Efficient estimation for a multiple matrix sample design

Harvey Goldstein and **Anthony N. James**

A generalized least squares method is proposed for estimation in multiple matrix sampling designs. It is shown that this provides efficient estimates, is flexible, and makes fewer assumptions than other procedures.

1. Introduction

Multiple matrix sampling (MMS) was suggested by Lord (1962), among others, in order to make efficient use of item responses where not all individuals responded to all items. Thus a long test, considered as a collection of items, might be split into, possibly overlapping, subtests and each subject given only one subtest in order to avoid lengthy administration. More generally, the items may be grouped into categories; for example a collection of mathematics items may be categorized into those which are algebra, geometry, number, etc. Thus, each subtest would be made up of groups of items from some of the categories, and the subtests will also have items in common. We wish to make separate inferences about each category.

Figure 1 shows a subject by items data matrix with an MMS design for three overlapping subtests administered to three groups of subjects, and the items divided into three categories. Figure 2, which is more convenient for present purposes, shows the way in which the subtests incorporate the categories.

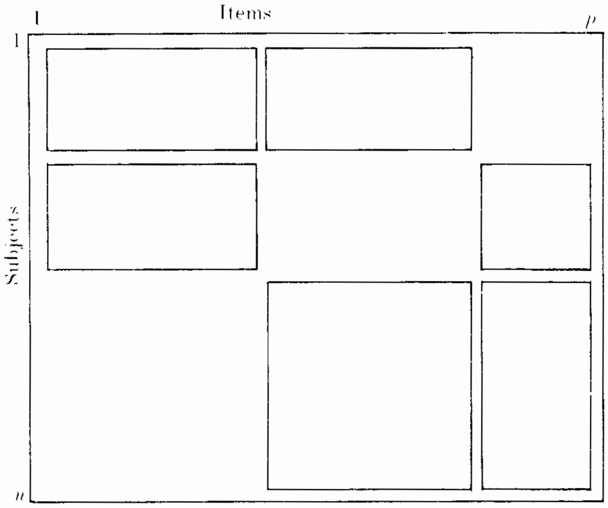

Figure 1.

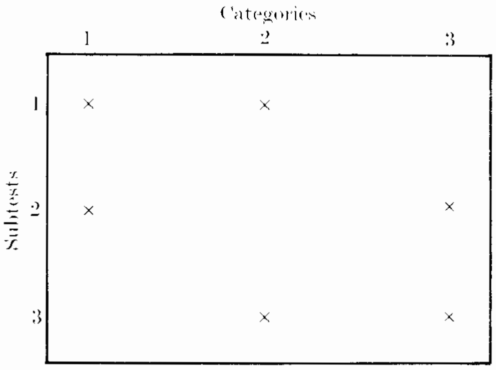

Figure 2.

2. Models for MMS estimation

A model for MMS designs is described by Sirotnik & Wellington (1977), who give detailed computational procedures, etc. Essentially, for a single item pool it is as

0007 1102/83/020167 08 $02.00/0 © 1983 The British Psychological Society

follows

$$z_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, \tag{1}$$

where

z_{ij} is the (0, 1) response of the ith subject to the jth item,
α_i is a parameter (ability) for the ith subject,
β_j is a parameter (facility) for the jth item,
$(\alpha\beta)_{ij}$ is an interaction,
ε_{ij} is a random error term.

Since in general each subject responds to each item only once, $(\alpha\beta)_{ij}$ is confounded with ε_{ij} and so is not separately estimable, and is subsumed under ε_{ij}.

The usual assumptions are:

ε_{ij} are independent random variables (local independence assumption) with variance σ^2;
α_i, β_j are random variables with variance $\sigma_\alpha^2, \sigma_\beta^2$.

Clearly (1) can be extended readily to incorporate terms for groups, covariates, and different item pools or categories.

Thus equation (1) is a two-way random effects analysis of variance model with a (0, 1) response applied to an incomplete data matrix.

Because of the finite range of the dependent response variable, efficient maximum likelihood estimation based on a binomial error model is problematical, and the usual procedures are based on equating expected mean squares using generalized symmetric means. It is worth noting that if instead of (1) we write

$$\text{logit}\,(P_{ij}) = \mu + \alpha_i' + \beta_j', \tag{2}$$

where $P_{ij} = E(z_{ij})$ and α_i', β_j' are random variables, then this is the random effects version of the so called one-parameter latent trait or 'Rasch' model. Thus (1) may be regarded as a random effects latent trait model which uses an identity rather than a logit 'link function'.

In common with other latent trait models which operate at the item level, (1) makes certain assumptions. In particular it assumes that the errors are independent and the between-individual space is one-dimensional. The model could in principle be extended to more than one dimension, although this seems not to have been done. Likewise there seems to be little work on the consequences of violations of these assumptions or on the question of efficiency. Furthermore, as pointed out by Goldstein (1980), the form of link function will in general affect the parameter estimates obtained. Finally, the model implies a random sampling of items, which will rarely be the case, although this could be overcome by requiring the β_j' to be fixed effects in which case (2) becomes the mixed effects model studied by Bock & Aitkin (1981).

The following section introduces an alternative model which makes fewer strong assumptions than (1) and (2) while providing efficient estimates.

3. Generalized least squares estimation

Since we wish to make inferences at the category level rather than the individual item level, latent trait models which incorporate item parameters are strictly unnecessary. Instead, since the basic unit of an MMS design is the set of category items which appear together within subtests, statistics derived from the responses to these sets of items can form the basis of estimation and inference procedures. In particular, in what follows we shall choose the 'raw score', or the number of correct responses if the items are binary. In order to make inferences about the totality of items for each category, we shall assume that we can find a simple transformation of the score for each set of items from a category, which is an unbiased estimator of the overall category score. In general this overall category score will be the mean raw score for the complete set of items in that category. If this is the case each item should be given the same probability of occurrence within a set. (If, instead, we wished to weight some items differentially then these would be given probabilities of occurrence proportional to the weights.) Thus, for example, in the simplest case, if items to be equally weighted are allocated with equal probability at random to sets with the same number of items per set, then the set raw scores will each estimate the required category score.

Referring to Fig. 2, denote the score for the set of items from category t for subtest i on subject j as $x_{ij}(t)$, and the sample mean for subtest i and category t as $x_i(t)$, where in general $i = 1, \ldots, q$ and $t = 1, \ldots, r$.

In the extreme case each category would contain a single item. For the present we suppose that this is not the case and that each category contains several items yielding a 'psuedo-continuous' score, which will very often be true in practice. We shall return in the discussion to the case of single item categories. We also assume that individuals are a simple random sample from a population and subtests allocated randomly to sample units. The case of complex sample designs will be taken up in the discussion.

We now need to make a further assumption. Consider a given category, say t. We wish to estimate the mean score of the items in category t, and perhaps for different groups of subjects. In the simplest case described above, for each chosen set of category t items we would have

$$E(x_i(t)) = \theta(t), \quad \text{say.}$$

More generally we can write

$$E(a_i(t) + c_i(t)x_i(t)) = \theta(t), \tag{3}$$

where the $a_i(t)$ and $c_i(t)$ are constants. In many applications we may be able to assume $a_i(t) = 0$. For example, if the category t items for a subtest are a random sample from a pool, then $a_i(t)$ can be taken as zero and $c_i(t)$ as inversely proportional to the number of category t items in subtest i. The expectation operator in (3) could then be assumed to apply to repeated selections of items for subtests. In this case, however, although the following estimation procedures can be used, a model which incorporates random components reflecting item sampling might be more appropriate (see Discussion). More generally, however, items will be selected systematically, and care will be needed with assumption (3) in any practical application, although in the case of subgroup comparisons the requirement to prespecify $a_i(t)$, $c_i(t)$ precisely can be relaxed (see Discussion). It should be noted that this assumption is effectively the usual assumption of congeneric test scores on which a great deal of classical test score theory is based (Lord & Novick, 1968). A similar assumption is also implicit in the traditional MMS estimation procedures, although perhaps less easily recognized.

Finally, equation (3) assumes that there are no 'context effects', i.e. that $E(x_i(t))$ is independent of the particular subtest in which the set of items occurs. This assumption is also made by the traditional MMS estimation procedures and can often be verified empirically. Write

$$y_{ij}(t) = a_i(t)+c_i(t)\,x_{ij}(t);$$
$$y_i(t) = a_i(t)+c_i(t)\,x_i(t).$$

We have then

$$y_i(t) = \theta(t)+\varepsilon_i(t),$$

where ε_i is a random error term. We refer to $y_i(t)$ as the elementary estimate for category t from subtest i. Write

$$\sigma_{it}^2 = \mathrm{var}\,(\varepsilon_i(t)) = \mathrm{var}\,(y_{ij}(t))/n_i,$$

where n_i is the number of individuals responding to subtest i. In general we cannot assume that $\sigma_{it}^2 = \sigma_t^2$ for all i, but where such an assumption is reasonable or verified empirically, then pooled estimates can be used in the following equations.

Thus we have for the design in Fig. 2

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\theta}+\mathbf{E}, \tag{4}$$

where

$$\mathbf{Y}^T = \{y_1(1), y_1(2), y_2(1), y_2(3), y_3(2), y_3(3)\},$$
$$\boldsymbol{\theta}^T = \{\theta(1), \theta(2), \theta(3)\},$$
$$\mathbf{E}^T = \{\varepsilon_1(1), \varepsilon_1(2), \varepsilon_2(1), \varepsilon_2(3), \varepsilon_3(2), \varepsilon_3(3)\},$$

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus $\mathbf{U}$ is an $m\times r$ incidence matrix which associates the appropriate θs with the $y_i(t)$.

We require an efficient and unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and we obtain immediately the generalized least squares estimator

$$\hat{\boldsymbol{\theta}} = (\mathbf{U}^T\mathbf{K}^{-1}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{K}^{-1}\mathbf{Y}, \tag{5}$$

where $\mathbf{K} = \mathrm{var}\,(\mathbf{E})$ and

$$\mathrm{var}\,(\hat{\boldsymbol{\theta}}) = (\mathbf{U}^T\mathbf{K}^{-1}\mathbf{U})^{-1}. \tag{6}$$

In the present case we have

$$\mathbf{K} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{11,12} & & & & \\ \sigma_{11,12} & \sigma_{12}^2 & & & & \\ & & \sigma_{21}^2 & \sigma_{21,23} & & \\ & & \sigma_{21,23} & \sigma_{23}^2 & & \\ & & & & \sigma_{32}^2 & \sigma_{32,33} \\ & & & & \sigma_{32,33} & \sigma_{33}^2 \end{bmatrix}.$$

where $\sigma_{ij,kl} = \text{cov}(\varepsilon_i(j), \varepsilon_k(l))$. In general $\mathbf{K}$ is a block diagonal matrix and here,

$$\mathbf{K} = \begin{bmatrix} \mathbf{k}_1 & & \\ & \mathbf{k}_2 & \\ & & \mathbf{k}_3 \end{bmatrix}.$$

So

$$\mathbf{K}^{-1} = \begin{bmatrix} \mathbf{k}^{-1} & & \\ & \mathbf{k}_2{}^{-1} & \\ & & \mathbf{k}_3{}^{-1} \end{bmatrix}.$$

where $\mathbf{k}_i$ is the error covariance matrix for the subtest i elementary estimates. The variances and covariances $\sigma_{ij,kl}$ can be estimated from the within-subtests sample covariances. Thus, in the present case,

$$\hat{\sigma}_{11,12} = \text{cov}(y_1(1), y_1(2))/n_1, \quad \text{etc.}$$

It should also be noted that an estimate of the variance of any linear combination of category parameters can be obtained using (6).

4. Linear modelling

In addition to estimating the population means $\theta(t)$ we may also be interested in the dependence of these on further explanatory variables such as social class, sex, etc. Thus we can specify linear models of the form

$$y_i(t) = \beta_0(t)x_{0i} + \beta_1(t)x_{1i} + \ldots\ldots + \beta_p(t)x_{pi} + \varepsilon_i(t); \quad i = 1, \ldots, q, \quad t = 1, \ldots, r. \tag{7}$$

where, typically, $x_{0i} = 1$ and the x_{ki} are the independent variable means for subtest i.

Model (4) now becomes

$$\mathbf{Y} = \mathbf{WB} + \mathbf{E}, \tag{8}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{U}_1 & \otimes & \mathbf{W}_1{}^T \\ \vdots & \vdots & \vdots \\ \mathbf{U}_m & \otimes & \mathbf{W}_m{}^T \end{bmatrix},$$

$\mathbf{U}_j$ is the jth row of $\mathbf{U}$, where j indexes the set of items from subtest i and category t.

$$\mathbf{W}_j{}^T = (x_{0i}, \ldots, x_{pi}),$$

$$\mathbf{B}^T = (\beta_0(1), \ldots, \beta_p(1), \ldots, \beta_0(r), \ldots, \beta_p(r)).$$

Generalized least squares can be applied to obtain estimates for this model, and here the covariance matrix of errors, for either simple or complex sampling schemes, can be obtained from the within-subtest regression analyses of $y_{ij}(t)$ on the x_{rij}. Hypotheses about the coefficients $\mathbf{B}$ can be tested by standard methods (see, for example, Bhapkar, 1976).

5. Common items

In the above discussion each set of category items within a subtest has been treated as a separate entity. In practice, however, these will often have items in common and this information should be utilized. Suppose that we have two sets, each of which can be divided into subsets with no items in common. This will give three such subsets, say u, v, w, where the original sets are composed of $u+v$ and $v+w$ items, say. The analysis then proceeds as before but using the three distinct subsets.

6. Example

The following example uses data collected by the British Assessment of Performance Unit (APU) (Foxman *et al.*, 1982) which consist of scores on mathematics test items administered to 11-year-olds in England, Wales and Northern Ireland in 1980. The survey design used 26 subtests each of which contained three sets of between 14 and 20 items from three of 13 subcategories of mathematics. For illustration, three subtests have been used and three categories have been chosen: symmetry, transformation and coordinates (B), computations of whole numbers and decimals (G) and Generalized Arithmetic (M). The data relate to boys only.

The sample design involved stratification and clustering and the first stage in the estimation process was to calculate the elementary estimates, their variances and covariances taking into account the sample design. These are shown in Tables 1 and 2 together with the sample sizes. In fact the off-block diagonal terms in Table 2 are not zero since the subtests are administered across clusters. These covariances are relatively small however and assumed zero for purposes of illustration. Since the items were chosen so that the average difficulty of items in each subtest would be approximately the same, we have assumed $a_i(t) = 0$, and $c_i(t)$ as the inverse of the number of items.

Table 1. Mean category scores (per item) for subtests (numbers of test items in parentheses)

		Category		
Test	B	G	M	Subsample size
1	0·536 (17)	0·549 (18)		273
2	0·496 (16)		0·428 (20)	263
3		0·560 (18)	0·392 (18)	268

Table 2. Upper triangle covariance matrix of means ($\times 10^3$)

$$\begin{bmatrix} 0{\cdot}233 & 0{\cdot}157 & 0 & 0 & 0 & 0 \\ & 0{\cdot}280 & 0 & 0 & 0 & 0 \\ & & 0{\cdot}314 & 0{\cdot}185 & 0 & 0 \\ & & & 0{\cdot}203 & 0 & 0 \\ & & & & 0{\cdot}396 & 0{\cdot}241 \\ & & & & & 0{\cdot}224 \end{bmatrix}$$

Table 3 gives the efficient generalized least squares estimates together with their standard errors and the equivalent estimates derived by forming weighted averages of the columns of Table 1, using the reciprocals of the variances as weights. The gain in efficiency of the generalized least squares estimates ranges from 18 to 37 per cent, indicating a worthwhile gain.

Table 3. Estimates and standard errors

Category	Generalized least squares (SE)	Simple (SE)	Efficiency
B	0·512 (0·0107)	0·519 (0·0116)	1·18
G	0·561 (0·0116)	0·553 (0·0128)	1·22
M	0·413 (0·0088)	0·411 (0·0103)	1·37

Note. Goodness-of-fit $\chi^2 = 19{\cdot}6$, d.f. = 3, $P = 0{\cdot}0002$.

A goodness-of-fit test, equivalent to testing jointly the equality of the means of each pair of item sets, is highly significant, thus providing evidence either for context effects, or for an incorrect choice of $a_i(t)$, $c_i(t)$. Of course, even with careful selection of subtest items one would not expect the $y_i(t)$ for a category to have exactly the same population means, so this goodness-of-fit test should not be taken too literally.

7. Discussion

We have shown how efficient estimates can be obtained, using minimal assumptions, from a multiple matrix design. In the limited example given in the paper the gains in efficiency are clearly worthwhile and preliminary results from applying the method to the full APU dataset indicate that gains of 30–40 per cent are readily obtainable.

An important question is that of sample design. Clearly, for a given total sample size, the way in which subtests are formed from combinations of elementary item sets should depend on the within-subtest covariances and estimates of these can be used to design efficient schemes. Work on this topic is currently being pursued and will be the subject of a forthcoming paper.

We have already suggested that care is needed in 'standardizing' the $x_i(t)$ using values of $a_i(t)$ and $c_i(t)$. In comparative studies, however, we can largely avoid this problem by incorporating in (8) adjustment terms representing the departure of each elementary item set mean from the category mean. If we make the assumption that these adjustments are the same for every subgroup, that is there is no interaction, then valid subgroup comparisons can be carried out. Of course, this can be done with the $x_i(t)$ directly, so avoiding the need to specify values for $a_i(t)$ and $c_i(t)$. In this case, however, the assumption of no interactions may be implausible since subgroup differences might be expected to depend on the number of items in a set and its overall difficulty. As well as incorporating adjustment terms in (8), therefore, it would seem preferable to carry out an initial standardization, using values from other sources or, as in the example above, from internal considerations, even if these are only approximations. In cases, such as our example, where items are designed to have the same average difficulty for each subtest, this is equivalent to assigning prior values of $c_i(t)$ and estimating the $a_i(t)$. In any case a test for interaction can always be carried out. If random sampling of items is assumed then the model (4) would contain a random component reflecting item sampling and the adjustment terms in (8) likewise would represent a random component, and suitable estimation procedures should be used.

There is one important case where this issue does not arise at all, namely where estimates are required separately for each elementary estimate, which could be a single item. This would arise for example when the elementary sets of items form cohesive curriculum elements of interest in their own right rather than as part of a larger category. This is clearly relevant to so called 'profile' reporting of educational attainments. Another situation is where we have a 'battery' of fixed tests. a subset of which is administered to any one subject. In all these cases we will still need to recognize possible context effects. although these will often be regarded as of interest for their own sake rather than merely nuisance factors.

As in our example. many if not most data arise from complex sampling schemes. Where there is stratification and clustering the covariance matrix will no longer be diagonal and it needs to be estimated taking into account the sample design. Estimation details for such samples. together with a program description are given. for example. by Hidiroglou (1981).

Acknowledgements

We wish to thank the Assessment of Performance Unit for permission to use the data given in the example.

References

Bhapkar. V. P. (1976). Generalized least squares estimation and testing. *The American Statistician*. **30**. 73 74.

Bock. R. D. & Aitkin. M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*. **46**. 443 458.

Foxman. D. D.. Ruddock. E. J.. Badger. M. E. & Martini. R. M. (1982). *Mathematical Development: Primary Survey Report No. 3*. Assessment of Performance Unit. London: HMSO.

Goldstein. H. (1980). Dimensionality. bias. independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*. **33**. 234 246.

Hidiroglou. M. A. (1981). Computerization of complex survey estimates. *Proceedings of the Statistical Computing Section of the American Statistical Association*. 1 7.

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

Lord. F. M. & Novick. M. (1968). *Statistical Theories of Mental Test Scores*. Reading. MA: Addison-Wesley.

Sirotnik. K. & Wellington. R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*. **14**. 343-399.

Received 7 September 1982; revised version received 8 November 1982

Requests for reprints should be addressed to Harvey Goldstein. Department of Mathematics. Statistics and Computing. Institute of Education. 20 Bedford Way. London WC1H 0AL. U.K.

Anthony N. James was at the National Foundation for Educational Research. Slough and is now an honorary associate of the Institute of Education.