RELATIVE EFFICIENCY OF FOUR MULTIPLE MATRIX SAMPLING MODELS IN ESTIMATING AGGREGATE PERFORMANCE FROM PARTIAL KNOWLEDGE OF EXAMINEES' ABILITY LEVELS

Dr. Sunday Oche Emaikwu and Prof. B.G. Nworgu

Abstract

The existence of incomplete data matrix in educational assessment has been a source of worry to many educationists. Sequel to this, some psychometric experts have advocated the use of multiple matrix- sampling models as one of the appropriate statistical techniques for treating incomplete data matrix.

This studs investigated the relative efficiency of four multiple matrix- sampling models in estimating aggregate performance from partial knowledge of examinee's ability levels. The design of this study was quasi- experimental research design. The data for the study were collected using a 90-item cognitive test of ability in mathematics (CO I'AM) administered on 600 examinees. To carry out the study, three research questions were answered and three hypotheses tested. The results indicated that the estimated mean square errors of estimates did not vary significantly over the four multiple matrix-sampling models used. The overall findings failed to support the superiority of one model over another. Observably, test results arrived at by the systematic process of multiple matrix sampling techniques could provide reliable and valid index of examinees' true scores.

Recommendations were made based on these findings.

Introduction and Statement of the Problem

Education is a dynamic process which changes with the needs and aspirations of the society, for education to he functional there must be a constant and careful evaluation of educational system. Testing is one of the instruments for evaluation and accreditation in schools. Regular testing of students' educational achievements is necessary for determining learning difficulties and levels of mastery of examinees. The essence of testing is to reveal the latent ability of examinee (Emaikwu. 2004). Ability testing has always been an important part of the school system that even the habitual absentees usually turn-up to school and present themselves for testing on examination days.

MacDonald and Sampo (2002) maintain that ability connotes the characteristics of the examinees that the test is intended to measure. It includes factual knowledge, specific skills as well as more general skills. For an examinee's true ability to be estimated, the examinee has to respond to a sample of questions. A test score based on this sample of questions will be an approximate indicator of examinee's true ability (Nworgu, 2003). Test as an instrument of evaluation is often feared in schools even by the most brilliant students. This fear emanates as aresult of the nagging effects and consequences of failure. Failure to most people is taken asa complete economic, social and psychological disaster in life (Emaikwu and Eba. 2001).

from item response theory (IRT). Wilcox (1999) observes that ability is fairly invariant within an examinee for a given period of time. Hence if an examinee's ability has been established through his responses to certain carefully calibrated test items, then it is possible to estimate his ability in a similar situation in which lie is not present. On the contrary, empirical investigations often conducted in applied setting are frequently hindered by incomplete data. Sometimes students' personnel records arc unusable for research purpose because some students may have incomplete school performance records and there is little advice concerning the appropriate method for dealing with such problem. Often times, some students might have winked hard on their class work and assignments at a particular period but have failed to complete a given programme in school due to reasons

such as sudden illness, transfer of parents, war. accidents, etc. In fact such students may even be made to lose some academic sessions as results of these unforeseen and unfortunate circumstances. There could even be high probabilities based on their previous performance that such students would have passed the tests if they had taken part completely in the programme. Under this kind of situation, what then should educational practitioners do to facilitate smooth evaluation? I he elimination of

cases with incomplete data records or substitution of missing values with variable mean may not be good approaches (Raymond and Roberts. 1987).

I he two main consequences of missing data are (a) decrease in statistical power due to loss of information and (b) the possibilities of biased estimates for parameters since most statistical procedures require complete data for a case to be included in analysis. Hence the existence of incomplete data matrix in educational assessment has been a source of *worry* to many psychometric experts. In Nigeria currently, there have been no ideal and acceptable statistical procedures available in schools whereby students' aggregate performance could be estimated on the basis of partial information about their test performance. Sequel to this, some psychometric experts have advocated the use of multiple matrix sampling models as one of the appropriate statistical techniques for treating incomplete data matrix in educational measurement.

Multiple matrix sampling model is a statistical procedure in which a set of K-items is subdivided into t-subsets containing k-items each, with each subtest being administered to n-examinees selected randomly from the population of N-examinees. Although each examinee tested is administered only a portion of the K-items, the results from each subtest may be used to estimate the statistics of the universe scores which would have been obtained by administering all the K-items to all the N-examinees (Gressard and Loyd, 1991; Childs, 2003). What is meant by multiple matrixsampling models is explained easily through considering N by K data matrix of items scores which would have been obtained by administering all K-items in a giving universe to each of the N-examinee constituting the examinee population. The goal of multiple matrix-sampling model is to estimate the attributes of this N by K data matrix using a subset of scores selected randomly from complete matrix (Lord, 1980). Precisely, multiple matrix-sampling models is therefore a statistical procedure in which a domain of test items is subdivided into several test forms, with each form being administered to a certain number of examinees selected randomly from the examinee population. Hence in multiple matrix-sampling models, each examinee is presented with a sample of items from the total test; on the basis of his performance on that item sample, his score of the remainder of the total test, the composite of the items with which he was not presented is predicted.

Multiple matrix sampling could make testing less labourious, uses minimal inputs to maximize output, makes testing more interesting and test result more meaningful as well as having much bearing on the characteristics being measured (Kleinke, 1983; Shoemaker, 1980; Gressard and l.ovd. 1991). Multiple matrix sampling models allows for the generalization about the domain of items without having to consider' the whole universe of item domain (Bunda. 1986). It employs the use of sampling techniques and regression anal) sis for investigating statistical relationship. One of the major contributions of this statistical model is that it involves testing examinees on a portion of the test items in the total pool and yet the parameters of the universe scores can be estimated quite accurately.

The theoretical basis of multiple matrix-sampling models is item response theory (IRT). Item response theory is a hybrid of latent trait measurement model. IRT is a mathematical function which specifies the relationship between observable examinee test performance and the unobservable trait or ability assumed to underlie performance on the test (Warm, 1978; Hambleton, 1999). An appealing feature of these models is that once an examinee's ability has been established through his responses to certain carefully calibrated test items, then it is possible to determine the probability of a correct response to an item the examinee has never taken assuming that certain item parameters have already been determined (Wilcox, 1999). With multiple matrix-sampling models, it is possible to estimate score that examinee would make on items to which they do not respond from scores that they made on items to which they have responded to. There exist numerous alternatives multiple matrix sampling models which could be used for treating incomplete data matrix in educational measurement and these include; Kleinke model. Jaeger model. Bunda model. Raseh model, Sachar-suppes model, etc.

Kleinke (1983) offers a method for predicting tola I test scores from partial scores using nonoverlapping item samples and linear prediction approach for generating the estimated total test score distribution. With this approach, the total test score of the examinees may be considered as a composite of two tests. X. consisting of items presented to the examinees and Y, consisting of items not presented to the examinees. The obtained score on X is used to predict the score on Y; so that the

predicted total test score is the sum X+ Y, where, $Y' = y_{xv}("/,_v)|\ X, - X\ ] + V$

Jaeger (1984) offers an approach of item sampling and estimates of total test scores that multiplies the examinees' scores by a constant recollecting the proportion of items from the total scores, which were administered to the examinees in order to obtain the total test scores. Using this model, if there were K items on the total test which were administered and an examinee gets the score X, on both subsets consisting 2k of these items, then his predicted total test score is Y i = Kx,/2k.

Sacliar and Suppes (1985) affirm that Rasch developed and empirical formula for predicting ability estimates using the relationship given by B, = H + X [log{'/(.,}] where, X = {I + $W^2/2.89$ |[1'. B is the person's ability measure , U is the mean difficulty - of the test items, Was the standard dev iation of these item difficulty, all in logic, r is the person's raw score, and U is the number of items in the test.

Bunda (1986) proposes a model for estimating total test scores using multiple linear regression equation whose coefficients are found from item total covariance matrix and item mean which is given by the formula Y, = a + b.v, + car,, where a, b and c are regression constants and x, and x,, are the scores of the examinees on different item samples.

More over, some of these identified multiple matrix sampling models are not of equal efficacy in estimating aggregate scores from incomplete data matrix. Based on this context, the present study was aimed at assessing the relative efficiency of four multiple matrix sampling models in estimating aggregate performance from partial knowledge of examinees' ability levels. The study specifically investigated the extent to which the total test scores of examinees would differ with different multiple matrix sampling models. It equally examined how comparable to traditional testing procedures are scores which could be arrived at by using multiple matrix sampling procedures.

Research Questions

In pursuance of this study, the following research questions were formulated and answered.

1. To what extent do the estimates of the examinees total test scores over the different samples of test items vary?
2. To what extent do the estimates of the examinees' total test scores using the four multiple matrix sampling models (namely: Kleinke, Jaeger, Bunda, and Rasch models) vary?
3. Mow comparable to traditional testing procedures are scores which could be obtained by using multiple matrix sampling procedures?

Hypotheses

The following hypotheses were formulated to guide the study and were tested at 5% level of significance.

I. The estimates of the examinees' total test scores, given the raw scores, do not vary significantly over the different samples of test items.

2 The estimated mean square errors of estimates do not v ary significantly over the various multiple matrix sampling models (Kleinke. Jaeger. Bunda and Rasch models) used.

3. There is no statistical significant difference between the mean scores obtained by using the conventional assessment approach and that based on multiple matrix sampling techniques.

Research Method

7'he design of tins study was quasi-experimental research design. Specifically this study adopted the logic of balanced incomplete block design commonly used in quasi-experimental research in the estimation of test item statistics. T his design is often used when an experimenter considers a case where observations are made for only selected subsets of treatments. The data for the study were collected using a 90-item cognitive test of ability in mathematics (COTAM) comprising 6 subsections of 15-item each patterned according to the multiple matrix sample specifications with triple notations ***t/k/n.*** The items of the instrument were constructed by the researchers using senior secondary school mathematics WAEC syllabus. Two specialists in measurement and evaluation as well as three mathematics educators carried out the preliminary validation of the instrument. Subsequently, empirical item analvsis was carried out on the items of the instrument. Only items w ith satisfactory statistical qualities were included in the final version of the instrument. The reliability coefficient of the entire instrument using Ruder - Richardson formula 20 was 0.86. T he sampling

technique used in this study was multi-stage stratified random sampling. The items of the instrument were administered on a sample of 600 randomly selected candidates from a population of 6000 Senior Secondary three students in Zone A Senatorial district of Benue State. The research questions posed were answered using descriptive statistics namely: mean, standard deviation and standard errors of estimate. The hypotheses formulated were tested at 5% level of significance using inferential statistics namely: repeated measure analysis of variance (ANOVA) model and dependent t-test statistics. The figure below shows the incomplete block design with 8-examinee sample assigned to 6-subtests of the instrument.

11 e m S a m n 1 c

	Type-A item			Type B item Sample		
Examinee .Sample	A,	A;	A;	B,	B,	IT
1						
2						
3						
4						
5						
6						
7						
8						

Key :
The sample of block of test items taken by examinees.
The sample of block of test items not taken by examinees.

Fig. I

Data Analysis and Results

Research Question l

To what extent do the estimates of the examinees' total test scores over the different samples of test items vary? To answer this research question, each of the item samples was scored for all examinees and analyzed for mean and standard deviation to get a complete data matrix and the result was presented in Table I.

Table 1: Means and Standard Dev iations of Scores of Examinees on Six Sub-Test of COTAM

Subtest	Case	Mean	Standard Deviation
A,	600	7.3167	2.4047
A,	600	7.3767	2.3800
A,	600	7.2850	2.6575
B,	600	7.3033	2.7417
B,	600	7.5017	4.0407
IT	600	7.2150	4.5446

From table I, relatively very small differences |less than 0.3 taken pair-wise] exist among the mean scores of the several test forms of the instrument. To ascertain whether these observed differences in means are statistically significant, there was therefore the need to test the corresponding hypothesis.

Hypothesis 1

The estimates of the examinees' total test scores, given the raw scores, do not vary significantly over the different samples of test items. To test this hypothesis, a one-way repeated measure analysis of variance (ANOVA) model was used to test for statistical significant difference in means and the result was presented in table 2.

Table 2: One-way Repeated Measure ANOVA Model on the Performance of Examinees on Six Sub-Tests of COTAM

Sources of Variation	Sum of Squares	Degree of Freedom	Mean Square	*F-ratio*	F-critical
Subjects	1 1464.3303	599	19.1391		
Between groups	55.8514	5	1 1.1703	2.01	2.21
Residual	16643.8567	2995	5.5572		
Total	28164.0384	3599			

from Table 2 above, the F-test statistic yielded F-ratio value of 2.01. which is less than the critical F-value of 2.21 for 5 and 2995 degrees of freedom, therefore the result of F-test statistic is not

significant at 5% level of significance. Therefore the null hypothesis is iicecptod. Precisely, the estimated total test scores of examinees from any given subtests of the instrument were statistically equivalent.

Research Question 2

To what extent do the estimates of the examinees' total test scores using the different multiple matrix-sampling models (Jaeger, Kleinke, Bunda and Rascli models) vary? To answer this research question, the means, standard deviations, mean square errors and standard errors of estimates using the four models were computed and presented in Table 3.

Table 3: Means and Standard Deviations of Scores Estimated Using the Four Models

Variable	Cases	Mean (x)	Standard Deviation (S.D)	Mean Square Error	Stand; Error Estim;
Observed raw score	600	43.78	10.28		
Jaeger Model	600	44.17	10.74	74.0109	3.74 ;
.Kleinke Model	600	44.09	1 1.43	17.2615	4.15 i
Rasch Model	600	43.59	1 1.97	14.6855	3.8'2 i
Bunda Model	600 | 43.73		11.12	19.1362	4.37 is

from the fable 3 above, the computed statistic for the four models yielded very close values.

Nevertheless, to investigate whether the noticeable difference in the estimated mean square errors statistically significant, the corresponding hypothesis was therefore tested.

Hypothesis 2

The estimated mean square errors of estimates do not vary significantly over the various multiple matrix-sampling models (Jaeger, Bunda, Kleinke and Rasch Models). To test this hypothesis, the observed total test scores of the examinees were compared with the estimated total test scores for each model using the mean square error statistic. A one-way repeated measure analysis of variance (ANOVA) model was used and the result was presented in Table 4.

Table 4: One-Way Repeated Measure Analysis of Variance (ANOVA) of the Residuals of the four Models

Sources of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F-ratio | F- | j C1 llu
Subjects	4214060.517	599	7035.1595	
Between groups	7646037.466		2548679.155	0.75 *1* i - *Ini*
Residual	6046904924	I 797	3364999.958	
Total	6058765022	2399		

From Table 4 above, since the f-ratio value of 0.7574 was less than t-hc ! : 2.60; therefore the f-tesl statistic was not significant. For this reason, the nub n;y accepted. This implies that the estimated mean square error of estimates did not vary sigmi:
over the Various multiple matrix sampling models. This non noticeable statistical significant dillcrence m the estimated mean square error statistic as well as the residuals could be an indication of the effectiveness of the four multiple matrix sampling models.

Research Question 3

How comparable to traditional testing procedures are estimated scores, which could be obtained by using multiple matrix sampling procedure? To answer this research question, the Pearson product moment correlation coefficient, coefficient of determination as well as the means and standard deviations of conventional testing procedures and multiple matrix sampling techniques were computed and presented in Table 5.

Table 5: Correlation Coefficient and Coefficient of Determination of Scores Based on Matrix - Sampling and Conventional Techniques

Method	Mean	S.D	EX	EY	EX^2	EY^2	EXY	TXY	r^2	N
(X)	43.72	14.32	26816	26251	1319860			0.71	.504	**600**
(Y)	44.71	10.29				1212065	1235505			

X: Conventional Assessment Method **Y: Multiple Matrix Sampling Technique**

From Table 5 above, the value of the Pearson Product moment correlation coefficient was 0.71. This shows that there exists high positive correlation coefficient between the scores obtained using multiple matrix sampling techniques and conventional assessment procedures. Moreover, the value of coefficient of determination, which is the square of the correlation coefficient, was 0.504. This can be interpreted to mean that 50.4% of the variations in the scores obtained by using
conventional assessment technique could be accounted for or predicted using multiple matrix sampling techniques. Nevertheless, there exists slight variation in the means and standard deviations based on the two approaches. To investigate whether the difference in the respective means was statistically significant, the corresponding hypothesis was therefore tested.

Hypothesis 3

There is no significant difference between the mean scores obtained by using the conventional assessment approach and that based on multiple matrix sampling technique. To investigate this hypothesis, a dependent or correlated t-test statisti
conventional assessment method andthat based on multiple matrix sampling techniques. The
correlated t-test statistic was adopted because of the fact that the same subjects were used under both treatment conditions. The result of the computed dependent t-test statistic was presented in Table 6.

Table 6: Dependent t-test Statistic Betw een the Means of Conventional Assessment and Multiple

Method	Mean	S.D	D	*y\f*	N	D	t-cal.	t-critical
Conventional assessment	43.72	14.32	688	35984708	600	1.1467	0.1 146	1.96
Matrix sampling technique	44.71	10.29						

From fable 6 above, the dependent t-test statistic yielded a l-value of 0.1 146, which is less than the t-critical value of 1.96 at 5% level of significant, lienee the null hypothesis of no significant difference in means is therefore accepted. This implies that test results arrived at by the systematic process of multiple matrix sampling technique could provide accurate and therefore valid index of an examinee's true scores.

Discussion of the Result of the Findings

The analysis of the research question I and corresponding hypothesis I revealed that the estimates of the examinee's total test scores, given the raw, did irot vary significantly over the

different samples of test items. As observed from Table 1, no noticeable differences were apparent amongst the means of scores of several test forms of the instrument. In addition, the result of the repeated measure analysis of variance (ANOVA) model as presented in Table 2 indicated that the estimated total test scores of examinees from the given calibrated subsets of the instrument were statistically equivalent.

This finding is consistent with the earlier research results that ability is fairly invariant especially on parallel test forms. Shoemaker (1980) as cited by Wilcox (1999) observes that if an examinee's ability has been established through his responses to certain carefully calibrated items, then it is possible to estimate his ability in a similar situation even though he is not present, lie observes that the estimated test parameters on parallel test items are always statistically equivalent given the invariant nature of ability. These statements agree firmly with item response theory in terms of measurement of ability. Shoemaker (1980) maintains that from item response theory, it is possible to estimate scores that examinees make on items to which they do not respond from scores that they make on the items to which they have responded. He used his result to illustrate that if examinee Z is administered items 1 and 2 but not 3. If one actually knows the three item characteristic curves, obviously one could estimate the probability of passing item 3 from the examinee Z's position on the latent attribute.

The results from the analysis of research question 2 and the corresponding hypothesis 2 indicated that the means and standard deviations of the total test scores and the estimates based on the four multiple matrix sampling models showed slight and negligible variations as presented in Table 3. The obtained statistics were highly homogeneous using the four models. The result shown in Table 4 indicated that the estimated mean square errors of estimates did not vary significantly over the four multiple matrix sampling models. This non-significant difference in the means could be an indication of the effectiveness of the four multiple matrix sampling models in the treatment of incomplete data matrix. This result agrees with Sachar and Suppes' (1985) study that all regression based models yield fairly good estimations of the total test scores of the examinees. In the same way, Raymond and Roberts (1987) maintain that no uniform best techniques exist among the procedures for handling missing data. Since the estimated total test scores obtained using the four models were statistically equivalent, one can accept without contradiction that the four models yielded good predictions of the total scores from partial scores. The overall findings failed to support the superiority of one model over another. The measurement frameworks of the four multiple matrix sampling models used do not provide a firm basis for preferring one model to another. They all have considerable intuitive appeal. Hence the four multiple matrix sampling models were equally efficient in the treatment of incomplete data matrix.

The analysis of research question 3 and the corresponding hypothesis 3 indicated that there was no statistical significant difference between the mean scores obtained by using the conventional assessment technique and that based on multiple matrix sampling technique. From table5, it could be observed that there was no apparent difference in the mean of scores based on multiple matrix sampling technique and that based on the conventional assessment approach. Moreover there exists high positive correlation coefficient between the scores obtained by using multiple matrix sampling techniques and conventional assessment procedures. The computed values of the Pearson product moment correlation coefficient and coefficient of determination between scores based on multiple matrix sampling technique and conventional assessment technique were 0.71 and 0.504 respectively. This therefore implies

that at least, 50.4% of the scores obtained by using conventional assessment technique could be accounted for or predicted using multiple matrix sampling techniques. The result of the correlated t-test statistic of fable 6 is also a confirmation of the non-significant statistical difference in the mean scores of the two approaches. The results from this study have shown that tests for measuring classroom achievement based on multiple matrix sampling technique could provide nearly precise and therefore valid, information as could tests of the same length constructed in the traditionally manner. This result agrees with earlier findings which demonstrated that test results arrived at by a systematic process of multiple matrix sampling technique could provide nearly accurate and therefore valid index of an individuals true-scorcs (Shoemaker, 1980). Childs (2003) further asserts that multiple matrix-sampling techniques could be employed for classroom evaluation without fear that the procedure itself will detract from some estimate of an individual's performance.

Conclusion and Recommendations

The four multiple matrix sampling models were equally efficient in the treatment of incomplete data matrix. It can be concluded that the basic characteristic of the four multiple matrix sampling models do not provide a firm basis for preferring one model to another, as they all have considerable intuitive appeal. Multiple matrix sampling models could yield accurate and stable index of examinee's achievement level during evaluation relative to indices that could be arrived at by traditional method of assessing the examinee's knowledge and understanding. Test results arrived at by the systematic process of multiple matrix sampling cotdd provide accurate and therefore valid index of an examinee's true scores.

It can therefore be said that the use of multiple matrix-sampling techniques promises to lend support to the ever-increasing demands of continuous assessment in our school system. Based on multiple matrix sampling, scores could be earned, by prediction, on test items not taken, from the items that have been taken. This implies that an examinee that took 3 instead of 4 tests in a continuous assessment may earn scores on the fourth test with the application of multiple matrix sampling technique.

The techniques of multiple matrix sampling designs could be effectively utilized by national examination bodies like WATC, NECO, NABTEB, and JAMB in the standardization of achievement test and in the treatment of incomplete data matrix.

Based on the findings of this study, the following recommendations were made:

(i) The four multiple matrix sampling models (namely: Kleinke, Jaeger, Bunda, and Rasch models) be integrated into our system of educational assessment as this will go a long way to improving the educational fortunes of examinees as well as monitoring the quality of educational assessment especially where there exists incomplete data matrix about examinee's test performance.

(ii) Since multiple matrix sampling techniques could help the education industry to cope with the demands of continuous assessment in school, it is therefore recommended that teachers need to be taught the techniques of multiple matrix sampling so as to incorporate it in their testing, measurement and evaluation programmes.

(iii) If multiple matrix sampling technique is to be used more widely, computational formula which incorporate all uniform item-scoring procedures must be available and in the form easy to compute. Hence there should be a step in such direction.

References

Bunda, M. A. (1986). An investigation of an extension of item sampling which yields individual *scores. Journo/ of Educational Measurement 10* (2) 117-130.

Childs, A.R. (2003). Matrix sampling of items in iarge-scale assessments. *A peer-review Electronic Journal of Practical Assessment. Research and Evaluation. 8* (16) 1-12.

Emaikwu, S.O. (2004). Relative efficiency of four multiple matrix sampling models in estimating aggregate performance from partial knowledge of examinees' ability levels. An unpublished Pli.D. thesis submitted at the University of Nigeria, Nsukka.

Emaikwu, S.O. & Nvvorgu, B.G. (2004). Item response theory: A tool for democratization of assessment in schools. A paper accepted for publication in the *Journal of Nigerian Academic Eorum.*

CJressard. R. P. & l oyd, B. II. (1991). A comparison of item sampling plans in the application of multiple matrix sampling. *Journal of Educational Measurement. 28* (2). I I 9 - 130.

Ilambleton, R. K. (1999). Principles and selected applications of item response theory. *Journal of Educational Measurement. 14* (3) 75-96.

Jaeger, R. M. (1984). Estimation of individual test scores from balanced item samples. Paper presented at the National Council on Measurement in Education Annual Meeting held in Chicago. 6"'-9^{lh} April.

Kleinke. D. .1. (1983). A linear-prediction approach to developing test norms based on matrix sampling. *Journal of-Educational and Psychological Measurement. 30* (2). 7 5 - 84.

Lord, f. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jessey: Lrbaum Press.

MacDonald, P & Sampo. V.P. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Journal of Educational and Psychological Measurement,* 7A(5), 1040-1050.

Nworgu, B. G. (2003). *Educational measurement and evaluation: Theory and practice.* (3rd edition) Awka: Hallman Publishers.

Plumblee, L. B. (1996). An empirical check on Lord's item - sampling technique. *Journal of Educational and Psychological Measurement, 24* (4), 623 - 630.

Raymond, M. R. & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement. 47* (2), 13 — 24.

Sachar. .1. & Suppes, P. (1985). estimating total test scores from partial scores in a-matrix sampling design. *Journal of Educational and Psychological Measurement,* 7(1), 68- 79

Shoemaker. D. M. (1980). *Principles and procedures of multiple matrix sampling.* Cambridge: Ballinger Publishing Company.

Warm, T. A. (1978). *A primer of item response theory.* Technical Report with No. 941078. Oklahoma City: USA Coast Guard institute.

Wilcox. R. R. (1999). Some empirical and theoretical results on an answer-until-correct scoring procedure. *British .Journal of Mathematics and Statistical Psychology. 35* (2), 57-70.