DOCUMENT RESUME

ED 104 939 TM 004 397

AUTHOR	Myerberg, N. James
TITLE	The Effect of Item Stratification in Multiple Matrix Sampling.
PUB DATE	[75]
NOTE	14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	Computer Programs; Item Analysis; *Matrices; *Measurement Techniques; Response Style (Tests); *Sampling; Statistical Analysis
IDENTIFIERS	*Multiple Matrix Sampling

ABSTRACT

The effect of stratified sampling of items on the estimation of test score distribution parameters by multiple matrix sampling was studied. Item difficulty and/or interitem correlations were the bases of stratification. Various item iniverses were created by computer simulation and sampled according to several plans. The results indicate that stratification of items does not consistently improve the stability of parameter estimation. The results also show that the variance estimate used in many studies is biased for some item universes when difficulty stratification is used. A variance estimate developed in the current study removes this bias. (Author)

THE EFFECT OF ITEM STRATIFICATION IN MULTIPLE MATRIX SAMPLING

by

N. James Myerberg
Montgomery County (Md.) Public Schools

Paper presented at the annual meeting of
The American Educational Research Association

Washington
March 30 - April 3, 1975

The Effect of Item Stratification

in Multiple Matrix Sampling[1]

Matrix sampling consists of a sample of _n_ examinees responding to a subtest of _m_ items. The results of this subtest administration are used to estimate parameters of the test score distribution that would result if the population of _N_ examinees responded to the universe of _M_ items. When several n-by-m samples are used for this estimation the procedure is called multiple matrix sampling. The mean of the estimates from the several matrix samples is presented as the estimate of the test score distribution parameter.

The usefulness of multiple matrix sampling has been demonstrated by several authors including Lord (1962) and Plumlee (1964). Once the efficacy of this method was shown, one of the main questions that needed to be answered was which sampling plan produced the most stable parameter estimates. Shoemaker (1970, 1971) investigated this question by varying the sizes and numbers of the item and examinee samples. Defining an observation as one examinee's response to one item he concluded that increasing the number of observations improved the stability of the estimates. He also stated that in estimating the mean it was best to use many small item samples.

In other studies of sampling plans Kleinke (1969, 1972) tried item stratification to improve the stability of estimation from multiple matrix sampling. He stratified items on the basis of content, difficulty, and a combination of both, and concluded that stratification did not improve the stability of the estimates of the mean and variance from the stability attained using simple random sampling of items. However, Kleinke sampled from only one data base. He suggested that stratified sampling of item universes with a variety of combinations of item difficulties and interitem correlations be investigated before a conclusion is reached concerning item stratification in multiple matrix sampling. This investigation was carried out in the study presented below.

[1]This paper is based on the author's Ph.D. dissertation submitted to the faculty of the Graduate School of Syracuse University.

Theoretical Framework

Rajaratnam, Cronbach, and Gleser (1965) derived equations for estimating the coefficient of generalizability from a model that took stratification into account. Two studies (Cronbach, Schonemann, and McKie, 1965; and Shoemaker and Osburn, 1968) have demonstrated that for stratified sampling of items these equations produced more accurate estimates of the coefficient than equations that were derived from a model that did not consider stratification. In the present study an equation for estimating the variance of a test score distribution based on stratification of items was derived. The development of this equation followed closely the method used by Sirotnik (1970) in deriving the equation for estimating the variance in matrix sampling without item stratification. Sirotnik based his derivation on a two-way analysis of variance model, the two factors being examinees and items. Through algebraic manipulation of the expected mean squares from this model he derived the equation

$$\hat{\sigma}^2_Y = \frac{n(N-1)}{NM(n-1)(m-1)}\left[m(M-1)s^2_y - (M-m)\bar{s}^2_J\right] \qquad [1]$$

where: $\hat{\sigma}^2_Y$ = estimated variance of test score distribution of proportion correct scores;

s^2_y = sample variance of examinee proportion correct scores;

$\bar{s}^2_j$ = mean of sample item variances.

This equation had earlier been derived by Lord (1960) using a method based on bipolykays.

If the items are stratified , the appropriate analysis of variance model is a split-plot design with items nested within strata and completely crossed by examinees. The equation for estimating the variance from this model is

$$\hat{\sigma}^2_{Ys} = \frac{n(N-1)}{NM(n-1)(m-H)}\left[M(m-H)s^2_y - (M-m)\bar{s}^2_j + \frac{(M-m)\sum_h^H m_h s^2_{y(h)}}{m}\right] \qquad [2]$$

where: $\hat{\sigma}^2_{Ys}$ = the appropriate estimator of the variance of the test score distribution of proportion correct scores when items are stratified sampled;

m_h = the number of items sampled from stratum _h_;

$s^2_{y(h)}$ = sample variance of examinee proportion correct scores within stratum h;

H = number of strata in sample or universe.

Methodology

Computer simulation of examinees responding to dichotomously scored items produced the universes that were investigated. The study was carried out using programs written in Fortran IV and run on an IBM System/370 computer. The distributions of item difficulties and interitem tetrachorics were manipulated to produce a variety of universes. The tetrachorics were used to simulate content strata.

Three distributions of item difficulties were used - rectangular, normal, and negatively skewed. The negatively skewed distribution was a reflection of a chi-square curve with three degrees of freedom. All three distributions were limited to difficulties between .1 and .9. Item difficulties for each item universe were pseudo-randomly sampled from these distributions. The difficulty strata were established by ranking the difficulties from low to high and then dividing this ranking into quarters. It should be noted that the item universes that were created by simulation did not exactly meet the specifications discussed above because each examinee population consisted of only 1000 examinees. The first four moments of the distributions that were produced were well within the expected range of error. The accuracy of some of these moments could not be determined precisely because the curves for the normal and skewed distributions had closed, not infinite, tails.

Each distribution of difficulties was paired with each of three sets of interitem tetrachorics to form different universes. Twenty-four universes were studied. The within strata and among strata tetrachorics for the three sets were, respectively, .3 and .3, .5 and .3, and .5 and 0. The last set of correlations is not likely to be found on a mental test but was included in the study to determine the effect of such pure content strata on the stability of parameter estimation in multiple matrix sampling. There were four content strata in each universe.

In generating examinee responses to items the assumption was made that the underlying ability of examinee i on the skill being measured by item j within stratum h is represented by the linear model for the split-plot analysis of variance design. This model is

$$X_{ij(h)} = \mu + \psi_i + \delta_h + \pi_{j(h)} + \delta\psi_{hi} + \psi\pi_{ij(h)} \qquad [3]$$

where: $X_{ij(h)}$ = the ability level of examinee i on the skill being measured by item j within stratum h;

μ = general effect, equivalent to the matrix population mean;

ψ_i = the effect of examinee i;

δ_h = the effect of stratum h;

$\pi_{j(h)}$ = the effect of item j within stratum h;

$\delta\psi_{hi}$ = the interaction between examinee i and stratum h;

$\psi\pi_{ij(h)}$ = the interaction between examinee i and item j within stratum h.

To produce items with the desired tetrachorics it was necessary to generate these underlying ability levels for the items as multivariate normal variables with product-moment correlations equal to the specified tetrachorics. The correlation of abilities tested by any two items, j and j*, is represented by the correlation between $X_{ij(h)}$ and $X_{ij*(h*)}$ (h = h* and j = j* may be true) across the examinees. This will be indicated by r_{jj*}. This correlation between two sums is affected by the examinee related components of equation 3 - ψ, $\delta\psi$, and $\psi\pi$. The covariances of δ_h and δ_{h*} and of $\pi_{j(h)}$ and $\pi_{j*(h*)}$ are equal to zero since these factors are constant for all values of i. The ψ and $\delta\psi$ factors were generated as normally distributed variables with means equal to zero and variances of one. The covariance of the ψ factors was equal to one since the effect of examinee i was the same for all items. Since the $\delta\psi$ factors were fixed effects it was determined that the covariance of $\delta\psi_{hi}$ and $\delta\psi_{h*i}$ was equal to 1/3 for all $h \neq h*$. This has been proven by Searle (1971, pp. 400-402). Using the values discussed above and setting the value of $\sigma_{\psi\pi}$ equal to five, the equation for r_{jj*} could be solved for the values of $r_{\psi\pi}$ (the correlation between $\psi\pi_{ij(h)}$ and $\psi\pi_{ij*(h*)}$) needed to produce the desired values of r_{jj*}. Establishing $\sigma_{\psi\pi}$ equal to five was necessary to produce values of $r_{\psi\pi}$ that would form a proper correlation matrix. The $\psi\pi$ terms were normally distributed with means equal to zero. The sum of the three examinee related terms produced the examinee ability level. These ability levels were standardized to normal variables with means equal to zero and variances equal to one.

The pseudo-random generation of multivariate normal deviates was performed by first generating a row vector, $\underline{U}$, of independent normal deviates using a method outlined by Meyer (1969). This vector was then transformed to $\underline{V} = \underline{UT}$, a vector that was, in effect, sampled from a population of vectors whose elements have specified correlations. The transformation matrix, $\underline{T}$, is the upper triangular matrix that results from the square-root decomposition of the correlation matrix of the variables being generated. This method of generation was outlined by Barr and Slezak (1972).

The continuous ability levels were dichotomized by comparing the examinee's ability score for item $\underline{j}$ to $Z_j = \Phi^{-1}(1 - P_j)$, where Φ is the standard normal distribution function and P_j is the difficulty of item $\underline{j}$. If examinee $\underline{i}$'s ability level on item $\underline{j}$ was greater than or equal to Z_j, examinee $\underline{i}$ was considered to be successful on item $\underline{j}$ and was given a score of one for that item. If the ability level was less than Z_j, the score was zero. This method produced items with resulting difficulties that were extremely close to the values that had been originally generated to form the item universes. About 85% of the items had difficulties that were within .02 of these original values. The tetrachorics that resulted from this method,for the universes of 1000 examinees, were generally close to the values specified. Eighty-four percent of a 2.5% pseudo-random sample of the correlations were within $\pm$ .07 of the specified values.

For each combination of difficulties , tetrachorics, and type of stratification three multiple matrix sampling plans were studied to see what effect differing item sample sizes would have on the estimation of the mean and variance. The three plans divided the 48-item universes into 3, 6, or 12 samples. The items were exhaustively sampled. The types of sampling studied were simple random, difficulty stratified, content (tetrachoric) stratified, and combined difficulty and content stratified. Each item sample was matched with an examinee sample of 16. In none of the plans were the examinees exhaustively sampled. Each sampling plan was replicated 500 times for each universe.

To judge which item sampling plan and which variance estimator produced the best estimates, the mean squared errors (MSEs) of the estimates were compared. Negative estimates of the variance were included in the computation of the MSEs because the main purpose of the study was to investigate the stability of the parameter estimates. Not including the negative estimates would have distorted the estimates of this stability.

Results and Discussion

The results indicate that stratification of items does not consistently improve the stability of the estimation of the mean and variance in multiple matrix sampling for the item universes and sampling plans studied. In a few cases there may be evidence favoring stratification. However, with one possible exception, no trend emerges to indicate that stratification of items should be recommended on statistical grounds.

Presented in Table 1 are the results for estimating the mean. The means ($\hat{\mu}$) and the MSEs for each distribution of 500 estimates are presented along with the means (μ) of the test score distributions. In 16 of the 24 universes studied, stratification of items on the basis of difficulty produced a smaller MSE than simple random sampling of items. However, no distribution of difficulties, no set of tetrachorics, nor any sampling plan had a systematically lower MSE for stratification. In only 7 of 18 cases was the MSE from item sampling with content stratification less than from simple random sampling of items. The only systematic improvement was found with simultaneous stratification of content and difficulty where stratification produced smaller MSEs for all 6 universes.

The results for estimating the variance are shown in Tables 2 and 3. The means ($\hat{\sigma}^2_Y$ or $\hat{\sigma}^2_{Ys}$) and MSEs for each distribution of 500 estimates are presented along with the variances (σ^2_Y) of the test score distributions. Difficulty stratification and simultaneous difficulty and content stratification were sometimes accompanied by a negative bias when $\hat{\sigma}^2_Y$ was used to estimate the variance. The bias increased as the difference between the inter-item correlations within strata and among strata increased. There

TABLE 1

Results for estimating the mean (μ) of test score distributions using multiple matrix sampling.

Item Univ.	# of Samples	μ	Random $\hat{\mu}$	Random MSE($\hat{\mu}$)	Diff. Str. $\hat{\mu}$	Diff. Str. MSE($\hat{\mu}$)	Cont. Str. $\hat{\mu}$	Cont. Str. MSE($\hat{\mu}$)	CD Str. $\hat{\mu}$	CD Str. MSE($\hat{\mu}$)
1.U,N[a]	3[b]	24.271	24.229[c]	2.207	24.209	2.411				
2.U,R	3	26.181	26.167	1.791	26.132	1.736				
3.U,S	3	33.915	33.905	2.011	33.865	1.982				
4.V,N	3	22.863	22.856	3.118	22.892	3.118	22.883	3.116	22.863	3.087
5.V,R	3	25.956	26.042	2.243	26.031	2.210	26.010	2.186	26.040	2.220
6.V,S	3	33.983	33.880	1.965	33.936	1.869	33.958	2.031	33.966	1.925
7.O,N	3	25.701	25.702	1.267	25.735	1.071	25.728	1.361	25.690	1.129
8.O,R	3	24.265	24.231	1.107	24.278	1.110	24.256	1.157	24.238	.964
9.O,S	3	34.812	34.841	1.118	34.779	1.158	34.842	1.170	34.804	1.077
10.U,N	12	23.963	23.992	.855	24.014	.953				
11.U,R	12	23.767	23.798	.800	23.736	.765				
12.U,S	12	33.722	33.661	.747	33.721	.824				
13.V,N	12	25.187	25.185	1.086	25.192	1.042	25.255	1.023		
14.V,R	12	22.333	22.296	.896	22.286	.793	22.298	.915		
15.V,S	12	31.738	31.733	.822	31.735	.691	31.741	.873		
16.O,N	12	23.588	23.521	.772	23.591	.634	23.606	.806		
17.O,R	12	24.000	24.004	.620	24.011	.476	23.993	.603		
18.O,S	12	34.625	34.688	.656	34.663	.459	34.684	.574		
19.V,N	6	24.001	23.976	1.490	24.073	1.405	24.103	1.466		
20.V,R	6	23.629	23.661	1.221	23.712	1.243	23.739	1.185		
21.V,S	6	31.787	31.770	1.126	31.806	1.217	31.771	1.137		
22.O,N	6	25.974	25.941	.956	25.999	.816	25.973	1.029		
23.O,R	6	25.088	25.041	.900	24.994	.833	24.984	.921		
24.O,S	6	34.124	34.139	.738	34.112	.651	34.145	.758		

[a]U=Unidimensional content, tetrachorics are about .3 within, .3 among strata;
V=Varied content, tetrachorics are about .5 within, .3 among strata;
O=Orthogonal content, tetrachorics are about .5 within, .0 among strata;
N=Normal distribution of difficulties; R=Rectangular distribution of difficulties;
S=Skewed distribution of difficulties.

[b]Each universe contained 48 items and was exhaustively sampled.

[c]$\hat{\mu}$= the mean of the distribution of 500 estimates of the test score distribution mean.

TABLE 2

Results for estimating the variance (σ^2_Y) of test score distributions using the equation developed by Lord (1960) for multiple matrix sampling.

Item Univ.	# of Samples	σ^2_Y	Random $\hat{\sigma}^2_Y$	Random MSE $[\hat{\sigma}^2_Y]$	Diff. Str. $\hat{\sigma}^2_Y$	Diff. Str. MSE $[\hat{\sigma}^2_Y]$	Cont. Str. $\hat{\sigma}^2_Y$	Cont. Str. MSE $[\hat{\sigma}^2_Y]$	CD Str. $\hat{\sigma}^2_Y$	CD Str. MSE $[\hat{\sigma}^2_Y]$
1.U,N [a]	3 [b]	101.89	101.88 [c]	429.9	101.86	394.7				
2.U,R	3	77.89	77.70	267.3	77.41	286.0				
3.U,S	3	81.73	80.73	356.4	81.79	367.7				
4.V,N	3	112.58	112.82	476.3	109.47	440.0	112.54	531.4	110.15	466.9
5.V,R	3	101.40	102.12	414.6	99.85	422.2	101.98	461.2	99.47	427.9
6.V,S	3	83.83	83.49	431.2	81.69	439.8	83.80	458.2	81.47	412.2
7.O,N	3	49.27	49.32	202.1	43.79	206.8	49.12	214.6	43.87	185.2
8.O,R	3	42.08	41.39	148.3	36.54	158.3	41.93	166.9	36.88	151.4
9.O,S	3	36.69	36.83	120.5	31.91	126.4	37.10	121.0	32.48	112.1
10.U,N	12	102.75	102.89	404.4	102.96	342.2				
11.U,R	12	76.40	76.31	310.4	74.32	271.4				
12.U,S	12	80.41	79.33	297.5	78.24	314.5				
13.V,N	12	126.50	126.41	475.5	108.06	779.6	128.30	508.3		
14.V,R	12	105.09	105.75	350.0	89.97	532.1	105.97	330.6		
15.V,S	12	89.68	89.65	345.3	74.63	502.3	90.31	342.5		
16.O,N	12	50.24	49.47	361.1	11.33	1757.9	53.34	383.9		
17.O,R	12	40.01	40.31	275.5	10.14	1075.6	42.24	309.7		
18.O,S	12	39.41	38.77	246.2	9.36	1057.6	41.28	258.3		
19.V,N	6	114.23	115.30	361.6	109.99	330.9	115.94	366.2		
20.V,R	6	88.47	88.98	256.5	82.45	282.2	88.19	285.6		
21.V,S	6	83.00	83.79	285.9	78.43	285.1	82.70	301.7		
22.O,N	6	48.95	49.17	259.5	34.15	357.2	49.84	266.6		
23.O,R	6	41.54	42.39	199.6	27.66	326.5	42.64	203.3		
24.O,S	6	37.29	37.46	154.9	25.16	253.7	38.48	164.7		

[a] U=Unidimensional content, tetrachorics are about .3 within, .3 among strata; V=Varied content, tetrachorics are about .5 within, .3 among strata; O=Orthogonal content, tetrachorics are about .5 within, .0 among strata; N=Normal distribution of difficulties; R=Rectangular distribution of difficulties; S=Skewed distribution of difficulties.

[b] Each universe contained 48 items and was exhaustively sampled.

[c] $\hat{\sigma}^2_Y$ = the mean of the distribution of 500 estimates of the test score distribution variance.

TABLE 3

Results for estimating the variance (σ^2_Y) of test score distributions using the equation appropriate for stratified item sampling in multiple matrix sampling.

Item Univ.	# of Samples	σ^2_Y	Diff. Str. $\hat{\sigma}^2_{Ys}$	Diff. Str. MSE $[\hat{\sigma}^2_{Ys}]$	Cont. Str. $\hat{\sigma}^2_{Ys}$	Cont. Str. MSE $[\hat{\sigma}^2_{Ys}]$
1.U,N[a]	3[b]	101.89	101.95[c]	395.3		
2.U,R	3	77.89	77.55	287.7		
3.U,S	3	81.73	81.87	370.0		
4.V,N	3	112.58	112.07	427.9	112.35	530.2
5.V,R	3	101.40	102.05	419.0	101.85	461.5
6.V,S	3	83.83	83.69	435.9	83.69	457.4
7.0,N	3	49.29	49.19	172.7	48.72	215.2
8.0,R	3	42.08	41.39	125.8	41.60	167.8
9.0,S	3	36.69	36.19	103.5	36.79	121.8
19.V,N	6	114.23	116.34	310.2	115.56	368.1
20.V,R	6	88.47	88.31	249.4	87.59	291.5
21.V,S	6	83.00	83.70	270.4	82.47	306.3
22.0,N	6	48.95	49.48	135.9	48.89	279.0
23.0,R	6	41.54	41.88	131.2	41.59	207.0
24.0,S	6	37.23	37.17	107.7	37.73	170.1

[a]U=Unidimensional content, tetrachorics are about .3 within, .3 among strata; V=Varied content, tetrachorics are about .5 within, .3 among strata; O=Orthogonal content, tetrachorics are about .5 within, .0 among strata; N= Normal distribution of difficulties; R=Rectangular distribution of difficulties; S=Skewed distribution of difficulties.

[b]Each universe contained 48 items and was exhaustively sampled.

[c]$\hat{\sigma}^2_{Ys}$= the mean of the distribution of 500 estimates of the test score distribution variance.

was no bias when the within and among correlations were equal. The bias was removed by $\hat{\sigma}^2_{Ys}$. There was not any bias in the variance estimates of $\hat{\sigma}^2_{Y}$ when content stratification was used.

When the derivations of $\hat{\sigma}^2_{Y}$ and $\hat{\sigma}^2_{Ys}$ are compared, the algebraic representation of the bias in $\hat{\sigma}^2_{Y}$ can be seen. Sirotnik (1970) showed that

$$\sigma^2_Y = \frac{E[MS(\text{exam.})]}{m} - \frac{\left(1-\frac{m}{M}\right) E\left[MS(\text{exam. by items})\right]}{m} \qquad [4]$$

when the two-way analysis of variance design is used. When the split-plot design is appropriate, the second term on the right side of equation 4 becomes

$$\frac{\left(1-\frac{m}{M}\right) E\left[MS(\text{exam. by items within strata})\right]}{m} . \qquad [5]$$

The first term remains the same. The relationship between expression 5 and the second term on the right side of equation 4 can be determined from the equality

$$SS(\text{exam. by items}) = SS(\text{exam. by strata}) + SS(\text{exam. by items within strata}). \qquad [6]$$

After determining the expected values of both sides of equation 6, the bias in $\hat{\sigma}^2_{Y}$, when stratified sampling is appropriate, can be shown to be

$$\frac{m(H-1)}{M(m-1)}\sigma^2_{\psi\pi} - \frac{m-m_h}{m-1}\sigma^2_{\delta\psi} . \qquad [7]$$

As the sampling fraction for items decreases the second term in expression 7 becomes dominant, increasing the negative bias. The results in Table 2 verify this statement. The reasons that certain interitem correlations affect this bias have not been determined. Future investigation of this problem is needed.

In general, stratified sampling is beneficial compared to simple random sampling when it establishes a sampling plan that can force similarity among samples and thereby control a large portion of the variance across the samples. Item stratification, as done in the present study, does not do this. It would be possible to control more variance across samples if examinees were also stratified. However, the complexity of such a sampling plan may make it impractical.

Another possible way to improve the stability of estimation in multiple matrix sampling might be to sample more items from strata with larger variances of item difficulties. Cochran (1963, p. 96) has shown that in the usual one dimensional sampling, larger samples should be taken from strata wi[illegible]r variances. However, the results of the present study seem to indicate that this probably will not reduce MSEs in multiple matrix sampling. The normal and skewed distributions of items had strata with unequal variances of item difficulties. If these unequal variances had an effect on the MSEs of the estimates from universes with normal and skewed distributions of difficulties the evidence presented by Cochran indicates that stratified sampling would have produced consistently larger MSEs than simple random sampling for these universes. The results did not show this. The proportion of universes in which stratified sampling produced smaller MSEs was about the same for all three distributions of difficulties. For example, in estimating the mean for universes with rectangular distributions, 11 of 16 sampling plans favored stratification. For the skewed and normal distributions there were 8 of 16 and 10 of 16 plans, respectively, that favored stratification.

The conclusion that item stratification does not improve the stability of parameter estimation in multiple matrix sampling is consistent with the conclusion presented by Kleinke (1972). However, as he pointed out, there may be practical considerations that indicate stratification should be used. One such consideration is the time needed to administer each sample of items. Certainly most principals would not want to have a test used in their school that would cause some students to finish long before others. There is always going to be some variance in testing time for examinees but stratified sampling of items can help to minimize this variance. Item stratification does not hurt the stability of estimation when the proper variance estimation is used. Thus, if practical problems can be solved by item stratification, it certainly should be used.

REFERENCES

Barr, D. R., and Slezak, N. L. A comparison of multivariate normal generators. Communications of the Association for Computing Machinery, 1972, 15, 1048-1049.

Cochran, W. G. Sampling Techniques. (second edition) New York: John Wiley and Sons, 1963.

Cronbach, L. J., Schonemann, P., and McKie, D. Alpha coefficients for stratified parallel tests. Psychometrika, 1965, 30, 291-312.

Kleinke, D. J. A linear prediction approach to developing test norms based on item sampling. Unpublished doctoral dissertation, State University of New York at Albany, 1969.

Kleinke, D. J. The accuracy of estimated total test statistics. (U. S. Department of Health, Education, and Welfare, Project No. IB070, Grant No. OEG-2-710070) Syracuse New York, 1972.

Lord, F. M. Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. Psychometrika, 1960, 25, 325-342.

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

Meyer, D.L. Methods of generating random normal numbers. Educational and Psychological Measurement, 1969, 29, 193-198.

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

Rajaratnam, N., Cronbach, L. J., and Gleser, G. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-56.

Searle, S. R. Linear Models. New York: John Wiley and Sons, 1971.

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

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

Shoemaker, D. M., and Osburn, H. G. An empirical study of generalizability coefficients for unmatched data. British Journal of Mathematical and Statistical Psychology, 1968, 21 (part 2), 239-246.

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