DOCUMENT RESUME

ED 075 502 TM 002 604

AUTHOR Shoemaker, David M.
TITLE A Note on Allocating Items to Subtests in Multiple Matrix Sampling and Approximating Standard Errors of Estimate with the Jackknife.
PUB DATE 25 Nov 72
NOTE 17p.; Revised; 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; *Error Patterns; *Item Sampling; *Matrices; Technical Reports; *Test Construction
IDENTIFIERS Jackknife

ABSTRACT

Investigated empirically through post mortem item-examinee sampling were the relative merits of two alternative procedures for allocating items to subtests in multiple matrix sampling and the feasibility of using the jackknife in approximating standard errors of estimate. The results indicate clearly that a partially balanced incomplete block design is preferable to random sampling in allocating items to subtests. The jackknife was found to better approximate standard errors of estimate in the latter item allocation procedure than in the former. These and other results are discussed in detail. (Author)

Revised Manuscript:
25 November 1972

ABSTRACT

A NOTE ON ALLOCATING ITEMS TO SUBTESTS IN MULTIPLE MATRIX SAMPLING AND APPROXIMATING STANDARD ERRORS OF ESTIMATE WITH THE JACKKNIFE

DAVID M. SHOEMAKER

Southwest Regional Laboratory for Educational Research and Development

Investigated empirically through post mortem item-examinee sampling were the relative merits of two alternative procedures for allocating items to subtests in multiple matrix sampling and the feasibility of using the jackknife in approximating standard errors of estimate. The results indicate clearly that a partially balanced incomplete block design is preferable to random sampling in allocating items to subtests. The jackknife was found to better approximate standard errors of estimate in the latter item allocation procedure than in the former. These and other results are discussed in detail.

A NOTE ON ALLOCATING ITEMS TO SUBTESTS IN MULTIPLE MATRIX SAMPLING AND APPROXIMATING STANDARD ERRORS OF ESTIMATE WITH THE JACKKNIFE

DAVID M. SHOEMAKER

Southwest Regional Laboratory for Educational
Research and Development

Multiple matrix sampling or, more popularly, item-examinee sampling, is a procedure in which a set of K test items is subdivided randomly into t subtests containing k items each with each subtest administered to n examinees selected randomly from the population of N examinees. Although each examinee receives only a proportion of the K test items, the equations given by Hooke (1956) and Lord (1960) permit the researcher to estimate parameters of the test score distribution which would have been obtained by testing all N examinees over all K test items. Because numerous combinations of t, k, and n are feasible in any investigation, the researcher must come to grips with several questions about how the procedure should be implemented. "How should items be allocated to subtests?" is one important question requiring an answer and is the one addressed specifically herein; concomitantly, the feasibility of using the jackknife procedure for approximating standard errors of estimate in multiple matrix sampling is considered in some detail.

A basic requirement in multiple matrix sampling is that k items from the K-item population are allocated randomly to each subtest. However, in constructing the t subtests, four general item allocation procedures are possible -- each of which is described more appropriately as <u>restricted random sampling</u>. The four procedures and concomitant restrictions are listed in Table 1 and an example of each procedure is given in Table 2 for $k = 3$ and $K = 7$.

Please insert Tables 1 and 2 about here.

Procedures 1, 2 and 3 are implemented easily in practice; Procedure 4, however, is more difficult and the degree of difficulty increases with increases in $\underline{K}$. Within the context of the design of experiments, Procedures 3 and 4 are referred to, respectively, as a "partially balanced incomplete block" design (PBIB) and a "balanced incomplete block" design (BIB). That which is "partially balanced" or "balanced" by each design is the item pairings. In the BIB design, all possible item pairings occur among subtests and they occur with equal frequency; in the PBIB design, item pairings do not occur with equal frequency and, indeed, some item pairs may be excluded completely. A BIB design is often difficult to implement because, for a given $\underline{K}$, no design may exist, or, if there is a design, the number of subtests required is excessively large. This limitation is most serious when $\underline{K}$ exceeds 50 even permitting minor adjustments in $\underline{K}$ to fit an available design. For example, when $\underline{K} = 91$ and $\underline{k} = 10$, 91 subtests would be required; for $\underline{K} = 97$ and $\underline{k} = 10$, 4656; and, for $\underline{K} = 199$ and $\underline{k} = 10$, 19701. The first of these three BIB designs is cited and illustrated by Cochran and Cox (1957); the other two are given by Ramanujacharyulu (1966) and cited by Knapp (1968a). Although BIB designs have been used on a few occasions (e.g., Knapp, 1968a, 1968b) when $\underline{K}$ was small (i.e., 43, 29 and 13 with Knapp), such designs are ill-suited to large item populations. This point is of no minor import because one of the major reasons for using multiple matrix sampling is its potential for dealing with large item populations. Because of this, it is expected that the majority of item allocation procedures in multiple matrix sampling will involve Procedures 1, 2 or 3.

It should be noted that, in practice, Procedures 1, 2, and 3 are implemented typically in conjunction with item stratification, that is, a stratified-random sampling procedure is used with the stratification being on item content, item difficulty level or both item content and item difficulty level. The relative merits of such stratification procedures have been discussed previously (i.e., Shoemaker and Osburn, 1968; Kleinke, 1971) and are not considered here.

Of principal interest in this investigation were the relative merits of Procedures 1 and 3. Procedure 2 was excluded because it is used rarely in practice. The metric by which these two item allocation procedures were contrasted was the standard error of estimate.

METHOD

The research design was one of post mortem item-examinee sampling with the required data bases generated through a computer simulation model described previously by Shoemaker (1971). In post mortem item-examinee sampling, various samples of items and examinees are selected randomly from a data base (an item by examinee matrix) and used to estimate parameters of the base from which they have been sampled. The researcher acts as if only certain examinees have been tested over certain items knowing all the while the results obtained by testing all examinees over all items.

Parameters of the data base manipulated systematically were: (a) the number of test items ($\underline{K}$ = 40, 60), (b) variance of the item difficulty indices (σ_p^2 = .00, .05), (c) reliability of total test scores (α = .80, .90), and (d) degree of skewness in the normative distribution (distributed normally, markedly negatively-skewed). When the distribution of test scores was negatively-skewed, only σ_p^2 = .00 was used. The selection of parameters was not unrelated to that encountered frequently in practice. It is well-known that when items are scored dichotomously the variance of the item difficulty indices for most standardized achievement tests (whose test scores are frequently distributed approximately normally) ranges typically from .04 to .08 and the corresponding value for markedly-skewed distributions of test scores (e.g., those resulting from pretests, posttests, and "criterion-referenced" tests) is approximately zero. The reliability coefficients selected are not unusual and span a familiar range. The procedure used in this investigation to generate data bases was costly and, for this reason, data bases having 40 and 60 items were generally used. However, to determine the degree of generalizability of results obtained using these data bases, several additional sampling plans were used on bases having 100 items ($\underline{K}$ = 100).

The nine item-examinee sampling plans used on data bases having 40 and 60 items are listed in Table 3. For several of these sampling plans, the number of examinees per subtest was varied systematically ($\underline{n}$ = 10, 20, 30 and 40) to determine the degree of generalizability of results obtained when $\underline{n}$ = 50 to other values of $\underline{n}$. A PBIB design was used only when $\sigma_p^2 > 0$ for a given data base. When $\sigma_p^2 = 0$, all items are statistically parallel and Procedures 1 and 3 produce equivalent results (and all differences observed between the two procedures would be due to the sampling of examinees.)

The parameters estimated were μ_1' (the mean test score), μ_2, μ_3, μ_4 (the second through fourth central moments) and σ_p^2. Estimating moments of the test score distribution is important in multiple matrix sampling because they are the required statistics in graduating the normative distribution -- one of the major objectives of multiple matrix sampling.

The equations used to estimate the moments of the test score distribution were those given by Lord (1960); σ_p^2 was estimated through a components of variance analysis. The results of each sampling plan were replicated 50 times.

The Jackknife Procedure

Of additional concern in this investigation was examining the feasibility of a statistical procedure known as the "jackknife" in approximating standard errors of estimate in multiple matrix sampling. A good description of the jackknife is given by Mosteller and Tukey (1968) and some preliminary results in applying the procedure to multiple matrix sampling are given by Shoemaker (1972a). In general, the jackknife operates on a data base which has been divided into subgroups of data and produces a mean estimate of the parameter and approximates the standard error of estimate associated with this statistic. The basic component of the jackknife is the pseudovalue associated with each subgroup which is the weighted difference between the statistic computed on all the data and the statistic computed on the body of data which remains after omitting that subgroup. Because the pseudovalues behave as though they were independent of each other, the standard error of the statistic is computed according to the well-known formula for the standard error of a sample mean. When the jackknife is applied to multiple matrix sampling there are t subgroups of data but only one score (the estimated parameter) for each subgroup with that statistic weighted according to the number of observations tk acquired by that subtest. The jackknife operates on the statistics obtained from <u>one</u> set of t subtests and approximates the variability of the pooled estimates which would have been observed over repeated replications of the design.

The computations involved in the jackknife are relatively simple. Let

t = the number of subgroups (subtests),

y_{all} = the statistic computed on all the data, and

$y_{(j)}$ = the statistic computed on all the data left after removing subgroup j.

The pseudovalues, y_{*j}, are then equal to

$$y_{*j} = ty_{all} - (t - 1)y_{(j)} \quad \text{for } j = 1, 2, \ldots , t.$$

The jackknifed estimate of the parameter is equal to

$$y_* = (y_{*1} + y_{*2} + \ldots + y_{*t})/t$$

with an estimate of its variance given by

$$s_*^2 = \frac{\sum^t (y_{*j} - \overline{y_{*j}})^2}{t(t - 1)} .$$

The procedure used in this investigation for testing the jackknife was relatively straight-forward. Because each sampling plan was replicated _r_ times, _r_ estimates of each parameter were produced as well as _r_ estimates of the jackknifed standard error for each parameter. At the end of _r_ replications, _two_ estimates of the standard error of estimate for each parameter for each sampling plan were computed. The first estimate was obtained by computing the standard deviation of the _r_ estimates of each parameter; the second, by computing the mean of the _r_ jackknifed standard errors for each parameter. The jackknife is justified to the degree that the two standard errors agree.

RESULTS

The interrelations among standard errors obtained when α = .80 were very similar to those obtained when α = .90 and, for this reason, only those results obtained when α = .80 are reported in detail in Tables 3 and 4. The only difference observed between the two data sets was that, result for result, the standard errors of estimate per item-examinee sampling plan were generally larger for the higher reliability. This increase was not unexpected and was consistent with previous results reported by Shoemaker (1972b). Concomitantly and to conserve space, only results obtained for $\hat{\mu}_1'$ and $\hat{\mu}_2$ are tabulated. There is no loss of information here because results similar to $\hat{\mu}_2$ were obtained for $\hat{\mu}_3$, $\hat{\mu}_4$ and $\hat{\sigma}_p^2$. Although not reported in detail here, the results obtained using data bases having 100 items (K = 100) and item-examinee sampling plans involving examinee subgroups of size 10, 20, 30 and 40 suggest strongly that the conclusions drawn here are generalizable to a variety of data bases and to a variety of item-examinee sampling plans.

Please insert Tables 3 and 4 about here.

The entries in Tables 3 and 4 are interpreted similarly and only those for one sampling plan in Table 3 need be described in detail to explain both tables. The first three entries in the first row of Table 3 give the parameters of the data base. In this case, the item population consisted of 40 items, the variance of the item difficulty indices (p = proportion answering the item correctly) was equal to 0 and the test scores were distributed normally. Using a ($\underline{t}$ = 4/$\underline{k}$ = 10/$\underline{n}$ = 50) item-examinee sampling plan with random allocation of items to subtests (Procedure 1 in Table 1) and replicating the sampling plan 50 times, the standard deviation of the 50 pooled estimates of the mean test score on the 40-item test was equal to .4695. Fifty jackknifed estimates of the standard error of the mean were produced. Their mean was equal to .4793; their standard deviation, .2445. If the items for each subtest had been allocated using a PBIB design (Procedure 3 in Table 1), corresponding results would have appeared under 'PBIB' in the first row. None are given there because $\sigma_p^2 = 0$ and the two item allocation procedures are equivalent.

Looking at all results for SE(R), it was generally the case that, for each sampling plan, the standard error of estimate was less when a PBIB design was used. The relative magnitude of this discrepancy was greater for the mean test score and decreased sharply for successively higher central moments. Because several combinations of $\underline{t}$ and $\underline{k}$ (for a given $\underline{tk}$) occurred among sampling plans, it was possible to examine the effect of certain combinations on the standard error of estimate. For a given $\underline{tk}$, an increase in $\underline{t}$ resulted in a decrease in SE(R) when estimating the mean test score; for the second through fourth central moments, an increase in $\underline{k}$ resulted in a decrease in SE(R); and, for σ_p^2, no trend was discernable.

Regarding the jackknife, the results indicate that on the average it did approximate well standard errors of estimate. A major exception, and one noted previously by Shoemaker (1972a), was found in estimating the standard error of the mean test score using a PBIB design where the jackknife consistently and markedly overestimated SE(R). However, the jackknife did approximate well the standard error here when a random sampling design was used to allocate items to subtests. Looking at the results across parameters, it was generally found that, when a PBIB design was used, the jackknife overestimated standard errors of estimate. This did not occur when a random sampling design (Procedure 1 in Table 1) was used. The relative discrepancy was most marked for the mean test score and decreased in magnitude for successively higher central moments. In a manner similar to SE(R), the standard deviation of the jackknifed estimates of the standard error SD(J) decreased with increases in _t_ when estimating the standard error of the mean test score and decreased generally with increases in _k_ when estimating the standard errors of the higher central moments for a given _tk_.

DISCUSSION

The results support the conclusion that the procedure for allocating items to subtests in multiple matrix sampling is an important consideration. Specifically, a partially balanced incomplete block design is preferable to a random allocation for sampling plans having the same _tk_. The superiority of the PBIB is most apparent in estimating the mean test score and becomes less apparent in estimating higher central moments. This reinforces a conclusion made by Lord and Novick (1968) that in estimating the mean test score omitting even one item has a drastic effect on the standard error of estimate. In this investigation, a PBIB design guaranteed that each of the $\underline{K}$ items was included in some subtest. Such was not the case with a random allocation of items where it was quite possible for certain items to be omitted completely (as happened to item 2 in Procedure 1 in Table 2). The results indicate that the Lord and Novick conclusion is applicable to higher central moments but the expected discrepancies are not as drastic as those expected with the mean test score.

Of additional interest in this investigation was the use of the jackknife in approximating standard errors of estimate in multiple matrix sampling. The results reinforce the conclusion drawn by Shoemaker (1972a) that the jackknife can be used for this purpose and also shed light on a problem mentioned therein. Shoemaker noted that the jackknife overestimated the standard error of the mean test score when $\sigma_p^2 = .05$ and items were allocated to subtests using a PPIB design. The results in Table 3 suggest that the inability of the jackknife to perform well in this case was a function of the item allocation procedure. For the jackknife to be appropriate, the pseudovalues must behave as though they are independent and the results suggest that this requirement is violated with a PBIB design. Regarding this violation, the jackknife is not as robust when estimating the standard error of the mean test score as it is in estimating standard errors of higher central moments. The conclusion seems warranted that, when σ_p^2 departs significantly from zero and a PBIB design is used to allocate items to subtests, the jackknife will approximate conservatively the standard error of estimate in multiple matrix sampling. It works quite well for all other cases.

REFERENCES

Cochran, , Cox, G. M. Experimental designs. (2nd Edition) New York: Wiley, 1957.

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

Kleinke, D. J. The accuracy of estimated total test statistics. Final Report: Project No. IB070, Grant No. OEG-2-710070, U. S. Department of Health, Education and Welfare, Office of Education, National Center for Educational Research and Development, 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. (a)

Knapp, T. R. BIBD vs. PBIBD: an example of a priori item sampling. Unpublished manuscript, University of Rochester, 1968. (b)

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

Mosteller, F. & Tukey, J. W. Data analysis, including statistics. In Lindzey, G. & Aronson, E. (Eds.) The handbook of social psychology. (2nd Edition) Reading, Mass.: Addison-Wesley, 1968.

Ramanujacharyulu, C. A new general series of balanced incomplete block designs. Proceedings of the American Mathematical Society, 1966, 17, 1064-1068.

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

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Southwest Regional Laboratory for Educational Research and Development, Technical Report No. 34, 1971.

Shoemaker, D. M. & Okada, M. The communication skills program and concomitant spelling proficiency. Southwest Regional Laboratory for Educational Research and Development Technical Note, 1970.

Shoemaker, D. M. A general procedure for approximating standard errors of estimate in multiple matrix sampling. Southwest Regional Laboratory for Educational Research and Development Technical Memorandum, 1972. (a)

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. (b)

TABLE 1

Procedures for Allocating Items to Subtests in Multiple Matrix Sampling

Item Allocation Procedure	Restrictions On tk	Restrictions On Sampling Of Items
1. Random Sampling	None	Without replacement within each subtest With replacement among subtests
2. Partially Balanced Incomplete Block Design (not all items tested)	$tk < K$	Without replacement within each subtest Without replacement among subtests
3. Partially Balanced Incomplete Block Design (all items tested)	$tk \geq K$ $tk = rK$ (r integer)	Without replacement within each subtest Each of the K items appears with equal frequency (r) among subtests
4. Balanced Incomplete Block Design	$tk \geq K$ $tk = rK$ (r integer) $tk = \frac{K(K - 1)\lambda}{k - 1}$ (λ integer)	Without replacement within each subtest Each of the $K(K - 1)/2$ item pairings appears with equal frequency (λ) among subtests

TABLE 2

Examples of Subtests Resulting From the Four Item Allocation Procedures Described in Table 1 Using k = 3 and K = 7

Subtest Number	Procedure 1	Procedure 2	Procedure 3	Procedure 4
1	1 3 5	1 2 3	1 2 3	1 2 4
2	3 4 5	4 5 6	4 5 6	2 3 5
3	1 3 5		7 1 2	3 4 6
4	1 4 7		3 4 5	4 5 7
5	4 5 6		6 7 1	5 6 1
6	3 4 6		2 3 4	6 7 2
7	3 6 7		5 6 7	7 1 3

Table 3

Standard Errors Of Estimate For μ_1' As A Function Of K, σ_p^2 And Degree Of Skewness In Normative Distribution For Selected Sampling Plans Using Two Item Allocation Procedures (α = .80)

Degree of Skewness In Normative Distribution	K	σ_p^2	Sampling Plan (t/k/n)	Random Allocation SE(R)	Random Allocation MN(J)	Random Allocation SD(J)	PBIB SE(R)	PBIB MN(J)	PBIB SD(J)
Normal	40	.00	(04/10/50)	.4695	.4793	.2445			
			(08/10/50)	.3421	.3487	.1238			
			(12/10/50)	.3141	.3242	.0890			
			(10/04/50)	.3871	.3948	.1258			
		.05	(04/10/50)	1.5477	1.2802	.6016	.4030	1.4623	.4959
			(08/10/50)	.8547	.9318	.2176	.3688	.9339	.2476
			(12/10/50)	.8539	.7474	.1509	.2963	.8454	.1506
			(10/04/50)	1.4660	1.3748	.3469	.3475	1.5271	.2657
	60	.00	(06/10/50)	.6474	.5889	.2319			
			(12/10/50)	.3999	.4002	.1159			
			(18/10/50)	.2815	.3057	.0628			
			(10/06/50)	.5043	.4595	.1479			
			(10/18/50)	.4831	.3853	.1019			
		.05	(06/10/50)	1.7968	1.7133	.5550	.6181	1.8427	.4875
			(12/10/50)	1.1683	1.1441	.2121	.3550	1.1650	.2622
			(18/10/50)	1.0170	.9547	.1710	.3001	.9328	.1566
			(10/06/50)	1.6451	1.7862	.4223	.5444	1.7093	.3135
			(10/18/50)	.9840	.8627	.2195	.3607	1.3160	.2781
Negatively Skewed	40	.00	(04/10/50)	.4454	.3565	.1808			
			(08/10/50)	.3082	.2611	.0914			
			(12/10/50)	.2753	.2399	.0570			
			(10/04/50)	.3657	.3188	.1183			
	60	.00	(06/10/50)	.4077	.4186	.1742			
			(12/10/50)	.3062	.3204	.0807			
			(18/10/50)	.2646	.2836	.0656			
			(10/06/50)	.4213	.3823	.1186			
			(10/18/50)	.3588	.3251	.0967			

Table 4

Standard Errors Of Estimate For μ_2 As A Function Of K, σ_p^2 And Degree Of Skewness In Normative Distribution For Selected Sampling Plans Using Two Item Allocation Procedures (α = .80)

Degree of Skewness In Normative Distribution	K	σ_p^2	Sampling Plan (t/k/n)	Random Allocation			PBIB		
				SE(R)	MN(J)	SD(J)	SE(R)	MN(J)	SD(J)
Normal	40	.00	(04/10/50)	10.3190	11.1584	5.7984			
			(08/10/50)	8.4280	8.1964	2.4210			
			(12/10/50)	7.2476	6.6786	1.9780			
			(10/04/50)	25.0248	23.0188	7.1526			
		.05	(04/10/50)	7.9539	8.9085	3.4973	7.1495	7.8056	3.1068
			(08/10/50)	6.6268	5.9696	1.9546	4.3141	5.5723	1.6081
			(12/10/50)	5.3812	5.1389	1.5443	3.7071	4.9847	1.1616
			(10/04/50)	9.7860	10.7747	2.7205	8.6920	11.4619	3.8258
	60	.00	(06/10/50)	25.6763	22.1921	8.6872			
			(12/10/50)	19.6631	18.4521	6.5854			
			(18/10/50)	16.6729	14.1373	3.9119			
			(10/06/50)	33.0618	32.5629	11.6935			
			(10/18/50)	10.0443	9.1222	2.7066			
		.05	(06/10/50)	12.5992	11.8214	4.7175	10.9600	13.3160	5.3499
			(12/10/50)	10.3995	9.7892	2.1409	7.6493	9.5244	2.7857
			(18/10/50)	8.6830	8.3365	1.9409	5.7198	8.3856	1.6831
			(10/06/50)	16.3602	16.3580	5.9341	15.1729	15.5876	5.8778
			(10/18/50)	7.2842	7.1261	2.0668	6.1346	10.4181	2.8439
Negatively Skewed	40	.00	(04/10/50)	8.1275	7.0332	2.9720			
			(08/10/50)	5.9043	5.2687	1.5521			
			(12/10/50)	3.8652	4.5914	1.0571			
			(10/04/50)	11.0677	11.4673	3.2327			
	60	.00	(06/10/50)	10.4518	12.6290	4.8652			
			(12/10/50)	7.7333	9.2427	2.2513			
			(18/10/50)	7.0408	7.9627	1.8957			
			(10/06/50)	15.4877	16.3723	4.4385			
			(10/18/50)	7.7647	7.5323	2.6283			