R&D NOTITIES 2000–1

ESTIMATING VARIANCE COMPONENTS IN UNBALANCED DESIGNS

N D. Verhelst

Cito groep

ESTIMATING VARIANCE COMPONENTS IN UNBALANCED DESIGNS

N D. Verhelst

Cito
Arnhem, oktober 2000

1 Introduction

The algebra to derive unbiased estimates of variance components in balanced designs is easy and straightforward. However, once one comes to unbalanced designs, things suddenly get more complicated, and often recourse has to be made to quite sophisticated software to get reasonable results.

The purpose of this report is to derive some unbiased estimators of the variance components for a two-way table or a three-way table, with one observation per cell as is usual in generalizability theory, but where some of the observations are missing. It will be assumed throughout that missing observations are missing completely at random. Furthermore, it will be assumed that the data are collected to produce a complete design, so that it is reasonable to expect that the missing values are incidental and that their number is not exorbitantly high. In the simulation studies therefore, the highest percentage of empty cells will be set to 16%.

In the next section a rather detailed derivation will be given for a two-way table. The subsequent section will give the results for a three-way table. In Section 4, some examples will be given which compare the present estimators with the REML estimators provided by BMDP3V.

2 Two-way tables

The basic observations consist of a rectangular table with I rows and J columns, with in each cell a single observation denoted by Y_{ij}. The model for the observations is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{1}$$

where μ is a constant and the other terms are considered as random variables with finite variance and expectation zero, and which are all mutually independent. Notice that the residual ε_{ij} is the sum of an interaction term and an error, which are confounded since we only have one observation per cell.

If we have an observation in each cell, the estimation of the variance components is easy (see, e.g., Veldhuijzen, Goldebeld & Sanders, 1993). If for a number of cells in the table, the observation is missing, the design becomes unbalanced and the estimation is no longer simple. We will treat this case by adding a $I \times J$ matrix U (with elements u_{ij}), defined as

$$u_{ij} = \begin{cases} 1 & \text{if there is an observation in cel } (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

In case $u_{ij} = 0$, the value of Y_{ij} is arbitrary. In the complete case, all entries of U equal one.

We will use the following notation:

$$\begin{aligned} m_i &= \sum_j u_{ij}, \\ n_j &= \sum_i u_{ij} \\ N &= \sum_i m_i = \sum_j n_j \end{aligned}$$

If we denote averages, we will always refer to weighted averages where the weights are contained in the matrix U. For averages we will use the dot-notation:

$$\begin{aligned} Y_{i.} &= \frac{1}{m_i} \sum_j u_{ij} Y_{ij}, \\ Y_{.j} &= \frac{1}{n_j} \sum_i u_{ij} Y_{ij}, \\ Y_{..} &= \frac{1}{N} \sum_i \sum_j u_{ij} Y_{ij} = \frac{1}{N} \sum_i m_i Y_{i.} = \frac{1}{N} \sum_j n_j Y_{.j}. \end{aligned} \tag{2}$$

We can always write:

$$Y_{ij} - Y_{..} = (Y_{i.} - Y_{..}) + (Y_{.j} - Y_{..}) + (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}). \tag{3}$$

In the complete case, the total sum of squares can be written as a sum of three sums of squares:

$$SS_{tot} = SS_{row} + SS_{col} + SS_{res}, \tag{4}$$

but with the definitions (5) through (8) below, equality (4) is not valid any more in unbalanced designs. In the general case (using the weights u_{ij}), the four sums of squares are defined as

$$SS_{tot} = \sum_i \sum_j u_{ij} (Y_{ij} - Y_{..})^2, \tag{5}$$

$$SS_{row} = \sum_i m_i (Y_{i.} - Y_{..})^2, \tag{6}$$

$$SS_{col} = \sum_j n_j (Y_{.j} - Y_{..})^2, \tag{7}$$

$$SS_{res} = \sum_i \sum_j u_{ij} (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2. \tag{8}$$

and simple algebraic manipulations lead to the following equations:

$$SS_{tot} = \sum_i \sum_j u_{ij} Y_{ij}^2 - N Y_{..}^2, \tag{9}$$

$$SS_{row} = \sum_i m_i Y_{i.}^2 - N Y_{..}^2, \tag{10}$$

$$SS_{col} = \sum_j n_j Y_{.j}^2 - N Y_{..}^2, \tag{11}$$

$$SS_{res} = \sum_i \sum_j u_{ij} Y_{ij}^2 - \sum_i m_i Y_{i.}^2 - \sum_j n_j Y_{.j}^2 - N Y_{..}^2 + 2 \sum_i \sum_j u_{ij} Y_{i.} Y_{.j} \tag{12}$$

Notice that (12) defies simplification in the general case of an unbalanced design.

Using p to denote one of the elements of $\{tot,\ row,\ col,\ res\}$, and using model (1), it is easily shown that

$$E(SS_p) = a_p \sigma_\alpha^2 + b_p \sigma_\beta^2 + c_p \sigma_\varepsilon^2 \tag{13}$$

Equating any three of the four empirical SS to their expected value as given in (13) yields a system of three linear equations, the solution of which is an unbiased estimator of the three variances. Therefore the problems which are to be solved are:

1. Find the coefficients a_p, b_p and c_p ($p \in \{tot,\ row,\ col,\ res\}$);
2. If one forms a system of four equations in three unknowns, either this system will be consistent or it will not be consistent. If the system is consistent, any three equations can be used to solve for the parameters, and the results will be unique. If the system is not consistent, there exist four different estimators, and the question arises which one to choose.

These problems will be treated in the next two subsections.

2.1 The coefficients a_p, b_p and c_p

A straightforward way to derive expected sums of squares is to expand the squares in (5) through (8) and to take expectations of each term. We demonstrate the technique by deriving the expected value of $\sum_i \sum_j u_{ij} Y_{ij}^2$. First, notice that (1) gives

$$Y_{ij}^2 = \alpha_i^2 + \beta_j^2 + \varepsilon_{ij}^2 + E_0 \tag{14}$$

where E_0 is a generic symbol to denote a sum of terms where each term is a constant or contains a factor which is a product of non-identical random

variables, which, due to the assumption of independence, has expectation zero. Next we multiply by u_{ij} and take sums :

$$\sum_i \sum_j u_{ij} Y_{ij}^2 = \sum_i m_i \alpha_i^2 + \sum_j n_j \beta_j^2 + \sum_i \sum_j u_{ij} \varepsilon_{ij}^2 + E_0. \tag{15}$$

Since the variables u_{ij}, or their sums m_i and n_j, are independent of the variables α, β and ε, it holds, for example, that

$$E[\sum_i m_i \alpha_i^2] = E[\sum_i m_i] \times E[\sum_i \alpha_i^2] \tag{16}$$

At present we will leave the expected value of the u_{ij}-variables and their sums as they are (without mentioning explicitly the expected value operator), and since $E(\alpha_i^2) = \sigma_\alpha^2$, we have that

$$E(\sum_i \sum_j u_{ij} Y_{ij}^2) = N\sigma_\alpha^2 + N\sigma_\beta^2 + N\sigma_\varepsilon^2. \tag{17}$$

Following the same way of reasoning we find, after some tedious algebra, that

$$E(\sum_i m_i Y_{i.}^2) = N\sigma_\alpha^2 + I\sigma_\beta^2 + I\sigma_\varepsilon^2 \tag{18}$$

$$E(\sum_j n_j Y_{.j}^2) = J\sigma_\alpha^2 + N\sigma_\beta^2 + J\sigma_\varepsilon^2, \tag{19}$$

$$E(NY_{..}^2) = \sigma_\alpha^2 \frac{\sum_i m_i^2}{N} + \sigma_\beta^2 \frac{\sum_j n_j^2}{N} + \sigma_\varepsilon^2, \tag{20}$$

and

$$E(\sum_i \sum_j u_{ij} Y_{i.} Y_{.j}) = J\sigma_\alpha^2 + I\sigma_\beta^2 + \sigma_\varepsilon^2 \sum_i \sum_j \frac{u_{ij}}{m_i n_j}. \tag{21}$$

Notice that the last term in (21) defies simplification.

Using (17) through (21), we find that

$$.E(SS_{row}) = \sigma_\alpha^2 \left(N - \frac{\sum_i m_i^2}{N} \right) + \sigma_\beta^2 \left(I - \frac{\sum_j n_j^2}{N} \right) + \sigma_\varepsilon^2 (I-1), \tag{22}$$

$$E(SS_{col}) = \sigma_\alpha^2 \left(J - \frac{\sum_i m_i^2}{N} \right) + \sigma_\beta^2 \left(N - \frac{\sum_j n_j^2}{N} \right) + \sigma_\varepsilon^2 (J-1), \tag{23}$$

$$E(SS_{res}) = \sigma_\alpha^2\left(J - \frac{\sum_i m_i^2}{N}\right) + \sigma_\beta^2\left(I - \frac{\sum_j n_j^2}{N}\right) + \sigma_\varepsilon^2\left(N - I - J - 1 + 2\sum_i\sum_j \frac{u_{ij}}{m_i n_j}\right), \quad (24)$$

$$E(SS_{tot}) = \sigma_\alpha^2\left(N - \frac{\sum_i m_i^2}{N}\right) + \sigma_\beta^2\left(N - \frac{\sum_j n_j^2}{N}\right) + \sigma_\varepsilon^2(N-1) \quad (25)$$

Notice that in a complete design, it holds that

$$m_i = J \text{ and } n_j = I$$

whence it is easy to check that the formulae (22) through (25) also hold for the complete case.

We will end this section with some remarks on the conditions where the foregoing results are valid. The key assumption is already given by an example in equation (16), but generally it means that the model parameters (α, β and ε) and the design variables u must be independently distributed. This allows us to write, for example,

$$E(\alpha_i^2|u_{ij} = 1) = E(\alpha_i^2|u_{ij} = 0) = \sigma_\alpha^2.$$

From this independence, we can quite accurately deduce when the results are valid in any practical situation. We give three examples. In the first example some observations are missing incidentally, but the design was planned as a complete design. If it is reasonable to assume that the mechanism that caused the not observed cells is totally unrelated to the model parameters, the foregoing results are valid. Notice that if the observations would be repeated, the value of the design variables u probably will not be identical across replications. The second example is a planned incomplete design, where two booklets of items (with some or no overlap) are administered to two groups of students. If the items are assigned randomly to the booklets and the students are assigned randomly to the groups, the results are valid. If on the contrary - and this is the third example- items are assigned to the booklets on the basis of difficulty (making an easy and a hard booklet) or students are allocated to the groups on the basis of some estimate of their ability, the results are no longer valid. This is easy to understand. Suppose the booklets have no overlap. The β_j-parameter in (1) is an index for the easiness of item j. If one constructs rather homogeneous booklets, the estimate $\widehat{\sigma}_\beta^2$ will estimate the within booklet variance which will understimate the variance of the itemparameters.

A related topic concerns the precise meaning of the expected value operator $E(.)$ in (22) through (25). Since in the right hand sides of these equations the variables u appear as they have been observed in the sample, all the expected values are only taken with respect to the model parameters and conditional

on the design variables as observed. With a planned incomplete design the distribution of the design variables is degenerate, i.e. the u have no variance by definition. With incidental missings, the scope of the research can be different. The narrow scope tries to find the variance of the effects and the residual given that the design variables take values as in a particular case, what is usually what one finds after the data collection. The broad scope supposes a distribution of the u-variables, and the observed value of these variables in a particular case is considered as a random draw from this distribution. For example, one might ask for the expected sum of squares if for each cell (independently) the probability of not having an observation equals some number π. For such an approach the formulae (22) through (25) are not appropriate, but should be replaced by similar formulae where all functions of the u variables (like $\sum_i \sum_j u_{ij}/(m_i n_j)$ in (25)) are replaced by their expected values in the distribution of u. Such an approach allows one to generalize over a whole population of design matrices, but the expected values of the design variables presuppose a model, and their computation will in general not be easy.

2.2 Choice of the estimator

The results of the derivations are given by the equations (22) through (25). Equating these expected values with the observed sums of squares yields a system of four linear equations in three unknowns. A simple example, with an arbitrary matrix of observations and a few missing values learns that this system is in general not consistent, i.e., there do not exist in general three variance components such that the observed sums of squares all equal their expected values.

Of course one could use any three of the aforementioned equations and equate them to the corresponding observed sums of squares, and the (unique) solution is a consistent and unbiased estimator of the variance components. (Of course, there may exist designs where there is no unique solution, because the matrix of the equation system is singular, but such a case is not considered further, because it is not to be expected to occur in practical cases where the number of missing values is low in comparison to the number of cells.)

To get an idea on the possible differences, two small simulation studies were set up with $I = 40$ and $J = 15$. The first study was concerned with continuous variables, while the second study used binary responses.

2.2.1 Study 1

The true variance components are displayed in Table 1.

Table 1: true variances in study 1

σ^2_α	σ^2_β	σ^2_ϵ
1	0.49	3

Considering the row elements as persons and the column elements as test items, the values in Table 1 correspond to a test of 15 items with a Cronbach's alpha of $5/6 = 0.83$, which is quite realistic. To construct an incomplete table, we proceeded in two steps. First for each row i a row effect was drawn from a normal distribution with mean zero and variance σ_α^2, and similarly for each column j a column effect β_j was drawn from $N(0, \sigma_\beta^2)$. For each cell (i, j) of the table the observation was defined as $\alpha_i + \beta_j + \varepsilon_{ij}$, with ε_{ij} independently drawn from $N(0, \sigma_\varepsilon^2)$. Second, for each cell independently a biased coin with success probability π was tossed, and on success the cell was designated as non-observed.

In the study, the value of π took 5 different values: 0, 0.02, 0.04, 0.08 and 0.16, where each value represents a condition. Within each condition 1000 tables were created and analyzed. For each table, the variance components were estimated in four different ways, by dropping each time one of the equations (22) through (25). In each condition the average estimate and the standard deviation of the estimates was computed across replications. The results can be summarized as follows:

- For the condition $\pi = 0$, the four estimation procedures give identical results (because of (4)). Means and standard deviations across 1000 replications are given in Table 2.

Table 2. Results for the complete condition in study 1

	σ_α^2	σ_β^2	σ_ε^2
mean	1.001	0.492	2.993
SD	0.272	0.216	0.183

The row SD can be considered as an estimate of the standard errors.

- Within each condition with incomplete data the mean and the SD across estimation methods did show a very slight variation in the order of magnitude of 0.001, which may be considered negligible when compared to the standard error of the estimates.

- Within each condition the correlations (across replications) between the estimates for each variance component were very high. The smallest correlation found was 0.987

- The means of the estimates in the incomplete conditions were as close to the true values as in the complete case, showing clearly that the estimators are unbiased.

- The SD for the three estimators and the five conditions are shown in Table 3.Although there is certainly a tendency for the SDs to increase with increasing values of π, the main result is that the increase is very small: even in the case with 16% of empty cells, the standard error is only slightly larger than in the complete case.

Table 3. SD across conditions in study 1

π	σ^2_α	σ^2_β	σ^2_ε
.00	0.272	0.216	0.183
.02	0.271	0.207	0.183
.04	0.275	0.221	0.178
.08	0.278	0.224	0.191
.16	0.286	0.221	0.203

In summary we can choose freely one of the four methods of estimation, since they all give unbiased estimates with the same accuracy, and the estimates themselves correlate very highly. Moreover, a very comforting finding is that the accuracy of the estimators decreases only very slowly with increasing number of empty cells.

2.2.2 Study 2

The setup of this study is similar to that of study 1. The only difference is the definition of the response variable. To make things clear, the two definitions are displayed together:

$$\text{study 1:} \qquad Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij},$$

$$\text{study 2:} \quad Y_{ij} = \begin{cases} 1 \text{ if } \alpha_i + \varepsilon_{ij} > \beta_j, \\ 0 \text{ otherwise.} \end{cases}$$

Of course, by using this discretization, it is not straightforward to predict the true values of the variance components. But for comparisons across conditions this is not very important, since we know that all estimators used are unbiased.

The results are summarized as follows:

- For each of the 1000 replications with complete tables, Cronbach's alpha was computed. The average alpha is 0.72; the standard deviation was 0.07.
- Means and standard deviations across 1000 replications for the complete tables are given in Table 4.

Table 4. Results ($\times 10$) for the complete condition in study 2

	σ^2_α	σ^2_β	σ^2_ε
mean	0.358	0.174	1.967
SD	0.094	0.077	0.102

- As in study 1, there was almost no variation across estimation methods in means and standard deviations, and the correlations between estimates from different estimation methods were very close to 1 (lowest value: 0.987).

- The SDs for the three estimators and the five conditions are shown in Table 5. As in study 1, there is an increase in the standard errors with increasing number of missing observations, but again, the increase is very slow

Table 5. SD ($\times 10$) across conditions in study 2

π	σ^2_α	σ^2_β	σ^2_ε
.00	0.094	0.077	0.102
.02	0.096	0.079	0.106
.04	0.098	0.078	0.105
.08	0.103	0.081	0.106
.16	0.106	0.080	0.115

In summary we can draw essentially the same conclusions from both studies: the estimation method does not matter very much. All four estimation methods yield unbiased, very highly correlated estimates which have much the same accuracy.

3 Three-way tables

3.1 The results

For three-way tables, with I rows, J columns and K layers, a similar approach as in the two-way case can be used. The model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijk} \qquad (26)$$

where ε_{ijk} is the sum of the highest-order interaction and the measurement error. There are seven variance components: three for the main effects, three for the first order interactions and one for the residual.

To derive the expected values of the sums of squares, we need a slightly more complicated notation than with the two-way design. The design variables u now have three indices, and we need univariate as well as bivariate totals. Therefore we define

$$m_i = \sum_j \sum_k u_{ijk}, \quad n_j = \sum_i \sum_k u_{ijk}, \quad s_k = \sum_i \sum_j u_{ijk}, \qquad (27)$$

and

$$(mn)_{ij} = \sum_k u_{ijk}, \quad (ms)_{ik} = \sum_j u_{ijk}, \quad (ns)_{jk} = \sum_i u_{ijk}. \qquad (28)$$

As before N denotes the total number of observations.

In the case of a three-way table, there are seven variance components, while there are eight sums of squares, with the result that any seven sums of squares

can be equated to their expected values. In the simulation studies with two-way tables it was found, however, that the different estimation equations gave very similar results. With three-way tables, the derivation of the expected residual sum of squares gives rise to quite cumbersome expressions, which will be omitted from the present report. In analogy with the equations (9) through (12), it is easily shown that

$$SS_{tot} = \sum_i \sum_j \sum_k u_{ijk} Y_{ijk}^2 - NY_{...}^2, \tag{29}$$

$$SS_{row} = \sum_i m_i Y_{i..}^2 - NY_{...}^2, \tag{30}$$

$$SS_{col} = \sum_j n_j Y_{.j.}^2 - NY_{...}^2, \tag{31}$$

$$SS_{row,col} = \sum_i \sum_j (mn)_{ij} Y_{ij.}^2 - \sum_i m_i Y_{i..}^2 - \sum_j n_j Y_{.j.}^2 - NY_{...}^2 + 2 \sum_i \sum_j (mn)_{ij} Y_{i..} Y_{.j.} \tag{32}$$

where it is clear that the expressions involving layers can be constructed easily by appropriate changes of subscripts. To derive the expected values, it showed useful to introduce the following generic quantities:

$$A_\alpha = \frac{\sum_i m_i^2}{N}, \tag{33}$$

$$B_{\alpha\beta} = \frac{\sum_i \sum_j (mn)_{ij}^2}{N}, \tag{34}$$

$$C_{\alpha\beta} = \sum_i \frac{\sum_j (mn)_{ij}^2}{m_i}, \tag{35}$$

$$D_{\alpha\beta} = \sum_i \sum_j \frac{(mn)_{ij}^2}{m_i n_j}, \tag{36}$$

$$E_{\alpha\beta} = \sum_i \sum_j \frac{(mn)_{ij}^3}{m_i n_j}, \tag{37}$$

$$F_{\alpha\beta} = \sum_i \sum_j \frac{(mn)_{ij}}{m_i n_j} \sum_k (ms)_{ik} (ns)_{jk}, \tag{38}$$

and

$$G_{\alpha\beta} = \sum_i \sum_j \frac{(mn)_{ij}}{m_i n_j} \sum_k (ms)_{ik} u_{ijk}. \tag{39}$$

The above definitions are generic in the following sense: the subscript α in the left hand side corresponds to the subscript i and the frequencies m in the right hand side; the subscript β corresponds to j and n, while the subscript k and the frequency s do not have a corresponding subscript in the left hand symbol, because they refer to the remaining dimension of the table. From this convention, it should be clear that, for example,

$$A_\gamma = \frac{\sum_k s_k^2}{N} \text{ and } G_{\beta\gamma} = \sum_j \sum_k \frac{(ns)_{jk}}{n_j s_k} \sum_i (mn)_{ij} u_{ijk}$$

Notice, furthermore, that some of the above quantities are symmetric, and others are not. For example, it is easy to check that

$$B_{\alpha\beta} = B_{\beta\alpha},$$

and that the same relation holds for $D_{\alpha\beta}$, $E_{\alpha\beta}$ and $F_{\alpha\beta}$, but not for $C_{\alpha\beta}$ and $G_{\alpha\beta}$.

Using the equations (29) to (32) and the generic definitions (33) through (39), the following table of coefficients can be constructed (Table 6). The value in a cell is the coefficient of the column variance component in the expected value of the row.

Table 6. Coefficients of variance components of 11 expected values

Expected value	σ^2_α	σ^2_β	σ^2_γ	$\sigma^2_{\alpha\beta}$	$\sigma^2_{\alpha\gamma}$	$\sigma^2_{\beta\gamma}$	σ^2_ϵ
$E\left(\sum_i \sum_j \sum_k u_{ijk} Y^2_{ijk}\right)$	N	N	N	N	N	N	N
$E\left(NY^2_{...}\right)$	A_α	A_β	A_γ	$B_{\alpha\beta}$	$B_{\alpha\gamma}$	$B_{\beta\gamma}$	1
$E\left(\sum_i m_i Y^2_{i..}\right)$	N	$C_{\alpha\beta}$	$C_{\alpha\gamma}$	$C_{\alpha\beta}$	$C_{\alpha\gamma}$	I	I
$E\left(\sum_j n_j Y^2_{.j.}\right)$	$C_{\beta\alpha}$	N	$C_{\beta\gamma}$	$C_{\beta\alpha}$	J	$C_{\beta\gamma}$	J
$E\left(\sum_k s_k Y^2_{..k}\right)$	$C_{\gamma\alpha}$	$C_{\gamma\beta}$	N	K	$C_{\gamma\alpha}$	$C_{\gamma\beta}$	K
$E\left(\sum_i \sum_j (mn)_{ij} Y^2_{ij.}\right)$	N	N	IJ	N	IJ	IJ	IJ
$E\left(\sum_i \sum_k (ms)_{ik} Y^2_{i.k}\right)$	N	IK	N	IK	N	IK	IK
$E\left(\sum_j \sum_k (ns)_{jk} Y^2_{.jk}\right)$	JK	N	N	JK	JK	N	JK
$E\left(\sum_i \sum_j (mn)_{ij} Y_{i..} Y_{.j.}\right)$	$C_{\beta\alpha}$	$C_{\alpha\beta}$	$F_{\alpha\beta}$	$E_{\alpha\beta}$	$G_{\alpha\beta}$	$G_{\beta\alpha}$	$D_{\alpha\beta}$
$E\left(\sum_i \sum_k (ms)_{ik} Y_{i..} Y_{..k}\right)$	$C_{\gamma\alpha}$	$F_{\alpha\gamma}$	$C_{\alpha\gamma}$	$G_{\alpha\gamma}$	$E_{\alpha\gamma}$	$G_{\gamma\alpha}$	$D_{\alpha\gamma}$
$E\left(\sum_j \sum_k (ns)_{jk} Y_{.j.} Y_{..k}\right)$	$F_{\beta\gamma}$	$C_{\gamma\beta}$	$C_{\beta\gamma}$	$G_{\beta\gamma}$	$G_{\gamma\beta}$	$E_{\beta\gamma}$	$D_{\beta\gamma}$

Table 6 is the main result, because the expected values of the seven sums of squares needed (total, three main effects and three first order interactions), are easily found as linear combinations of the rows of Table 6.

If the design is complete, Table 6 reduces to Table 7

Table 7. Coefficients of variance components of 11 expected values (complete design)

Expected value	σ^2_α	σ^2_β	σ^2_γ	$\sigma^2_{\alpha\beta}$	$\sigma^2_{\alpha\gamma}$	$\sigma^2_{\beta\gamma}$	σ^2_ε
$E\left(\sum_i \sum_j \sum_k u_{ijk} Y^2_{ijk}\right)$	N	N	N	N	N	N	N
$E\left(NY^2_{...}\right)$	JK	IK	IJ	K	J	I	1
$E\left(\sum_i m_i Y^2_{i..}\right)$	N	IK	IJ	IK	IJ	I	I
$E\left(\sum_j n_j Y^2_{.j.}\right)$	JK	N	IJ	JK	J	IJ	J
$E\left(\sum_k s_k Y^2_{..k}\right)$	JK	IK	N	K	JK	IK	K
$E\left(\sum_i \sum_j (mn)_{ij} Y^2_{ij.}\right)$	N	N	IJ	N	IJ	IJ	IJ
$E\left(\sum_i \sum_k (ms)_{ik} Y^2_{i.k}\right)$	N	IK	N	IK	N	IK	IK
$E\left(\sum_j \sum_k (ns)_{jk} Y^2_{.jk}\right)$	JK	N	N	JK	JK	N	JK
$E\left(\sum_i \sum_j (mn)_{ij} Y_{i..} Y_{.j.}\right)$	JK	IK	IJ	K	J	I	1
$E\left(\sum_i \sum_k (ms)_{ik} Y_{i..} Y_{..k}\right)$	JK	IK	IJ	K	J	I	1
$E\left(\sum_j \sum_k (ns)_{jk} Y_{.j.} Y_{..k}\right)$	JK	IK	IJ	K	J	I	1

3.2 Study 3

Analogously to study 1, a simulation study was carried out with continuous response variables in a three way design. In Table 8 the true values of the variance components are displayed.

Table 8: true variances in study 3

σ^2_α	σ^2_β	σ^2_γ	$\sigma^2_{\alpha\beta}$	$\sigma^2_{\alpha\gamma}$	$\sigma^2_{\beta\gamma}$	σ^2_ε
1	0.49	0.36	0.25	0.16	0.09	3

As in studies 1 and 2, there are five conditions with a probability of non-observation of 0%, 2%, 4%, 8% and 16% respectively, and in each condition 4000 data sets with $I = 40$ rows, $J = 15$ columns and $K = 10$ layers were constructed from which the variance components are estimated using the formulae derived in the preceding section. Notice that only one estimation procedure was applied since the expected value of the residual sum of squares was not derived.

The results are not substantially different from the ones in study 1.

- The average estimates and the standard deviations for the complete design ($\pi = 0$) are displayed in Table 9. Notice that the estimates of the interaction components have a much smaller standard error than the estimates of

the components of the main effects, which demonstrates the paradoxical situation that the components which are usually of the greatest interest, the main effects, are more difficult to estimate than the components of lesser interest, the interactions. Searle (1971, pp. 415-417) gives expressions for the standard errors of the variance component estimates. These values are computed for the present study and given in the row labeled SE. If the effects are normally distributed, the expressions given by Searle are exact, and as can be seen from Table 9, the empirical SDs (computed on a sample of 4000 tables) are indeed in very close correspondence to the theoretical values.

Table 9: summary of results for the complete case in study 3

	σ^2_α	σ^2_β	σ^2_γ	$\sigma^2_{\alpha\beta}$	$\sigma^2_{\alpha\gamma}$	$\sigma^2_{\beta\gamma}$	σ^2_ϵ
mean	.9955	.4895	.3602	.2498	.1607	.0898	3.0007
SD	.2399	.1940	.1806	.0341	.0272	.0206	.0608
SE	.2384	.1938	.1768	.0338	.0275	.0208	.0605

- For the other conditions, the mean estimates were virtually equal to the means in the complete conditions, and the standard deviations showed a small increase with increasing percentage of empty cells. The tables are very similar to the Tables 3 and 4, and will not be reproduced here.

In summary we can say that the results of study 3 are much the same as the results of study 1: the moment estimator is unbiased, and its standard error is of the same order as the standard error in the complete case.

3.3 Study 4

Although the method presented above is easy to use in practical settings, two theoretical issues should be discussed. The first concerns the efficiency of the method, and the second has to do with the estimation of the standard errors in the case of highly discrete observations. We comment on these problems in turn.

As can be seen from the simulation studies, and especially from study 3, the standard errors (as estimated from an empirical distribution) are quite large, especially for the main effects. Of course, the number of observations in a single table in the studies is not overwhelmingly large, but in the context of generalizability theory, with one facet being items and the other raters, the numbers used in the simulation studies will in most contexts be larger than is feasible in any practical application. Therefore, it might be useful to try other estimation methods which may be more efficient than the proposed moment method. Serious candidates of course are estimation methods which are based on maximizing some likelihood function. A standard statistical package like BMDP (1992) allows for estimation with maximum likelihood (ML) and restricted ML

(REML) of variance components in unbalanced designs. Therefore, we planned a study to compare our estimation method with these two ML-methods.

But this brings us immediately to the second problem. To use ML one needs a model for the distribution of the observations. In the standard packages this model is always the normal distribution, i.e. all effects are normally distributed, but in many applications the observations are discrete, or even binary, and it is largely unknown how the use of a model for continuous observations will perform with discrete data.

To shed light on these two problems, 1000 40×15 tables were generated with binary data and with a missing probability of 16%, i.e., this corresponds to the fifth condition of study 2: the same values of the variance components were used, and the same method of dichotomization. Each one of the 1000 tables were used to estimate the variance components of the two main effects and the residual with three estimation methods: the moment method proposed in this report, the ML and the REML estimation method implemented in BMDP3V. (Many thanks to Niels Veldhuijzen for running 2000 BMDP jobs, and for collecting the six numbers I needed from each the 2000 abundant BMDP output files.) The results of the study can be summarized as follows:

- For each of the three variance components, the correlations were computed between the three estimation methods. The minimum correlation found was 0.994, which shows that the three methods give virtually the same estimates, a part from possibly a scale factor or a shift.

- In Table 10 the average estimates are given. The moment method and REML give the same results, while the ML method gives somewhat lower values for the main effects, possibly pointing to a slight bias, which is not important for practical purposes.

Table 10. mean estimates ($\times 10$) in study 4

	σ^2_α	σ^2_β	σ^2_ε
moment	.3532	.1761	1.9717
REML	.3533	.1761	1.9718
ML	.3459	.1655	1.9726

- Some results concerning standard errors are displayed in Table 11. The standard deviations from the moment method were equal to the ones from the REML procedure. It can be seen from the table that the ML-method is more efficient than the REML and the moment method in estimating the main effects. Of course, when estimating variance components in practice, there is little possibility of doing replications with one (incomplete) data set. (It is even not clear how one could apply the bootstrap method.). But an estimate of the (asymptotic) standard error can be derived on theoretical grounds, and the program BMDP3V gives such an estimate for ML as well as for REML. In the present study we took the square root of the mean squared (estimated) standard errors as an estimate of

the asymptotic standard error. These values are displayed in Table 11 in the two rows labeled SE. Although asymptotic standard errors are usually smaller than the real standard errors in finite samples, here we see the opposite relation: the theoretical standard errors (SE) are systematically larger than the empirical ones (estimated by SD), and the reason for this is undoubtedly the fact that normal theory has been applied to binary data. From the table it is seen that using normal theory overestimates the error variances with about 30% (see the rows labeled SE^2/SD^2).

Table 11. SD and SE of the estimates ($\times 10$) in study 4

	σ^2_α	σ^2_β	σ^2_ε
REML-SD	.1045	.0832	.1149
REML-SE	.1185	.0943	.1317
SE^2/SD^2	1.29	1.28	1.31
ML-SD	.1022	.0782	.1152
ML-SE	.1162	.0884	.1375
SE^2/SD^2	1.29	1.28	1.42

4 Conclusion

In the present report formulae have been derived to estimate variance components in a two-way design and a three-way design with (when complete) one observation per cell, and where observations can be missing. It is assumed that the distribution of the missing values is independent of the distribution of the model parameters.

When a design is incomplete, it becomes unbalanced, and in unbalanced designs the property that the total sum of squares can nicely be decomposed as a sum of three (in a two way design) or seven (in a three-way design) sums of squares is lost. This is clearly demonstrated in Table 7, where the last three rows are equal to the second row in a complete design but this property is lost in an unbalanced design. Because of this, the moment estimators are no longer uniquely defined. In fact, Table 6 contains 11 rows, and any seven (linearly independent) rows can be used to estimate the seven variance components, and any choice gives an unbiased estimator, although the standard errors may be quite different. And the expected values used in Table 6 are by no means exhaustive. We could, for example, have derived also the expected value of $\sum_i \sum_j \sum_k u_{ijk} Y_{ij.} Y_{i.k}$ which is also a linear combination of the seven variance components.

The choice of particular linear combinations was inspired by two considerations. First, it was the purpose to choose the same sums of squares which are used in the classical analyses with balanced designs, and second, the choice was partially inspired by the wish to avoid lots of tedious algebraic derivations. In particular, the expected value of the residual SS in the three-way design was not derived.

In the first two simulation studies, it was investigated whether any choice of three SS from the set $\{SS_{row}, SS_{col}, SS_{res}, SS_{tot}\}$ would have an important influence on the resulting estimates. It appeared that this was not the case. The four methods gave virtually the same means and standard deviations across 1000 replications and the correlations between the estimates under the four estimation methods were all very close to one.

This led us to the conclusion not to investigate the effect of different estimation methods in the case of a three-way design, but to use only the total SS, the SS of the main effects and the SS of the first order interactions. The main result is a table of coefficients (Table 6) of eight sums of squares and three sums of cross products.

In the first three simulation studies it is clear that the estimators used are unbiased, and that the empirical standard errors increase with increasing number of missing cells (as should be expected.) It is, however, not easy to describe the pattern in the increase (see the Tables 3 and 5), because the increase is not monotonic in all cases, suggesting that 1000 replications is not enough to have stable estimates of the standard deviations.

The problem of the standard errors is quite complicated, as appeared from study 4 where incomplete two-way tables of binary observations with (on the average) 16% missing observations were analyzed with the moment method, the REML and the ML method. The finding that the estimates of the variance components correlate extremely high is comforting, but the estimation of the standard errors is problematic. The estimates of the asymptotic error variances for ML and REML appear to be about 30% larger than the empirical estimates, an effect which is ascribed to the fact that normal theory is applied to binary data. For the moment method this causes a kind of a dilemma. The results derived by Searle also use normal theory, especially the fact that, if $X \sim N(0, \sigma^2)$, then $E(X^3) = 0$ and $E(X^4) = 3\sigma^4$. Moreover, the basic model (1) cannot be true if the observed variables are binary and the effect variables are continuous, and at the same time independent. Therefore we can expect that the formulae developed by Searle will also be in error when applied to binary data. To check this we computed the standard errors using the formulae of Searle on the complete data of study 2 (see Table 4). The average parameter estimate was used as the value for the variance components, and the result of Searle's formulae, labeled SE were compared to the empirical SD, giving a ratio SE^2/SD^2 of 1.40, 1.20 and 1.35 for the row, column and residual components respectively. These results are comparable with the same ratios given by both ML procedures (see Table 11).

In view of these results, it seems not to be fruitful to adapt Searle's formulae to unbalanced designs (which is easy in principle but involves a lot of tedious algebra) since they also will probably be grossly in error when applied to binary data. In any case, the development of these formulae is beyond the scope of the present report.

5 References

BMDP (1992). *BMDP Statistical Software*. Los Angeles.

Searle, S.R. (1971). *Linear Models*. New York: Wiley

Veldhuijzen, N.H., Goldebeld, P. & Sanders, P.F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. In T.J.H.M. Eggen & P.F. Sanders, *Psychometrie in de praktijk* (pp. 33-82). Arnhem: Cito.