DOCUMENT RESUME

ED 076 614 TM 002 619

AUTHOR Brandenburg, Dale C.; Forsyth, Robert A.
TITLE The Use of Multiple Matrix Sampling to Approximate Norm Distributions: An Empirical Comparison of Two Models.
PUB DATE 73
NOTE 17p.; Paper presented at American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; *Matrices; *Models; *National Norms; *Sampling; Scores; Standardized Tests; Statistical Analysis; Technical Reports; Test Results

ABSTRACT

Multiple matrix sampling (MMS) procedures were utilized to determine the necessary parameters of a Pearson Type I curve. Empirical norms distributions were approximated by both the Type I model and the negative hypergeometric model. Four existing ITED norms distributions, two subtests and two grades, were approximated by the MMS procedures. Two sampling designs for each test-grade combination were studied. Comparison of approximations obtained for the Type I curve and the negative hypergeometric curve supported the use of the Type I curve for determining test score distributions of large populations. (Author)

THE USE OF MULTIPLE MATRIX SAMPLING TO APPROXIMATE NORM DISTRIBUTIONS: AN EMPIRICAL COMPARISON OF TWO MODELS

Dale C. Brandenburg and Robert A. Forsyth
University of Illinois University of Iowa

INTRODUCTION

Lord (1962) and others (Kleinke, 1972; Plumlee, 1964; Shoemaker, 1970) have contended that more representative samples of students could be obtained for the national norms of standardized achievement tests if less examinee time were requested. They proposed an item sampling plan or a multiple matrix sampling (MMS) plan to reduce the amount of time needed per examinee. To use matrix sampling for this purpose, it is assumed that normative distributions conform closely to one or another theoretical probability distribution. The normative testing enables the testing agency to estimate the parameters of the assumed theoretical probability model. Then, an estimate of the entire norms distribution (i.e., the distribution obtained when all examinees take all items) is derived from this theoretical model.

Past research in this area has primarily involved the use of the negative hypergeometric distribution as the model for the estimated norms distribution (Lord, 1962; Shoemaker, 1970). Recently, however, Brandenburg and Forsyth (1973) have found that the empirical norms distributions of certain types of standardized tests can be approximated more adequately by using a Pearson Type I curve (Pearson and Johnson, 1968) rather than the negative hypergeometric model. The Brandenburg and Forsyth (1973) study was not an MMS or item sampling study. They

utilized the entire norms distribution for each test to compute the necessary parameters of both the negative hypergeometric (first two moments required) and the Pearson Type I model (first four moments required).

When the two theoretical cumulative distributions were reproduced using the moments of the empirical data, the Type I model was found to fit the observed data more closely. Since the Pearson Type I model requires estimates of the first four moments of the norms distribution and since these higher moments of a distribution are known to have a high degree of sampling error, it seemed reasonable to investigate the superiority of the Pearson Type I model under MMS conditions. The primary purpose of this study was to compare the adequacy of these two probability models for approximating entire norms distributions when MMS procedures were utilized to estimate the distribution parameters.

PROCEDURES

Multiple Matrix Sampling

A multiple matrix sampling experiment consists of administering samples of items (i.e., subtests) from a pool of items (or a test) to samples of examinees from some well-defined population. The random samples of examinees are given either completely non-overlapping sets of items (i.e., the items are sampled without replacement) or potentially overlapping sets of items (i.e., the items are sampled without replacement for each subtest but with replacement between subtests).

The basic purpose of MMS is to make inferences about the scores of the population of examinees on the population of items. For example, in curriculum evaluations, the evaluator may be interested in the estimating

of the mean score for the population on a number of different criterion measures. Rather than give the entire set of instruments to all students, he may use an MMS approach to estimate the means.

In order to make inferences about the total population of items and examinees, it is necessary to extrapolate the information in each matrix sample. Thus, for example, if in the curriculum evaluation project mentioned above, ten samples (subtests) of items were given to ten samples of students, the information from each sample is utilized in some way to provide an estimate of the mean of all examinees on all items. For a more detailed discussion of the mechanics of these operations, the reader is referred to Shoemaker (1971a) and Knapp (1972).

Theoretical Models Utilized

This study was primarily concerned with the use of the MMS concept to estimate the parameters of two theoretical probability models: Pearson Type I and negative hypergeometric. With the numerical value of its parameters, thus specified, each model would then serve as an approximation for the entire norms distribution.

The Pearson Type I model requires the estimation of four parameters (mean, variance, a skewness index and a kurtosis index). The estimation of these quantities was accomplished primarily through the use of Lord's (1960) formulas for estimating the moments of a K-item test from the moments of a k-item test (K > k).

The first around zero moment, ($\hat{\mu}_1'$) and the next three central moments ($\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$) were further adjusted to obtain unbiased estimates of the population parameters when examinee sampling is also assumed. These formulas are shown on the following page:

Mean:
$$\hat{\hat{\mu}}_1' = \hat{\mu}_1' \tag{1}$$

Variance:
$$\hat{\hat{\mu}}_2 = \frac{n}{n-1}\,\hat{\mu}_2 \tag{2}$$

Third Central Moment (Cramér, 1946):
$$\hat{\hat{\mu}}_3 = \frac{n^2}{(n-1)(n-2)}\,\hat{\mu}_3 \tag{3}$$

Fourth Central Moment (Cramér, 1946):
$$\hat{\hat{\mu}}_4 = \frac{n(n^2-2n+3)}{(n-1)(n-2)(n-3)}\,\hat{\mu}_4 - \frac{3n(2n-3)}{(n-1)(n-2)(n-3)}\,(\hat{\mu}_2)^2 \tag{4}$$

To specify a particular Pearson Type I curve, the four moments must be converted to coefficients of skewness and kurtosis. For a single matrix sample, the coefficients of skewness and kurtosis could be estimated as follows (population coefficients are given also):

	Population		Matrix Sample	
Skewness:	$\sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$	(5)	$\sqrt{\hat{\beta}} = \hat{\hat{\mu}}_3/\hat{\hat{\mu}}_2^{3/2}$	(6)
Kurtosis:	$\beta_2 = \mu_4/\mu_2$	(7)	$\hat{\beta}_2 = \hat{\hat{\mu}}_4/\hat{\hat{\mu}}_2^{2}$	(8)

However, when several sets of matrix sample data are available for estimating a given parameter, there exists at least two ways of combining this data. The moments estimated from each matrix sample could be combined to estimate the population second, third and fourth moments. From these estimates, skewness and kurtosis indices could be computed (average moment method). Or alternatively, estimates of the skewness and kurtosis indices could be obtained from the moments of each matrix sample and combined to yield overall estimates of the indices

(average ratio method). From the data obtained in the MMS experiments, both methods appeared to yield equally good estimates of the coefficients. However, when the Type I and negative hypergeometric curves were constructed from the two sets of coefficients, the "average moment" method gave better results. The results reported in this study were those associated with the "average moment" method.

Data Sources

The available score distributions for this study were based on the Iowa Tests of Educational Development (Lindquist and Feldt, 1970). Thirty-six distributions of scores (9 tests at 4 grades) obtained from the 1971 Iowa high school testing program were available. From this pool of data, the scores from two grades, 9 and 12, and two tests, Quantitative Thinking (Q) and Use of Sources of Information (SI), were chosen for study. These four distributions represented extremes in skewness and a sizeable range in the kurtosis index. The descriptive data related to these distributions are given in Table 1.

(Insert Table 1 about here)

Item and Examinee Sampling Designs

This investigation was a post mortem or post hoc experiment. That is, the norms distributions were known, and from these distribution items and examinees were selected using several MMS designs.

Four restrictions were placed on the sampling designs. First, the items and examinees were sampled without replacement for all matrix samples. This restriction was made to limit the scope and cost of the study, and in recognition of Shoemaker's results (1971a, 1971b,

and 1972). Shoemaker has shown only small differences between overlapping and nonoverlapping item sampling. Secondly, it was arbitrarily decided that the number of items in each matrix should be no more than 1/4 of the total number of items. This means that the number of matrix samples would be at least four. Thirdly, within each design the number of items (k) and the number of examinees (n) were constant for each matrix sample. Finally, the number of examinees for each matrix sample was set at 500. Although most MMS studies have not utilized sample sizes that large, such a number of examinees per matrix sample did not seem unreasonable since the purpose was to approximate entire norms distributions.

Given these restrictions, the two tests chosen for this study exhibit different problems in choosing item sampling designs. The Quantitative test is composed of 36 items which is easily divisible into the following k by t (number of items by number of subtests) designs: 4 x 9, 6 x 6, 9 x 4. Following Shoemaker's suggestion (personal communication, 1972), the 4 x 9 plan for the Quantitative test was eliminated. The other two plans were implemented for grade 9 and grade 12 populations.

The 46-item Sources of Information test did not lend itself to a similar variety of simple designs. If it is required that every item be placed in one or another matrix, a test of this length permits only two sampling plans: 2 x 23 and 23 x 2. The first of these has too few items per matrix to make possible the necessary estimates of moments, and the second has more than the maximum allowable proportion of items per subtest. On the basis of published results (Shoemaker, 1971b), it was concluded that the random exclusion of one or two items from all subtests would not seriously affect the accuracy of the estimates of the moments. Thus, two sampling plans were adopted: four subtests of

eleven items each (11 x 4) and five subtests of nine items each (9 x 5). It should be noted that despite the exclusion of one or two items, the distribution moments that were estimated from the data were for a 46-item test. A summary of the designs for both tests is presented in Table 2.

(Insert Table 2 about here)

Five replications of each sampling design were carried out in order to provide an indication of the variability of the moment estimates and the corresponding variability of the approximations to the norms distribution. These five replications of each of two sampling plans produced 10 approximations of each norms distribution. Since four norms distributions had been chosen for study, there was a total of 40 replications of the MMS experiment.

Evaluation of Approximations

After the necessary parameters were estimated from the MMS procedure, the resultant Pearson Type I and negative hypergeometric curves were compared to the empirical norms distributions. Four measures of the discrepancy between the theoretical and empirical distributions were calculated: a) the maximum absolute difference in the relative frequency for any score interval; b) the mean absolute difference in relative frequency for all intervals; c) a chi-square type index calculated on relative frequencies (Lord's D index, 1962); and d) the maximum absolute difference in the relative cumulative frequency (rcf). For the purposes of this study, the last of these indexes (referred to as MDC) was considered to be the most accurate representation of the results. As a consequence, it is the only index discussed here. Results for the

other indices are quite similar, and they may be found in Brandenburg (1972).

The index MDC is defined as follows:

$$MDC = \text{Maximum overall empirical score points of } \left| \text{rcf at } X_i \text{ for empirical distribution} - \text{rcf at } X_i \text{ for theoretical distribution} \right|.$$

Thus, MDC represents the maximum ordinate discrepancy between the theoretical and empirical ogives. Except for the location of the decimal point, MDC equals the maximum PR difference overall score points between the theoretical curve and the "true" curve.

RESULTS AND DISCUSSION

The MDC indices obtained from the five replications for each test-grade-sampling design combination and each theoretical curve are presented in Table 3. In parentheses preceding the data for each set of ten replications is the MDC index obtained when the moments of entire norms distribution were used to define the theoretical ogive, and this model was compared to the empirical ogive. Each value is designated as an "original MDC."

(Insert Table 3 about here)

Two observations about these original MDC indices should be made before additional results are discussed. First, each original Type I MDC index is less than the corresponding negative hypergeometric index. Second, the differences between the two original indices are greater for the Q-distributions than for the S-distributions.

Given the above observations, it was not surprising to find that

in 19 of 20 replications related to the Q-test, the Type I MDC index was less than the corresponding negative hypergeometric index. In 13 of these 19 replications, the difference between the observed Type I index and observed negative hypergeometric was smaller than the difference between the original indices. However, the median difference was still approximately -0.027 for Q-9 and -.021 for Q-12. Thus, under the sampling design restrictions of this investigation for the Q-test, the results of the MMS experiment strongly support the utilization of the Type I model for approximating norms distributions rather than the negative hypergeometric model.

The results related to the S-test are not nearly as conclusive. For distribution S-9, the original difference in MDC indices was -.004108. Thus, for all practical purposes, both models were providing similar approximations. In only 5 of the 10 replications for this test was the Type I model better. Given the size of the original difference, such a result was not unexpected. However, it does provide some evidence that, under the sampling conditions of this study, the estimation of four moments rather than two does not introduce an excessive amount of error in the approximations.

For distribution S-12, 6 of the 10 replications yielded better MDC indices for the Type I model. Since, the original MDC difference (-.105498) was in favor of the Type I model, this result was expected also.

In summary, the empirical data of this study seem to support the utilization of the Type I model. Of course, generalizations beyond the restrictions of this study are difficult to make. The present study concerned a particular type of achievement test data.

Without resort to item sampling, it had been established that the norms distributions for these tests were better approximated by a Type I model. Also, it is possible that the negative hypergeometric model (which requires estimates of only two parameters) may provide "good" approximations with smaller sample sizes and hence decrease the cost of obtaining the MMS data. It is possible that other sampling designs may produce different results. Finally, the distributions studied here were somewhat "atypical" in terms of the original MDC values. In a study of 90 empirical distributions (4 of which are used here), Brandenburg and Forsyth (1973) found that the median MDC index (i.e., original MDC) for the Type I fit was .015, and the median MDC for the negative hypergeometric fit was .033. Thus, the four distributions examined in this study had above "average" MDC indices. Perhaps other distributions with smaller original MDC indices would not produce similar results. The effect of the above factors must, of course, be examined before any general conclusions regarding the Type I model can be made. Nevertheless, the results of the present study do indicate that future investigations into the use of MMS procedures for the purpose of approximating norms distributions should include the Type I model.

ADDITIONAL COMMENTS

General conclusions involving test differences or grade difference from Type I approximations are difficult to make from Table 3. There are, however, noticeable differences between designs within each test-grade combination. For Q, the 9 x 4 design yielded better results than the 6 x 6 design. This does not substantiate Shoemaker's (1971a) statement that all sampling designs with equal values of the product (t)(k)(n) give essentially the same standard error for estimated parameters. Shoemaker,

however, based his inference on mean and variance estimation, whereas our MDC data involved the estimation of four parameters.

Shoemaker (personal communication, 1972) indicated that a greater number of items per subtests may be used to better estimate the higher-order moments. This is true for MMS results for Q. However, it is not true for the MMS results for S; better results (lower MDC values) were achieved for the 9 x 5 design compared to the 11 x 4 design. But the interpretation of this reversal is confounded somewhat by factors affecting these results and not the Q results. Although the 11 x 4 design uses 2 more items per subtest, it also omits 2 items, whereas the 9 x 5 design only omits 1 item. Furthermore, the 9 x 5 design has 500 more observations per replication.

Also, it may be observed from Table 3 that 7 of the 40 MMS-derived Type I approximations and 18 of the 40 MMS-derived negative hypergeometric approximations yielded MDC indices less than their respective original MDC indices obtained from approximations using the norms distribution moments. This means that the use of population moments does not guarantee a "best-fitting" curve. In general, however, the original MDC index was for all practical purposes a lower bound for the obtained MDC indices from MMS.

The Type I approximations of the four norms distributions from the MMS experiments had MDC indices about .015 larger than their corresponding original MDC indices. Thus, it might be hypothesized that even for relatively good original MDC indices (say, less than .015) the MDC indices from the MMS technique would be near .030. If this hypothesis is assumed to be true, and if it is also assumed that the possibility of good norms approximations are greatest when a post mortem-type design is utilized, the MMS results may not be very encouraging. On the other hand, if biased populations are obtained via the traditional standardization procedures, then these results

may be viewed quite positively.

Also, it should be noted that the MDC indices were computed on the basis of the raw score norms distributions before any smoothing had been undertaken. Since most test publishers would smooth the obtained raw score distributions before assigning percentile ranks, it is possible that the norms distribution [illegible]mated from MMS techniques would approximate the smoothed distributions better than the unsmoothed distributions.

Table 1

Descriptive Data For the Four Tests

Test & Grade	No. of Items	N	Mean(μ_1)	Variance(μ_2)	Skewness($\sqrt{\beta_1}$)	Kurtosis(β_2)
Q-9	36	16,867	12.143	33.551	0.8559	3.4573
Q-12	36	11,581	17.820	66.489	0.3402	2.1100
SI-9	46	16,867	22.304	62.949	0.2027	2.2319
SI-12	46	11,581	29.058	74.220	-0.4918	2.3925

Table 2

MMS Sampling Designs (t/k/n)

Test	Grade	
	9	12
Quantitative (Q)	6/6/500* 4/9/500	6/6/500 4/9/500
Sources of Information (SI)	5/9/500 4/11/500	5/9/500 4/11/500

*The first number represents subtests, the second the number of items on each subtest, and the third the number of examinees taking each subtest.

Table 3

MDC Values

Test-Grade	k x t Design		MDC-Type I (.021232)*	MDC-Neg. Hyp. (.047654)	(TI-NH) DIFF (-.026422)
Q-9	6 x 6	R_1	.029929	.058785	-.028850
		R_2	.035248	.042731	-.007483
		R_3	.045532	.062194	-.016662
		R_4	.044109	.034803	+.009306
		R_5	.028778	.060987	-.032209
	9 x 4	R_1	.019368	.053258	-.033890
		R_2	.039317	.067945	-.028628
		R_3	.014036	.043052	-.029016
		R_4	.028100	.054013	-.025913
		R_5	.033013	.052195	-.019182
Q-12	6 x 6		(.021541)	(.050193)	(-.028652)
		R_1	.036305	.046241	-.009936
		R_2	.039912	.064887	-.024975
		R_3	.030392	.058305	-.027913
		R_4	.028969	.057338	-.028369
		R_5	.031347	.048560	-.017213
	9 x 4	R_1	.020191	.050880	-.030689
		R_2	.053794	.056605	-.002811
		R_3	.027561	.051635	-.024074
		R_4	.033359	.049260	-.015901
		R_5	.032023	.047143	-.015120

*Numbers in parentheses are the "original" MDC values calculated when population moments and the given models were used to fit the empirical norms.

Table 3 (cont.)

MDC Values

Test-Grade	k x t Design		MDC-Type I (.042927)*	MDC-Neg. Hyp. (.047035)	(TI-NH) DIFF (-.004108)
S-9	11 x 4	R_1	.066313	.068373	-.002060
		R_2	.071241	.055280	+.015961
		R_3	.047555	.053200	-.005645
		R_4	.052957	.040040	+.012917
		R_5	.036337	.032900	+.003437
	9 x 5	R_1	.040228	.046565	-.006337
		R_2	.046792	.042747	+.004045
		R_3	.036583	.041477	-.004894
		R_4	.044829	.043297	+.001532
		R_5	.040741	.043557	-.002816
S-12			(.017199)	(.032697)	(-.015498)
	11 x 4	R_1	.031203	.047869	-.016666
		R_2	.043915	.043399	+.000516
		R_3	.026028	.029613	-.003585
		R_4	.034280	.055394	-.021114
		R_5	.034280	.028068	+.006212
	9 x 5	R_1	.019118	.034215	-.015097
		R_2	.035228	.034104	+.001124
		R_3	.050286	.042598	+.006688
		R_4	.017334	.032830	-.015496
		R_5	.025612	.030752	-.005140

*Numbers in parentheses are the "original" MDC values calculated when population moments and the given models were used to fit the empirical norms.

REFERENCES

Brandenburg, Dale C. The Use of Multiple Matrix Sampling in Approximating an Entire Empirical Norms Distribution. Unpublished Ph. D. Thesis, University of Iowa, 1972.

Brandenburg, Dale C. and Forsyth, Robert A. Approximating Standardized Achievement Test Norms with a Theoretical Model. Paper to be presented at the 1973 NCME Annual Meeting.

Cramér, H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.

Kleinke, D. J. A Linear Prediction Approach to Developing Test Norms Based on Matrix Sampling. Educational and Psychological Measurement, 1972, 32, 25-84.

Knapp, J. R. Item Sampling. Unpublished manuscript, University of Rochester, 1972.

Lindquist, E. F. and Feldt, L. S. Iowa Test of Educational Development: Form X-5. Chicago: Science Research Associates, 1970.

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 Psycholog Measurement, 1962, 22, 259-267.

Pearson, E. S. and Johnson, N. L. Tables of the Incomplete Beta-Function (2nd ed.). Cambridge, London: University Press, 1968.

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.

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. Principles and Procedures of Multiple Matrix Sampling. Technical Report 34, Southwest Regional Laboratory for Educational Research and Development, August, 1971. (a)

Shoemaker, D. M. Further Results on the Standard Errors of Estimate Associated with Item-Examinee Sampling Procedures. Journal of Educat Measurement, 1971, 8, 215-220. (b)

Shoemaker, D. M. Personal interview. April 3, 1972.