Journal Title
XX(X):1–27
©The Author(s) 2021
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Designing, understanding and modelling two-phase experiments with human subjects

Christopher James Brien[1][2]

Abstract

In a recent paper, Jarrett, Farewell and Herzberg discussed a strategy for developing the analysis of a previously published two-phase experiment that investigated the effect of training on pain rating by occupational and physical therapy students. Here, their example is used to illustrate how a multi-step factor-allocation paradigm can be employed (i) to design an experiment, (ii) to understand the confounding in the design and (iii) to formulate linear mixed models, called prior allocation models, for the design. These models are intended as starting models for the analysis of the data, when it becomes available. An understanding of the confounding intrinsic to a design is achieved through an anatomy of the design presented in an analysis-of-variance-style table that can be obtained using functions from the R package dae. The analysis of the pain-rating experiment is re-examined and it is recommended that conclusions be based on a model with heterogeneous residual variances, in addition to the previously proposed block-treatment interactions. The paradigm is also used in producing an alternative design, taking into account the results of the re-analysis.

Keywords

analysis of variance, ANOVA, design anatomy, block-treatment interaction, human experiments, intertier interaction, laboratory phase, linear mixed models, multiple randomizations, two-phase experiments

[1] University of South Australia; [2] University of Adelaide

Corresponding author:

[1] UniSA STEM, University of South Australia, GPO Box 2471, Adelaide, South Australia. 5001.
Email: chris.brien@unisa.edu.au

1 Introduction

As noted in Section 4.3.6 of Brien,[1] Walwyn and Roberts[2] appear to have been the first to recognise the need for multiphase experiments in research on human subjects. Also, Brien and his collaborators use variations on an athlete training experiment as examples of some of the issues involved in designing two-phase experiments on human subjects.[1,3,4] Recently, Jarrett, Farewell and Herzberg[5] published an insightful paper in which a pain-rating experiment reported by Solomon et al.,[6] that involved clinicians rating patients, was identified as being a two-phase experiment. Jarrett, Farewell and Herzberg,[5] by extending the previous analysis of Farewell and Herzberg,[7] demonstrated that block-treatment interactions, not included in the model for the original analysis, should not be ignored in formulating a model for such experiments. They discussed an approach, based on formulating various analyses of variance, whose objective was to ensure that the appropriate block-treatment interactions were included in the model.

Two-phase designs are members of the more general class of multiphase designs. They are a very rich class of designs as Brien et al.[3] describe in Section 4 and is evident from the review in.Brien[1] However, there are many experiments that are run on human subjects that are two-phase, or even multiphase, without this aspect of the experiment being recognized and a design employed for each phase. There is no reason why the use of experimental designs should be limited to the first phase, as often happens; there is every reason to believe that experimentation would be improved by employing designs for all phases. In this paper, lower case ‘phase’ is used to distinguish the use of phases in this context from that of ‘Phases’ in clinical trials. Here phases are different parts of a single experiment.[1] Each phase involves different units and produces an outcome. The outcome can be material for processing in the next phase, or values for response variables, or both. The phase is the period of time during which a set of units are engaged in producing their outcome. Only the final phase need have a response variable.

The Solomon et al. experiment used an orthogonal, two-phase design, a class of design described by Brien et al.[3] However, it has some features not found in the athlete training experiments.[1,3,4] The present paper appears to be the first to consider, in a single paper, the whole statistical story for a two-phase experiment on human subjects from its design to the results of the analysis of its data. The factor-allocation paradigm outlined by Brien[1] is applied to the design of and the model formulation for the pain-rating example. Then the resulting linear mixed model is revised, fitted and further revised in analyzing the data. Also considered are the implications of the analysis results for designing similar experiments. While the paradigm uses an analysis-of-variance (ANOVA)-like procedure, designers are urged to employ this procedure when designing an experiment i.e. before any data is available. Thus, the paradigm has in common with Jarrett et al.[5] that it stresses the use of ANOVA to understand the properties of a design.

The paradigm starts by hypothesizing a model for the effects anticipated to occur in this experiment, then chooses a design that allocates some factors to other factors in a manner that takes into account the hypothesized model and finally formulates a linear mixed model for analyzing the experiment when the data is available. It requires a careful and detailed consideration of the effects that are important and the consequent allocations that are made using the design for the experiment. This contrasts with the approach taken by Jarrett et al.,[5] who use what Brien et al. (Section 3.2)[3] termed the single-set description and which omits some aspects of the allocations. The practical benefit from our characterization of the allocations of factors in a design is that the ANOVAs produced display the confounding between different sources of variation in the experiment and so an understanding of the nature of the relationships between these sources of variation is gained, something that is missing when the single-set description is used.

In Section 2 it is imagined that the Solomon et al.[6] experiment has not yet been run and the factor-allocation paradigm is used to develop a design for the Farewell and Herzberg subset of it.[7] In Section 3, it is conjectured that the need for block-treatment interactions has now been realized and so the linear mixed model is revised to incorporate them and its properties are assessed; then, the analysis of the data for the Farewell and Herzberg subset, based on the new model, is investigated. Section 4 uses the factor-allocation paradigm in an exploration of alternative designs for the pain-rating scenario. A discussion is provided in Section 5. Further resources, including a glossary of terms, are available on the multitiered web site.[8]

2 Designing the pain-rating experiment for the Farewell and Herzberg example

The example that Farewell and Herzberg[7] describe is a subset of a pain-rating experiment reported by Solomon et al.[6] that is a two-phase experiment. The first phase is a self-assessment phase in which patients self-assess for pain while moving a painful shoulder joint. The second phase of this experiment is an evaluation phase in which occupational and physical therapy students (the raters) are evaluated for rating the pain that patients from the first phase are apparently feeling.

The self-assessment phase involved eight patients, each of whom was observed on two occasions. Four of the patients had been judged to be expressive patients, having high levels of pain, and the other four had been judged to be unexpressive patients, having low levels of pain. Two motions are to be allocated to the two occasions for each patient: active motion performed by the patient without assistance and passive motion in which a therapist guided the patient's limb through its range of movement. The outcomes from this phase are 16 videos made from the 8 patients on the two occasions and a pain rating made by each patient on each of the occasions that they underwent a shoulder motion.

The evaluation phase involved 74 raters (students) who each rated the 16 videos at 16 video viewings. In this phase, half of the raters were randomly selected to be trained in identifying the expression of pain via facial movements; the other half were untrained. The outcome of this phase is a set of pain ratings made by the raters over 16 viewings.

To illustrate the factor-allocation approach[1] to design and model formulation, we suppose that a design for this experiment is yet to be developed and that the need for block-treatment interactions terms has not been recognized. The process for this approach is outlined in Figure 1, an extension of Brien's[1] Figure 2 to explicitly allow for decomposition tables for both the initial and homogeneous allocation models.

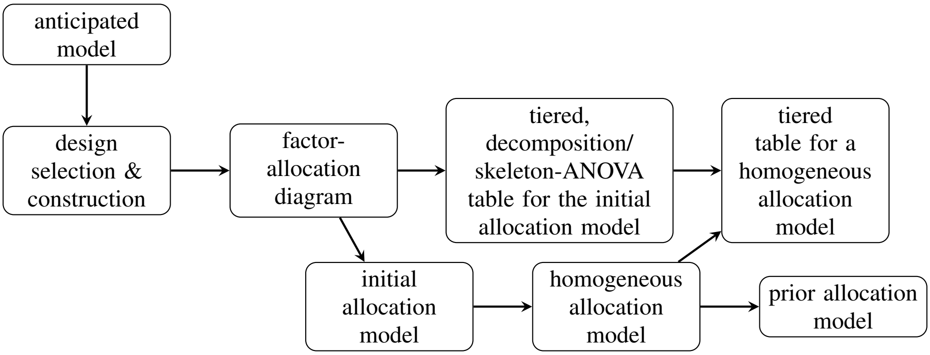
anticipated model
design selection & construction
factor-allocation diagram
tiered, decomposition/ skeleton-ANOVA table for the initial allocation model
tiered table for a homogeneous allocation model
initial allocation model
homogeneous allocation model
prior allocation model

Figure 1. Flowchart of items produced in using the factor-allocation paradigm to design an experiment.

The anticipated model. The process begins with the determination of the anticipated model as a means to identifying the effects that are to be taken into account in the experimental design. It is preferable that this is done in concert with the researchers. It can be facilitated by considering all the factors in each phase and identifying those that are the unit factors and those that are the allocated factors for the phase. The unit factors are, potentially at least, recipient factors in that they are to receive factors that are allocated to them. In single-phase experiments, the allocated factors are the treatment factors; in standard two-phase experiments[3] both the treatment factors and the unit factors from the first phase are allocated, as Brien and Bailey explain.[9] In considering the terms for effects to be taken into account, it is useful to identify the inherent crossing and nesting relationships between the recipient factors and to decide whether the allocation of factors in the design being developed should respect any crossing between factors or that, for the purposes of the design, some crossed factors are to be treated as nested.

The factors in the first phase of the pain experiment are Expressiveness, Patients, Occasions and Motions, with Motions to be allocated to the Occasions for each Patient. So Expressiveness, Patients and Occasions are designated as recipient factors and Motions is the sole allocated factor. The recipient factors uniquely index the observations for this phase. In considering a design for this experiment, the factor Expressiveness is an inherent characteristic of each patient and so Patients is nested within Expressiveness. The factor Occasions is crossed with Expressiveness and Patients because, for all eight patients, the first occasion has the consistent property that it occurs before the second occasion. The crucial question is whether an Occasions main effect should be included in the anticipated model because there is likely to be a systematic Occasions effect that is similar for all patients. If the answer is yes, then, to maximize the precision, a design is needed that allows for the Occasions effect to be removed, such as a row-column design. Otherwise, Occasions can be regarded as nested within Patients and Motions randomized to the two Occasions within each Patient. However, for the example, the motions were completed in the standard clinical order of active then passive.[10] That is, they are systematically allocated to Occasions and effectively any order effect of

Occasions is absorbed into the clinical procedure; Occasions will be regarded as nested within Patients and Expressiveness. Also, because an interaction between Motions and Expressiveness is considered possible, a term for it is included. The anticipated model, using just the initial capital letters of the factor names, for the first-phase is:

$$\mathrm{M} + \mathrm{E} + \mathrm{M{:}E} \quad | \quad \mathrm{P{:}E} + \underline{\mathrm{O{:}P{:}E}}, \tag{1}$$

where terms before the '|' are fixed and those after it are random; the underlined term, called an identity term, uniquely indexes the observational units, the videos, from this phase. The effects for each of the random terms are assumed to be independently and identically distributed (IID) with zero expectation. This is a symbolic model that corresponds to a formal model in which there is a set of effects for each term (see Supplemental Section B for formal models).

The factors unique to the second phase are Raters, Viewings and Trainings. The factors Raters and Viewings are recipient factors and Trainings is an allocated factor; Trainings is to be allocated to Raters. In addition, the first-phase factors must be allocated to the second-phase recipient factors; they are to be allocated to Viewings. In this phase Raters and Viewings are inherently crossed, in the same way that Occasions is in the first phase. Again, the question that arises is whether the order of viewing is likely to have an effect such that the main effect for Viewings should be included in the anticipated model. If the answer to this questions is no, then a design in which videos are randomized for each rater would be appropriate. Here, it is assumed that there is a possible order effect and so the main effect will be included in the anticipated model. In the example, for practical reasons, the order of watching the videos is the same for all raters.[7] So a randomization available in these circumstances is to randomize the order of the videos to the Viewings. It would appear that this was not done, given that it is stated that 'Formal randomization occurs only for the training classification'.[7] However, the viewing order is not given and I have arbitrarily specified the viewing order to be the same as in Table 1 of Farewell and Herzberg,[7] i.e. in standard order for Expressiveness, Patients and Motions. Nonetheless, some form of serial correlation between Viewings cannot be ruled out and the anticipated model should make provision for it. It is not included here only because it cannot be estimated given that the viewing order is unknown. The full anticipated model is obtained by adding to the first-phase anticipated model in (1), terms derived from the second-phase factors and the interactions between first- and second-phase factors that are of potential interest. In this case, the following terms might be added to the first-phase model:

$$\mathrm{T} + \mathrm{T{:}M} + \mathrm{T{:}E} + \mathrm{T{:}M{:}E} \quad | \quad \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}V}},$$

A difference between the factor-allocation approach and those used in previous discussions of this example, including Jarrett et al.,[5] is the inclusion of the factors Occasions and Viewings in describing the experiment. These are needed to complete the description of the first- and second-phase units and to describe the allocations in the experiment.[3] Their inclusion makes explicit the order of the Motions and Videos. Overlooking them is analogous to leaving out Plots in the description of a randomized complete-block design, Treatments being randomized to Plots within Blocks.

Design identification. The next step in Figure 1 is design selection and construction. In this case, given the practical restrictions that have been described, the design selection is straightforward in that factors, either singly or in combination, are allocated either systematically or randomly, as already described. A plot of the layout generated using the R[11] package dae[12] is in Supplemental Figure 1.

The factor-allocation diagram. Figure 1 suggests that the allocations are exhibited in a factor allocation diagram. Figure 2 does this for the example. It consists of four panels, each of which contains the factors that index the, possibly conceptual, objects that label the panels; lower case labels are used for objects to distinguish them from factors whose names use an initial capital letter. The set of factors in a panel is called a tier of factors by Brien and Bailey.[9] There are three allocations: the allocations of motions to videos and videos to (rating) episodes, which are composed allocations,[9], and the allocation of trainings to episodes, which is an independent randomization.[9] The first two allocations are composed in that the allocation of motions to episodes is achieved by combining the result of the allocation of motions to videos with the result of the allocation of videos to episodes. The motions and videos panels show the systematic allocation of Motions to Occasions for the first phase; Occasions is not nested here because Motions were allocated in the same order for each patient. The episodes panel shows the second-phase recipient factors to which all other factors in the design are allocated. The factors in different panels (tiers) differ in their roles in the allocations: (i) the factor in the motions panel is allocated in both the first and second phases; (ii) the factor in the trainings panel is allocated in just the second phase; (iii) the factors in the videos panel are recipient factors in the first phase and allocated factors in the second phase; and (iv) the factors in the episodes panel are only ever recipient factors in the second phase.

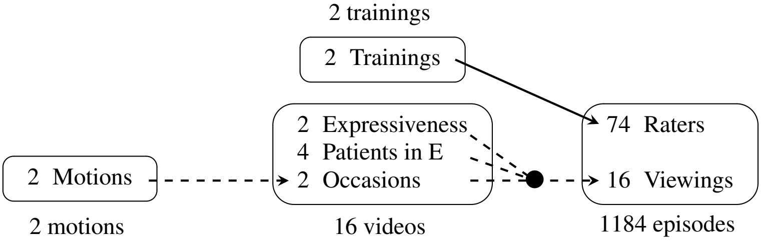

Figure 2. Factor-allocation diagram for the Farewell and Herzberg example, an orthogonal two-phase clinical experiment: motions are allocated to videos and both trainings and videos are allocated to episodes; the '•' with the dashed lines leading to it indicate that the combinations of the levels of Expressiveness, Patients and Occasions are allocated together; the dashed arrows indicate that the allocation of motions and of the combinations of the videos factors are systematic; the solid arrow indicates that the allocation of trainings is random; E = Expressiveness.

The designs for both phases are repeated measurements designs, the time factor in the first phase being Occasions and that in the second phase being Viewings. The units in the first phase are the videos and in the second phase have changed to the episodes. The outcomes of the first phase are the videos and pain ratings by the patients for each of their videos. The outcome of the second phase is the pain ratings by the raters. The first-phase design is a systematic design and the second-phase design is a plaid-square design,[13] where the rows are Raters and the columns are Viewings. A particular feature of this experiment is that 'treatment' factors, specifically Motions and Trainings, are allocated in both phases, with most interest in the experiment focused on Trainings that is allocated in the second phase. Also, Expressiveness is a recipient factor of interest in the first phase, whereas factors of interest are more often allocated factors.

The initial allocation model. In keeping with Figure 1, the initial allocation model is now derived from the factor allocation diagram. Terms are formed by taking all combinations of the factors within a panel, subject to the restriction that a factor that is nested cannot occur without its nesting factors. Terms with factors that were only ever allocated are designated as fixed. The other terms are designated as random and each has a canonical (covariance) component, a component that can be be negative, associated with it. The initial allocation model for the example is:

$$\begin{aligned} \mathrm{T} + \mathrm{M} \quad | \quad & \mathrm{O} + \mathrm{E} + \mathrm{P{:}E} + \mathrm{O{:}E} + \mathrm{O{:}P{:}E} + \\ & \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}V}}. \end{aligned} \tag{2}$$

The fixed terms in first line of (2) are derived from the motions and trainings panels; the random terms in the first line are the terms derived from the videos panel and the terms in the second line are the random terms derived from the episodes panel. A compact form of (2), using Wilkinson and Rogers[14] notation, is $\mathrm{T} + \mathrm{M} \mid \mathrm{O} * (\mathrm{E} \,/\, \mathrm{P}) + \mathrm{R} * \mathrm{V}$.

The homogeneous and prior allocation models. The factor allocation paradigm in Figure 1 acknowledges that the initial allocation model may need to be modified and allows for this in two steps. In the first, consideration is given to modifying the initial allocation model by changing the designation of terms as fixed or random and by adding interactions between terms derived from factors in different panels. The ensuing model is referred to as the homogeneous allocation model because, like the initial allocation model, it is a set of fixed terms and a set of IID random terms. Terms resulting from factors from different panels (tiers) are called intertier interactions by Brien and Bailey (Section 7.1).[9] They discuss different types of such interactions and the need for them; most commonly, they are block-treatment interactions, but can be interactions between treatment factors from different phases or between units factors from different phases. The second step is to modify the homogeneous allocation model to form the prior allocation model; the modifications envisaged involve changes to the form of the terms. Thus, one might specify autocorrelation between Viewings within Raters or heterogeneity of residual variance; perhaps, it is suspected that the variance in individual episodes will vary between levels of the fixed factors, for example Expressiveness.

Here, the initial allocation model differs from the full anticipated model in that: (i) Expressiveness is designated as random, (ii) there are no interactions of factors from different panels, and (iii) Occasions is treated as crossed. The initial allocation model is modified to reflect the first two of these so that the homogeneous allocation model is:

$$\mathrm{T} * \mathrm{M} * \mathrm{E} \quad | \quad \mathrm{O} + \mathrm{P{:}E} + \mathrm{O{:}E} + \mathrm{O{:}P{:}E} + \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}V}} \tag{3}$$

In this case, no changes to the homogeneous allocation model are thought to be needed in forming the prior allocation model.

Tiered decomposition and skeleton-ANOVA tables. A highlight of the factor-allocation paradigm is the production of a tiered decomposition table for the design, as a way of checking and understanding the design; adding expected mean squares (EMSs) to this table results in a tiered skeleton-ANOVA table. Because sources for different tiers are put into in separate columns in these tables, the confounding intrinsic to the design can be investigated. Brien[1] refers to this as displaying the anatomy of the design; the anatomy is the structure built on the allocations in the design and is expressed in the confounding relationships between the sources derived from the terms in a model for the design. All designs involve confounding, but this is not widely recognized. This is because units factors like Viewings and Occasions are often omitted from the description of an experiment, preventing a complete description of the confounding. Here confounding is used in its original sense in the context of experimental design: the confounding of allocated (treatment) sources with recipient (block) sources as described by Fisher.[15]

It is strongly advocated by Brien et al.[1,3] that tiered decomposition tables should be routinely generated when producing the design for an experiment. Software for producing it has been provided in the R package dae via the designAnatomy function. This function requires a layout for the experiment, in the form of the values for the factors involved in the design, and a set of model formulae; no data values are needed. A decomposition table depicts the anatomy of the design corresponding to the terms in a model, ignoring whether the terms are fixed or random. Usually, it is based on either the initial or the homogeneous allocation model.

The tiered decomposition for the terms in the initial allocation model is produced straightforwardly by using the factor-allocation diagram to establish a formula for each panel (tier) that specifies the sum of the terms derived from the factors in the panel as described for the initial allocation model. The formula order is important: the formula for the units terms from the second phase is considered first, followed by that for the unit terms from the first phase. Finally, formulae for the first-phase and second-phase allocated factors are considered. In this case, it is convenient to combine the two allocated factors into a single formula. For the example, the three formulae are $\mathrm{R} * \mathrm{V}$, $\mathrm{O} * (\mathrm{E} \,/\, \mathrm{P})$ and $\mathrm{T} + \mathrm{M}$.

The tiered decomposition for the homogeneous allocation model extends that for the initial allocation model by adding any extra terms that it includes. For intertier interactions, these are normally included in the last formulae that contains a factor from the term. Thus, the third formula would become $\mathrm{T} * \mathrm{M} * \mathrm{E}$. In this case, the initial and homogeneous models are quite similar and so only the tiered decomposition table for the homogeneous allocation model is presented, in Table 1. The Supplemental R script FHdesign.r uses the designAnatomy function to produce the sources in this table. The table contains a source for each term in the model, fixed sources for fixed terms and random sources for random terms. A '#' separates factors that interact and '[...]' contains the generalized or joint factor comprised of the combinations of the nesting factors, if any. Sources and terms differ in their dimensions. The dimension of a term is the number of observed combinations of the levels of its factors, which is the number of parameters

Table 1. Tiered, skeleton-ANOVA table for the homogeneous allocation model for the Farewell and Herzberg example: R = Raters; V = Viewings; E = Expressiveness; P = Patients; O = Occasions; T = Trainings; M = Motions; EMS = Expected Mean Square; DF = Degrees of Freedom.

episodes		videos		trainings-motions		EMSs†								
Source	DF	Source	DF	Source	DF	ϕ_{RV}	ϕ_{R}	ϕ_{V}	ϕ_{OPE}	ϕ_{PE}	ϕ_{OE}	ϕ_{O}	ϕ_{E}	$\theta_{\cdot}$
Mean	1	Mean	1	Mean	1	1	16	74	74	148	296	592	592	θ_{μ}
R	73			T	1	1	16							θ_{T}
				Residual	72	1	16							
V	15	O	1	M	1	1		74	74		296	592		θ_{M}
		E	1			1		74	74	148	296		592	
		P [E]	6			1		74	74	148				
		O # E	1	M # E	1	1		74	74		296			θ_{ME}
		O # P [E]	6			1		74	74					
R # V	1095			T # M	1	1								θ_{TM}
				T # E	1	1								θ_{TE}
				T # M # E	1	1								θ_{TME}
				Residual	1092	1								
Total	1184													

†Each ϕ is a canonical component that, except for ϕ_{RV}, is allowed to be negative. Their subscripts are comprised of the first letter of each factor in the corresponding term and the numbers in the table are the coefficients of the canonical components in the EMSs. Each θ is the same quadratic function of the expectation as the corresponding mean square is of the data.

for the term in the model. On the other hand, the dimension of a source is its degrees of freedom (DF) and is the dimension of the subspace of the data space for the source. A source has been made orthogonal to all the terms marginal to its term; a term is marginal to another term, derived from factors from the same panel, if the column space of the term for the marginal term is a subspace of that for the other term. When the subspace for a term is null after it has been orthogonalized to all its marginal terms, it is said to be aliased. As an example, consider the terms I, R, V and R:V, where I is the factor for the overall mean. Their sources are I, R, V and the interaction R # V. The terms I, R and V are marginal to R:V because the column spaces of the terms I, R and V are subspaces of the column space of R:V. Thus, R # V is orthogonal to the terms I, R and V. The method used by designAnatomy to construct Table 1 is described in Supplemental Section B.

So what does Table 1 show us has been confounded in this design? The confounding results from the relationships between sources derived from the different panels. In Table 1, the confounding is encapsulated in the horizontal alignment of sources from the different columns of sources. Now, a source is partitioned into a set of sources to its right, if any, starting with the first of these sources, and those sources below the starting source that do not have sources to their left. This set of sources to the right, except for a Residual source, are confounded with the partitioned source in that each is a subspace of the partitioned source. The Residual source is that part of the partitioned source that is orthogonal to the sources confounded with it. Table 1 shows (i) T is confounded with R, (ii) the videos sources O, E, P [E], O # E and O # P [E] are all confounded with V, (iii) M is confounded with O and M # E is confounded with O # E, all of which are also confounded with V, and (iv) T # M, T # E and T # M # E are confounded with R # V. There are two Residual sources: (i) one for R, it being the subspace of the R source that is orthogonal to the T subspace, and (ii) the other for R # V, it being the subspace of R # V that is orthogonal to the subspaces for the three interactions confounded with R # V. Without Occasions and Viewings being included in the models for the experiment and a tiered decomposition, this confounding is not displayed. Brien et al. (Section 3.2)[3] canvas this and other consequences of omitting such factors in describing an experiment.

Evident in this table is a difference in the relationship between T and R as compared to the relationship between E and P. Because the level of Expressiveness is inherent to a Patient, Patients are intrinsically nested within Expressiveness and this results in the nested source P [E] for Patients. On the other hand, because Trainings is randomly assigned to Raters, Trainings is not an intrinsic grouping of Raters; it is only known post-randomization which level of Trainings is associated with a Rater. So, in the analysis presented, ϕ_{R} represents the variability of all 74 Raters, irrespective of the Trainings group to which they belong. Of course, this variability is estimated using the differences between Raters within Trainings that is provided by the Residual variance, but it is the Rater variability that is being measured.

The anatomy produced using designAnatomy shows that the design is orthogonal and so EMSs have been added to Table 1 using the rules outlined in Supplemental Section C, the result being a skeleton-ANOVA table. For each random model term, there is a canonical component (ϕ) and, for each row of the table with a fixed source, there is a fixed contribution (θ) to the EMS. The EMSs reflect the confounding exhibited in the table.

A highlight of this table, coming from the confounding it displays, is that it exposes pseudoreplication. Pseudoreplication manifests in a tiered decomposition table as a recipient source being exhaustively confounded by one or a set of allocated sources. It is seen that the videos sources exhaustively confound V, there being no Residual for V. Thus, the combinations of the allocated factors E, P and O are pseudoreplicated and, as the EMSs show, there is no test available for testing O # P [E]. This was partially recognized by Farewell and Herzberg,[7] in that the confounding of Patients, but not Occasions, with Viewings was noted. Even more importantly, allocated source M exhaustively confounds its recipient source O showing that M is pseudoreplicated and so no test of M available. Similarly, M # E exhaustively confounds O # E. The pseudoreplication of M and M # E appears not to have been previously documented.

3 Analysis of the Farewell and Herzberg pain-rating example

Having run the experiment and obtained the data, which in this case is available online,[16] the analysis of the design given in Section 2 can be utilized to start the data analysis. However, suppose that it was realized, after the experiment had begun, that interactions between allocated factors and the recipient factors from both phases (e.g. block-treatment interactions) may have occurred in the experiment, as Jarrett et al.[5] did. So the homogeneous, and hence prior, allocation models need extending to include such interactions. Of course, if these interactions had been anticipated when the experiment was being designed, the factor-allocation paradigm allows for them to be included in the homogeneous allocation model at the design stage. In general, interactions between allocated factors and the recipient factors in the tier to which they are allocated are limited to those that either do not involve terms for recipient sources with which the allocated factors are confounded, or the term containing all factors in the recipient tier. If this restriction is not observed, the source for the new term will be aliased with the recipient source whose term is part of the new term. Thus, interactions between T and either R or R:V are not considered.

3.1 Revising the homogeneous and prior allocation models

There are four tiers and so, potentially, six pairs of tiers to contribute intertier interactions. The terms for the three tiers, motions, trainings and videos, can be combined into one model formula that includes all of their intertier interactions: $\mathrm{T} * \mathrm{M} * (\mathrm{E} \,/\, \mathrm{P})$; it does not include any terms with O, these being aliased with the terms from this formula that include M. The intertier interactions between episodes and videos factors can be added to the videos terms in (3) to yield the formula $\mathrm{R} * \mathrm{O} * (\mathrm{E} \,/\, \mathrm{P})$. There is one intertier interaction between trainings and episodes factors, T:V, and one between motions and episodes, R:M. Adding the intertier interaction terms to the homogeneous allocation model in (3) produces the revised homogeneous allocation model:

$$\begin{aligned} \mathrm{T} * \mathrm{M} * \mathrm{E} \quad | \quad & \mathrm{O} + \mathrm{P{:}E} + \mathrm{O{:}E} + \mathrm{O{:}P{:}E} + \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}V}} + \\ & (\mathrm{T} * \mathrm{M})\mathrm{{:}P{:}E} + \mathrm{T{:}V} + \mathrm{R{:}M} + \underline{\mathrm{R{:}(O} * (\mathrm{E} \,/\, \mathrm{P}))} \end{aligned} \tag{4}$$

In this model there are, in order, (i) the fixed effects, including intertier interactions, for the factors of interest, (ii) the random effects on the first line corresponding to the units in each phase, and (iii) on the second line, the random intertier interactions. Apart from block-treatment interactions, there are intertier interactions between treatment factors, e.g. T # M, and between block factors from different phases, e.g. R # O.

To understand the effects of the revisions made to the homogeneous allocation model, the anatomy in Table 1 is extended to include sources for all of the terms in (4), producing the new anatomy displayed in Table 2. The sources in the decomposition can be produced using the code in the Supplemental R script in the file FHdesign.r, as can the corresponding model terms. It uses the four model formulae $\mathrm{R} * \mathrm{V}$, $\mathrm{T} * \mathrm{V}$, $\mathrm{R} * \mathrm{O} * (\mathrm{E} \,/\, \mathrm{P})$ and $\mathrm{T} * \mathrm{M} * (\mathrm{E} \,/\, \mathrm{P}) + \mathrm{M} * \mathrm{R}$. The new anatomy shows that the decomposition is orthogonal and so the rules in Supplemental Section C were used to add the EMSs to produce the tiered skeleton-ANOVA table. Further, it confirms that the sources specified by $\mathrm{R} * \mathrm{O} * (\mathrm{E} \,/\, \mathrm{P})$ give a complete decomposition of the data space. A major difference between Table 2 here and Table 4 of Jarrett et al.[5] is that Table 2 displays the confounding of treatment and intertier sources with the unit sources. Another difference is that Table 2

Table 2. Tiered skeleton-ANOVA for the revised homogeneous allocation model for the Farewell and Herzberg example: $\mathrm{R} = \mathrm{Raters}$; $\mathrm{V} = \mathrm{Viewings}$; $\mathrm{E} = \mathrm{Expressiveness}$; $\mathrm{P} = \mathrm{Patients}$; $\mathrm{O} = \mathrm{Occasions}$; $\mathrm{T} = \mathrm{Trainings}$; $\mathrm{M} = \mathrm{Motions}$; $\mathrm{EMS} = \mathrm{Expected\ Mean\ Square}$; $\mathrm{DF} = \mathrm{Degrees\ of\ Freedom}$.

episodes		trainings-episodes		videos-episodes		motions-trainings-videos-episodes		EMSs[†]															
Source	DF	Source	DF	Source[‡]	DF	Source	DF	ϕ_{RV}	ϕ_{R}	ϕ_{V}	ϕ_{TV}	ϕ_{OPE}	ϕ_{PE}	ϕ_{ROPE}	ϕ_{ROE}	ϕ_{RPE}	ϕ_{RE}	ϕ_{RO}	ϕ_{RM}	ϕ_{TMPE}	ϕ_{TPE}	ϕ_{MPE}	$\theta.$
Mean	1	Mean	1	Mean	1	Mean	1	1	16	74	37	74	148	1	4	2	8	8	8	37	148	74	θ_{μ}
R	73	T	1					1	16		37			1	4	2	8	8	8	37	148		θ_{T}
		Residual	72					1	16					1	4	2	8	8	8				
V	15			O	1	M	1	1		74	37	74		1	4			8	8	37		74	$\theta_{\mathrm{O,M}}$
				E	1			1		74	37	74	148	1	4	2	8			37	148	74	θ_{E}
				P [E]	6			1		74	37	74	148	1		2				37	148	74	
				O # E	1	M # E	1	1		74	37	74		1	4					37		74	$\theta_{\mathrm{OE,ME}}$
				O # P [E]	6	M # P [E]	6	1		74	37	74		1						37		74	
R # V	1095	T # V	15	R # O$_1$	1	T # M	1	1			37			1	4			8	8	37			θ_{TM}
				R # E$_1$	1	T # E	1	1			37			1	4	2	8			37	148		θ_{TE}
				R # P [E]$_1$	6	T # P [E]	6	1			37			1		2				37	148		
				R # O # E$_1$	1	T # M # E	1	1			37			1	4					37			θ_{TME}
				R # O # P [E]$_1$	6	T # M # P [E]	6	1			37			1						37			
		Residual	1080	R # O$_2$	72	M # R	72	1						1	4			8	8				
				R # E$_2$	72			1						1	4	2	8						
				R # P [E]$_2$	432			1						1		2							
				R # O # E$_2$	72			1						1	4								
				R # O # P [E]$_2$	432			1						1									
Total	1184																						

[†] Each ϕ is a canonical component that, except for ϕ_{RV} and ϕ_{ROPE}, is allowed to be negative. Their subscripts are comprised of the first letter of each factor in the corresponding term and the numbers in the table are the coefficients of the canonical components in the EMSs. Each θ is the same quadratic function of the expectation as the corresponding mean square is of the data.

[‡] Sources with the same label but different subscripts are orthogonal subspaces of the space for the full source that is their sum.

has more intertier interactions than their Table 4, but, as will be revealed using our table, the extra interactions are exhaustively confounded. The result is that the decompositions for the two tables are equivalent. Three things are apparent from Table 2.

Firstly, as Jarrett et al.[5] point out, valid hypothesis tests for some sources cannot be achieved as the ratio of one mean square to another. For example, a test for T # M # E requires a ratio of two linear combinations of the mean squares because there is no mean square whose EMS only differs from that for T # M # E in not having a fixed contribution θ.

Secondly, the fixed sources O and M are confounded, as are O # E and M # E. To address this, remove O and O:E from the model, as was done for the anticipated model.

Thirdly, the random sources V, R # V, T # V, O # P [E] and R # O_2 are exhaustively confounded; they have no Residual source, in contrast to R. Consequently, the variance matrix for the homogeneous allocation model in (4) is singular. Examination of the confounding or the EMSs reveals that it is impossible to separately estimate (i) ϕ_{V}, ϕ_{OPE} and ϕ_{MPE}, (ii) ϕ_{RV} and ϕ_{ROPE}, (iii) ϕ_{TV} and ϕ_{TMPE}, and (iv) ϕ_{RO} and ϕ_{RM}; the remaining canonical components are estimable. To achieve a nonsingular variance matrix, all but one of the terms for the canonical components in each of these four sets of components need to be removed from the model, resulting in what Brien and Demétrio[17] term a 'model of convenience': a model that does not include all the possible terms at play in the experiment but which is nonsingular and can be fitted. I would remove the intertier interaction terms M:P:E, R:O:P:E, T:V and R:M, as well as the term O:P:E. This retains, as far as is possible, the terms that are germane to the allocations used in the design. However, which are removed makes no difference to the fit of the model. The important point is that, whichever terms are retained, their estimates are the sums of the confounded components. That is, ϕ_{V}, ϕ_{RV}, ϕ_{TMPE} and ϕ_{RO} are in reality estimates of $\phi_{\mathrm{V}} + \phi_{\mathrm{OPE}} + \phi_{\mathrm{MPE}}$, $\phi_{\mathrm{RV}} + \phi_{\mathrm{ROPE}}$, $\phi_{\mathrm{TV}} + \phi_{\mathrm{TMPE}}$ and $\phi_{\mathrm{RO}} + \phi_{\mathrm{RM}}$. Having the extra intertier interactions and the anatomy has the advantage that it makes explicit that there can be multiple explanations of an intertier variance, should an analysis detect one.

The nonsingular prior allocation model, derived from the homogeneous allocation model in (4) by removing the suggested terms, is:

$$\begin{aligned} \mathrm{T} * \mathrm{M} * \mathrm{E} \quad | \quad & \mathrm{P{:}E} + \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}V}} + \\ & \mathrm{T{:}P{:}E} + \mathrm{T{:}M{:}P{:}E} + \mathrm{R{:}O} + \mathrm{R{:}E} + \mathrm{R{:}O{:}E} + \mathrm{R{:}P{:}E} \end{aligned} \tag{5}$$

Overall, adding intertier interactions after doing the experiment has, in this case, not compromised the design for the experiment.

3.2 Analyzing the data

While skeleton-ANOVA tables are used to understand designs, we prefer to analyze data from designed experiments by fitting and comparing linear mixed models using software that provides the appropriate tests and estimates automatically. To be suitable, such software must (i) be able to fit models with other than IID residuals, (ii) calculate approximate denominator DF using Kenward-Rogers,[18] or equivalent, approximations, (iii) produce predicted means and their variance matrices, and (iv) ideally, allow negative estimates of canonical/variance components. An R[11] package with these capabilities is asreml[19] and so it was used, in conjunction with the R package asremlPlus,[20] to analyze the data. The Supplemental R script in the file FHanal.r was used to achieve the analysis.

The response to be analyzed is the difference between a student rating and the corresponding patient rating. The general model-fitting strategy is to (i) fit the nonsingular prior allocation model using mixed model software, (ii) examine residual-versus-fitted-values and residual-QQ plots to assess the need for model changes, and, if any are, modify and refit the model, (iii) determine what changes made in (ii) are required and abandon any that are not, (iv) decide whether random intertier interactions (intertier variances) improve the model fit and remove any that do not, and (v) use hypothesis tests to identify how the fixed factors affect the response. Terms in the initial allocation model are retained, even if nonsignificant, unless they are random terms estimated to be zero.

Models that differ in their random terms are compared using the AIC[21] to balance Type I error and power in choosing between models as advocated by Matuschek et al.;[22] the model with the smallest AIC is selected. Their argument that likelihood ratio tests based on $\alpha = 0.05$ impose a strong penalty on model complexity is accepted. When it comes to fixed effects, Wald F-statistics, with Kenward-Rogers approximations[18] to the denominator DF, are computed and used in hypothesis tests with α set to 0.05. The only nonsignificant fixed terms that are omitted are nonsignificant post-hoc covariate terms. This avoids biasing the estimates of random terms with fixed effects for which a Type II error has been made. Further, the hypothesis testing for fixed effects respects the marginality (or hierarchical) relationships between the fixed factors e.g. a main effect is not tested if it occurs in a significant interaction, or a significant two-factor interaction is not tested if a three-factor interaction that involves both of the factors in the two-factor interaction is significant. Nonetheless, estimated marginal means[23] (EMMs) for the combinations of the fixed factors that conform to the model chosen for these factors are obtained. For example, suppose that in a two-factor experiment involving treatment factors A and B, the two-factor interaction is not significant, but that the A and B main effects are significant. The term A:B is not omitted from the model. Instead, EMMs are obtained for each combination of the levels of A and B that are additive in A and B; that is, EMMs that conform to the two-factor additive model $\mathrm{A} + \mathrm{B}$. This can be achieved by using a function in asremlPlus[20] to take the linear transformation of the simple means for the combinations of the levels of A and B that projects the simple means into the additive model space. Comparing the EMMs for the same level of one of the factors will have greater precision than comparing the corresponding simple means for the combinations of A and B because, in essence, main effects are being compared.

The following four model fits, each beginning with a different model derived from the nonsingular prior allocation model in (5) but with variance parameters constrained to be nonnegative and proceeding as described, are compared:

a) no intertier variances, equal residual variances: Remove the random intertier interactions in the second line of (5), fit the model and examine the plots of standardized residuals (Figures 3A and B, upper left quadrants);

b) no intertier variances, unequal residual variances: Remove the random intertier interactions in (5) and replace R:V with R: $\mathrm{idh}(\mathrm{M{:}E})$:P, where $\mathrm{idh}(\mathrm{M{:}E})$ specifies

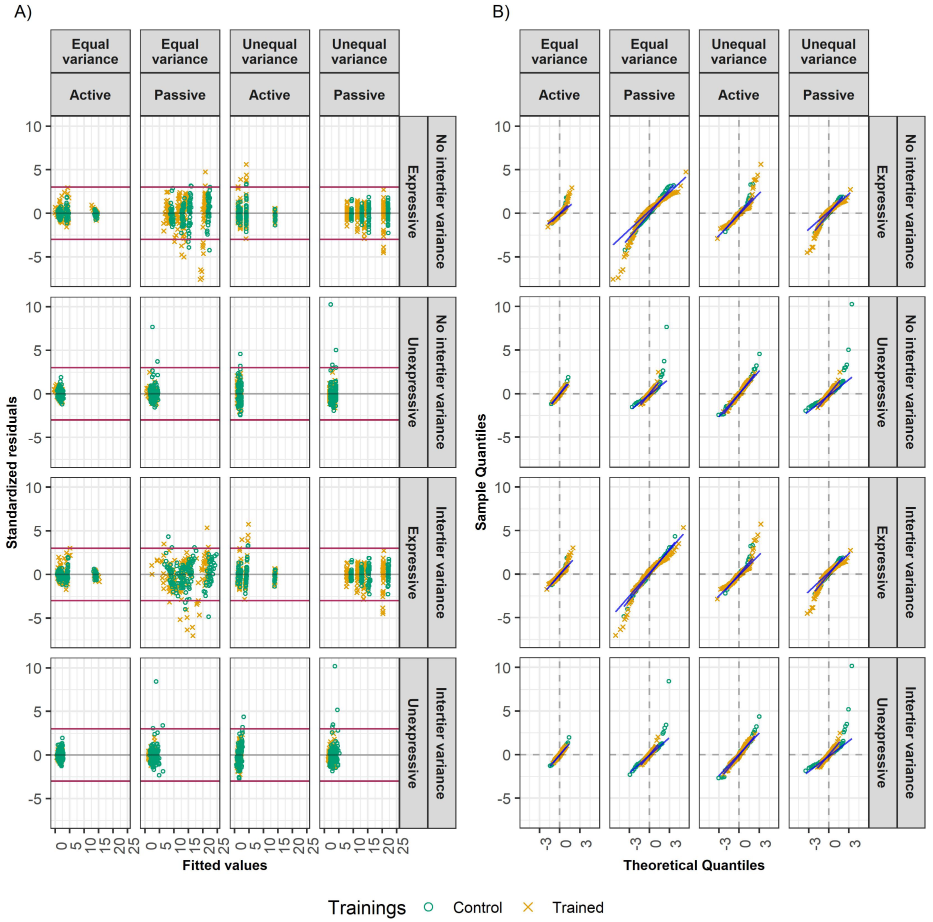

Figure 3. A) Standardized-residuals-versus-fitted-values plots and B) standardized-residual-QQ plots for four models fitted to the data[16] for the Farewell and Herzberg example. The solid horizontal lines at ± 3 standard deviations in A) are used to suggest that points not lying between them are potential outliers.

that R:M:E:P is to have independent errors with variances heterogeneous between the combined levels of Motions and Expressiveness; fit the new model and examine the plots of standardized residuals (Figures 3A and B, upper right quadrants);

c) intertier variances, equal residual variances: Fit the model in (5), change to allow negative estimates of the canonical components, i.e. the model equivalent to that of Jarrett et al.;[5] examine the plots of standardized residuals (Figures 3A and B, lower left quadrants); the R:E and R:E:P canonical components had small negative estimates;

d) intertier variances, unequal residual variances: Include a term in (5) for different residual variances for the combinations of Motions and Expressiveness and fit the model; the canonical components for R:E, R:O and R:O:E were estimated to be zero. Use the AIC to remove, one at a time, any random intertier interactions that did not improve the fit; only the term T:P:E was removed; examine the plots of standardized residuals (Figures 3A and B, lower right quadrants).

Table 3. Results from the analysis of the data[16] for the Farewell and Herzberg example under four models. The models differ as to whether or not random intertier interactions (intertier variances) and whether equal or unequal residual variances were fitted.

A) AIC and log-likelihood for models a)–d).

Model	No. of variance parameters	AIC	REML log-likelihood
a) no intertier variances, equal residual variances	4	3041.31	-1516.65
b) no intertier variances, unequal residual variances	7	2291.71	-1138.85
c) intertier variances, equal residual variances	10	2965.91	-1472.95
d) intertier variances, unequal residual variances	9	2283.26	-1132.63

B) Variance parameter estimates for the model with intertier variances and unequal residual variances, model d).

Term	R	V	P:E	R:P:E	T:M:P:E
Component	0.0018	1.3985	14.5637	0.1868	0.0179

Term	R: idh(M:E):P			
Expressiveness	Expressive		Unexpressive	
Motion	Active	Passive	Active	Passive
Variance	1.0536	13.6863	0.5031	2.1762

C) Wald F-statistics, denominator degrees of freedom (DF) and p-values for the fixed effects under models a)–d).

	Model											
	a) no intertier, equal residual			b) no intertier, unequal residual			c) intertier, equal residual			d) intertier, unequal residual		
Source	DF	F	p	DF	F	p	DF	F	p	DF	F	p
(Mean)	6.0	18.91	0.0048	6.0	18.07	0.0054	6.0	18.90	0.0048	6.0	18.07	0.0054
E	6.0	7.36	0.0350	6.0	7.09	0.0374	6.0	7.34	0.0352	6.0	7.09	0.0374
M	6.0	66.02	0.0002	5.9	65.70	0.0002	6.5	63.31	0.0001	5.9	65.70	0.0002
M # E	6.0	37.77	0.0009	5.9	39.27	0.0008	6.4	36.46	0.0007	5.9	39.30	0.0008
T	72.0	11.38	0.0012	71.9	13.65	0.0004	13.7	6.42	0.0242	9.2	6.89	0.0271
T # E	1092.0	4.77	0.0292	624.1	0.03	0.8695	12.9	1.36	0.2645	9.9	0.32	0.5821
T # M	1092.0	10.22	0.0014	544.6	2.96	0.0860	12.7	2.95	0.1101	11.9	2.58	0.1346
T # M # E	1092.0	8.44	0.0037	481.0	7.96	0.0050	11.5	2.57	0.1357	27.5	5.43	0.0274

The general model-fitting strategy would start with fitting model c) then proceed to model d); models a), b) and d) might be the model fitting sequence if intertier interactions are not posited prior to the analysis. The evidence in Figure 3 is that the residuals from models with unequal variance display a more even vertical spread and so are a better fit. The AIC values in Table 3A indicate that the best fit is obtained for model d). Indeed, the change in the AIC when unequal variances are added to the model is much greater than when intertier interactions are added; the smallest change in the AIC is between models b) and d) that differ in whether intertier interactions are in the model. However, Figure 3 also shows evidence of nonnormality for model d); given the number of observations, this departure from normality should not unduly affect the analysis results. It appears that heterogeneous residual variances should be added to the model proposed by Jarrett et al.[5] The final fitted model, with confounded terms bracketed, is:

$$\begin{aligned}\mathrm{T} * \mathrm{M} * \mathrm{E} \quad | \quad & \mathrm{R} + [\mathrm{V} + \mathrm{O{:}P{:}E} + \mathrm{M{:}P{:}E}] + \mathrm{P{:}E} + \mathrm{R{:}P{:}E} + [\mathrm{T{:}V} + \mathrm{T{:}M{:}P{:}E}] + \\ & [\underline{\mathrm{R{:}V} + \mathrm{R{:}\,idh(M{:}E){:}P}}].\end{aligned} \tag{6}$$

All of the terms that include P:E, except T:P:E, have been retained in the model. For the two sets of confounded terms, it is unclear whether the variation is due to order effects associated with Viewings and/or Occasions or with Motions per se. Even so, the estimates of the variance parameters in Table 3B show that (i) there is considerable variability between patients within each Expressiveness level, and (ii) the variability in the score differences of expressive patients performing passive motions is considerably greater than that of other Expressiveness-Motions combinations. An analysis of the patients' pain ratings from the first phase would allow their contribution to the variability of the rating differences to be gauged.

The Wald F-statistics, their denominator DF and p-values differed between the models (Table 3C). The inclusion of the random intertier interactions had a dramatic effect on the denominator DF. The results under model c) in Table 3C are within rounding errors of those in Table 6 of Jarrett et al.[5] That is, as expected, using mixed model software to fit model c) results in Wald F-statistics, denominator DF based on the Kenward-Rogers approximation[18] and p-values that are equivalent to using combinations of mean squares from an ANOVA and approximate DF based on the Satterthwaite approximation.[24] The only model for which $\mathrm{T} \# \mathrm{M} \# \mathrm{E}$ is not significant is model c). For model c), the fixed-effects model that describes how the Trainings, Motions and Expressiveness affect the response is $\mathrm{T} + \mathrm{M{:}E}$. That is, the Trainings effects do not depend on Motions or Expressiveness. A simulation study is reported in Supplemental Section E; it compares the power of models c) and d) under (6), using values for (i) the expectation similar to the EMMs for model d) and (ii) the variance parameters similar to those in Table 3B. The estimated power to detect $\mathrm{T} \# \mathrm{M} \# \mathrm{E}$ was 0.262 and 0.476 under models c) and d).

Figure 4 exhibits the EMMs for the chosen fixed effects models under models a)–d) and Supplemental Figure 3 shows heat maps of p-values for all pairwise comparisons between them. For all fitted models, except c), the EMMs are the simple means for the eight combinations of Trainings, Motions and Expressiveness. For model c), the simple means were linearly transformed so that they conform to the fixed effects

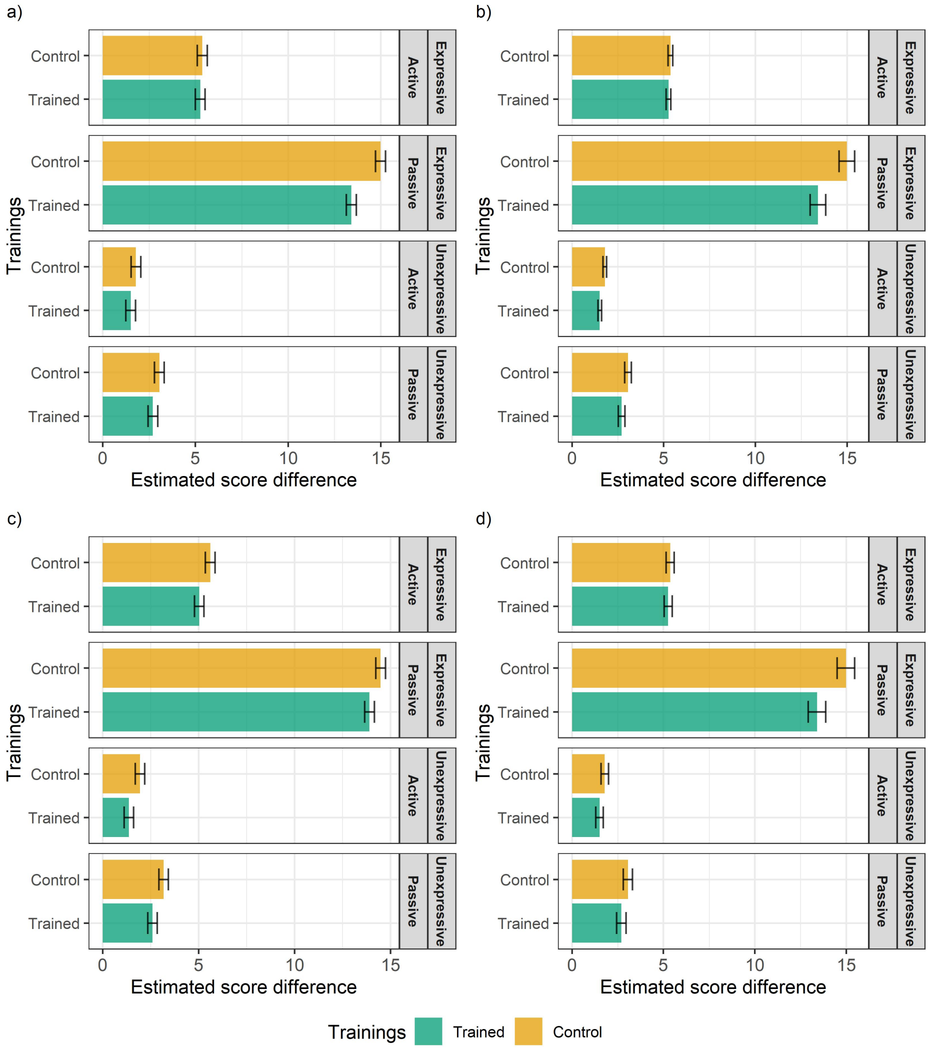

Figure 4. Estimated marginal means for the chosen fixed model under model a) no intertier variances and equal residual variances, b) no intertier variances and unequal residual variances, c) intertier variances and equal residual variances, and d) intertier variances and unequal residual variances, each fitted to the data[16] for the Farewell and Herzberg example. The error bars are ± 0.5 least significant difference for $\alpha = 0.05$ for comparisons within a panel; nonoverlapping bars in a panel indicate that the Trainings are significantly different. Heat maps of the p-values for all pairwise comparisons of the estimated marginal means are in Supplemental Figure 3.

model T + M:E, in a similar manner to that described above for the two-factor additive model A + B. Here, this transformation has the advantage that the standard error of the Trainings difference changes from 0.45 to 0.23. However, the pattern in the EMMs for model c) differs from those for the other models; the Trainings difference is the same for all combinations of Motions and Expressiveness. Further, the widths of their error intervals are the same for all combinations of Motions and Expressiveness, in contrast to the error intervals for model d), the model identified as best for the data. Thus, the conclusions from the EMMs under model c) and d) differ. From Figure 4d and the heat map of p-values for model d), the conclusions are that the EMMs for both untrained and trained raters rating expressive patients undergoing passive motions differs significantly from all other EMMs and were themselves significantly different. However, the Trainings difference was relatively small. It is also the case that the rating for expressive patients undergoing passive motions were the most variable.

4 An alternative design for the second-phase of the pain-rating experiment

In Section 2 it was noted that the plaid design proposed by Farewell and Herzberg[7] involved pseudoreplication. While Farewell and Herzberg[7] acknowledged that there was confounding that would compromise the 'reasonableness of some assumptions', they chose to ignore the confounding. It seems that practical considerations meant that only a single ordering of the videos on a videotape was feasible. For the purposes of investigating the design options, it is assumed that these practical considerations can now be overcome and that the technology to change the presentation order of the videos for each rater is available.

With variation in the presentation order being possible, the fundamental question that will determine the type of design that is best is whether or not there is likely to be an appreciable viewing-order effect. If such an order effect is anticipated then the main effect for Viewings needs to be included in the anticipated model, as it was in the anticipated model in Section 2; otherwise, it would be excluded. An equivalent question is whether or not it is best to honour the intrinsic crossing of Raters and Viewers in the design. An alternative split-unit design, taking into account what has been found from the analysis of the Farewell and Herzberg example, is now explored for the pain-rating scenario, with the factor-allocation paradigm used in deciding and understanding the design.

The anticipated model. As information about a Viewings-order effect is unavailable from the analysis of Farewell and Herzberg example, we continue to anticipate both Ratings and Viewings main effects. However, the Raters component is small and so less Raters can be tolerated. On the other hand, there is considerable patient variability and so increasing the number of patients is likely to be beneficial. To do this, without increasing the number of videos to be rated by each rater, only a randomly chosen subset of the patients are to be rated by a rater. Also the intertier interactions described in Section 3.1 are expected. Because patients are to be randomly divided into subsets, extra variability between patient groups is unlikely. However, as suggested in Section 2, serial correlation between Viewings cannot be ruled out and, unlike for Occasions, the number of viewings is sufficient to estimate it. Thus, (4) gives the anticipated model, except that R:V is replaced by R: $\mathrm{sc}(\mathrm{V})$, where sc means serial correlation such as first-order autoregressive correlation.

Design identification. Given that both Ratings and Viewings main effects are included in the anticipated model, respecting the intrinsic crossing of the two factors will maximize the precision and so a row-column design is appropriate. There will be eight patients, randomly divided into two groups, for each Expressiveness level and the treatments are to be the 16 videos for four patients in a group from each of two Expressiveness levels that are observed on two Occasions. Given 16 treatments, 64 raters is a convenient number of raters. Then, a 16×16 Latin square design for each of the 16 raters in each training level that are to rate the same group of patients can be used to assign videos. While this design has not been optimized for serial correlation, it is a good design even if serial correlation should occur. To randomize this design, its full 64 rows are randomized, along with the associated Trainings levels, to the Raters; its columns are randomized to the Viewings across all Raters. This randomization is the same as for a plaid-square design, but plaid-square designs allocate only one video to each column. A randomized layout generated using the package dae is in Supplemental Figure 2.

The factor-allocation diagram. The factor-allocation diagram in Figure 5 differs from that in Figure 2 in proposing a randomized, nonorthogonal design for allocating the videos.

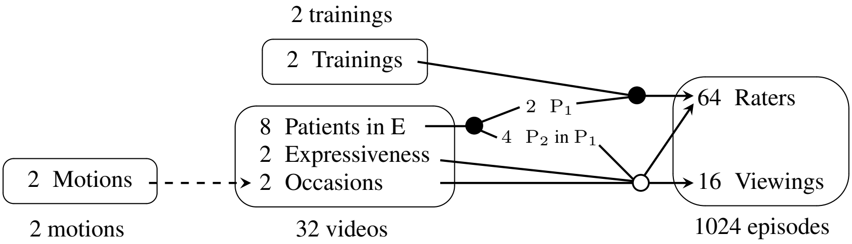

Figure 5. Factor-allocation diagram for a pain-rating experiment in which some raters assess different patients and that employs a row-column design in the second-phase: motions are allocated to videos and both trainings and videos are allocated to episodes; the line running from Patients in E to the solid circle (●) and the two lines running to P_1 and P_2 indicate that Patients has been split into two pseudofactors, P_1 and P_2, to create two groups of four patients each; the lines running from Trainings and P_1 to the solid circle (●) and the arrow running to Raters mean that the combined levels of Trainings and P_1 are randomized to Raters; similarly the lines running from P_2, Expressiveness and Occasions to the open circle (○) and the arrows running to Ratings and Viewings imply that P_2, Expressiveness and Occasions are jointly randomized to the combinations of Ratings and Viewings using a nonorthogonal design; the dashed arrow indicates that the allocation of motions is systematic; E = Expressiveness.

The initial and homogeneous allocation models. The initial allocation model remains the model in (2). However, the presence of intertier interactions in the anticipated model prompts for their inclusion in the homogeneous allocation model. Thus, the homogeneous allocation model is taken to be the same as the anticipated model in (4).

Tiered decomposition and skeleton-ANOVA tables. Given the difference in the initial and homogeneous models for this design, the anatomy for the simpler initial allocation model is obtained before that for the homogeneous allocation model. The Supplemental R script in the file AltRowColDesign.r uses designAnatomy to produce both tiered decomposition tables, as well as tables for some component designs.[4]

Table 4. The tiered skeleton-ANOVA table, under the initial allocation model, for a two-phase design for a pain-rating experiment that employs a nonorthogonal row-column design in the second-phase: R = Raters; V = Viewings; E = Expressiveness; P = Patients; O = Occasions; T = Trainings; M = Motions; ; EMS = Expected Mean Square; DF = Degrees of Freedom; Eff = A-efficiency criterion.

episodes		videos			trainings-motions		EMSs†								
Source	DF	Eff†	Source‡	DF	Source	DF	ϕ_{RV}	ϕ_{R}	ϕ_{V}	ϕ_{OPE}	ϕ_{PE}	ϕ_{OE}	ϕ_{O}	ϕ_{E}	$\theta.$
Mean	1	1.00	Mean	1	Mean	1	1	16	64	32	64	256	512	512	θ_{μ}
R	63	1.00	$\mathrm{P}\,[\mathrm{E}]_1$	1	T	1	1	16			64				
		1.00	Residual	62	T	1	1	16							θ_{T}
					Residual	61	1	16							
V	15	0.50	$\mathrm{P}\,[\mathrm{E}]_2$	1			1		64	$\frac{32}{2}$	$\frac{64}{2}$				
		0.50	$\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_2$	1			1		64	$\frac{32}{2}$					
			Residual	13			1		64						
R # V	945	1.00	O	1	M	1	1			32		256	512		θ_{M}
		1.00	E	1			1			32	64	256		512	
		0.50	$\mathrm{P}\,[\mathrm{E}]_2$	1			1			$\frac{32}{2}$	$\frac{64}{2}$				
		0.50	$\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_2$	1			1			$\frac{32}{2}$					
		1.00	$\mathrm{P}\,[\mathrm{E}]_{\vdash}$	12			1			32	64				
		1.00	O # E	1		1	1			32		256			
		1.00	$\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_1$	1			1			32					
		1.00	$\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_{\vdash}$	12			1			32					
		1.00	Residual	915	T # M	1	1								θ_{TM}
					Residual	914	1								
Total	1024														

† Each Eff is the A-efficiency criterion, being the harmonic mean of the canonical efficiency factors, for the videos source when it is confounded with the episodes source aligned with or immediately to the left and above it. No efficiencies are shown for trainings-motions because the efficiencies with videos are all one and those with episodes are the same as for videos.

‡ The 14-dimensional sources $\mathrm{P}\,[\mathrm{E}]$ and $\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]$ are each split into three orthogonal sources: two sources with 1 DF and one with 12 DF. $\mathrm{P}\,[\mathrm{E}]_1$ and $\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_1$ are the one-dimensional subspaces that are the contrast between the two groups of Raters and its interaction with Occasions. $\mathrm{P}\,[\mathrm{E}]_2$ ($\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_2$) is the one-dimensional subspace that is the interaction of the group contrast with Expressiveness (and with Occasions). $\mathrm{P}\,[\mathrm{E}]_{\vdash}$ is the 12-dimensional subspace orthogonal to the two one-dimensional $\mathrm{P}\,[\mathrm{E}]$ subspaces; it corresponds to the differences between the 4 patients within each of the two patient groups for each level of Expressiveness and so its DF are calculated as $(4-1) \times 2 \times 2$. $\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]_{\vdash}$ is the subspace of $\mathrm{O} \# \mathrm{P}\,[\mathrm{E}]$ orthogonal to its one-DF subspaces; its DF are $(2-1) \times (4-1) \times 2 \times 2$.

The anatomy for the proposed design, for the terms in the initial allocation model in (2), is presented in Table 4. This design is not orthogonal and so is in the class of multiphase designs covered by Brien.[4] However, it is structure-balanced because, as the anatomy obtained using the R script shows, the number of unique, nonzero efficiency factors, their order, is one for all videos sources. For this to occur, the pseudofactor indexing the groups of Patients had to be included in the formula for videos-episodes. Being structure-balanced, EMSs have been added to Table 4 using the rules outlined in Supplemental Section C. This anatomy shows that one DF of P [E] (P $[\mathrm{E}]_1$) is confounded with Raters, it being the contrast between the two patient groups assigned to the different Raters. Another one DF of P [E] (P $[\mathrm{E}]_2$) is partially confounded with both V and R $\#$ V and the remaining 12 DF is confounded with R $\#$ V. The partial confounding of P $[\mathrm{E}]_2$ results from the incorporation of the term V into the model and would not occur if Viewings was nested within Raters, but with the consequence that a Viewings-order effect would not be taken into account. The confounding of T remains orthogonal and M exhaustively confounds O, there being no Residual for O. Nonetheless, the design has very good properties; in particular, viewing order little affects the conclusions about the factors of interest.

The anatomy for the proposed design, for the terms in the homogeneous allocation model in (4), is in Table 5. To produce this anatomy, the pseudofactor P_1 was omitted because pseudofactors do not represent real sources of random variation. Thus, sources involving P [E] are not always partitioned; for example, the three sources for O $\#$ P [E] in Table 4 that are confounded with the R $\#$ V Residual are combined in Table 5. For the row-column design, as compared to the plaid-square design, P [E] and O $\#$ P [E] have more DF, but the intertier interactions with R have less DF because (i) there are less raters and (ii) raters do not rate all patients. Further, the interactions with R have lost 30 of their 930 DF because of the need to accommodate the 30 DF for the video sources within in the 930 DF for the R $\#$ V Residual. The smallest A-efficiency for these interactions, when confounded with the R $\#$ V Residual, is for R $\#$ P [E] at 0.73. Even so, the properties of the design continue to be good; all fixed effects are estimated with high efficiency and will be little affected by Viewing order and its variation between Trainings. While the Residual DF for the random intertier interactions are reduced, they remain large and so the reduction is manageable. The results of the power simulation study confirm the superiority, for detecting T $\#$ M $\#$ E, of the row-column design analyzed with model d) over the other design and model combinations, when V is a source of variability; its estimated power is 0.590. This, in spite of having 10 less raters, and so 160 less observations, than the plaid design.

The exhaustive confounding of fixed sources in this design and the design in Section 2 are the same: M with O and M $\#$ E with O $\#$ E. However, for the random sources, the magnitude of the V and the T:V variability are now estimable. Yet, the variance matrix for the homogeneous allocation model is still singular because the following pairs of sources continue to be inextricably confounded: (i) M $\#$ P [E] with O $\#$ P [E], (ii) O $\#$ R $\#$ P [E] with R $\#$ V and (iii) M $\#$ R with O $\#$ R.

The prior allocation model. Removing the terms M:P:E, R:O:P:E and R:M from the homogeneous allocation model, which is the same as (4), results in the nonsingular prior

Table 5. The tiered decomposition table, under the homogeneous allocation model, for a two-phase design for a pain-rating experiment that employs a nonorthogonal row-column design in the second-phase: R = Raters; V = Viewings; E = Expressiveness; P = Patients; O = Occasions; T = Trainings; M = Motions; DF = Degrees of Freedom; Eff = A-efficiency criterion.

episodes		training-episodes		videos-episodes			trainings-motions-videos-episodes		
Source	DF	Source	DF	Eff†	Source‡	DF(Max¶)	Eff†	Source‡	DF
Mean	1	Mean	1	1.00	Mean	1	1.0000	Mean	1
R	63	T	1						
		Residual	62	1.00	P [E]$_1$	1			
					Residual	61		T # P [E]$_1$	1
								Residual	60
V	15			0.50	P [E]$_2$	1			
				0.50	O # P [E]$_1$	1	0.50	M # P [E]$_1$	1
				1.00	O # R$_1$	1	**1.00**	M # R$_1$	1
				0.15	R # P [E]$_1$	12	1.00		
R # V	945	T # V	15	1.00	O # R$_2$	1	1.00	M # R$_2$	1
				0.50	E # R$_1$	2	0.50	T # P [E]$_2$	1
							0.50	T # M # P [E]$_1$	1
				0.15	R # P [E]$_2$	12			
		Residual	930	**1.00**	O	1(1)	**1.00**	M	1
				1.00	E	1(1)			
				0.93	P [E]$_3$	13(14)			
				1.00	O # E	1(1)	**1.00**	M # E	1
				0.93	O # P [E]	14(14)	**0.93**	M # P [E]	14
				1.00	O # R$_\vdash$	60(62)	**1.00**	T # M	1
							1.00	T # M # P [E]$_2$	1
							1.00	M # R$_\vdash$	58
				0.97	E # R$_2$	61(62)	**1.00**	T # E	1
							0.50	T # P [E]$_2$	1
							0.50	T # M # P [E]$_1$	1
								Residual	58
				0.73	R # P [E]	372(372)	**1.00**	T # P [E]$_\vdash$	12
								Residual	360
				1.00	O # E # R$_\vdash$	59(62)	**1.00**	T # M # E	1
								Residual	58
				1.00	O # R # P [E]$_\vdash$	348(372)	**1.00**	T # M # P [E]$_\vdash$	12
								Residual	336
Total	1184								

† Each Eff is the A-efficiency criterion, being the harmonic mean of the canonical efficiency factors, for a source when it is confounded with the sources to its left. Those for the videos-episodes sources are obtained from an anatomy that omits the last column of sources. All efficiencies for trainings-episodes are one. Bolded efficiencies have the greatest portion of the information for the full source.

‡ Sources with the same label, but different subscripts, are different subspaces of the full source; a source with the subscript $\vdash$ is the subspace of the full source orthogonal to other subspaces with the same label and other sources with which it is aliased; if a full source appears in the table, it is not subscripted.

¶ The maximum DF for each source confounded with the R # V Residual. For R # P [E], the interaction between R and P can only be estimated from within each patient group and Expressiveness level and so the maximum DF are $(32-1)\times(4-1)\times 2\times 2$, there being 32 Raters that rate each patient group.

allocation model in (7). It includes the terms O:P:E and T:V, along with serial correlation, that were not in (5) because they are inestimable with the Farewell and Herzberg design.

$$\mathrm{T} * \mathrm{M} * \mathrm{E} \quad | \quad \begin{aligned} &\mathrm{P{:}E} + \mathrm{O{:}P{:}E} + \mathrm{R} + \mathrm{V} + \underline{\mathrm{R{:}\,sc(V)}}+ \\ &\mathrm{T{:}P{:}E} + \mathrm{T{:}M{:}P{:}E} + \mathrm{T{:}V} + \mathrm{R{:}O} + \mathrm{R{:}E} + \mathrm{R{:}O{:}E} + \mathrm{R{:}P{:}E} \end{aligned} \tag{7}$$

If it happens that, in analyzing data from an experiment utilizing this design, it becomes apparent that variances that vary with Motions and Expressiveness and serial correlation between Viewings are required, then it may be difficult to identify software that is capable of specifying such a model. It is possible with asreml.[19]

5 Discussion

The purpose of the paradigm in Figure 1 is to provide a framework for designing experiments, understanding them and formulating potential models for them. It begins with the anticipated model that informs the choice of a design and that, in turn, leads via increasingly complicated models to the formulation of a linear mixed model, called the prior allocation model. The call by Jarrett et al.[5] to include all ”appropriate error terms” is accommodated in the paradigm by inviting the user of it to consider, even at the design stage, the need to modify the initial allocation model so that the homogeneous allocation model includes intertier interactions. The advantage of considering this need at the design stage is that, if they are likely, then they can be allowed for in choosing a design, this being most beneficial when a nonorthogonal design is likely to be required.

Also recommended is that a design's anatomy, as exhibited in a tiered decomposition or skeleton-ANOVA table, be examined at the design stage, i.e. before the data is available, firstly for the relatively simple, initial allocation model and then for the homogeneous allocation model, with terms grouped into formulae according to the tiers/panels with which they are associated. Admittedly, some effort is required to become accustomed to the use of the paradigm, but the practical benefits that accrue from the examination of a design's anatomy is the dividend. Also, the dae[12] functions, by deriving the decomposition tables, lessen the effort. A key adjustment is to always include all of the recipient factors in an experiment, rather than just those that are associated with the blocking. The practical benefits include (i) it allows the designer to check the randomized layout for a design, (ii) it provides an understanding of the confounding inherent in the design, which can be used to assess the appropriateness of the design for an experiment, (iii) the inclusion of all factors relevant to the design in its description gives the designer confidence that all terms relevant to the design have been incorporated into the model and reveals their effects on the analysis, (iv) it can result in a richer insight into the sources of variability that contribute to the variation in the experiment, as in (6), and (v) using the paradigm ahead of deploying a design often raises issues that might otherwise be overlooked. On the other hand, the commonly used single-set description,[3] because the confounding relationships between sources of variation accounted for in design is not described, lacks these benefits. Useful insights into the properties of component designs[4] can also be gleaned from their anatomies. For example the anatomy of the first-phase design is relevant when the first-phase will yield data.

In seeking to understand a design via ANOVA, the paradigm used here is similar to the approach taken by Jarrett et al.[5] However, it differs from their approach in that they formulate an initial model based on a single set of factors that the analyst proposes to include in the analysis, with the terms formed from the nesting and crossing relations between these factors; this contrasts with our approach of founding model formulation on the allocations by dividing the factors into sets, called tiers. Also, they prefer to begin by deriving a simpler model for the data collapsed over a suitably chosen factor, Motions in their example. This model is then expanded to a model for the full data set. Our approach is to start with a model without the intertier interaction terms and then extend this to add these interactions. While the two approaches differ, they lead to equivalent decompositions of the data space for the Farewell and Herzberg example, albeit with different labelling and availability of confounding information.

The need for intertier interaction raises the question of whether they should be included in the homogeneous allocation model. Two extreme positions are: (i) always assume intertier additivity and include no intertier interactions, and (ii) include all possibly estimable intertier interactions as a matter of course. The advantage of the first position is that the models obtained are, when all allocations are randomizations, randomization models and randomization tests of hypotheses are possible. The approach that has been taken here is the latter because, as Jarrett et al.[5] argue, when intertier interactions occur in an experiment they can affect the results of its analysis. The strategy used in the analysis here was to omit random intertier interactions that did not improve the model by lowering the AIC. On the other hand, terms in the initial allocation model are not omitted from the model, except when they are random terms that are estimated to be zero. Such omissions can be minimized by allowing negative estimates of canonical components. This retains any terms justified by the allocations, including randomizations, in the model. Also, fixed terms were never omitted from a model to protect against the effects of Type II errors, but EMMs were made to conform to the chosen fixed model.

The data analysis reported here has shown that allowing for heterogeneous residual variance can be even more important than the inclusion of intertier interactions. It may also be informative to analyze the first-phase self-assessed ratings of the patients, so as to characterize their variability. Potential design improvements are to (i) if possible, reorder the videos for different raters, using a row-column design for the second-phase when viewing-order effects are anticipated and a nested design when they are not, and (ii) increase the patient numbers by assigning different groups of patients to the raters.

Declaration of conflicting interests

The author declares no conflict of interest.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of data and code for replication

The data for analysis can be accessed from the figshare repository at https://doi.org/10.25909/13122095. R scripts and their output for producing the designs and analyzing the data are available in the Supplemental material.

Supplemental material

Supplemental material is available online.

References

1. Brien CJ. Multiphase experiments in practice: A look back. *Aust N Z J Stat* 2017; 59: 327–352. URL https://doi.org/10.1111/anzs.12221/.
2. Walwyn R and Roberts C. Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. *Stat Methods Med Res* 2010; 19: 291–315. URL https://doi.org/10.1177/0962280209105017/.
3. Brien CJ, Harch BD, Correll RL et al. Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *J Agric Biol Environ Stat* 2011; 16: 422–450. URL https://doi.org/10.1007/s13253-011-0060-z/.
4. Brien CJ. Multiphase experiments with at least one later laboratory phase. II. Northogonal designs. *Aust N Z J Stat* 2019; 61: 234–268. URL https://doi.org/10.1111/anzs.12260/.
5. Jarrett RG, Farewell VT and Herzberg AM. Random effects models for complex designs. *Stat Methods Med Res* 2020; 29: 3695–3706. URL https://doi.org/10.1177/0962280220938418.
6. Solomon PE, Prkachin KM and Farewell V. Enhancing sensitivity to facial expression of pain. *Pain* 1997; 71: 279–284. URL https://www.sciencedirect.com/science/article/pii/S0304395997033770/.
7. Farewell VT and Herzberg AM. Plaid designs for the evaluation of training for medical practitioners. *J Appl Stat* 2003; 30: 957–965. URL https://doi.org/10.1080/0266476032000076092/.
8. Brien CJ. Multitiered experiments web site, 2001–21, accessed April 4, 2021. URL http://chris.brien.name/multitier/.
9. Brien CJ and Bailey RA. Multiple randomizations (with discussion). *J R Stat Soc Series B Stat Methodol* 2006; 68: 571–599. URL https://doi.org/10.1111/j.1467-9868.2006.00557.x/.
10. Prkachin KM and Mercer SR. Pain expression in patients with shoulder pathology: validity, properties and relationship to sickness impact. *Pain* 1989; 39: 257–265. URL https://doi.org/10.1016/0304-3959(89)90038-9/.
11. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021, accessed March 22, 2021. URL https://www.r-project.org/.
12. Brien CJ. *dae: functions useful in the design and ANOVA of experiments. R package*

version 3.1-37, 2021, accessed March 19, 2021. URL https://CRAN.R-project.org/package=dae/.

13. Yates F. The design and analysis of factorial experiments. *Imperial Bureau of Soil Science Technical Communication* 1937; 35.
14. Wilkinson GN and Rogers CE. Symbolic description of factorial models for analysis of variance. *J R Stat Soc Ser C Appl Stat* 1973; 22: 392–399. URL https://doi.org/10.2307/2346786/.
15. Fisher RA. *The Design of Experiments*. 1st ed. Edinburgh: Oliver and Boyd, 1935.
16. Brien CJ. *Data for the Farewell and Herberg example of a two-phase experiment using a plaid design*, 2021. URL https://doi.org/10.25909/13122095.
17. Brien CJ and Demétrio CGB. Formulating mixed models for experiments, including longitudinal experiments. *J Agric Biol Environ Stat* 2009; 14: 253–280. URL https://doi.org/10.1198/jabes.2009.08001/.
18. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; 53: 983–997. URL https://doi.org/10.2307/2533558/.
19. Butler DG, Cullis BR, Gilmour AR et al. *ASReml-R reference manual. Version 4*, 2020, accessed September 27, 2020. URL https://asreml.org/.
20. Brien CJ. *asremlPlus: Augments 'ASReml-R' in Fitting Mixed Models and Packages Generally in Exploring Prediction Differences. R package version 4.2-32*, 2021, accessed March 22, 2021. URL https://CRAN.R-project.org/package=asremlPlus/.
21. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974; 19: 716–723. URL https://dx.doi.org/10.1109/TAC.1974.1100705/.
22. Matuschek H, Kliegl R, Vasishth S et al. Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 2017; 94: 305–315. URL https://doi.org/10.1016/j.jml.2017.01.001/.
23. Searle SR, Speed FM and Milliken GA. Population marginal means in the linear model: an alternative to least squares means. *Am Stat* 1980; 34: 216–221. URL https://doi.org/10.2307/2684063/.
24. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics* 1946; 2: 110–114. URL https://doi.org/10.2307/3002019/.