that were as good as those from examinee sampling. The examinee sampling

provided more precise estimates of population variances than the MMS plans in which each subtest included less than .25 of the items. It should be noted that most standardized tests have item discrimination values less than .55 (Chartock, 1980).

On the basis of the above studies one can conclude that when the parameter of primary importance is the population mean, small sized subtests should be used; when the parameter of primary importance is the population variance attention should be given to the level of item discrimination of test items. With tests of low item discriminations (e.g., biserial correlation coefficients in the range of .05 to .50) larger subtests should be used, specifically subtests containing .25 to .50 of the items. With tests of high item discrimination (e.g., biserial correlation coefficient over .50) small sized subtests should be used.

(d) Another area of research with regard to MMS is related to the assumption of context effect. In MMS, it is assumed that the response of an examinee to an item is independent of the context in which the item is presented. In other words, it is assumed that the examinee would respond to the items in a sample, in the same way as if they had been embedded in the universe of items (Sirotnik, 1974). The research which has been conducted to test this assumption has generally resulted in favorable conclusions, that is, estimates of population parameters are nearly the same whether or not items are responded to as part of the total test or a subtest (French and Greer, 1964; Hill, 1975; Huck

and Bowers, 1972; Marso, 1970; Novak, 1974; Owens and Stufflebeam, 1969; Sax and Cromack, 1966; Sirotnik, 1970a; Sirotnik and Wellington, 1974).

Additional studies in MMS have been conducted to examine the effect of stratification of items by item difficulty, by inter-item correlation and by item content regardless of context, on the estimation of the mean and variance of the test score.

Shoemaker and Osburn (1968) found that stratification by item content does not result in a smaller standard error of the mean, but stratification by item difficulty results in greatly improved estimates of the mean when MMS is used. Subsequent research by Feldt and Forsyth (1974), Kleinke (1972) and Myerberg (1979) showed no improvement in the stability of parameter estimation when items are stratified either by item difficulties or by item homogeneity (i.e., inter-item correlation).

On the basis of the above studies one can conclude that it appears that taking test items out of their original context does not seem to systematically affect the estimation of the test mean and variance.

(e) Use of MMS has been generally confined to estimation of group parameters such as the test mean and variance in the context of achievement testing. Some researchers have extended the MMS methodology in other areas. Shoemaker (1971a), for example, has investigated the application of MMS to scaling attitudes. Results indicate that scaled values determined by the method of paired comparison can be

approximated satisfactorily by MMS procedures.

Peterson and Anderson (1971) have compared the factor structure of data collected on a student attitude scale using MMS (MMSWR design) and examinee sampling. They found that the factor structure of an attitude scale obtained using the two sampling procedures was somewhat similar but not equivalent. One possible explanation for this difference may be the fact that the sampling procedures were used on different samples of students and also the data were collected at different times of the year.

Klienke (1972) and Bunda (1973) have extended MMS methodology to the estimation of individual scores. Such procedures are mathematically very complicated and at the present time are not very meaningful (Sirotnik, 1974).

On the basis of the above review, one can conclude that MMS has many potential applications in numerous fields. However, to date, it has primarily been used for estimating group parameters (such as test means and variances) associated with a distribution of test scores.

Traditionally test developers have estimated item parameters by using one group of examinees to respond to all the items in the item pool. Because of the amount of time required per examinee and the little perceived benifit to the examinees or institutions, it has been difficult to obtain a sample of examinees which is both representative and large enough. However if one uses MMS for collecting data on items, it may be easier to get a more representative sample of examinees because with MMS the testing time per examinee

is much shorter. It remains to be seen if the MMS procedure does in fact provide good estimates of item parameters.

This study was designed to explore the use of MMS procedures for item analysis and test development. The problem explored in the study is stated in the following section.

Statement of the Problem

In this study a comparison was made between MMS and examinee sampling with respect to the estimation of item parameters and characteristics of the final test developed using item analysis data obtained from the two sampling procedures. More specifically, the three main questions explored in this study are as follows:

(1) How does the MMS procedure compare with the traditionally used examinee sampling procedure when estimating item parameters such as item difficulty and item discrimination?

(2) How do the final test forms, that are produced using item analysis data obtained from MMS and examinee sampling methods, compare with respect to the test reliability and validity?

(3) How do the number of observations, item subtest size and the characteristics of the items affect the estimates of item parameters and the reliability and validity of tests produced?

CHAPTER II

RESEARCH DESIGN

This study was primarily concerned with the comparison of examinee sampling and MMS plans with respect to the estimation of item parameters and the test reliability and validity.

This chapter includes a discussion of the following: The approach used to collect the data for the study; assumptions made in the study; the characteristics of item sets used and the underlying rationale for selecting these item sets; sampling plans used in the study; the simulation model applied for data collection; and the comparison of the examinee sampling with the MMS sampling plans with respect to estimation of the item parameters and the test reliability and validity.

Approach to Data Collection

A Monte Carlo approach was used to construct item pools from different item sets[1] and to simulate data for the study. The reasons for using a Monte Carlo approach were as follows: One can simulate parameters for different item pools with defined characteristics; examinee scores for these item pools

1. In the study an item set is defined as a set which theoretically may have an infinite number of items with certain characteristics. An item pool is defined as a finite pool of items sampled from an item set.

can be simulated for a random sample of examinees taken from a defined distribution of examinees; sampling methods, examinee sample sizes and the item subtest sizes can easily be varied. Thus this approach allows a comparison to be made between examinee sampling and MMS, and also among different MMS plans, with respect to the estimation of item parameters and the test characteristics across a wide variety of conditions. Furthermore each condition can be repeated any number of times which in turn makes it possible to estimate the stability of results.

Basically in this study, item difficulty and item discrimination indices for each item pool were simulated by having the computer use random number generators. The examinees' ability scores on the trait being measured were also simulated by the computer. The examinees ability scores were normally distributed. For each examinee the item score for each item in the item pool was computed using a linear transformation method. Two different sampling techniques (examinee sampling and MMS) and a number of plans within each technique were used for estimating item parameters and the test reliability and test validity. Results obtained from different sampling plans under different conditions, were compared.

Assumptions

The following assumptions were made in the study: (a) The items in an item set or test measure a single trait and the tests are the norm referenced maximum performance type.

Anastey (1966), Lumsden (1976) and Messick and Jackson (1958) have stated that scores of a test are more meaningful when it is known that only one continuum is involved. If unidimensionality is not evident, summation scoring is illogical as individuals having the same total may possess characteristics which are markedly different. Norm referenced tests were selected here because the item analysis procedures used in the study are most appropriate for norm referenced tests. (b) Items are scored dichotomously but an item response continuum underlies the binary item variate. McNemar (1962) has stated that the knowledge required to answer an item is generally continuously distributed. (c) If guessing is not a factor, this continuum and the trait measured, jointly follow a bivariate normal distribution in the population of examinees from which the sample is drawn. (d) All subjects who do not know the correct answer to an item guess, and do so with a probability of .25 of guessing correctly. The guessing probability is set at .25 to represent an average guessing probability found for examinees guessing randomly in tests containing items with four choices. Although it may be unrealistic to assume that examinees guess randomly, it is difficult to provide a more realistic model for simulation. Zimmerman and Williams (1965) have argued that the results obtained from models which assume random guessing differ minimally from those which do not. (e) The context in which an item occurs does not influence the response, i.e., an individual responds independently to each item. (f) All examinees have the

opportunity to attempt each item in the test or subtest, i.e., speed of response is not a consideration. (g) Thirty items in the final form of the test are sufficient to measure a single trait.

Characteristics of Item Sets

Three types of item sets were used in the study. Each item in a set of items is characterized by two values, its difficulty and discrimination. The item sets varied in the range and the shape of the distribution of difficulties. The distributions of difficulties used and the argument for selecting them is given in the following paragraphs.

For the general purpose of testing, one of the following three distributions of test scores is generally applicable. (a) Normal distribution: it is considered useful for most statistical analyses of results. This distribution is generally obtained when items are homogeneous and item difficulties are spread out and evenly distributed over the range (Lord and Novick, 1968). It was mentioned in the previous chapter that Brogden (1946), Davis (1951), Guttman and Loevinger (Bowers, 1972) and Henrysson (1971) have supported arguments in favor of difficulties which are spread out in a unidimensional test. (b) Rectangular distribution: it provides higher test variance than any other distribution for a given range of ability. It is generally obtained when items are homogeneous and item difficulties are concentrated around .5 (Anstey, 1966). As mentioned in the previous chapter, tests with item difficulties near .5 have been

advocated by authors such as Cronbach and Warrington (1952), Ebel (1972), Gullikson (1945), and Thurstone (1932). (c) Platykurtic distribution: this distribution provides more discrimination among the examinees who fall in the middle of the distribution than does a normal distribution . That is why it is preferred by some researchers (Anstey, 1966). It is generally obtained when items are homogeneous and item difficulties are spread out and normally distributed, i.e., there are more items of middle difficulties (Anstey, 1966).

In order to obtain the normal ,uniform and platykurtic distributions of test scores the following ranges of item difficulties and distributions were used respectively. Item difficulties varying approximately between .16 and .84 and uniformly distributed; item difficulties varying between .40 and .60 and uniformly distributed; item difficulties varying between .16 and .84 and normally distributed. It was assumed in the study that the above ranges and distributions of item difficulties would approximate those of most norm referenced tests. For all item sets the range of item discrimination varied between .05 and .60 and discriminations were uniformly distributed. Theoretically, item discrimination can vary anywhere from -1.00 to 1.00. However, in most published tests, item discrimination rarely exceeds .60 to .70 (Chauncey and Frederkson, 1951; Chartock, 1980). It was expected that the above range of item discrimination would allow some poor items (i.e., items with very low item discrimination) as well as good items (i.e., items with high item discrimination) in the initial item pool.

Using the given combination of difficulties and discriminations, three types of item sets were used in the study. They are shown in Table 1. From each item set 50 item pools with 60 items in each item pool were simulated. These item pools were used for 50 replications done under each condition which are discussed later. Nunnally (1967) has stated that in order to have sufficient room to discard those items which are poor, the pretest should have two to three times as many items as the finished test. Thus if one wishes to have 30 items in the final test form, one should have 60 to 90 items in the initial pool of items. In this study each item pool consisted of 60 items. Although 60 items are fewer than what one would desire to have in an initial pool of items, this number was used because of constraints on computer time.

Sampling Plans

The examinee sampling and the MMS plans used in this study are described in this section.

Examinee Sampling:

For examinee sampling, each examinee responded to all items in the item pool. The three examinee sample sizes (N) used in the study were N=120, N=300 and N=600. They correspond to 7200, 18000, and 36000 observations respectively. Nunnally (1967) has stated that in order to have fairly stable estimates of item parameters, items should

TABLE 1

Characteristics of Item Sets Used in the Study

Item Set	Range of Difficulty	Distribution of Difficulties	Range of Discrimination Indices	Distribution of Discrimination Indices	Expected Test Score Distribution
A	.16 to .84	Uniform	.05 to .60	Uniform	Normal
B	.40 to .60	Uniform	.05 to .60	Uniform	Rectangular
C	.16 to .84	Normal	.05 to .60	Uniform	Platykurtic

be pretested on a large number of subjects, preferably 5 to 10 times as many subjects as items. The sample size N=120 was included here to determine if examinee sample sizes smaller than 5 to 10 times the number of items can give as stable estimates of item parameters as examinee sample sizes equal to 5 to 10 times the number of items.

Multiple Matrix Sampling:

The partially balanced incomplete block design (PBIB) with some restrictions was used to divide items into subtests. The restrictions were that each item was repeated an equal number of times and each item was paired with every other item at least once. This design was chosen because a full matrix of item covariances was needed for estimating item discriminations.

Three item subtest sizes investigated were: one half of the item pool (i.e., 30 items in each subtest), one third of the item pool (i.e., 20 items in each subtest) and one fifth of the item pool (i.e., 12 items in each subtest). The exact procedure used for forming subtests for the three different subtest sizes was as follows: items in the test were first randomly divided into sections equal to one half the size of the subtest required. For example, for subtests of sizes 30, 20 and 12, items were randomly divided into sections of 15, 10 and 6 items, respectively. Then all possible combinations of sections two at a time were taken to form subtests. Subtest sizes smaller than one fifth were not used because

the number of subtests (t) constructed for each test would have been too large for practical purposes. Although it would be preferable that all item pairs be repeated an equal number of times, it is generally not possible to obtain this with PBIB designs. One could use BIB designs to satisfy this condition but BIB designs are impractical to use because the number of subtests required for each test is excessive (Shoemaker, 1973a).

The number of examinees used for each MMS plan (i.e., subtest sizes) was chosen in such a way that the number of observations in different MMS plans corresponded to number of observations in examinee sampling. This was done to keep an equal number of observations for each comparison between examinee sampling and MMS. Shoemaker (1970a) has stated that the number of observations is an important factor in estimating test parameters. An observation is defined as an examinee's response to an item. In examinee sampling, the number of observations is equal to the number of items (K) in the item pool times the number of examinees (N) taking the test. In a MMS plan, the number of observations is found by multiplying the number of subtests (t)by the number of items per subtest (k) by the number of examinees taking each subtest (n). The sampling plans, examinee sample sizes, number of observations for each plan and total number of examinees used for each sampling plan are presented in Table 2.

TABLE 2

Number of Observations, Examinee Sample Sizes and MMS Plans

Number of Examinees Used in Examinee Sampling (N)	Number of Observations 0	MMS Plans t/k/n[a]	Total Number of Examinees Used For Each MMS
120	7200	6/30/40	240
	7200	15/20/24	360
	7020	45/12/13[b]	585
300	18000	6/30/100	600
	18000	15/20/60	900
	17820	45/12/33[b]	1485
600	36000	6/30/200	1200
	36000	15/20/120	1800
	35640	45/12/66[b]	2970

[a] Code: t = Number of subtests
k = Number of items per subtest
n = Number of examinees per subtest

[b] Number of observations for these plans are not equal to the total number of observations as there is no whole number which, after multiplying 45 and 12, results in the total number of observations.

The Simulation Model

The data base for this study was generated on the computer. The item difficulty and item discrimination indices for 50 item pools representing each item set, described in Table 1, were simulated by the computer. The examinee scores for each item pool were also simulated by the computer for both examinee sampling and MMS plans. The procedure for simulating item indices and examinee scores is described below. In order to estimate the stability of results 50 replications were made under each sampling condition and item set. The item pool was varied in each replication. However the same 50 item pools were used with all sampling plans in a item set.

Simulation of Item Indices:

In order to simulate item indices for item pools representing different item sets two pseudo-random[1] number generators RAND and NORM were used at various steps: RAND was designed to generate a variate uniformly (rectangularly) distributed between 0 and 1; and NORM was designed to generate a normal variate with a mean of 0.0 and a standard deviation of 1.0. A further description of these random number generators can be found in Appendix A.

1. In simulatiom studies using a digital computer the generated values are not truely randoms but rather pseudo-random (Hammersley and Handscomb, 1967).

The 60 item difficulties for each item pool representing item set A (associated with a normal distribution of test scores) and item set B (associated with a rectangular distribution) were generated with RAND. The values generated were mapped into specified ranges by Formula 1.

$$(UL - LL) \times RN + LL \qquad (1)$$

Where:

UL = Upper limit of the range of item difficulties.

LL = Lower limit of the range of item difficulties.

RN = The generated pseudo-random number between 0 and 1.

The 60 item difficulties for each item pool representing item set C (associated with a platykurtic distribution of test scores) were generated with NORM. The values obtained from NORM were mapped into a normal distribution with mean of .5 and standard deviation of .113 as specified in Formula 2. This provided a range of values varying approximately between .16 and .84.

$$.5 + .113(Z) \qquad (2)$$

Where:

Z = The generated pseudo-random variate from NORM.

The 60 discrimination indices for all item pools representing different item sets were generated with RAND. They were mapped into the specified range by Formula 1.

Using the above procedure the item indices for 50 item pools representing each item set were generated. They were

stored in three separate files, one file for each type of item set (see Flow chart A in Appendix B). These files were used for 50 replications done under each testing condition.

Simulation of Examinee Scores:

The examinee scores for all sampling plans were generated by the following steps. In Step 1, a given number of examinees (see Table 2) of varying ability on the trait being measured was generated using NORM, for each testing condition of examinee sampling and MMS. These examinees represented a pseudo-random sample of examinees taken from a normal population of examinees on the trait being measured.

In Step 2, a continuous item score (Y_g) for item i was computed for examinee g using the following linear transformation:

$$Y_g = d_i X_g + (\sqrt{1-d_i})Z \qquad (3)$$

Where:

Y_g= Standard score of examinee g on item i.

d_i= Discrimination parameter of item i (i.e., Pearson Product Moment correlation between the trait and item score.

X_g= Ability of examinee g in standard score.

Z = A pseudo-random variate generated by NORM.

Once the Y_g score was obtained, it was dichotomized using the difficulty parameters. If Y_g was greater than the difficulty of item i (item difficulty was converted into a standard score by locating the point in the normal

distribution which corresponded to the difficulty level of the item), a score of "1" was assigned to the examinee for that item. If Y_g was less than the item difficulty, the examinee was allowed to guess with a probability of .25 of guessing correctly. A pseudo-random variate between zero and one was generated using the computer program RAND. If this variate was equal to or less than .25, it was assumed that the examinee guessed the item correctly and a score of "1" was assigned, otherwise a score of "0" was assigned. Step 2 was repeated until scores for each item in an item set were generated for each of the given number of examinees. At the end of Step 2, a matrix of examinee by item scores has been generated. Since the number of examinees was varied in different examinee sampling and MMS plans a different size matrix of examinee by item scores was generated for each of the examinee and MMS plans stated in Table 2. For example, for the examinee sampling where N=120 a matrix of 120 by 60 of examinee by item scores was generated whereas for the MMS shown as 6/30/40 in Table 2 a matrix of 240 by 60 of examinee by item scores was generated.

For each examinee sampling plan the matrix of examinee by item scores was used to estimate item difficulties and item discriminations. For each MMS plan examinees and items were sampled from the examinee by item matrix according to the plan. For example, for the MMS plan 6/30/40 (given in Table 2) first a matrix of 240 by 60 of examinee by item scores was generated then for the first 40 examinees (n), item scores for the first subtest were selected, for the next 40

examinees, item scores for the second subtest were selected and so on. In step 3 estimates of item parameters and other statistics were calculated using these sample data. Following this, in step 4, 30 items with the highest estimated biserial r were selected for the final form of the test, from each sampling plan. The final form of the test made up of 30 items with the highest estimated item discrimination is referred to as "final test" in the study. The test reliability (internal consistency) and test validity of the final test were estimated. In order to obtain estimates of test reliability and validity of the final test, the responses to these 30 items were simulated by computer for a random sample of 300 examinees. The procedure used for computer simulation of responses was the same as that described earlier. Three hundred examinees is ten times the number of items. Nunnally (1967) has suggested that in order to have stable estimates of parameters one should have a sample of examinees 5 to 10 times as large as the number of items. In addition to the above, 30 items were also randomly selected from the pool of items. The responses to these items were also simulated by computer for a sample of 300 examinees, and test reliability and validity were estimated. This was done to allow comparison between relibility and validity of final tests in which items were carefully selected with those tests in which items were randomly selected.

Steps 1 to 4 were repeated 50 times, giving 50 replications for each type of item set under each of the 12 sampling conditions described in Table 2.

Estimation of Item Parameters and Test Reliability and Validity

Estimation of Item Parameters:

The two item parameters estimated were item difficulty and item discrimination. The difficulty of an item was estimated by calculating the proportion of subjects among those who responded to the item, that answered it correctly. The item discrimination was estimated by calculating the biserial correlation between an item score and total test score. This measure of discrimination is considered relatively stable and independent of item difficulty (Pyrezak, 1973; Lord and Novick, 1968). The exact procedures used for estimating item difficulty and item discrimination for examinee and MMS sampling plan are as follows:

Examinee Sampling:

The formulae used for estimating item difficulty ($\hat{p}_i$) and biserial correlation ($\hat{d}_i$) between an item and total test score are as follows:

$$\hat{p}_i = \frac{\sum_{j=1}^{N} X_{ij}}{N} \qquad (4)$$

where:

$\hat{p}_i$ = estimate of difficulty of item i

$\sum_{j=1}^{N} X_{ij}$ = sum of scores of N examinees on item i

$$\hat{d}_i = \hat{\rho}_{Ti} \left(\frac{\hat{p}_i \hat{q}_i}{Y} \right) \qquad (5)$$

Where:

$\hat{d}_i$ = estimate of r-biserial between item i and total test score.

$\hat{\rho}_{Ti}$ = Point biserial correlation between item i total test score (T).

$\hat{p}_i$ = defined previously.

$\hat{q}_i$ = $(1-\hat{p}_i)$

Y= the ordinate of the dividing line between the proportions $\hat{p}_i$ and $\hat{q}_i$ in a unit normal curve.

Multiple Matrix Sampling:

The formula used for estimating item difficulty and biserial r between an item and total test score is

$$\hat{p}_i = \frac{\sum_{j=1}^{M} X_{ij}}{M} \qquad (6)$$

Where:

$\hat{p}_i$ = estimate of difficulty of item i.

$\sum_{j=1}^{M} X_{ij}$ = sum of scores of M examinees on item i. Where M is the number of examinees who responded to item i.

In the MMS procedure, since each examinee did not respond to every item in the test, total test scores were not available. Thus in order to calculate biserial r, first the matrix of covariances among the items was obtained. The size of this matrix was 60 by 60. When

estimating the covariance between two items only those item scores were used for which data were available on both items. One could also estimate the covariance between items using the generalized symmetric mean (gsm) approach proposed by Knapp (1979), which uses all the data available on items. However Knapp has stated that although the gsm approach gives an unbiased estimate of population covariance, it has a larger standard error when there is missing data (which is the case in MMS) than the covariance which is obtained using only those pairs of measurements for which complete data are available. Since all items were not paired an equal number of times, some covariance estimates were based on a large number of observations where others were based on a smaller number of observations.

The following formula was used for estimating biserial r between an item and total test score.

$$\hat{d}_i = \left(\frac{\sum_{j=1}^{K} cov_{ij}}{\sqrt{\hat{\sigma}_T^2 \, \hat{\sigma}_i^2}} \right) \left(\frac{\sqrt{\hat{p}_i \hat{q}_i}}{Y} \right) \qquad (7)$$

Where:

$\hat{d}_i$ = estimate of biserial r for item i.

$\sum_{j=1}^{K} cov_{ij}$ = sum of all the elements in row i of the matrix of inter item covariance. This is equal to item total covariance (Magnusson, 1966).

$\hat{\sigma}_T^2$ = estimate of test variance.

$\hat{\sigma}_i^2$ = estimate of item variance. It is equal to $\hat{p}_i$ times $\hat{q}_i$.

The estimate of test variance was obtained using the gsm approach described by Sirotnik & Wellington (1977). They have shown that this approach gives an unbiased estimate of the population variance, and that the standard error of this variance in MMS designs is less than the standard error obtained using any other approach.

Estimation of Test Reliability and Validity:

The test characteristics were measured by calculating the reliability and validity of the final test. The main aim of a test constructor is to construct a test which is reliable and valid. When constructing a test designed to measure a single underlying trait, one of the statistical criteria often used by test constructors to select or reject items for the final version of the test, is an estimate of item discrimination. The items with highest item discrimination are selected for the final test (Ghiselli, Campbell and Zedeck, 1981; Lord and Novick, 1968). According to Ghiselli, Campbell and Zedeck (1981) and Lord and Novick (1968) for tests constructed in the above manner the internal consistency of the test provides a useful approximation to the reliability of the test. In this study the 30 items with the largest estimated item discrimination were selected to be included in the final test. The reliability of the final test was estimated using the Kuder Richardson Formula 20 (e.g., Magnusson, 1966).

$$r_{xx} = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum_{i=1}^{K} \hat{\sigma}_i^2}{\hat{\sigma}_T^2}\right) \qquad (8)$$

Where

r_{xx}= reliability of the test.

$\sum_{i=1}^{K} \hat{\sigma}_i^2$= sum of item variances.

K= number of items (i.e., 30).

$\hat{\sigma}_T^2$= estimate of test variance.

The criterion validity of the final test was obtained by calculating the Pearson correlation coefficient between the total test score for each examinee and the corresponding ability score. It should be noted however that in practice one can not estimate validity this way, as the ability scores on the trait being measured are not known. In addition to the above estimates of reliability and validity under each replication of each sampling plan and condition, the reliability and validity estimates were also obtained for 30 items where items were randomly selected from the pool of items. This was done to determine if the reliability and validity of a final test composed of the "best" (items with largest item discrimination) items are in fact better than those of a test in which items are randomly selected.

The mean, standard deviation and range of the 50 estimates of reliability and validity of the final tests, as well as reliability and validity of tests composed of 30 randomly selected items, were obtained for each sampling plan and condition.

Comparison of Sampling Methods

The sampling plans were compared with respect to item parameters, test reliability and validity. Three dependent measures were selected to compare sampling plans with respect to item parameters. In this section description of the dependent measures, their estimation procedure and the methods applied for comparing results are discussed.

Dependent Measures:

The following three dependent measures were selected to measure the efficiency of estimates of item parameters.

(a). The mean squared error between estimated and true item parameters. The estimates which on the average have smaller mean square error are considered better than the estimates which on the average have larger mean square error.

(b). The average correspondence between estimated and true parameters. The correspondence is measured using Spearman rank order correlation. If the rank order correlation is high, there is a high correspondence, i.e., the estimated values are in an order similar to the true values. If one method consistently underestimates or overestimates an item parameter more than the other method, Spearman rank order correlation will not be affected whereas the mean squared error will be affected. The estimates which on the average have a larger Spearman rank order correlation are considered better than those which on the average have a smaller correlation.

(c). The proportion of misclassification of items.

This is estimated by calculating the following two proportions and adding them together. The proportion of items with true discrimination index higher than a critical value of .30 and estimated discrimination index lower than the critical value. The proportion of items with a true discrimination index lower than the critical value and an estimated discrimination index higher than the critical value. The item discrimination index is considered a fairly valid indicator of item quality (Ebel, 1972; Pyrczak, 1973). Ebel (1972) has suggested that any item with a discrimination index less than .30 is a marginal item and generally should be improved. Berk (1978) has also stated that in item selection procedures items with discrimination indices below .30 are often rejected in favour of those with greater discrimination power. Often when pretesting a pool of items, the test constructor is not really interested in the exact value of the true item discrimination of each item; what he is interested in is knowing whether this value is above or below a critical value (which is chosen to be equal to .30). Accordingly, the accuracy of sample statistics (estimates) may be measured, not in terms of the distance between the statistic (estimate) and the parameter of an item but in terms of the proportion of misclassifications of the items in the pretest pool. An item will be misclassified if, for instance, the real item discrimination is .40 and its estimate in the pretest sample is .28. On the other hand, an item with true discrimination of .55 that appears in the pretest sample as having a discrimination of .7 will not be

misclassified. Thus this measure is designed to estimate the proportion of misclassifications. The estimates of item discrimination which on the average have a smaller proportion of misclassified items are considered better than those which have a larger proportion of misclassified items.

Estimation of Dependent Measures:

(a). Mean square error (MSE):

The mean square errors for difficulty and discrimination item parameters were obtained by comparing estimated parameters with true parameters . The formula used for calculating the MSE for the difficulty parameter is as follows:

$$MSE\ (\hat{p}) = \sum_{i=1}^{K} \frac{(p_i - \hat{p}_i)^2}{K} \qquad (9)$$

Where:

$MSE(\hat{p})$= MSE for difficulty parameter.

p_i= true difficulty for item i.

$\hat{p}_i$= estimated difficulty for item i.

K= number of items (i.e., 60).

The 50 estimates of MSE(p) (one for each run) for each sampling plan under each condition were obtained. The mean, range and standard deviation of these 50 estimates were calculated for each sampling plan and condition.

The formula used for calculating MSE for biserial r [MSE(d)] is as follows:

$$MSE(\hat{d}) = \sum_{i=1}^{K} \frac{(d_i - \hat{d}_i)^2}{K} \qquad (10)$$

Where:

$MSE(\hat{d})$= MSE for discrimination parameter.

d_i= true discrimination parameter for item i.

$\hat{d}_i$= estimated discrimination parameter for item i.

The mean, range and standard deviation of the 50 estimates obtained for each sampling plan and condition were calculated.

(b). Spearman rank order correlation coefficient:

The Spearman rank order correlation between estimated and true parameters for each sampling plan and condition was obtained to measure average correspondence between true and estimated parameters for both difficulty and discrimination parameters. The Spearman rank order correlation between estimated and true item difficulties and between estimated and true item discriminations are referred to as $\rho(p\hat{p})$ and $\rho(d\hat{d})$ respectively. The mean, range and standard deviation of the 50 estimates of Spearman rank correlation coefficients were also computed for each sampling plan under each condition.

(c). Proportion of misclassification of items:

For each sampling plan and condition, after estimates of discrimination parameters were obtained, the proportion of items with $d_i > .30$ and $\hat{d}_i < .30$ was calculated. Similarily the

proportion of items with $d_i < .30$ and $\hat{d}_i > .30$ was also calculated. These proportions were added together to obtain the proportion of misclassifications. The mean, range and standard deviation of the 50 estimates of proportion of misclassification of items were obtained for each sampling plan and condition.

Comparison of Results:

This study was designed to answer the following questions: How does the MMS procedure compare with the traditionally used examinee sampling procedure when estimating item parameters such as item difficulty and item discrimination? How do the final tests, that are produced using item analysis data obtained from MMS and examinee sampling methods, compare with regard to test reliability and validity? How do the number of observations, item subtest sizes and the characteristics of item sets affect the estimates of item parameters and the reliability and validity of tests produced?

In order to answer the first question examinee sampling and MMS were compared with regard to the MSE($\hat{p}$), $\rho(p\hat{p})$, MSE($\hat{d}$), $\rho(d\hat{d})$ and the proportions of misclassified items. While comparing the sampling plans, the number of observations and type of item set were kept constant. The sampling plan which on the average had relatively smaller MSE($\hat{p}$) and larger $\rho(p\hat{p})$ was considered better than others for estimating item difficulty parameters. Similarly the sampling plan which on the average had relatively smaller

MSE($\hat{d}$) and proportions of misclassified items and larger $\rho(d\hat{d})$ was considered better than other sampling plans for estimating item discrimination parameters.

The second question was answered by comparing examinee sampling with MMS plans, for a given number of observations and item set with regard to the average reliability and validity of the final tests. The sampling plan which on the average provided relatively higher reliability was considered better in terms of reliability than the other sampling plans. Similarily the sampling plan which provided higher validity was considered better than the other sampling plans in terms of validity. In addition to comparing the sampling plans with regard to the average reliability and validity of the final tests, the average reliability and validity of the final tests were also compared with those of the 30 random item tests. These comparisons were done to see how the average reliability and validity of the final tests compared with those in which items were selected randomly. If the final tests were good tests, one would expect that their reliability and validity would be larger than those of tests in which items were selected randomly.

In order to answer the third question the results obtained from different sampling plans for varying numbers of observations were compared with regard to MSE($\hat{p}$), $\rho(p\hat{p})$, MSE($\hat{d}$), $\rho(d\hat{d})$, proportion of misclassified items, average reliability and average validity, keeping the item set constant. Similarly the results obtained from different MMS plans were compared with regard to the above measures while

keeping the number of observations and item sets constant but varying the item subtest sizes. The results obtained from different item sets were also compared with regard to the above measures, while keeping the number of observations and sampling plan constant.

CHAPTER III

RESULTS AND DISCUSSION

In this chapter the results are presented and discussed. The chapter is divided into two major sections. In the first section the results associated with estimates of item parameters obtained from the MMS plans and the examinee sampling plans under different conditions are discussed. In the second section the results associated with reliability and validity of the final test forms produced from MMS and examinee sampling methods, are discussed. The results associated with the number of observations, item subtest size and the characteristic of item sets are incorporated in the two sections described above.

Estimates of Item Parameters

The estimates of item parameters obtained from examinee sampling plans and MMS plans under different conditions were compared with respect to the following statistics: Mean square error for item difficulty [MSE($\hat{p}$)]; Spearman rank correlation between true and estimated item difficulties [$\rho(p\hat{p})$]; Mean square error for item discrimination [MSE($\hat{d}$)]; Spearman rank correlation between estimated and true item discrimination [$\rho(d\hat{d})$]; the proportion of items that are misclassified. The results obtained from the above measures are discussed below.

Mean Square Error for Item Difficulty [MSE(p)]:

The mean square error between estimated and true item difficulties was obtained for each replication, sampling plan and condition. Since 50 replications were made under each condition, the mean, standard deviation and range (minimum and maximum) of these 50 estimates of MSE($\hat{p}$) were calculated. The results are shown in Table 3.

The results show that for a given item set and number of observations, the means of MSE($\hat{p}$) were almost the same whether they were obtained using examinee sampling or MMS plans. The item difficulty of an item is defined as the proportion of correct responses among those who answered the item. Since the number of examinees who answered each item under each sampling plan of a test was kept constant for a given number of observations, one would expect that the estimates of item difficulties would be similar except for sampling fluctuations. The results shown tend to support this expectation.

The distribution of item difficulties in a item set and number of observations had some effect on the average MSE($\hat{p}$). For a given number of observations, on the average, MSE($\hat{p}$) for item set B is consistently lower than the average MSE($\hat{p}$) for item set C which in turn is consistently lower than MSE($\hat{p}$) for item set A. The reason for this variation could be the distribution of item difficulties in a item set. In item set B, item difficulties varied between.40 and .60, whereas item difficulties in item set A and item set C varied

TABLE 3

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of MSE ($\hat{P}$) for Item Sets A, B and C

Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plans	Item Set A				Item Set B				Item Set C			
		Mean	SD[a]	MIN[b]	MAX[c]	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.020	.003	.014	.026	.017	.002	.011	.022	.018	.002	.014	.025
	MMS (t/k/n)												
7200	6/30/40	.020	.003	.011	.027	.018	.002	.013	.023	.019	.003	.013	.026
7200	15/20/24	.020	.003	.015	.025	.018	.002	.013	.023	.019	.002	.013	.024
7200	45/12/13	.020	.003	.015	.026	.017	.002	.013	.021	.018	.002	.013	.024
18000	Examinee Sampling (N=300)	.019	.003	.013	.025	.016	.002	.011	.021	.017	.001	.014	.021
	MMS (t/k/n)												
18000	6/30/100	.019	.002	.014	.024	.016	.001	.013	.019	.017	.002	.013	.021
18000	15/20/60	.019	.002	.014	.024	.016	.001	.013	.019	.017	.002	.014	.022
17820	45/12/33	.019	.002	.015	.022	.016	.001	.013	.019	.017	.001	.015	.020
36000	Examinee Sampling (N=600)	.018	.002	.013	.023	.016	.001	.012	.018	.017	.001	.013	.020
	MMS (t/k/n)												
36000	6/30/200	.018	.002	.014	.022	.016	.001	.014	.018	.017	.001	.015	.021
36000	15/20/120	.018	.002	.014	.022	.016	.001	.014	.018	.017	.001	.015	.019
35640	45/12/66	.018	.002	.014	.022	.016	.001	.013	.018	.017	.001	.014	.019

Code: [a] SD = Standard deviation

[b] MIN = Minimum value

[c] MAX = Maximum value

between .16 and .84. In item set A, item difficulties were uniformly distributed over the range whereas in item set C they were normally distributed around the mean of .50. Thus there were more items of extreme (very high and very low) difficulties in item set A than in item set C, which in turn had more items of extreme difficulty than item set B. One would assume that the examinees would be more likely to guess when an item is relatively more difficult than when it is easy. Thus an estimate of an item parameter will fluctuate more from the true item parameter when the item is relatively more difficult than when it is easy. Since the differences between estimated and true item parameters are squared in the calculation of MSE, tests consisting of a relatively large number of very difficult items would inflate (increase) MSE more than those consisting of a relatively small number of very difficult items. There were more items of extreme difficulty in item set A than in item set C, which in turn had more items of extreme difficulty than item set B. Thus one would expect largest MSE($\hat{p}$) for item set A and smallest MSE($\hat{p}$) for item set B. The results tend to support this expectation.

For both examinee sampling and MMS plans, as the number of observations increased from 7200 to 18000 the mean MSE($\hat{p}$) decreased consistently. On the average, for item sets B and C, any further increase in the observations made no difference in MSE($\hat{p}$). For item set A, as the number of observations increased from 18000 to 36000, the average MSE($\hat{p}$) decreased still further. It may be due to the

distribution of item difficulties in the set. In item set A there were relatively more items of extreme difficulty than in item sets C and B. Since the guessing factor inflates the estimation of item difficulties more when items are relatively more difficult than when they are easy, one would expect that a relatively larger number of observations will be required to get stable results under such conditions.

The maximum standard deviation of 50 estimates of MSE($\hat{p}$) for all item sets and sampling plans is .003. Thus one can infer that estimates of item difficulties tend to be fairly consistent from sample to sample.

On the basis of the above results, one can conclude that for a given item set and number of observations, there tends to be little difference in mean MSE($\hat{p}$) obtained from examinee sampling and different MMS plans. On the average, MSE($\hat{p}$) tends to be smaller for item sets where most items are of medium difficulty than for item sets where item difficulties are spread out. This difference may be caused by the presence of a larger number of very difficult items in item sets in which item difficulties were spread out. As the number of observations increases, average MSE($\hat{p}$) decreases for all tests and sampling plans. However, one would require larger numbers of observations to obtain stable estimates of item difficulties when the item difficulties in a test are spread out (possibly to account for larger amount of guessing associated with very difficult items) than when most items in a test are of medium difficulty.

The Spearman Rank Correlation between Estimated and True Item Difficulties [$\rho(p\hat{p})$]:

The Spearman rank correlation between true and estimated item difficulties was obtained for each sampling plan and condition to measure correspondence between estimated and true item difficulties. Since 50 replications were made under each condition, the mean, standard deviation and range (minimum and maximum value) of 50 estimates of $\rho(p\hat{p})$ were calculated. The results are reported in Table 4.

Most of the results given in Table 4 are consistent with results in Table 3, except for findings associated with the different item sets.

For a given item set and number of observations, the mean of $\rho(p\hat{p})$ is approximately the same for examinee sampling and MMS plans. As the number of observations increased on the average $\rho(p\hat{p})$ also increased for all sampling plans and item sets.

The results in Table 4 show that for all sampling plans, for a given number of observations the mean $\rho(p\hat{p})$ is lower for the item set B than for the item set C, which in turn is lower than the item set A. On the basis of this, one would assume that correspondence between estimated and true item difficulties is lowest for item set B. Contrary to this finding, results in Table 3 show that on the average $MSE(\hat{p})$ is lowest for item set B and highest for item set A. The reason for this discrepancy of results seemed to be the susceptibility of rank correlation to the range of item difficulties in a test. In item set B, item difficulties

TABLE 4

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of $\rho(P\hat{P})$ for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plans	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.956	.015	.901	.978	.617	.088	.434	.777	.844	.060	.569	.911
	MMS (t/k/n)												
7200	6/30/40	.952	.010	.932	.970	.608	.082	.427	.786	.838	.043	.721	.913
7200	15/20/24	.953	.012	.919	.975	.623	.084	.448	.795	.847	.844	.713	.931
7200	45/12/13	.949	.015	.912	.977	.612	.083	.431	.787	.841	.044	.753	.931
18000	Examinee Sampling (N=300)	.978	.007	.954	.989	.780	.063	.543	.864	.922	.023	.861	.961
	MMS (t/k/n)												
18000	6/30/100	.978	.006	.960	.990	.783	.052	.653	.906	.925	.022	.868	.971
18000	15/20/60	.979	.004	.969	.988	.780	.046	.701	.885	.924	.025	.802	.960
17820	45/12/33	.977	.006	.960	.989	.785	.054	.621	.906	.926	.022	.870	.959
36000	Examinee Sampling (N=600)	.987	.003	.980	.993	.878	.024	.810	.933	.955	.018	.875	.979
	MMS (t/k/n)												
36000	6/30/200	.987	.004	.975	.992	.864	.032	.791	.925	.954	.014	.912	.983
36000	15/20/20	.988	.003	.980	.993	.870	.039	.745	.949	.957	.012	.932	.979
35640	45/12/66	.987	.005	.965	.994	.864	.033	.758	.916	.953	.019	.901	.980

were clustered near .5 and varied between .40 and .60. Whereas in item set A and C, item difficulties were spread out over the range of .16 to .84. However, the standard deviation of item difficulties was lower in item set C than in item set A, because the distribution of item difficulties was normal around .50 in item set C whereas the distribution of item difficulties was uniform over the range in item set A. Statisticians have shown that the size of the correlation coefficient is influenced by the size of the standard deviation or the range of scores being correlated (Nunnally, 1975; Shavelson, 1981).

One can conclude that the dependent measures MSE($\hat{p}$) and $\rho(p\hat{p})$ used to evaluate the estimates of item difficulties obtained from examinee sampling and MMS plans tend to provide similar conclusions, except for the results associated with different item sets. The conclusions are as follows: There is little difference in the accuracy with which item difficulties are estimated by examinee sampling and MMS plans; as the number of observations increases the accuracy with which item difficulties are estimated also increases.

Mean Square Error for Item Discrimination [MSE($\hat{d}$)]:

In order to see what effect different sampling plans (examinee sampling versus MMS plans), the number of observations, item subtest sizes in MMS plans and different characteristics of item sets have on estimation of item discrimination, the mean square error between estimated and true item discriminations was obtained for all item sets

under each sampling plan and condition. Since 50 replications were made under each of the above conditions, the mean, standard deviation and range (minimum and maximum value) of 50 estimates of MSE($\hat{d}$) were calculated. The results are reported in Table 5.

The results show that for a given item set and number of observations, the MSE($\hat{d}$) is lower for examinee sampling than for MMS plans. However, the difference between MSE($\hat{d}$) of the examinee sampling plan and the MMS plans decreased as the number of items per subtest increased while keeping the number of observations constant. Comparisons among different MMS plans for a given number of observations show that on the average MSE(d) decreased as the number of items in each subtest increased from one-fifth of the total number of items to one-half of the total number of items. Two factors that may have influenced the above results, could be the precision with which total test variance and inter item covariances are estimated. Barcikowski (1974) and Moy and Barcikowski (1974) have shown that the most precise (lowest standard error) estimates of population variance are obtained from those MMS plans which contain more items per subtest (i.e., MMS plans which contain one-quarter to one-half of the total items in each subtest) than from MMS plans which contain fewer items per subtest. For a given item set and number of observations one would expect that the estimate of total test variance would become more precise as the number of items in each subtest would increase. For a given number of observations and item sets, the minimum number of times each item was

TABLE 5

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of MSE ($\hat{d}$) for Item Sets A, B and C

Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.018	.003	.013	.027	.015	.002	.010	.020	.017	.003	.010	.027
	MMS (t/k/n)												
7200	6/30/40	.026	.006	.014	.040	.022	.005	.012	.034	.021	.004	.014	.036
7200	15/20/24	.037	.008	.023	.061	.029	.007	.016	.047	.031	.007	.017	.053
7200	45/12/13	.055	.015	.029	.095	.045	.012	.028	.084	.046	.010	.027	.065
18000	Examinee Sampling (N=300)	.011	.002	.007	.016	.009	.001	.006	.012	.009	.002	.005	.017
	MMS (t/k/n)												
18000	6/30/100	.014	.003	.009	.020	.011	.002	.008	.017	.012	.002	.008	.021
18000	15/20/60	.017	.003	.011	.025	.014	.002	.010	.021	.016	.003	.010	.024
17820	45/12/33	.025	.005	.016	.040	.020	.003	.014	.030	.022	.004	.013	.028
36000	Examinee Sampling (N=600)	.008	.002	.005	.011	.006	.001	.004	.008	.007	.001	.004	.010
	MMS (t/k/n)												
36000	6/30/200	.009	.002	.005	.012	.007	.001	.004	.010	.008	.002	.005	.013
36000	15/20/120	.011	.002	.007	.017	.009	.002	.006	.013	.009	.002	.007	.013
35640	45/12/66	.015	.003	.009	.020	.012	.002	.009	.018	.013	.002	.008	.020

paired with other items was highest for the examinee sampling and it decreased as the number of items in each subtest of MMS plans decreased. Statistically the number of pairs used for estimating a population correlation coefficient influences the stability with which the correlation coefficient is estimated. Since the biserial r between an item and the total test score is affected by inter-item covariance and test variance (see Formula 7), anything that influences estimation of inter-item covariance and test variance may also influence estimation of biserial r.

The results also show that for each sampling plan and item set, the MSE(d) decreased as the number of observations increased from 7200 to 36000. Lord and Novick (1968), Neill and Jackson (1970) and Nunnally have stated that one should have a large sample of examinees (preferably five to ten times the number of items) to have good estimates of item discriminations. The results of this study tend to support the above authors. The standard deviation of 50 estimates of MSE($\hat{d}$) also decreased as the number of observations increased also as the number of items in each subtest increased. Thus one can say that estimates of item discrimination not only become more accurate they also become more stable as the number of observations increases and as the number of items in each subtest increases.

Comparison of results of different item sets for examinee sampling plans and MMS plans show that for a given number of observations, on the average MSE($\hat{d}$) for item set B is consistently lower than MSE($\hat{d}$) for item set C which in

turn is consistently lower than MSE($\hat{d}$) for item set A. The reason for this difference in MSE($\hat{d}$) could be the distribution of item difficulties in different item sets. The influence of item difficulty on guessing and MSE has been described earlier. Ashler (1979) has also shown that biserial r underestimates the correlation coefficient in the presence of guessing. This influence is substantial when items are difficult. Kendall and Stuart (1967) have shown that the standard error for a fixed biserial r is smallest for items of .50 difficulty and it increases as item difficulty deviates from .50. On the basis of above findings one would expect that the estimates of item discriminations (biserial r) would be most stable for item set B because most items are of medium difficuly, and least stable for item set A because there are relatively larger number of difficult items in this item set than in other item sets, also item difficulties in item set A are uniformly spread out over a wider range. The results tend to support this expectation.

On the basis of above results and discussion one can conclude that if the number of observations are kept constant, MSE($\hat{d}$) is lower for examinee sampling than for MMS plans. However, this difference decreases as the number of observations and the number of items in each subtest of MMS plans increases. For a given number of observations and sampling plans, MSE($\hat{d}$) is lower for item sets where most items are of medium difficulty than for item sets where item difficulties are spread out. This is more likely due to differential influence of guessing and the standard error of

biserial r on items of different difficulty.

Spearman Rank Correlation Between True and Estimated Item Discrimination [$\rho(d\hat{d})$]:

In order to determine the correspondence between estimated and true item discrimination, Spearman rank correlation between true and estimated item discriminations was obtained for each sampling plan, under all conditions. Under each condition, 50 replications were done. The mean, standard deviation and range (minimum and maximum value) of 50 estimates of $\rho(d\hat{d})$ were calculated. The results are reported in Table 6.

The results shown in Table 6 are consistent with the results in Table 5. For a given number of observations and item set $\rho(d\hat{d})$ was higher for examinee sampling than for MMS plans. However, this difference in $\rho(d\hat{d})$ decreased as the number of observations and number of items per subtest increased. The average $\rho(d\hat{d})$ is higher for item set B than for item set C, which is higher than the average $\rho(d\hat{d})$ for item set A. This may be because the estimates of biserial r tend to be most stable for medium difficulty items.

Proportion of Misclassification of Items:

As discussed earlier the accuracy with which item discrimination is estimated can be determined in terms of proportion of misclassifications of items with regard to a certain critical value (i.e., .30) which was used for selecting items for the final test. An item is misclassified if (a) the true item discrimination (d_i) is greater than the critical value and estimated item discrimination ($\hat{d}_i$) is less

TABLE 6

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of $\rho(d\hat{d})$ for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.640	.084	.421	.793	.699	.072	.458	.804	.657	.088	.376	.821
	MMS (t/k/n)												
7200	6/30/40	.532	.101	.266	.710	.599	.099	.378	.768	.603	.096	.281	.767
7200	15/20/24	.461	.097	.217	.611	.526	.097	.301	.750	.501	.106	.276	.771
7200	45/12/13	.366	.121	.054	.588	.432	.129	.155	.666	.421	.109	.211	.643
18000	Examinee Sampling (N=300)	.785	.060	.613	.872	.834	.035	.727	.890	.814	.054	.699	.915
	MMS (t/k/n)												
18000	6/30/100	.716	.065	.554	.863	.770	.064	.545	.871	.750	.064	.479	.835
18000	15/20/60	.654	.068	.481	.782	.701	.057	.578	.800	.678	.077	.469	.858
17820	45/12/33	.563	.099	.294	.706	.621	.086	.408	.773	.596	.073	.453	.714
36000	Examinee Sampling (N=600)	.865	.031	.757	.910	.905	.021	.846	.944	.888	.028	.814	.935
	MMS (t/k/n)												
36000	6/30/200	.818	.044	.714	.894	.865	.029	.794	.907	.843	.039	.747	.914
36000	15/20/120	.763	.054	.651	.851	.818	.040	.710	.887	.812	.045	.701	.889
35640	45/12/66	.680	.069	.514	.837	.751	.062	.588	.866	.729	.060	.555	.832

than the critical value or (b) if the true item discrimination is less than the critical value and the estimated item discrimination is greater than the critical value.

For each replication of an item set and sampling plan, the proportion of items with $d_i > .30$ and $\hat{d}_i < .30$ and also the proportion of items with $d_i < .30$ and $\hat{d}_i > .30$ was calculated. The sum of the above two proportions was computed to obtain the proportion of misclassified items. Since 50 replications were made under each condition, the mean, standard deviation and range of 50 estimates of the above proportions of misclassified items were computed. The results are reported in Table 7.

The results given in Table 7 are consistent with findings obtained from results in Tables 5 and 6. For a given item set and number of observations, the proportion of items that were misclassified was less for examinee sampling than for MMS plans. However, this difference in the proportion of misclassified items decreased as the number of items in each subtest increased also as the number of observations increased. For a given number of observations and a given sampling plan, the proportion of items that were misclassified was consistently lower for item set B than for item set C, which was consistently lower than item set A. This could be due to the different distribution of item difficulties in item sets A, B, and C. The influence of guessing and standard error of biserial r on estimation of item discrimination of items of different difficulty has been

TABLE 7

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of Proportion of Items that are Misclassified for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plans	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.249	.064	.100	.400	.223	.048	.117	.317	.244	.061	.117	.383
	MMS (t/k/n)												
7200	6/30/40	.291	.059	.133	.417	.249	.058	.117	.367	.266	.068	.133	.433
7200	15/20/24	.337	.058	.200	.517	.301	.061	.167	.417	.306	.070	.167	.450
7200	45/12/13	.377	.067	.217	.533	.336	.078	.183	.517	.332	.069	.117	.433
18000	Examinee Sampling (N=300)	.181	.046	.067	.300	.151	.049	.067	.267	.162	.050	.067	.300
	MMS (t/k/n)												
18000	6/30/100	.227	.049	.133	.333	.187	.057	.067	.317	.193	.054	.083	.383
18000	15/20/60	.249	.054	.156	.333	.225	.046	.317	.184	.233	.053	.100	.383
17820	45/12/33	.273	.061	.133	.383	.250	.047	.167	.350	.264	.061	.150	.417
36000	Examinee Sampling (N=600)	.143	.046	.033	.217	.108	.039	.017	.183	.112	.042	.017	.217
	MMS (t/k/n)												
36000	6/30/200	.170	.048	.033	.283	.145	.036	.083	.217	.150	.041	.050	.250
36000	15/20/120	.193	.058	.083	.317	.161	.044	.033	.250	.167	.052	.033	.267
35640	45/12/66	.226	.049	.117	.350	.203	.043	.100	.300	.205	.047	.133	.317

described earlier.

With regard to estimation of item discrimination, three dependent measures, MSE($\hat{d}$), $\rho(d\hat{d})$ and proportion of misclassification, were used to determine the accuracy with which examinee sampling and MMS plans estimate item discrimination. The results obtained from each of these three dependent measures lead to the same conclusions, which are as follows: When the number of observations and characteristics of items in an item set are kept constant, examinee sampling provides consistently better estimates of item discrimination than MMS plans. However, this difference becomes smaller as the number of observations increases to 36000 and the number of items in each subtest of MMS plan increase from one-fifth to one-half of the total items in the test. For both examinee sampling and MMS plans, the accuracy with which item discrimination is estimated, increases as the number of observations increases from 7200 to 36000. Among the different MMS plans, plans with a large number of items per subtest (one-third to one-half of the total number of items in each subtest) provide more accurate estimates of item discriminations than plans with a small number of items (one-fifth of the total test items) per subtest. For both sampling plans, the estimates of item discrimination tend to be more accurate for item sets where most of the items are of medium difficulty than for item sets where item difficulties are spread out.

Estimates of Reliability and Validity of Final Tests

In this study, the 30 items with the highest estimated item total test score biserial correlation were selected from each replication of an item set under each sampling plan and number of observations, to be included in the final test form. For these items scores of 300 examinees were simulated using the procedure described in the previous chapter. An estimate of reliability (internal consistency) and an estimate of validity were obtained from these scores. Kuder Richardson formula 20 was used to calculate the reliability of the test. The validity was estimated by correlating ability scores of examinees with their total test scores. Since under each condition 50 replications were made, the mean, standard deviation and range (minimum and maximum value) of 50 estimates of reliability and validity were calculated.

In addition to the 30 items with the highest estimated item-total biserial correlation, another test of 30 items was randomly selected under each condition. Item response data for these items were simulated for a sample of 300 examinees. Reliability and validity of these tests were calculated. The mean, standard deviation and range of 50 estimates of reliability and validity of such tests under each condition were calculated. This was done to determine if on the average estimates of reliability and validity of final tests (consisting of 30 items with the highest item total biserial)

obtained under each condition is in fact better than the estimates of reliability and validity of any 30 item test.

The results associated with reliability and validity of final tests are discussed below.

Estimates of Reliability of final tests:

The results associated with the reliability of the final tests under different sampling plans and conditions are reported in Table 8. The results associated with the reliability of 30 item tests, where items were randomly selected from the pool of items, under each sampling plan and condition are reported in Table 9.

Ebel (1972) has stated that most expertly constructed unidimensional tests often yield reliability coefficients (internal consistency) around .90. The results in Table 8 show that for all tests and sampling conditions, the average reliability estimates of 30 item tests constructed with the highest estimated item discrimination are rather low. The mean reliabilities varied between .578 and .713. However, a comparison between results given in Tables 8 and 9 shows that under each sampling plan and condition, on the average the estimates of reliability of the final tests are higher than the average reliability of any 30 item test where items were randomly selected from the pool of items. The comparison of results of Table 8 and 9 also shows that for each sampling plan and condition the maximum reliability coefficient of 30 randomly selected items is higher than the minimum reliability coefficient of the final test form. If an item

TABLE 8

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of Reliability of Final Tests for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.643	.041	.523	.723	.681	.034	.613	.752	.660	.035	.534	.726
	MMS (t/k/n)												
7200	6/30/40	.626	.046	.488	.707	.666	.041	.546	.743	.655	.036	.579	.722
7200	15/20/24	.609	.051	.509	.712	.644	.041	.564	.730	.635	.042	.549	.746
7200	45/12/13	.578	.054	.376	.667	.633	.046	.483	.722	.628	.035	.554	.711
18000	Examinee Sampling (N=300)	.668	.038	.579	.742	.702	.034	.622	.774	.695	.028	.627	.753
	MMS (t/k/n)												
18000	6/30/100	.652	.038	.572	.758	.691	.037	.594	.771	.684	.033	.584	.747
18000	15/20/60	.643	.040	.551	.728	.676	.035	.602	.753	.668	.032	.578	.735
17820	45/12/33	.628	.043	.537	.726	.668	.038	.587	.780	.657	.034	.560	.716
36000	Examinee Sampling (N=600)	.680	.034	.611	.767	.713	.029	.652	.788	.701	.027	.635	.751
	MMS (t/k/n)												
36000	6/30/200	.668	.037	.592	.729	.703	.029	.631	.757	.693	.028	.643	.751
36000	15/20/120	.665	.043	.517	.750	.697	.034	.614	.774	.689	.030	.641	.761
35640	45/12/66	.651	.042	.573	.746	.685	.036	.565	.752	.677	.028	.606	.753

TABLE 9

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of Reliability of 30 Random Item Tests for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.501	.074	.267	.656	.537	.075	.330	.682	.546	.062	.368	.668
	MMS (t/k/n)												
7200	6/30/40	.519	.060	.377	.661	.552	.054	.439	.678	.551	.063	.383	.687
7200	15/20/24	.504	.078	.302	.662	.542	.078	.388	.672	.538	.066	.330	.660
7200	45/12/13	.493	.093	.287	.694	.533	.088	.340	.683	.534	.070	.344	.677
18000	Examinee Sampling (N=300)	.487	.074	.284	.640	.528	.070	.375	.664	.539	.072	.350	.677
	MMS (t/k/n)												
18000	6/30/100	.511	.077	.282	.639	.549	.074	.338	.664	.536	.063	.365	.643
18000	15/20/60	.505	.075	.328	.658	.547	.077	.346	.688	.527	.067	.380	.693
17820	45/12/33	.507	.080	.304	.636	.545	.074	.363	.661	.547	.072	.333	.698
36000	Examinee Sampling (N=600)	.510	.077	.246	.629	.548	.079	.259	.659	.538	.065	.348	.636
	MMS (t/k/n)												
36000	6/30/200	.497	.077	.288	.621	.535	.079	.295	.659	.538	.075	.356	.680
36000	15/20/120	.510	.076	.267	.618	.549	.068	.377	.646	.542	.069	.369	.649
35640	45/12/66	.508	.074	.328	.648	.549	.070	.362	.669	.536	.059	.407	.670

selection procedure is good, one would expect that the reliability of selected items (i.e., final test) would be better than the reliability of 30 randomly selected items in each replication. The reason why the maximum reliability of randomly selected items exceeded the minimum reliability of the final tests may be as follows: The 50 replications were made under each sampling plan and condition. The item parameters in each replication were varied and they were randomly generated within some constraints. It is possible that by chance some of these item pools consisted of a significantly large numbers of good items (i.e., items with $d_i > .30$). Similarily some of these item pools consisted of significantly fewer good items. It is quite likely that one may get higher reliability of 30 items randomly selected from the former item pools than the reliability of 30 items carefully selected from item pools which originally contained fewer number of good items. Furthermore, by comparing the standard deviation of 50 repetitions under each conditions, one can say that estimates of reliability of final tests tend to be more stable from sample to sample than the estimates of reliabilities of a test made up of 30 items selected randomly.

The comparison of reliabilities of final tests shows that for a given item set and number of observations, the mean reliability of final tests produced using item analysis data from examinee sampling is higher than the reliability of tests produced using item analysis data from MMS plans. However, this difference decreased as the number of items in

MMS plans increased from one-fifth to one-half of the total items. The comparison among different MMS plans shows that for a given item set and number of observations, on the average, reliability of final tests decreased as the number of items in each subtest (during the item analysis) decreased from one-half to one-fifth of the items. Furthermore, the average reliability of the final tests increased as the number of observations during item analysis increased from 7200 to 36000, for both examinee sampling and MMS plans. Gullikson (1950) has stated that there is a relationship between test reliability (internal consistency) and item discrimination. According to him the higher test reliability will result from items with higher item discrimination. On the basis of this one would assume that sampling plans that estimate item discrimination more accurately will more likely enable the test constructors to select items that have higher true item discrimination. Selection of relatively more good items (i.e.,items with relatively high discrimination) will result in higher reliability. The results of this study tend to be consistent with this assumption. The mean reliability for final tests was higher under conditions in which item discrimination was estimated relatively more accurately.

On the whole, one can conclude that the results related to the reliability of the final tests are quite consistent with results related to estimates of item discrimination. The reliabilities are relatively higher under conditions

where estimations of discrimination are relatively more accurate. Comparing reliabilities of different item sets, for a given sampling plan and number of observations, it appears that on the average, the reliability for item set B is higher than that for item set C, which is higher than that for item set A. The differences in reliabilities of different item sets may have been caused by differences in total test variance in different item sets. Formula 8, given in the previous chapter indicates that reliability depends on three quantities: K, the number of items in the tests; Σpq, the sum of item variance; and σ_T^2, the test variance. Ebel (1967) has stated that these three quantities are related. But, for tests of equal length (K), it is the variation from test to test in the value of test variance that is responsibile for most of the observed variations in test reliability.

Estimates of Validity of Final Tests:

Validities of final tests obtained under each sampling plan, item set and number of observations are reported in Table 10. Validities of 30 item tests, where items were randomly selected from the pool of items, obtained under different conditions are reported in Table 11.

The results with regard to average validity of final tests under different conditions are quite similar to the results of average reliability under different conditions. For a given item set and number of observations, validities of final tests are slightly higher for tests that were

TABLE 10

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of Validity of Final Tests for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.804	.027	.696	.862	.822	.021	.775	.865	.812	.021	.744	.849
	MMS (t/k/n)												
7200	6/30/40	.791	.030	.721	.844	.813	.025	.749	.858	.808	.021	.739	.845
7200	15/20/24	.783	.032	.710	.846	.804	.025	.759	.857	.800	.026	.745	.862
7200	45/12/13	.764	.036	.610	.817	.796	.029	.719	.848	.791	.021	.748	.844
18000	Examinee Sampling (N=300)	.815	.019	.779	.862	.836	.020	.783	.872	.830	.017	.786	.861
	MMS (t/k/n)												
18000	6/30/100	.809	.022	.757	.866	.830	.023	.777	.870	.825	.020	.776	.863
18000	15/20/60	.802	.026	.741	.850	.820	.022	.762	.870	.816	.018	.776	.849
17820	45/12/33	.792	.027	.728	.849	.815	.025	.757	.876	.810	.021	.752	.854
36000	Examinee Sampling (N=600)	.823	.021	.779	.869	.842	.017	.804	.880	.837	.017	.791	.866
	MMS (t/k/n)												
36000	6/30/200	.815	.023	.765	.853	.834	.020	.783	.879	.830	.018	.791	.866
36000	15/20/120	.813	.025	.742	.858	.833	.018	.783	.873	.829	.017	.797	.869
35640	45/12/66	.805	.025	.739	.848	.825	.023	.766	.859	.820	.018	.778	.852

TABLE 11

Mean, Standard Deviation and Range (Minimum and Maximum Values) of 50 Estimates of Validity of 30 Random Item Tests for Item Sets A, B and C Under Different Sampling Plans and Number of Observations

Number of Observations	Sampling Plan	Item Set A				Item Set B				Item Set C			
		Mean	SD	MIN	MAX	Mean	SD	MIN	MAX	Mean	SD	MIN	MAX
7200	Examinee Sampling (N=120)	.707	.049	.530	.809	.734	.048	.597	.819	.738	.043	.612	.822
	MMS (t/k/n)												
7200	6/30/40	.718	.041	.615	.812	.743	.039	.630	.824	.744	.040	.626	.836
7200	15/20/24	.708	.057	.566	.804	.736	.056	.588	.818	.735	.044	.619	.805
7200	45/12/13	.702	.066	.573	.828	.730	.060	.587	.827	.732	.046	.616	.813
18000	Examinee Sampling (N=300)	.700	.051	.585	.795	.729	.050	.587	.814	.734	.050	.616	.829
	MMS (t/k/n)												
18000	6/30/100	.714	.054	.542	.804	.739	.051	.575	.813	.732	.044	.582	.794
18000	15/20/60	.711	.051	.575	.828	.736	.051	.580	.835	.730	.043	.637	.826
17820	45/12/33	.707	.055	.522	.801	.736	.049	.597	.812	.737	.055	.538	.823
36000	Examinee Sampling (N=600)	.712	.056	.522	.790	.736	.054	.556	.814	.733	.044	.585	.791
	MMS (t/k/n)												
36000	6/30/200	.707	.058	.506	.792	.736	.054	.542	.810	.734	.048	.601	.838
36000	15/20/120	.711	.052	.563	.778	.739	.052	.609	.814	.733	.049	.603	.814
35640	45/12/66	.710	.051	.563	.806	.736	.046	.594	.815	.734	.040	.645	.826

produced using MMS plans. Among different MMS plans, for a given item set and number of observations, validities of the final test increased as the number of items in each subtest increased from one-fifth of the total items to one-half of the total items. For all item sets and sampling plans, validity increased as the number of observations during item analysis increased. On the average, for a given number of observations and sampling plans, the validity for item set B is higher than that for item set C, which in turn is higher than that for item set A. The differences in validities among different item sets could be due to different distributions of item difficulties. The distribution of test scores for item set B tends to be rectangular over the range, the distribution of test scores for item set C tends to be platykurtic over the range whereas the distribution of test scores for item set A tends to be normal over the range. Validity of a final test in this study was estimated by correlating test scores with ability scores. Since correlation coefficients are influenced by the distributions of the variables being correlated, one would expect that the validity for item set B would be higher than the validity for item set C, which in turn would be higher than the validity for item set A. The results tend to support the above expectation.

The comparison of results of Table 10 and 11 shows that under all conditions, mean validities of final tests are on the average higher than the mean validities of any 30 item tests where items were randomly selected. Furthermore under

each condition, the standard deviation of validity estimates of final tests is lower than those of tests consisting of 30 randomly selected items. However, the maximum validity of 30 item tests where items were randomly selected from the pool of items is higher than the minimum validity of final test form. The reason why this may have occured is given in connection with the reliability. On the basis of the above results, one can conclude that in general when items with the highest item discriminations are selected for the final test form, they not only produce a more valid test, but estimates of these validities are more consistent from sample to sample, than when items are randomly selected for the final test.

One can conclude that results related to estimates of test reliability and validity are consistent with results associated with estimates of item discrimination. In conditions under which item discriminations are estimated relatively more accurately, relatively more reliable and valid tests are produced than in conditions in which estimates of item discriminations are less accurate. One would assume that when estimates of item discriminations are more accurate, one would be more likely to select a greater number of good items (item with true item discrimination greater than .30) for the final tests than when estimates of item discriminations are relatively less accurate. The reliability and validity of a test in fact is based on the quality of items. The conclusions drawn from this study are stated in the next chapter.

CHAPTER IV

CONCLUSIONS

In this chapter, the conclusions drawn from the study, the limitations of the study and suggestions for further research are discussed.

This study was primarily conducted to compare the examinee sampling and MMS sampling techniques of data collection for test development. More specifically, the following questions were explored: How does the MMS procedure compare with the traditionally used examinee sampling procedure when estimating item parameters such as item difficulty and item discrimination? How do the final tests that are produced using item analysis data obtained from MMS and examinee sampling methods, compare with respect to test reliability and validity? How do the number of observations, item subtest size and the characteristics of item sets affect the estimates of item parameters, the reliability and the validity of the tests produced?

In answer to the first question relating to the comparison of the examinee sampling and the MMS with regard to the estimation of item difficulties and item discrimination, the following conclusions can be drawn from the study: For a given item set and number of observations, there is no systematic difference in estimates of item difficulty obtained from the examinee sampling and the MMS plans. For a given item set and number of observations,

examinee sampling produces a more accurate estimate of item discrimination than the MMS plans. This difference is largest when the number of items in each subtest of a MMS plan is small and decreases as the number of items in each subtest increases from one-third to one-half of the total number of items. Furthermore, the difference between the accuracy with which MMS plans and examinee sampling plan estimate item discrimination also decreases as the number of observations increase from 7200 to 36000.

In answer to the second question relating to the comparison of the examinee sampling and the MMS plans with regard to the reliability and the validity of the final test, the following conclusions can be drawn from the study: For a given item set and number of observations, the average reliability and validity of the final tests produced using examinee sampling are higher than the average reliability and validity of the final tests produced using MMS plans. However, these differences decrease as the number of items in each subtest increases from one-fifth to one-half of the items, also as the number of observations increases from 7200 to 36000. These conclusions are similar to the conclusions drawn with regard to estimation of item discrimination. One would expect the correspondence between conclusions drawn about item discrimination and test reliability and validity because reliability and validity of a final test depend on the characteristics of the items. Higher test reliability results from items with higher item discrimination.

In answer to the third question relating to the effect

of the number of observations, item subtest size and the characteristic of item sets on the estimation of item parameters, and the test characteristics, the following conclusions can be drawn from the study: For each item set and sampling plan, the estimates of item difficulty, item discrimination, test reliability and validity improve as the number of observations increases. For a given item set and number of observations, the estimates of item discrimination, test reliability and validity improve as the number of items in each subtest increases. The variation in item subtest sizes tends to have no effect on estimation of item difficulty. For a given sampling plan and number of observations, the estimates of item difficulty, item discrimination, test reliability and validity appeared to be better for item sets containing medium difficulty items than for item sets in which item difficulties were spread out. This difference may have been caused by factors such as differential influence of guessing on items of extreme difficulty and variation in total test variance of different tests.

Finally, from this study, one can conclude that while developing an unidimensional norm referenced test, if one were to use a large number of examinees (i.e., 5 to 10 times the number of items) for pretesting items, one should prefer examinee sampling over MMS plans. However, when the number of items in a pretest is large and one can not get a representative sample of examinees of adequate size (i.e., 5 to 10 times the number of items) for a sufficient period of

time to complete the items in the pretest, MMS does provide an alternative. MMS procedures with large numbers of observations (i.e., number of observations corresponding to examinee sample size of 10 times the number of items) provide estimates of item parameters and test characteristics which tend to be nearly as good as those obtained from traditionally used examinee sampling. As to the conditions under which MMS is most effective for estimating item parameters and test reliability and validity, one can say that MMS plans with a large number of items in each subtest (preferably one-third to one-half of the total number of items per subtest) are better than MMS plans with a small number of items in each subtest. Thus one can conclude that if one has a choice between giving all items to a relatively small number of examinees and giving a small number of items to a large number of examinees, one should prefer the latter for a test development.

In addition to the above conclusions one can also infer from the study that while selecting number of observations for pretesting items either by examinee sampling or by MMS plans, attention should be given to the item difficulties in the item pool. In order to obtain the same precision when estimating item difficulties and item discriminations, one may need a larger number of observations when item difficulties in the item pool are spread out than when most of the items in the pool are of medium difficulty.

While the results are consistent, it should be remembered that the conclusions drawn from the study may not

hold if the item characteristics and/or number of items in the initial item pools and the final test forms vary greatly from those used in this study.

The MMS has been mainly used for estimating group parameters such as test means and variances. This study shows that the MMS can be used for estimating item parameters such as item difficulties and discriminations. The extension of MMS to the estimation of item parameters can be of a great help to test developers because of the reduction of testing time per examinee.

Limitations of the Study

The most serious limitation of this study, at the same time that which made it possible, is the use of Monte Carlo methods. Critics of the approach often point out that simulated data may bear no resemblance to data gathered in a "real-word" setting. Although, an attempt was made to keep the parameters of the simulation model used in the study within realistic limits no assurance can be provided regarding the correspondence between the simulated data used in this study and actual data collected from examinees. It is also quite possible that the different restrictions imposed on the distribution of item parameters and the distribution of examinee ability may have made the outcome of this study different from the outcome one might have had received from actual data. In Monte Carlo studies, selection of random samples from the defined population is no problem.

But this may not be the case in real life situations. However, when using MMS one might have a higher chance of getting a more representative sample (in real life situations) from the defined population than when using examinee sampling, because it may be easier to get people to cooperate when a small amount of their time is involved as is the case in MMS.

A second limitation of this study is related to the notion of guessing. It was assumed in the study that all examinees who do not know the correct answer to an item guess, and do so with a probability of .25 of guessing correctly. It may be unrealistic to assume that examinees guess randomly with a probability of .25.

An additional limitation of this study is the number of items used in the initial item pool and in the final test. If one wishes to have 30 items in the final test, 60 items in the initial pool may not be sufficient. Yet, as in most research studies, limitations are also placed on the researchers, particularly in terms of time and money (in this study computer time was of concern).

Suggestions for Future Research

Recommendations concerning future areas of research have come from a recognition of the limitations of the present investigation. Some of these recommendations are as follows:

An empirical research with actual test data can be conducted to validate the findings of this research. This

research could be of the a posteriori type. However, with this type of research it would be impossible to calculate mean square error for different parameters, unless one has a very large data set. If the data set is very large one can draw a number of samples from it and can get estimates of mean square error for different parameters.

In this study, estimates of item parameters were found to be better for item sets containing medium difficulty items than for item sets in which item difficulties were spread out. It was speculated that these differences may have been caused by different amounts of guessing associated with items of different difficulty. One can further investigate the effect of guessing on different item sets using a simulation model which allows no guessing. One might also investigate the optimum number of observations for both examinee sampling and MMS plans in order to obtain stable estimates of item difficulties and item discriminations for item pools with different item characteristics.

REFERENCES

Aiken, L.R. Readings in psychological and educational testing. Boston: Allyn and Bacon, Inc., 1973.

Aleamoni, L.M. and Spencer, R.E. A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. Educational and Psycological Measurement, 1969, 29, 353-358.

Allen, M.J. and Yen, W.M. Introduction to measurement theory. Montery, California: Books/ Cole Publishing Company, 1979.

Anastasi, A. Psychological testing. (4th ed.). Toronto: MacMillan, 1976.

Anstey, E. Psychological tests. London: Nelson, 1966.

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4(4), 325-355.

Barcikowski, R.S. A Monte Carlo study of item sampling (versus) traditional sampling for norm construction. Journal of Educational Statistics. 1972, 9(3), 209-224.

Barcikowski, R.S. The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. Educational and Psychological Measurement, 1974, 34, 231-237.

Bashaw, W.L. A note on correcting item-total correlations. Journal of Educational Measurement, 1968, 5(3), 263-264.

Berk, R.A. Empirical evaluation of formula for correction of item-total point-biserial correlations. Educational and

Psychological Measurement, 1978, 38, 647-652.

Bowers, J. A note on comparing r biserial and r point biserial. Educational and Psychological Measurement, 1972, 32, 771-775.

Brogden, H.E. Variation in test validity with variation in the distribution of item difficulties, number of items and degree of their intercorrelation. Psycometrika, 1946, 11, 197-214.

Bunda, M.A. An investigation of an extension of item sampling which yields individual scores. Journal of Educational Measurement, 1973, 10(2), 117-130.

Burt, C. Mental and scholistic tests. London: P.S.King and Son, 1921.

Cahen, L.S., Romberg, T.A. and Zwirner, W. The estimation of mean achievement scores for schools by the item-sampling technique. Educational and Psychological Measurement, 1970, 30, 41-60.

Chartock, P.O. The effect of item discrimination on the precision of the estimates of the population mean and variance for nonoverlapping multiple matrix sampling plans. Doctoral Dissertation, Columbia University, 1980.

Chauncey, H. and Fredericksen, N. The functions of measurement in educational placement. In E.F. Lindquist (Ed.), Educational Measurement. Washington, D.C.: Council on Education, 1951.

Cook, D.L. and Stufflebeam, D.L. Estimating test norms from variable size item and examinee samples. Journal of

Educational Measurement, 1967, 4, 27-33.

Cooper, M.A. and Fiske, D.W. Undependability of construct validity patterns for tests and items. Educational and Psychological Measurement, 1976, 36, 631-637.

Cronbach, L.J. Essentials of psychological testing. New York: Harper, 1960.

Cronbach, L.J. and Warrington, W.G. Efficiency of multiple choice test as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-147.

Cureton, E.E. Corrected item-test correlations. Psychometrika, 1966, 31, 93-96.

Davis, F.B. Item selection techniques. In E.F. Lindquist (Ed.), Educational Measurement. Washington, D.C.: Council on Education, 1951.

Ebel, R.l. The relation of item discrimination to test reliability. Journal of Educational Measurement, 1967, 4(3), 125-128.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Engelhart, M.D. A comparison of several item discrimination indices. Journal of Educational Measurement, 1965, 2, 69-76.

Feldt, L.S. and Forsyth, R.A. An examination of context effect in item sampling. Journal of Educational Measurement, 1974, 11, 73-82.

Ferguson, G.A. On the theory of test discrimination. Psychometrika, 1949, 14, 61-68.

Forysth, R.A. Estimating means via multiple matrix sampling:

A note on the effect of selected data base characteristics. Educational and Psychological Measuement, 1976, 36, 275-282.

French, J.L., and Greer, D. Effect of test-item arrangement on physiological and psychological behavior in primary-school children. Journal of Educational Measurement, 1964, 1, 151-153.

Ghiselli, E.E., Campbell, J.P., and Zedeck, S. Measurement theory for the behavioral sciences. San Francisco: W.H. Freeman and Company, 1981.

Guilford, J.P. Psychometric methods. New York: McGraw-Hill, 1954.

Guilford, J.P. Fundamental statistics in psychology. New York: McGraw- Hill, 1965.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.

Henrysson, S. Correction of item -total correlations in item analysis. Psychometrika, 1963, 28, 211-218.

Henrysson, S. Gathering, analyzing and using data on test items. In R.L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: America Council on Education, 1971.

Hill, R. Minimizing context effect when using multiple matrix sampling. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Washington, D.C., March 31-April 12, 1975.

Hooke, R. Some applications of bipolykays to the estimation

of variance components and their moments. Annals of Mathematical Statistics, 1956, 27, 80-98. (a)

Hooke, R. Symmetric functions of a two-way array. Annals of Mathematical Statistics, 1956, 27, 55-79. (b)

Huck, S.W., and Bowers, N.D. Item difficulty level and sequence effects in multiple choice achievement tests. Journal of Eduational Measurement, 1972, 9, 105-111.

Hull, C.L. Apptitude testing. New York: World Book Co., 1928.

Humphreys, L.G. The normal curve and the attenuation paradox in test theory. Psychological Bulletin, 1956, 53, 472-476.

Johnson, M.C., and Lord, F.M. An empirical study of the stability of a group mean in relation to the distribution of test items among pupils. Journal of Educational and Psychological Measurement, 1958, 18, 325-329.

Kaplan, R.M. and Saccuzzo, D.P. Psychological testing principles, applications, and issues. Monterey, California: Books/Cole Publishing Co., 1982.

Kendall, M.G. and Stuart, A. The advanced theory of statistics (2^{nd} Ed.). London: Charles Griffin and Company Limited, 1967.

Kleinke, D.J. The accuracy of estimated total test statistics: Final report. Arlington, Va.: ERIC Document Reproduction Service, ED064-356, 1972.

Knapp, T.R. An application of balanced incomplete block designs to the estimation of test norms. Educational

and Psychological Measurement, 1968, 28, 265-272.

Knapp, J.R. Item sampling. Unpublished manuscript. University of Rochester, 1973.

Knapp, T.R. Using incidence sampling to estimate covariances. Journal of Educational Statistics, 1979, 4(1), 41-58.

Lord, F.M. An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

Lord, F.M. Estimating norms by item-sampling. Journal of Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F.M. Biserial estimates of correlation. Psychometrika, 1963, 28, 81-85.

Lord, F.M., and Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. Test theory. In M.R. Rosenzweig and L.W. Porter (eds.), Annual Review of Psychology, 1976, 27, 251-280.

McNemar, Q. Psychological Statistics. NewYork: John Wiley and Sons, Inc., 1962.

Magnusson, D. Test theory. Reading, Ma.: Addison-Wesley, 1966.

Marso, R. Test item arrangement, testing time, and performance. Journal of Educational Measurement, 1970, 7, 113-118.

Messick, S., and Jackson, D.N. The measurement of authoritarian attitudes. Educational and Psychological

Measurement, 1958, 18, 241-253.

Moy, M.L.Y., and Barcikowski, R.S. Item sampling: Optimal number of people and items. Journal of Experimental Education, 1974, 42, 46-52.

Myerberg, N.J. The effect of item stratification in multiple matrix sampling. Paper presented at the Annual Meeting of the American Educational Research Association. Washington, D.C., 1970.

Myerberg, N.J. The effect of item stratification on the estimation of the mean and variance of universe scores in multiple-matrix sampling. Educational and Psychological Measurement, 1979, 39, 57-68.

Neill, J.A. and Jackson, D.N. An evaluation of item selection strategies in personality scale construction. Educational and Psychological Measurement, 1970, 30, 647-661.

Novak, C.D. An empirical investigation of multiple matrix sampling in an elementary school testing. Arlington, Va.: ERIC, Document Reproduction Service, ED093997, 1974.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

Nunnally, J.C. Introduction to Statistics for Psychology and Education. New York: McGraw Hill Book Company, 1975.

Oosterhof, A.C. Similarity of various item discrimination indices. Journal of Educational Measurement, 1976, 13(2), 145-150.

Owens, T.R. and Stufflebeam, D.L. An experimental

comparison of item sampling and examinee sampling for estimating test norms. Journal of Educational and Psychological Measurement, 1969, 6, 75-83.

Pandey, T.N. and Shoemaker, D.M. Estimating moments of universe scores and associated standard errors in multiple matrix sampling for all item scoring procedures. Educational and Psychological Measurement, 1975, 35, 567-581.

Peterson, D.F. and Anderson, D.H. Closing the communications gap with item sampling. Paper presented at the annual meeting of the American Educational Research Association. New York, New York, February 1971.

Plumlee, L.B. Estimating means and standard deviations from partial data-an empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Pyrczak, F. Validity of the discrimination index as a measure of item quality. Journal of Educational Measurement, 1973, 10(3), 227-231.

Richardson, M.W. The relation between the difficulty and the differential validity of a test. Psychometrika, 1936, 1, 33-49. (a)

Richardson, M.W. The rational of item analysis. Psychometrika, 1936, 1, 69-76. (b)

Richardson, M.W. In E.F. Lindquist (Ed.), Educational measurement. Washington,D.C.: American Council on Education, 1951.

Sax, G. and Cromack, T.R. The effect of various forms of item arrangement on test performance. Journal of Educational Measurement, 1966, 3, 309-311.

Scheetz, J.P. Assigning items to subtests in multiple-matrix sampling:An empirical post-mortem investigation. Doctoral Dissertation, 1976.

Scott, W.A. The distribution of test scores. Educational and Psychological Measurement, 1972, 32, 725-735.

Shavelson, R.J. Statistical reasoning for the behavioral sciences. Boston: Allyn and Bacon, 1981.

Shoemaker, D.M. Allocation of items and examinees in estimating a norm distribution by item sampling. Journal of Educational Measurement, 1970, 7, 123-128.(a)

Shoemaker, D.M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970, 7, 255-262.(b)

Shoemaker, D.M. An application of item-examinee sampling to scaling attitudes. Journal of Educational Measurement, 1971, 8(4), 279-282. (a)

Shoemaker, D.M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8(3), 215-220. (b)

Shoemaker, D.M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. Educational and Psychological

Measurement, 1972, 32, 705-714.

Shoemaker, D.M. A note on allocating items to subtests in multiple-matrix sampling and approximating standard errors of estimate with the jacknife. Journal of Educational Measurement, 1973, 10, 211-219. (a)

Shoemaker, D.M. Principles and procedures of multiple-matrix sampling. Cambridge, Mass.: Ballinger Publ. Co., 1973. (b)

Shoemaker, D.M. and Osburn, H.E. An empirical study of generalizability coefficient for unmatched data. British Journal of Mathematical and Statistical Psychology, 1968, 7, 199-207.

Sirotnik, K. An investigation of the context effect in matrix sampling. Journal of Educational Measurement, 1970, 7, 199-207. (a)

Sirotnik, K. An analysis of variance framework for matrix sampling. Educational and Psychological Measurement, 1970, 30, 891-908. (b)

Sirotnik, K. Introduction to matrix sampling for the practitioner. In W.J. Popham (Ed.), Evaluation in education: Current Applications. Berkeley, Cal.: McCutrhan Publ. Corp., 1974.

Sirotnik, K. and Wellington, R. Incidence sampling: An integrated theory for "matrix sampling". Journal of Educational Measurement, 1977, 14, 343-395.

Sirotnik, K. and Wellington, R. Scrambling content in achievement testing: An application of multiple-matrix sampling in experimental design. Journal of Educational

Measurement, 1974, 11, 179-188.

Tate, R.F. Application of correlation models for biserial data. Journal of the American Statistical Association, 1955, 52, 205-216.

Thurston, L.L. The reliability and validity of tests. Ann Arbor, Mich.: Edwards Brothers, 1932.

Wellington, R. Extending Generalized symmetric means to arbitrary matrix sampling designs. Psychometrika, 1976, 41(3), 375-384.

Wolf, R. Evaluation of several formulae for correction of item-total correlations in item analysis. Journal of Educational Measurement, 4(1), 1967.

Zimmerman, D.W. and Williams, R.H. Effect of chance success due to guessing on errors of measurement in multiple-choice tests. Psychological Reports, 1965, 16, 193-196.

Zubin, J. The method of internal consistency for selecting test items. Journal of Educational Psychology, 1934, 25, 345-356.

APPENDIX A

Random Number Generators used in the Study

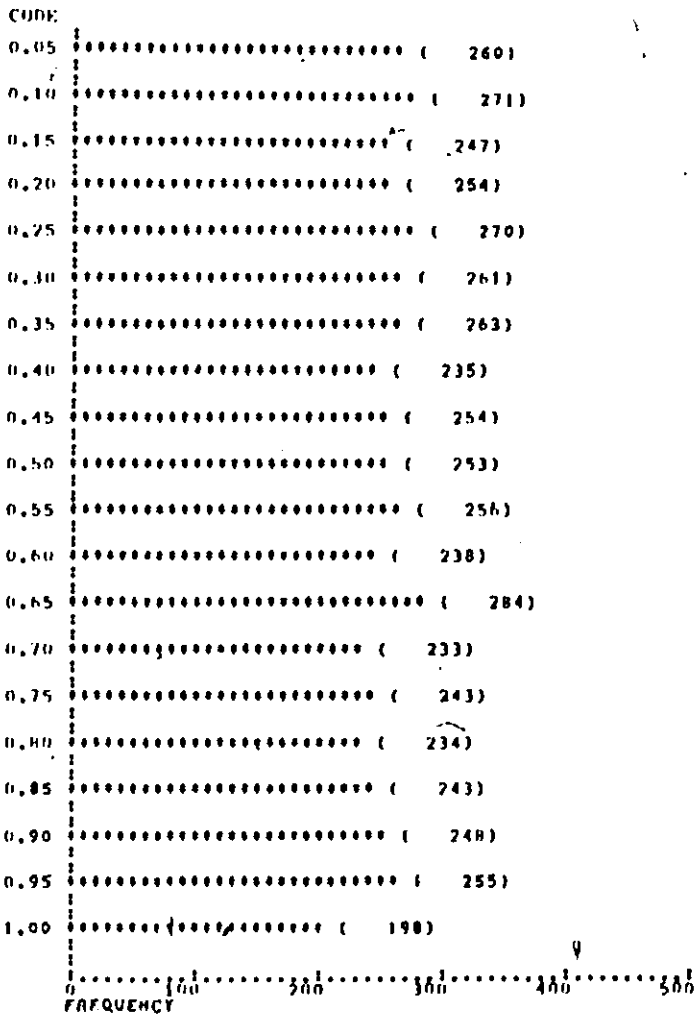
Two pseudo-random number generators were used in the study: RAND and NORM. The discription of these is given below.[1]

RAND: RAND is a linear congruential generator, it generates a rectangular distribution on the unit interval, with numbers ranging from 0.0 to 1.0. The generator has been tested and found satisfactory for uniformity. Uniformity was determined with a sample of 5,000 numbers, frequency distribution was obtained and histogram was ploted. The histogram is shown in Graph 1. The listing of the program for this generator is as follows:

```
RAND
REAL FUNCTION RAND(ISEED)
DOUBLE PRECISION Z,D2P31M
DOUBLE PRECISION RMOD,DMOD
DATA D2P31M/2147483647.D0/
DATA RMOD/2147483.D3/
Z=ISEED
Z=DMOD(16807.D0*Z.D2P31M)
ISEED=Z
RAND=Z/RMOD
RETURN
END
```

Figure 1

Histogram for 5,000 Scores Generated
from RAND, a Pseudo-random Number Generator



The chi-square goodness of fit statistic for comparing the above distribution to the uniform distribution is 23.852 with 19 degrees of freedom. This chi-quare value is not significant at .05 level of significance.

NORM: NORM generates numbers normally distributed with a mean of 0.0 and the standard deviation of 1.0. The generator has been tested and found satisfactory for normality. Ten samples of 5,000 numbers were generated and then the first four moments of each sample were calculated using the subprogram condescriptive of SPSS. The results of 10 samples are shown in Table 12. The listing of the program for this generator is as follows:

```
NORM
REAL FUNCTION NORM(ISEED)
U1=-2.0*ALOG(RAND(ISEED))
U1=SQRT(U1)
U2=COS(U2)
NORM=U1*U2
RETURN
END
```

1. Both RAND and NORM programs were received from A. K. Beuchert, University of Georgia.

TABLE 12

Mean, Standard Deviation, Skewness and Kurtosis of Ten Samples of 5,000 Numbers Generated from NORM, a Pseudo-Random Number Generator

Sample	Mean	Standard Deviation	Skewness	Kurtosis
1	-0.039	0.981	-0.027	0.1444
2	-0.000	1.021	-0.004	0.041
3	-0.048	0.992	0.054	0.043
4	0.000	0.999	-0.008	0.015
5	-0.009	1.001	0.058	-0.057
6	-0.014	1.003	0.029	0.009
7	0.001	0.993	-0.010	0.056
8	0.008	0.996	0.001	0.022
9	-0.032	1.014	0.038	0.126
10	-0.012	0.989	-0.048	-0.040
Column Sums	-.145	9.989	0.083	.359
Column Means	-.014	.998	0.008	.035

Note: The expected values for NORM are those of a normal distribution N(0,1).

Mean = .000
Standard Deviation = 1.000
Skewness = .000
Kurtosis = .000

APPENDIX B

Flow Charts for Computer Programs

Three flow charts are presented: Flow chart A was used for simulating item parameters, flow chart B for simulating scores for examinee sampling and flow chart C for simulating scores for MMS.

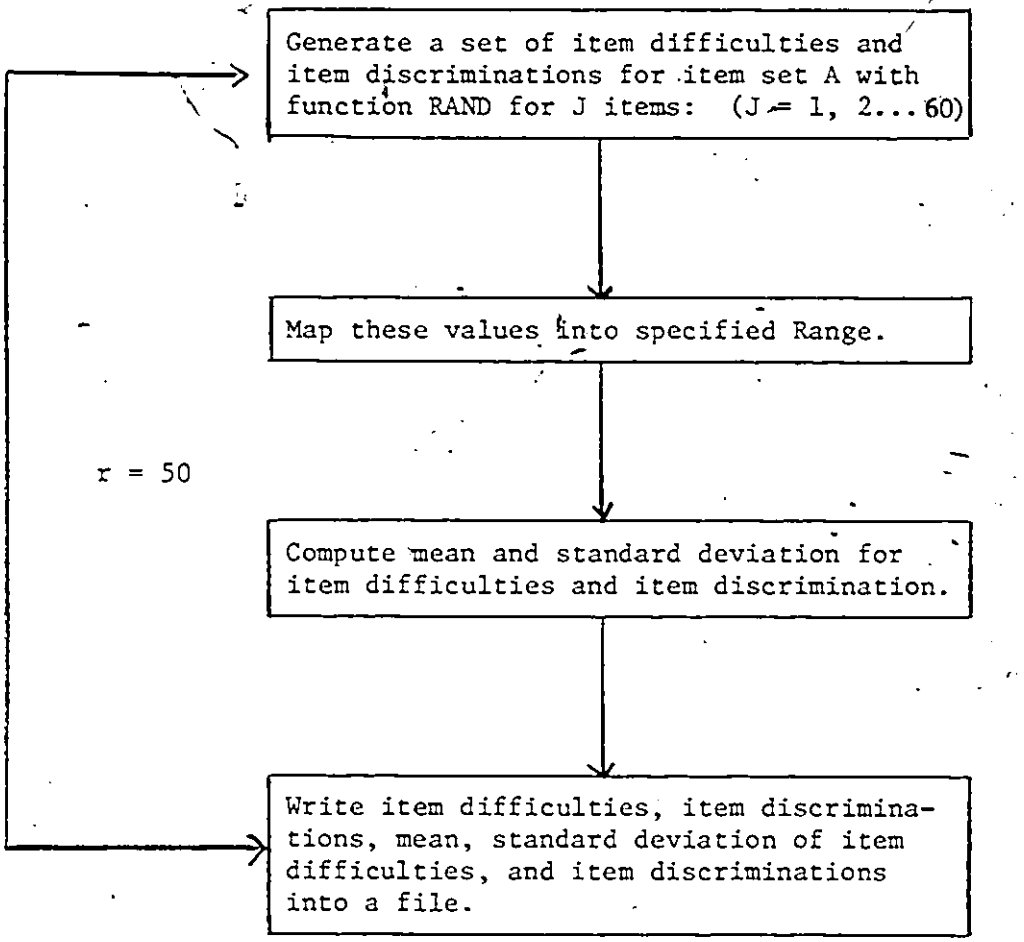
Flow chart A was used for simulating item parameters for item pools representing item sets A, B and C with the following variations: For item pools representig item sets A and B, function RAND was used to simulate item difficulties. For item pools representing item set C function NORM was used to simulate item difficulties. Item discriminations for all item pools were simulated using function RAND. Item difficulties and discriminations associated with different item sets were mapped according to the values given in Table 1.

Flow chat B was used for simulating item response data for all examinee sampling plans given in Table 2 with following variations: N was varied according to the number of examinees given in Table 2. ISEED number in the begining of the program was varied so that the same examinee scores were not repeated again and again. Item sets were varied according to the plans given in Table 1.

Flow chart C was used for simulating item response data for all MMS plans with following variations: N and sampling plans were varied according to the number of plans given in

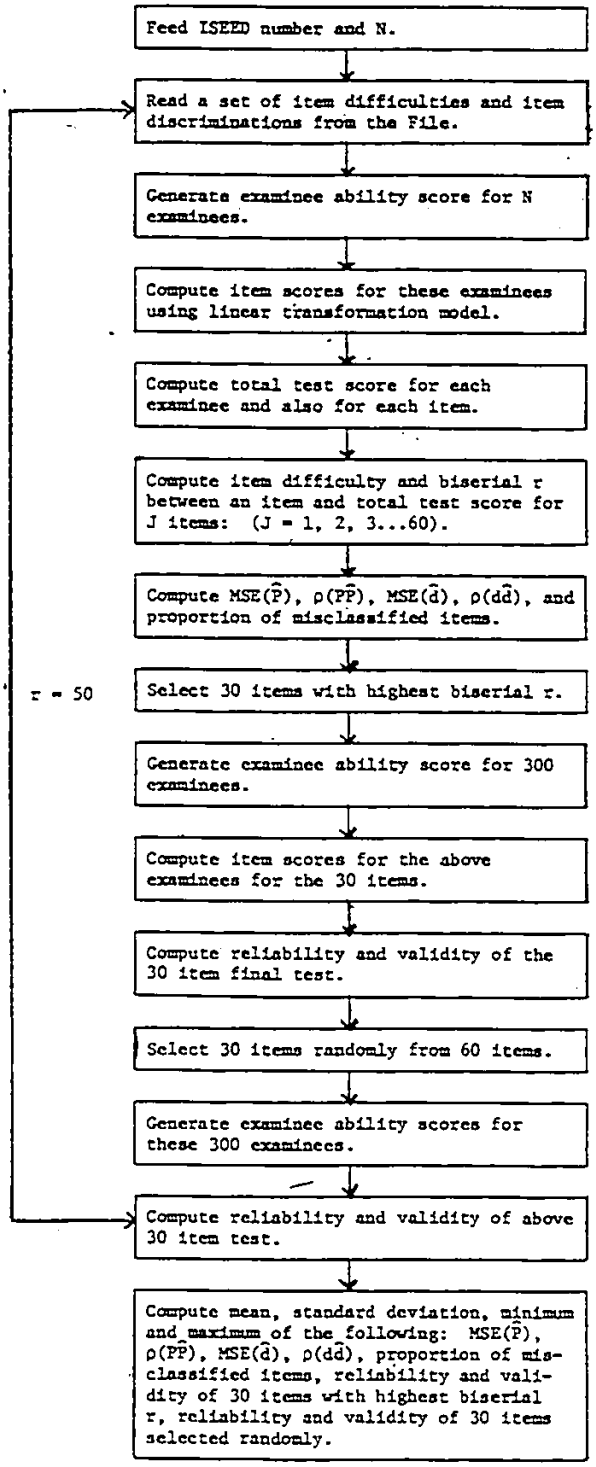
Table 2. ISEED number in the begining of the program was also varied so that same scores were not generated repeatedly. Item sets were varied according to the plans given in Table 1.

Flow Chart A

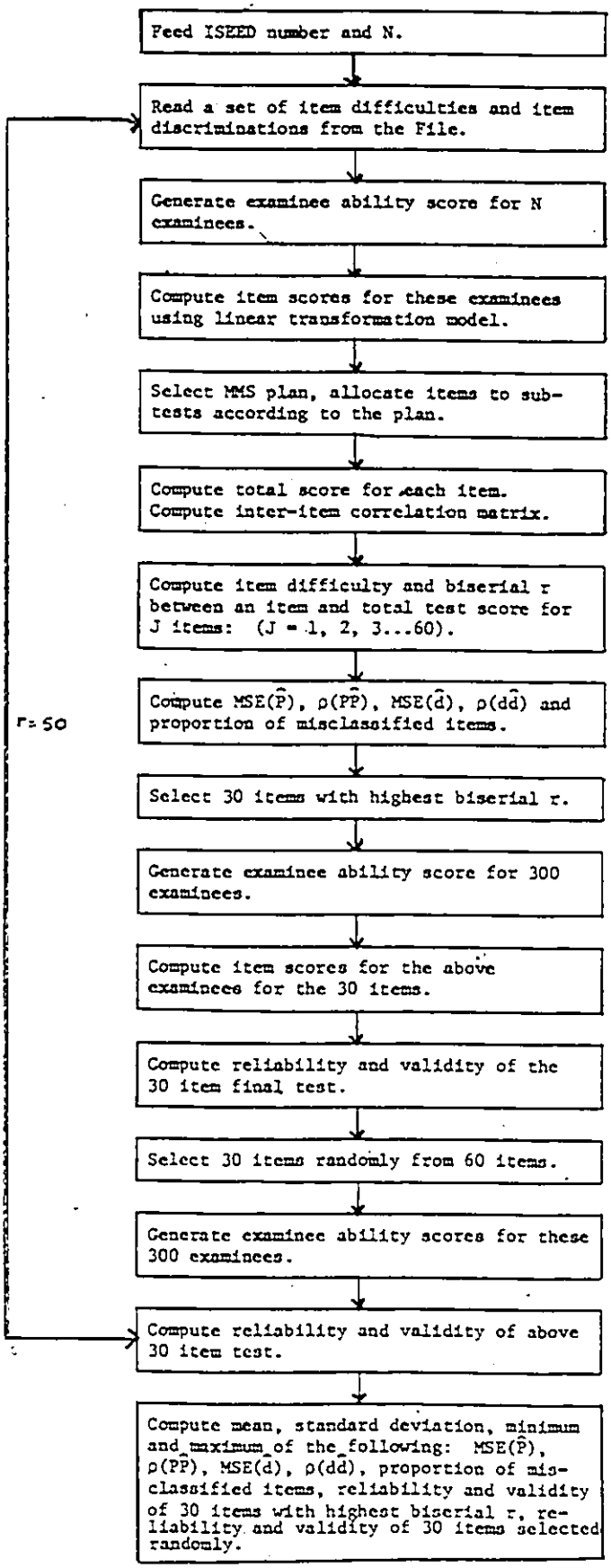


r = Number of replications.

Flow Chart B


Feed ISEED number and N.
Read a set of item difficulties and item discriminations from the File.
Generate examinee ability score for N examinees.
Compute item scores for these examinees using linear transformation model.
Compute total test score for each examinee and also for each item.
Compute item difficulty and biserial r between an item and total test score for J items: (J = 1, 2, 3...60).
Compute MSE(P̂), ρ(PP̂), MSE(d̂), ρ(dd̂), and proportion of misclassified items.
r = 50
Select 30 items with highest biserial r.
Generate examinee ability score for 300 examinees.
Compute item scores for the above examinees for the 30 items.
Compute reliability and validity of the 30 item final test.
Select 30 items randomly from 60 items.
Generate examinee ability scores for these 300 examinees.
Compute reliability and validity of above 30 item test.
Compute mean, standard deviation, minimum and maximum of the following: MSE(P̂), ρ(PP̂), MSE(d̂), ρ(dd̂), proportion of misclassified items, reliability and validity of 30 items with highest biserial r, reliability and validity of 30 items selected randomly.

Flow Chart C


Feed ISEED number and N.
Read a set of item difficulties and item discriminations from the File.
Generate examinee ability score for N examinees.
Compute item scores for these examinees using linear transformation model.
Select MMS plan, allocate items to subtests according to the plan.
Compute total score for each item. Compute inter-item correlation matrix.
Compute item difficulty and biserial r between an item and total test score for J items: (J = 1, 2, 3...60).
Compute MSE(P̂), ρ(PP̂), MSE(d̂), ρ(dd̂) and proportion of misclassified items.
r= 50
Select 30 items with highest biserial r.
Generate examinee ability score for 300 examinees.
Compute item scores for the above examinees for the 30 items.
Compute reliability and validity of the 30 item final test.
Select 30 items randomly from 60 items.
Generate examinee ability scores for these 300 examinees.
Compute reliability and validity of above 30 item test.
Compute mean, standard deviation, minimum and maximum of the following: MSE(P̂), ρ(PP̂), MSE(d), ρ(dd), proportion of misclassified items, reliability and validity of 30 items with highest biserial r, reliability and validity of 30 items selected randomly.

ABSTRACT

This study was designed to compare the multiple matrix sampling (MMS) technique of data collection with traditionally used examinee sampling for test development. More specifically, the three main questions explored in the study were as follows: How does the MMS procedure compare with the traditionally used examinee sampling procedure when estimating item parameters such as item difficulty and item discrimination? How do the final tests, that are produced using either MMS or examinee sampling methods, compare with respect to test reliability and validity? How do the number of observations, item subtest size and the characteristics of items in an item set affect the estimation of item parameters, test reliability and validity?

A monte Carlo approach was used to construct item pools for different item sets and to simulate data for the study. The variables in the study were sampling plans, item subtest sizes, number of observations and the characteristics of item sets. The sampling plans used were examinee sampling and MMS. MMS plans varied in terms of item subtest sizes. The three item subtest sizes used were one-fifth, one-third and one-half of the total number of items in each subtest. In each MMS plan items were divided into subtests using a partially balanced incomplete block design. For each sampling plan, the number of obsevations used were 7200, 18000 and 36000. Three types of item sets that varied in

their range of difficulties and the distribution of difficulties, were used to produce normal, rectangular and platykurtic distributions of test scores. The range and distribution of difficulties used in the three types of item sets were .16 -.84 uniformly distributed, .40 -.60 uniformly distributed, and .16 -.84 normally distributed. For all item sets, the range of item discriminations varied between .05 and .60 and they were uniformly distributed. Fifty item pools of 60 items were simulated for each item set with the specified characteristics. The item response data for each combination of sampling plans, number of observations and item sets were simulated using a simple linear transformation model. Examinees were generated with normally distributed levels of ability. Fifty replications were made under each combination of sampling plan, number of observations and item set. In each replication, estimates of item difficulties and item discriminations were obtained. Mean square error and Spearman rank order correlation between estimated and true parameters were computed to determine the precision with which different sampling plans estimated item parameters under different conditions. In each replication 30 items with the highest estimated item discrimination were selected to be included in the final test. For these 30 items, responses of 300 examinees were simulated using the above procedure. The reliability (internal consistency) and the validity of the 30 item final test were estimated. The mean and standard deviation of estimates of the above statistics (MSE, Spearman rank order correlation between true and

estimated item parameters, test reliability and validity), obtained from 50 replications under each condition, were computed. These means and standard deviations were used to answer the questions stated earlier.

In general, findings of the study were as follows: There were no systematic differences in estimates of item difficulty obtained from examinee sampling and MMS plans, while keeping the number of observations and item set constant. For a given item set and number of observations, examinee sampling plans provided more precise estimates of item discrimination, test reliability and test validity than MMS plans. However, this difference decreased as the number of observations and/or as the number of items in each subtest of a MMS increased. Comparison among different MMS plans showed that estimates of item discrimination, test reliability and test validity improved as the number of items in each subtest increased. For each sampling plan and item set as the number of observations increased estimates of item parameters, test reliability and the test validity improved. Estimates of item parameters, test reliability and validity appeared to be relatively higher for item sets containing medium difficulty items than for item sets in which item difficulties were spread out. This difference may have been caused by factors such as the differential influence of guessing on items of different difficulty, variation in standard error of discrimination for items of different difficulty and total test variance.