Rank Reducing Observation Sets

Timothy D. Knutson

Department of Mathematical Sciences
Montana State University

May 7, 2010

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Timothy D. Knutson

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5 May, 2011	Jim Robinson-Cox Writing Project Advisor
5 May, 2011	Steve Cherry Writing Projects Coordinator

1. Introduction

In an experiment, the design is the basis of everything that comes after. Consideration of questions like "what are the factors and levels of interest?" and "what comparisons address the objectives of this study?" help identify the design that is best suited for the goals of the study. For different circumstances, one experimental design can have advantages over competing designs. In some cases there are restrictions that are placed on a design. For example, if not all of the experimental runs can be done in a single day, then, by necessity, they must be run on multiple days. Or, it may be that there is not enough raw material in a single batch from a supplier to run the entire set of the experimental treatments. Situations like these and others lead to a need for blocking. Blocking is one way to deal with the variation in the data collected due to the differences between operators, batches of raw material, or day to day conditions that are uncontrollable.

Separating the experimental runs into blocks has the potential to cause problems. If all of the observations for one treatment formed a single block, then the effect due to the treatment would be indistinguishable from the effect of the block. This is easily avoided by a proper design taking the effect of blocking into account. If, however, there are observations that are missing after the experiment is completed, then we could end up in a situation very similar to that of a poorly designed experiment.

Missing data should always be avoided. If, however, there is a reason that observations are missing, we may be missing information that is useful to our study. For example, an experiment on the effects of a fertilizer may be affected by plants dying in the process. If the plants are growing so tall that they break because they needed support, we would not be able measure the effects that the fertilizer had on the yield of the plant. Or, maybe the plants on the boundary of the plot grow very tall because they get plenty of sun and water while the plants on the interior do not get enough sun to grow to maturity. Missing information like this is very case specific and not something that can be addressed simply by looking at the design.

Not all missing data have an assignable cause. Sometimes observations are lost due to random

chance. Some of the seeds in a batch may not germinate. An operator may drop a final product before it's measured. A miss-calibrated instrument could give a bad reading in a destructive test. These missing observations give either little or no information concerning the questions of interest, and they are often discarded from the data analysis. They may, however, cause major problems in the subsequent analysis of the experiment.

By removing these observations we are removing a row from the model matrix X corresponding to our design. By removing this row, we may change the column rank leading to a change in the set of functions having estimable parameters. If the functions that are now non-estimable were never of interest, then we have lost little. In the case of a blocked design, it is common for pairwise comparisons of treatment effects to become non-estimable which is problematic. Standard null and alternative hypotheses regarding treatment effects are frequently tested by multiple pairwise comparison methods. If one or more pairwise comparison is no longer estimable, then we cannot use this standard hypothesis testing approach.

Finding the rank reducing observation sets (RROSs) could help us decide which design to implement. If one design has fewer RROSs, then that design would be less likely to have a problem with a change in the estimability of functions vital to the study. It is also possible to examine the resulting design's estimable functions and be prepared for the analysis of the resulting design if RROS occur.

2. Estimability

This paper will use the notation of Godolphin (2004). Let the equation

$$E[QY] = \mu 1_n + X \begin{bmatrix} \tau \\ \beta \end{bmatrix} \tag{1}$$

define our linear model, where Y is an $n \times 1$ vector of the observations of interest, Q is a permutation matrix, 1_n is an n length vector of ones, and τ and β are vectors whose lengths correspond to the

number of treatments, say v and u. In this situation, μ is the overall mean for the model and X is the mean-free model matrix with a rank of $r < k$, where $k = u + v$. Let the τ vector correspond to the different treatments and their corresponding levels, while the β vector corresponds to the blocking effects and their levels.

When using a likelihood ratio test to determine significance of an effect (a common practice in many analyses), the null hypothesis can be written as

$$\Lambda \begin{bmatrix} \tau \\ \beta \end{bmatrix} = m \tag{2}$$

where Λ is a matrix whose rows correspond to linearly estimable functions and m is a vector of null values, often all zeroes. For example, if τ consisted of only one treatment, α, one may wish to test that all of the α_i treatments have the same effect. The Λ matrix would then be of the form

$$\Lambda = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & \cdots \\ 1 & 0 & -1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots \\ 1 & 0 & 0 & \cdots & -1 & 0 & \cdots \end{bmatrix}$$

would correspond to linear combinations of the form

$$\begin{bmatrix} \alpha_1 - \alpha_2 \\ \alpha_1 - \alpha_3 \\ \vdots \\ \alpha_1 - \alpha_v \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_v \end{bmatrix} = \vec{0}$$

The problem with RROSs is that they can reduce the number of estimable pairwise comparisons. If any of these pairwise comparisons were non-estimable then this test would not be viable.

To be estimable, a function must be able to be written as a linear combination of the rows of X. Another way to define an estimable function is that if $\lambda^T\beta$ is an estimable function, then there exists a linear unbiased estimator of $\lambda^T\beta$. That is, if $E[Y] = X\beta$, and there exist a vector a such that

$$E[a^TY] = a^TX\beta = \lambda^T\beta$$

then $\lambda^T\beta$ is estimable. In such a case, $\lambda = X^Ta$.

This relationship between X and λ shows that if the rank of X decreases, then the number of estimable functions would decrease.

3. Definition of Rank Reducing Observation Sets

Returning to our original linear model, let us assume that it is fully connected (i.e., all pairwise comparisons are estimable) in its current state. Suppose that running the experiment results in t missing observations. Let $Y_\#$ be the $(n-t) \times 1$ response vector of non-missing observations. All corresponding matrices and vectors will be labeled similarly to reflect the changes in their rows or columns that account for the missing observations. Our revised model can be expressed as

$$E[Q_\#Y_\#] = \mu 1_{n-t} + X_\# \begin{bmatrix} \tau \\ \beta \end{bmatrix} = \mu 1_{n-t} + X_{1\#}\tau + X_{2\#}\beta \tag{3}$$

If the changes in the $X_{1\#}$ and $X_{2\#}$ matrices cause an overlap in the column space such that $\mathcal{C}(X_{1\#}) \bigcap \mathcal{C}(X_{2\#})$ is larger than it was in the original design, then there is a reduction in the rank of X. Let the rank in the original X matrix be called r. Then $(X_\#) < r$, which means that the row space of $X_\#$ is less than that of X, i.e. $\mathcal{R}(X_\#) \subset \mathcal{R}(X)$. The set of observations that cause this reduction in rank is referred to as a Type I RROS.

There are two other RROSs that are easily identified in an experiment that involves blocking. First, if all of the observations that form a block are missing, then that block effect is no longer estimable. This is called a Type II RROS. This is not always a problem for the analysis. If there were no Type I RROSs in the block that is missing, then the analysis should be fine. Next, if all of the replicates

of a single treatment are missing, then the treatment effect is not estimable. This is called a Type III RROS. Similar to a Type II RROS, the loss of a single treatment can result in a design where the resulting analysis can still be performed. There may be some practical reason for this type of RROS to occur. For example, Godolphin (2006) indicated that this could be the result of a lethal dose being one of the experimental treatments.

The relationship between the design without missing observations and the design with missing observations that determines if a set of observations is rank reducing is uncovered in a comparison between the row spaces. If the row spaces are equal, then the set is obviously not rank reducing. All of the λ's that were estimable before are still estimable. However, a set of t observations is a RROS of size t and order s if

$$\dim\mathcal{R}(X_{\#}) = r - s \tag{4}$$

where s is bounded by r (the maximal rank of X) and t (the number of observations that are missing).

4. Methods of Determining Rank Reducing Observation Sets

Two methods are presented in the papers by Godolphin (2004, 2006) for identifying RROSs. The first method of Godolphin (2004) uses matrix $P = X(X^TX)^-X^T$ which is the perpendicular projection operator of X, and $(X^TX)^-$ which is the generalized inverse of the crossproduct of X. The method is simple and involves selecting columns corresponding to the observations of interest and then removing the rows. In Godolphin (2006), the author shows that the use of Theil's Z matrix to construct a matrix G whose rows are composed of a basis of the orthogonal complement to the column space of X can also be used to find RROSs. G also has the special property that

$$GX = G_*X_* + G_{\#}X_{\#} = 0 \tag{5}$$

where the rows of the matrices with $*$'s correspond to the missing observations and the rows of the matrices with #'s correspond to the observations that are still in the design.

There is no definitive discussion on which of the methods is better, but the author does mention that the G matrix is easier to use for the identification of RROSs of small sizes because the Z matrix is composed of integers, simple fractions or zeros. This is very useful for a quick comparison between designs. The first method requires the specification of a set of observations. This means that you must check each set individually to find all of the RROSs. Combining that with the need to use the P matrix, which is not as simple as the G matrix, can make it cumbersome. Even so, this method may be better suited to computer coding through some recursive algorithm to find RROSs. For proofs of the methods, refer to Godolphin (2006) and Godolphin (2004).

4.1. Perpendicular Projection Operator Method.

The first method is simple to apply. Construct a matrix consisting of the columns of P corresponding to the m observations of interest, $p_{t_1}, p_{t_2}, \ldots, p_{t_m}$. Remove the rows corresponding to all of the observations of interest, i.e. rows $t_1, t_2, \ldots, t_m$. If the rank of the resulting matrix is not full, then it is a RROS. As previously stated, this method is not easy to use in finding all RROSs of a design. For example, in a design with 24 observations, this procedure would need to be applied $\binom{24}{2} = 276$ times to find all the RROSs of size 2. This method also includes sets of observations that include a RROS. This means that if you find a RROS of size 2, then any set including those two observations will also be identified as rank reducing. The set is rank reducing, but it tends to hide unique RROSs when using an automated method.

4.2. G Matrix Using Theil's Z Matrix.

The G matrix of Godolphin (2006) is not unique. The only requirements for G is that it is composed of rows that span the orthogonal complement of $C(X)$. The author provides the following approach to defining a specific and convenient G matrix. Let the Q matrix from (1) rearrange our X matrix so that the first r rows of the X matrix are linearly independent. We can partition X as

$$X = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} = \begin{bmatrix} I_r \\ Z \end{bmatrix} X_0 \tag{6}$$

The X_0 matrix is $r \times k$, the X_1 is $(n-r) \times k$, and

$$Z = X_1 X_0^T (X_0 X_0^T)^{-1}. \tag{7}$$

The Z matrix leads to a convenient G matrix of the form

$$G = \begin{bmatrix} -Z & I_{n-r} \end{bmatrix} \tag{8}$$

It is readily apparent that this G has full rank of $(n-r)$ and its rows are a basis for the orthogonal compliment of $\mathcal{C}(X)$.

From (6) and (7) we get some useful results. A row of X_1 can be written as a linear combination of the columns of X_0. Equation (6) tells us that $X_1 = ZX_0$. The rows of this Z matrix can be associated with the rows of the X_1 matrix, while the columns of the Z matrix can be associated with the rows of the X_0 matrix. Godolphin concludes from this that "...there is a 1-1 correspondence between the columns of G and the rows of X. We say that the jth column of G corresponds to the jth row of X and that the jth column of G corresponds to the jth element of QY for $1 \leq j \leq n$." (Godolphin (2006))

This is where we begin to identify the RROSs. If the loss of a single observation causes a decrease in the rank of X, then the row must obviously be from the X_0. Looking back at (8) we can see that because this row is in the X_0 matrix, its corresponding column in G is part of the $-Z$ matrix. Theorem 1 from Godolphin (2006) states that if an observation set is rank reducing, then

the corresponding columns of G must have a rank less than t, the number of missing observations. Because this is a single missing observation, that means that its rank must be zero and that the corresponding column is a vector of zeroes. This makes identifying rank reducing sets of one very easy to spot.

Theorem 1 also tells us how to identify other RROSs. In the case where $t = 2$, the two columns of the G matrix must be multiples of each other. For $t \geq 3$, then one of the columns of G must be a linear combination of $t - 1$ other columns in G.

This theorem provides another interesting result. The G matrix does not differentiate between the different types of RROS. Therefore, a Type II or Type III RROS has the same properties as any other type. If the design has two replicates of each treatment then, the Type III RROSs are of size two. That means that the two columns of G corresponding to the replicates are multiples of each other. If one of these vectors is an element of a linearly dependent set, then replacing it with its replicate will also produce a RROS. This concept is stated as Theorem 4 in Godolphin (2006).

5. Examples

To help demonstrate the G matrix, an example from Godolphin (2006) using a John-Eccleston row-column α-design will be presented. After examining that case, the G matrix for several different Balanced Incomplete Block Designs (BIBDs) will be examined. The notation for this section differs from the original.

5.1. John-Eccleston row-column α-design.

The design considered by Godolphin (2006) involved 12 treatments replicated twice with two blocking variables. Call the column blocking variable β_i and the row blocking variable γ_j. Here is a table of the design.

	β_1	β_2	β_3	β_4	β_5	β_6
γ_1	1	5	9	4	8	12
γ_2	2	6	10	1	5	9
γ_3	3	7	11	6	10	2
γ_4	4	8	12	11	3	6

Godolphin describes a method for finding the X_0 matrix using the degrees of freedom for the residual variance, $n - r$. *Maple 11* was used to perform all of the following matrix calculations including finding all the linearly independent rows of the matrix. The X_0, X_1 and G matrices are the same.

Let the Q matrix reorder the observations in the following manner. Counting down each column of the design for the observation number, X_0 contains observations 1 - 17, 19, 21, and 24 while X_1 consists of observations 18, 20, 22, and 23. This yields the following G matrix.

$$G = \left[\begin{array}{cccccccccccc|cccccccccccc} 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & -1 & 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & -1 & 0 & 0 & 1 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \end{array}\right]$$

We can reduce this matrix because each treatment is replicated twice. We only need one column relating to each treatment, so we can just use the first half of G because it contains all of the different treatments.

$$G_2 = \left[\begin{array}{cccccccccccc} 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \end{array}\right]$$

A quick examination of the new G_2 matrix shows that there are no RROS of size 1. There are, however, several RROS of size 2: the observation sets corresponding to treatment pairs of {2,11}, {3,6}, and {7,10} are all rank reducing. Because any combinations of the two treatments with the corresponding two observations is rank reducing, this means that we have four RROSs from each pair. Along with the Type II RROSs that we identified previously, the total number of RROSs of size 2 is $12 + 3(4) = 24$. The number of possible observation sets of size 2 is $\binom{24}{2} = 276$ meaning that there are 252 observation sets that are not rank reducing. Godolphin uses a staircase partition diagram, which was discussed in Godolphin and Godolphin (2001), to show that these RROS cause 15 to 18 of the 66 pairwise comparisons to be non-estimable, while only a few of the remaining pairwise comparisons are possible without confounding blocking variables. The loss of one of these observation sets causes serious problems in the resulting analysis.

5.2. Balanced Incomplete Block Design (BIBD).

BIBDs are a class of commonly-used designs. A BIBD is very useful when all of the treatments cannot be observed in a single block. First, we will consider the following randomized complete block design (RCBD) with five treatments that are replicated across five blocks:

	β_1	β_2	β_3	β_4	β_5
A	x	x	x	x	x
B	x	x	x	x	x
C	x	x	x	x	x
D	x	x	x	x	x
E	x	x	x	x	x

where A, B, C, D, and E are the treatments while the β's are the block effects. The experimenter would randomize the run order, but the design would remain the same. From this we can see that there would be five Type II and Type III RROSs of size 5 corresponding to the replicates of the treatments and the number of observations in a block. The G matrix of this design is

$$G = \left[\begin{array}{cccccccccccccccccccccccc}
1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
1 & 2 & 3 & 4 & 5 & 6 & 11 & 16 & 21 & 7 & 8 & 9 & 10 & 12 & 13 & 14 & 15 & 17 & 18 & 19 & 20 & 22 & 23 & 24 & 25
\end{array}\right]$$

The last row of the matrix is the corresponding observations numbered in the same fashion as before, by counting down the columns of the design. The smallest RROSs are the Type II and Type III that were noted previously. This design is robust to missing observations. Any missing observation set of size $t \leq 7$ would at most cause a single block or treatment to be non-estimable in pairwise comparisons. A missing set of size seven would mean that you are missing over a quarter of your data. This could happen if the treatment (or application of the treatment) causes all of the observations to be lost or if all of the observations in a block were missing. There are some RROSs of size eight that are Type I. These involve removing all of the all of the treatments but one inside of a block and removing that treatment from all other blocks.

Consider the same situation except that only four observations can be made in a single block. Keeping with the idea of a BIBD, the design would look like the following.

	β_1	β_2	β_3	β_4	β_5
A	x		x	x	x
B	x	x		x	x
C	x	x	x		x
D	x	x	x	x	
E		x	x	x	x

There are now only 20 observations and each of the treatments has four replicates and appears in four of the blocks. Similarly to the RCBD, there are four Type II and Type III RROSs of size 4. It's G matrix is

$$G = \left[\begin{array}{cccccccccccccccccccc}
0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
1 & 2 & 3 & 4 & 5 & 8 & 9 & 13 & 17 & 6 & 7 & 10 & 11 & 12 & 14 & 15 & 16 & 18 & 19 & 20
\end{array}\right]$$

The results of this design are similar to the RCBD with the smallest RROSs being the Type II and Type III. The slight difference for this design is that instead of needing to remove eight observations to isolate a treatment, block combination (three more than a Type II or Type III RROS), you only need to remove six observations (two more than a Type II or Type III). Following from this, a missing observation set of less than six observations could at most remove a single treatment or block from the set of estimable pairwise comparisons. These sets would at most represent a quarter of the observations. The Type I RROSs for this design are of size six and following the same pattern as for the RCBD.

Reducing the number of treatments per block we get the following design and G matrix:

	β_1	β_2	β_3	β_4	β_5
A	x			x	x
B	x	x			x
C	x	x	x		
D		x	x	x	
E			x	x	x

$$G = \left[\begin{array}{cccccccccccccc} 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 2 & 3 & 4 & 6 & 7 & 9 & 10 & 13 & 5 & 8 & 11 & 12 & 14 & 15 \end{array}\right]$$

Again, the smallest Type II and III RROSs for this design are of a size of 3. The minimal Type I RROSs are size 4 for this model.

Reducing the maximum number of treatments per block again brings the design to the smallest it can be while still remaining connected.

	β_1	β_2	β_3	β_4	β_5
A	x				x
B	x	x			
C		x	x		
D			x	x	
E				x	x

$$G = \left[\begin{array}{cccccccccc} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \end{array}\right]$$

This G matrix does not look good. There are no RROSs of size one because none of the "columns" are zero, but any set of observations larger than one is rank reducing. Interestingly enough, the best RROSs in this design are the Type II and III because they would only make one treatment effect or one block effect nonestimable. A Type I RROS would make both a treatment effect and a block effect nonestimable.

An experimenter would hope to not end up using a design like this, but the choice of how many observations can be put in a single block is often not something that can be controlled. Obviously, the RCBD design that was looked at first would have been the best of these designs in terms of RROSs. It has other advantages because it has more replication. However, there could be reasons that a design like that would not be possible. Maybe there is not enough money to collect 25 observations. It could also be that there are physical limitations such that only two observations could be put in a block. If the main limitation was that only two observations could be put in a block, then the following design could be used.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
A	x	x	x	x						
B	x				x	x	x			
C		x			x			x	x	
D			x			x		x		x
E				x			x		x	x

This design has 20 observations and each treatment is observed four times. It also falls under Godolphin's Theorem 4 because it has Type II RROSs of size 2. This means that we can use the following reduced G_2 matrix instead of the full G matrix.

$$G_2 = \left[\begin{array}{cccccccccc} -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ \hline 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array}\right]$$

By inspecting the G_2 matrix, it becomes apparent that other than the Type II RROSs, there are no RROSs smaller than size 4. The Type II RROSs are possibly the most acceptable, especially if the β's are just a nuisance variable. This design is much more robust to RROSs, but it does have twice as many observations as the other BIBD that had only two observations per block.

6. Conclusion

In any field, sucessful experiments are not trivial. The time spent planning and implementing an experiment, along with any costs incurred in the process, can be substantial. Businesses invest considerable amounts of rescoures in experiments. The man hours and money spent in a study are an investment. It is expected that they will net a profit from the venture. Grants are not something to be wasted and a failed experiment does just that.

Steps are always taken to prevent experiments from failing to be useful. The loss of an observation is one step towards a failed experiment. An experiment is always limited in someway and a missing observation could be looked upon as a further limitation. A missing observation may just reduce the degrees of freedom for an analysis, but as this paper has shown, it could also cause a major

disrruption in estimability. This disconnectedness may damage or even ruin a study. Gaurding against disconnectedness is important to ensuring that a study will be useful.

The example of the John-Eccleston row-column α-design showed that the loss of just two observations can cause serious damage to an experiment. There are other examples of designs where a RROS can cause all of the pairwise comparisons to become non-estimable. The BBIDs that were examples in this paper become more vulnerable to RROSs as the number of observations per block is reduced. The number of observations needed to go missing in order for a RROS to occur decreased by two everytime there was a reduction in the block size.

It is easy to anticipate the effects of a Type II or Type III RROSs, and studies that end up with a missing block or set of treatments can still be useful. The occurance and effects of a Type I RROSs are not as obvious. The use of these methods developed by Godolphin can help. Identifying observations that may lead to lead to a disconnected design can help an experimenter decide between to different designs. The identifcation can be used in a physical sense in that once an observation is flagged as influential, special precautions could be taken to prevent it from going missing.

The use of RROSs is one of many tools that can be used to select an experimental design. Along with considerations of efficiency, power, and other design criterion, RROSs can help make a design robust and more likely to be successful.

References

Godolphin, J. "The specification of rank reducing observation sets in experimental design." *Computational Statistics & Data Analysis*, 51(3):1862 – 1874 (2006).

URL http://www.sciencedirect.com/science/article/B6V8V-4HVW6KB-1/2/909ce9107e2c80756ab2942ff101c6cc

Godolphin, J. and Godolphin, E. "On the connectivity of row-column designs." *UTILITAS MATHEMATICA*, 60:51–65 (2001).

Godolphin, J. D. "Simple pilot procedures for the avoidance of disconnected experimental designs." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):133–147 (2004).

URL http://dx.doi.org/10.1046/j.0035-9254.2003.05054.x