Introduction

A common goal in implementation research is to compare alternative implementation strategies across a range of clinical conditions. Block designs can be used to obtain estimates of effect size while controlling for nuisance variables [1]. A nuisance variable is correlated with the outcome of interest but not of direct interest to the researcher; it might be a characteristic of the participants or institutions under study. In general blocks correspond to different values of the nuisance variables.

Trietsch et al [2] discuss a particular form of block design—the balanced incomplete block (BIB)—and recommend against using it in implementation research. They also discuss a 2 by 2 block design which has been used in a number of implementation studies. They identify that this design is not a BIB and suggest describing it as a 2-arm trial. We suggest that "2 by 2 Latin square" is more appropriate. We describe the relationship between BIB and Latin square designs and show that contrary to the views expressed by Trietsch et al both have an important role in implementation research. We discuss the analysis of the 2 by 2 Latin square. Technical details supporting our arguments are given in the statistical appendix [SA].

The BIB design in relation to Latin squares

The BIB design is used to evaluate v different treatments (implementation strategies) in b blocks (usually groups of patients or health care professionals) where each individual treatment appears in r blocks and within each block k different treatments are implemented. Additionally each pair of treatments appears in exactly λ blocks. A set of criteria link the parameters v, b, r, k and λ. By relaxing the criterion that v should be greater than k to allow v to be greater than or equal to k we define a slightly more general family of designs that include both BIBs and Latin squares [SA].

Trietsch's contention that BIB design is invalid in implementation research

The utility of the BIB design depends upon several factors. The key issue is whether one can analyse the observed data to test the study hypothesis; can we specify a valid statistical model and then estimate the parameters of interest?

In discussing the applicability of the BIB design, Trietsch et al comment:

> "According to Cochran and Cox the BIB design is suitable for situations in which repeated testing of varieties will lead to the same result, as can be expected when conditions can be well controlled as in agricultural or laboratory sciences. Unfortunately in most types of clinical research patients will be permanently influenced by the intervention that is being evaluated and therefore repeated testing cannot be expected to lead to the same result. As a consequence the BIB design cannot be used for patient-centred research."

This is their justification for their main proposition—that the BIB design is invalid in implementation research. It raises two questions: firstly, is it correct that repeated testing in patient-centred research cannot lead to the same result; secondly, if so can we fit an appropriate statistical model and answer our research questions?

Repetition of an implementation strategy on the same set of patients or the same set of health care professionals is unlikely to have the same effect on outcome and we should not try to estimate a single effect across replications. In the context of a BIB design it is more usual to consider the repeatability of results from different blocks testing the same pair of implementation strategies. How reasonable is it to assume that the expected effect size will be the same in different groups of health care professionals? In medical research one often assumes that repeated testing in different groups of patients will yield the same result. For example, this assumption underpins the use of the t-test to compare the arms of a randomised controlled trial.

If repeated testing does not lead to the same result it is necessary to consider more complex statistical models [SA]. In some circumstances, by making assumptions about the distribution of treatment effects, we can obtain interval estimates of effect size using methods developed for the case where the residual errors distribution is a mixture of two different distributions.

The 2 by 2 Latin square design

Trietsch et al consider a 2 by 2 design used in several implementation studies to compare two alternative implementation strategies across four groups of patients (Table 1) and identify that it is not a BIB. Considered as two blocks of health care providers, the design is balanced but since both treatment strategies occur in each block it is not incomplete.

As each treatment occurs once in each row and once in each column it can be considered as a two by two Latin square although because of the small number of degrees of freedom associated with this design it cannot be analysed using classical methods developed for larger Latin squares [3]. In practice the implementation studies that have used this design have involved multiple centres in each block; by modelling differences between centres as random effects rather than fixed effects it is possible to obtain interval estimates of treatment effects [SA].

A key feature of this design is that information on the effectiveness of each implementation strategy comes potentially from two sources: the performance of centres randomised to strategy A compared with the performance of centres randomised to strategy B within each condition and the performance of strategy A against strategy B within centres (but across different conditions). Analysing the two conditions separately estimates effectiveness only through the first of these comparisons; analysing conditions simultaneously utilises more of the available information but requires a more restrictive set of assumptions about the observed data [4,SA].

If using a common measure of outcome across both conditions (for example a generic measure of quality of life such as the SF-36 measure [5]) then it may be reasonable to assume that the error variance is the same for each condition. Conversely, if one uses separate condition-specific measures of outcome for each condition, it is less likely that the errors will be identically distributed across conditions. Then one option is to transform each measure to a standard normal distribution. Alternatively one can analyse data from the rows separately; in the case of the 2 by 2 design each row functions as a cluster randomized trial with two arms.

Increased applicability of block designs in implementation research

BIB and Latin square designs have been used to evaluate alternative implementation strategies while controlling for one or more nuisance variables [6-14]. Historically these designs were particularly useful when observations were independent and normally distributed. Advances in statistical theory and computer processing power have led to increased flexibility.

Generalized linear models (GLM) [15,16] enable us to consider response variables with other error structures including binary, multinomial or Poisson distributions. Further developments allow for correlated observations. Generalized linear mixed models (GLMMs), also called multilevel or mixed models, include random effects in linear predictors yielding explicit probability models that explain the correlation [17]. One consequence is that we can randomize blocks of health care professionals between alternative strategies rather than individuals thereby ameliorating any contamination arising from communication between professionals.

An extension of GLMMs permits more complex variance component structures in which the observed errors follow a mixture of distributions [18]; potentially we can now address the issue of measurement error in outcome variables. Interval estimates of the effect of treatment strategies corresponding to a very wide range of assumptions that may reflect more realistic scenarios are now possible.

Block designs can be particularly useful in the environment faced by the implementation researcher where limited resources restrict both the number of strategies that can be investigated and the number of settings where each is implemented. In this context the 2 by 2 Latin square offers a basic design in which nuisance variables are balanced across study sites and, because both strategies are implemented in each site, there is less threat to the motivation of principals to participate in such research. However the design does have limitations: to what extent can we generalize results from two clinical conditions per study to the entire range?

Simultaneous analysis of both conditions yields a pooled estimate of the difference between implementation strategies; however this is rare in practice. Future research should assess the merit of this approach compared with the usual practice of analysing each condition separately. In either case the choice of conditions needs careful consideration. Implementation of a particular strategy for one condition should not influence (or 'contaminate') the outcome for patients with the other. For example the implementation of guidelines for diabetes may affect the treatment of patients with coronary heart disease if they advocate monitoring of blood pressure.

Conclusion

Both Latin square and balanced incomplete block designs can be used to evaluate alternative implementation strategies while controlling for nuisance variables such as differences between individuals or institutions. Advances in statistical theory and processing power mean that these designs are even more applicable than when the theory relating to their use was first developed.

References

1. Bailey RA. Design of Comparative Experiments. Cambridge: Cambridge University Press; 2008.
2. Trietsch J, Leffers P, van Steenkiste B, Grol R, and van der Weijden T. Incorrect use of the Balanced Incomplete Block Design for the evaluation of complex interventions. Journal of Clinical Epidemiology 2014 (to be published alongside this paper)
3. Armitage P and Berry G. Statistical Methods in Medical Research, second edition. Oxford: Blackwell Scientific Publications; 1987.
4. Steen IN. Application of a Latin square experimental design in health services research: estimation of the effects of setting clinical standards and performance review on the process and outcome of care in general practice: PhD thesis. Newcastle University 1997.
5. Ware JE, Kosinski M and Keller SD. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston MA: The Health Institute; 1994.
6. Gravenstein JS, Smith GM, Sphire RD, Isaacs JP and Beecher HK. Dihydrocodeine — further development in measurement of analgesic power and appraisal of psychological side effects of analgesic agents. New England Journal of Medicine 1956; 254: 877-885.
7. Yorkston NJ, Sergeant HGS and Rachman S. Methohexitone relaxation for desensitising agrophobic patients. The Lancet 1968; 292; 651-653.
8. Motolese M. Factorial design in a balanced incomplete block in the evaluation of the minimal active dose of a beta-blocking drug. Bollettino Chimico Farmaceutico 1970; 109: 363-368.
9. North of England Study of Standards and Performance in General Practice. Medical audit in general practice. I: effects on doctors' clinical behaviour for common childhood conditions. British Medical Journal 1992; 304: 1480-1484.

10. North of England Study of Standards and Performance in General Practice. Medical audit in general practice. II: effects on health of patients with common childhood conditions. British Medical Journal 1992; 304: 1484-1488.

11. Ramsay CR, Campbell MK, Cantarovich D, Catto G, Cody J, MacLeod AM et al. Evaluation of clinical guidelines for the management of end-stage renal disease in Europe: the EU BIOMED 1 Study. Nephrology Dialysis Transplantation 2000; 15: 1394-1398.

12. Catella-Lawson F, Reilly MP, Kapoor SC, Cucchiara AJ, DeMarco S, Tournier B et al. Cyclooxygenase inhibitors and the antiplatelet effects of aspirin. New England Journal of Medicine 2001; 345: 1809-1817.

13. Gilron I, Bailey JM, Tu D, Holden RR, Weaver DF and Houlden RL. Morphine, Gabapentin, or their combination for neuropathic pain. New England Journal of Medicine 2005; 352: 1324-1334.

14. Bardy GH, Smith WM, Hood MA, Crozier IG, Melton IC, Jordaens L et al .An entirely subcutaneous implantable cardioverter–defibrillator. New England Journal of Medicine 2010; 363:36-44.

15. McCullagh P and Nelder J. Generalized Linear Models, second edition. Boca Raton: Chapman and Hall; 1989.

16. Nelder JA and Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society Series A 1972; 135: 370–384.

17. Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993; 88: 9-25.

18. Aitkin M and Rocci R. A general maximum likelihood analysis of measurement error in generalized linear models. Statistics and Computing 2002; 12: 163–174.

Table 1: The ‘2 by 2 Latin square’ design

Clinical Condition	Blocks of health care providers	
	1	2
C1	A	B
C2	B	A

Statistical appendix

1 The balanced incomplete block (BIB) design in relation to Latin squares

The basic specification of the BIB design is mathematical. Given a finite set X of elements called points (e.g. strategies in implementation research) and a collection of non-empty subsets of X called blocks (e.g. groups of patients or professionals in implementation research), we define five integers:

- v = number of points in X
- b = number of blocks
- r = number of blocks containing a given point
- k = number of points in each block
- λ = the number of blocks in which each pair of points is present

Two equations connect these parameters:

$$bk = vr \qquad (1)$$

$$\lambda(v-1) = r(k-1) \qquad (2)$$

A design is a BIB if it meets the following conditions:

$$r > 0 \qquad (3)$$

$$\lambda > 0 \qquad (4)$$

$$v > k > 0 \qquad (5)$$

Equation (5) states that k, the number of implementation strategies in each block, is less than v, the total number of such strategies under evaluation; that is why those blocks are ‘incomplete’.

In this paper we consider a more general set of designs that relax this condition thus:

$$v \geq k > 0 \qquad (5a)$$

This family includes both Latin square and balanced incomplete block designs.

A good example of the use of these designs in practice is the North of England study of standards and performance in general practice. The aim of the North of England Study of Standards and Performance in General Practice was to evaluate five implementation strategies in British primary care [ref A1]. The study design comprised two replications of a five by five Latin square (Table A1).

Table A1: The replicated Latin square design of the North of England study of standards and performance in general practice and balanced incomplete block design of the postal outcome dataset

Symptomatic childhood condition	Blocks within replication 1					Blocks within replication 2				
	1	2	3	4	5	6	7	8	9	10
Acute cough	*A*	*B*	*C*	*D*	*E*	*A*	*E*	*D*	*C*	*B*
Itchy rash	*E*	*A*	*B*	*C*	*D*	*B*	*A*	*E*	*D*	*C*
Acute vomiting	*D*	*E*	*A*	*B*	*C*	*C*	*B*	*A*	*E*	*D*
Wheezy chest	*C*	*D*	*E*	*A*	*B*	*D*	*C*	*B*	*A*	*E*
Bedwetting	B	C	D	E	A	E	D	C	B	A

The letters A to E correspond to five different implementation strategies under investigation

Each replication evaluated five implementation strategies A, B, C, D and E in five blocks of practices. Each block evaluated a different implementation strategy for each of five clinical conditions. Within each replication, each strategy appears once in each row and once in each column, thus forming a Latin square [ref A2]. Combining both replicates, one possible formulation of this design is 5 implementation strategies ($v = 5$) evaluated in ten blocks of practices ($b = 10$) with each implementation strategy appearing in each block ($k = 5$, $r = 10$, $\lambda = 10$). This meets the criteria in equations (1) to (4), and that in (5a) since $v = k$, but not that in equation (5). Hence the design is not incomplete.

In the postal outcome one row was missing; not postal outcomes were collected for one of the conditions [ref A3]. The resulting data set can be considered as a design with parameters $b = 10$, $v = 5$, $k = 4$, $r = 8$ and $\lambda = 6$ which satisfy the criteria set out for a BIB design:

- $bk = vr = 40$
- $\lambda(v-1) = r(k - 1) = 24$
- $r = 8 > 0$
- $\lambda = > 0$
- $v = 5 > k = 4 > 0$

2 Repeatability across multiple implementations of an implementation strategy

Consider the case where we wish to compare two implementation strategies. If the results are repeatable if one compares these strategies over several instances, i, the **expected value** of the mean difference D between them should be the same for each instance. We can model this scenario thus:

$$E[D_i] = \mu \qquad (6)$$

where D_i is the observed difference between the two strategies specifically for instance i.

For each instance i the observed value of D_i differs from this expected value μ by a random error e_i.

$$D_i = \mu + e_i \qquad (7)$$

The standard method of analysis would then make assumptions about the distribution of these random errors so as to formulate a statistical model for the observed data and an appropriate estimation procedure is then used to generate an interval estimate of μ.

If the results are not repeatable equation (6) does not; the expected mean difference will not be the same for each comparison. Hence we specify a separate mean μ_i for each instance:

$$E[D_i] = \mu_I \qquad (8)$$

The observed difference differs from this mean by a random amount:

$$D_i = \mu_i + e_i \qquad (9)$$

In general a model that yields a different estimate of effect size for every instance is not useful. When implementing a strategy across a range of conditions, it is typical to assume that expected mean differences vary randomly about some overall mean μ_0:

$$\mu_i = \mu_0 + f_i \qquad (10)$$

Replacing μ_i in equation (9) by $\mu_0 + f_i$ gives

$$D_i = \mu_o + e_i + f_i \qquad (11)$$

which can be written as

$$D_i = \mu_0 + g_i \qquad (12)$$

where $g_i = e_i + f_i$. This resembles equation (7) except that the observed error distribution is a mixture of two distributions. Methods have now been developed for calculating standard errors for the estimates of μ_0 (for example the approach described by Aitkin 1999 [ref A4]).

3 The analysis of the 2 x 2 Latin square

The standard statistical model used to analyse a Latin square is to assume orthogonal row, column and treatment effects with no interactions (see for example Armitage and Berry ref A5). Applying this to the two by two Latin square gives

$$Y_{ij} = \alpha + \beta R_i + \gamma C_j + \delta T_{ij} + e_{ij} \qquad (13)$$

where

- Y_{ij} is the observation in row i and column j
- R_i is 0 in row 1 and 1 in row 2
- C_j is 0 in column 1 and 1 in column 2
- T_{ij} is 1 if the treatment in cell ij is strategy A and 0 otherwise
- e_{ij} is a random error
- α, β, γ and δ are parameters to be estimated

In fitting this fixed effects model, each of the row and column effects takes one degree of freedom and the treatment effect takes a third degree of freedom, thus leaving no degrees of freedom for error. Hence the model is not well defined and it is not possible to estimate the treatment effect δ while adjusting for row and column effects.

For this reason the smallest Latin squares considered in most statistical text books are three by three designs. In practice, however, the implementation studies using this design include many centres in each block. We can then represent the design as in Table A2 where each of the $n_1 + n_2$ columns represents a centre.

Table A2: The 2 by 2 Latin square design with multiple centres within each column

	Study centres in block 1				**Study centres in block 2**			
	1	**2**	**…**	**n_1**	**n_1+1**	**n_1+2**	**…**	**n_1+ n_2**
Condition 1	A	A	…	A	B	B	…	B
Condition 2	B	B	…	B	A	A	…	A

The letters A and B correspond to two different implementation strategies; there are n1 practices randomised to block 1 and n2 practices to block 2.

In equation (13) we replace γC_j by $\sum_{j=2}^{n1+\ n2} \gamma_j C_j$ where the C_js are a set of $n_1 + n_2$ -1 dummy variables such that C_j is 1 in column j and zero otherwise and the γ_js are the corresponding model coefficients:

$$Y_{ij} = \alpha + \beta R_i + \sum_{j=2}^{n1+\ n2} \gamma_j C_j + \delta T_{ij} + e_{ij} \qquad (14)$$

We can now fit this statistical model using classical methods assuming that the e_{ij} are independently and identically distributed (typically in a normal distribution) and using least squares or maximum likelihood estimation.

But before fitting this model it is important to consider whether the underlying assumptions hold. For example the model assumes that the difference between implementation strategies A and B is the same for conditions 1 and 2. In many practical situations this assumption is unrealistic. It is then necessary to fit an interaction between the row effect and the treatment effect:

$$Y_{ij} = \alpha + \beta R_i + \sum_{j=2}^{n1+\ n2} \gamma_j C_j + \delta T_{ij} + \phi R_i T_{ij} + e_{ij} \qquad (15)$$

Unfortunately in this design with fixed row and column effects the interaction term is confounded with the main effects; it is not possible to estimate separate effects of the implementation strategies for each condition. To resolve this issue we can treat the centres as a random sample from all possible centres. Rather than fitting a separate effect γ_j for each centre we assume that the centre effects vary randomly about some general mean γ_0; each column mean is equal to γ_0 plus a random error u_j. The resulting model now has two fixed effects, their interaction and two random errors:

$$Y_{ij} = \alpha^* + \beta R_i + \delta T_{ij} + \phi\ R_i T_{ij} + u_i + e_{ij} \qquad (16)$$

where we have absorbed γ_0 into the new constant α^*. This type of model is commonly called a mixed model as it includes both fixed and random effects.

Making appropriate assumptions about the characteristics of the random errors terms leads to methods for estimating the parameters. For example, if we assume a normal distribution for both errors, we can estimate the parameters by iterative generalized least squares as described by Goldstein [ref A6].

References

A1 North of England Study of Standards and Performance in General Practice. Medical audit in general practice. I: effects on doctors' clinical behaviour for common childhood conditions. British Medical Journal 1992; 304: 1480-1484.

A2 Cayley A. On Latin Squares. The Oxford Cambridge and Dublin Messenger of Mathematics 1890; 19: 135-137.

A3 North of England Study of Standards and Performance in General Practice. Medical audit in general practice. II: effects on health of patients with common childhood conditions. British Medical Journal 1992; 304: 1484-1488.

A4 Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 1999; 55: 117-128.

A5 Armitage P and Berry G. Statistical Methods in Medical Research, second edition. Oxford: Blackwell Scientific Publications; 1987.

A6 Goldstein, H. Multilevel Statistical Models, third edition. London: Edward Arnold; 2003.