frontiers in Neuroscience

REVIEW
published: 17 April 2019
doi: 10.3389/fnins.2019.00352

Modeling Group-Level Repeated Measurements of Neuroimaging Data Using the Univariate General Linear Model

*Martyn McFarquhar**

Division of Neuroscience & Experimental Psychology, University of Manchester, Manchester, United Kingdom

Group-level repeated measurements are common in neuroimaging, yet their analysis remains complex. Although a variety of specialized tools now exist, it is surprising that to-date there has been no clear discussion of how repeated-measurements can be analyzed appropriately using the standard general linear model approach, as implemented in software such as SPM and FSL. This is particularly surprising given that these implementations necessitate the use of multiple models, even for seemingly conventional analyses, and that without care it is very easy to specify contrasts that do not correctly test the effects of interest. Despite this, interest in fitting these types of models using conventional tools has been growing in the neuroimaging community. As such it has become even more important to elucidate the correct means of doing so. To begin, this paper will discuss the key concept of the *expected mean squares* (EMS) for defining suitable *F*-ratios for testing hypotheses. Once this is understood, the logic of specifying correct repeated measurements models in the GLM should be clear. The ancillary issue of specifying suitable contrast weights in these designs will also be discussed, providing a complimentary perspective on the EMS. A set of steps will then be given alongside an example of specifying a 3-way repeated-measures ANOVA in SPM. Equivalency of the results compared to other statistical software will be demonstrated. Additional issues, such as the inclusion of continuous covariates and the assumption of sphericity, will also be discussed. The hope is that this paper will provide some clarity on this confusing topic, giving researchers the confidence to correctly specify these forms of models within traditional neuroimaging analysis tools.

Keywords: repeated measurements, within-subject, flexible factorial, SPM, FSL, GLM

OPEN ACCESS

Edited by:
Jean-Baptiste Poline,
University of California, Berkeley,
United States

Reviewed by:
Gang Chen,
National Institutes of Health (NIH),
United States
Martin A. Lindquist,
Johns Hopkins University,
United States

***Correspondence:**
Martyn McFarquhar
martyn.mcfarquhar@manchester.ac.uk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *17 July 2018*
Accepted: *27 March 2019*
Published: *17 April 2019*

Citation:
McFarquhar M (2019) Modeling Group-Level Repeated Measurements of Neuroimaging Data Using the Univariate General Linear Model. Front. Neurosci. 13:352. doi: 10.3389/fnins.2019.00352

1. INTRODUCTION

The modeling of group-level repeated measurements is a common, yet complex, topic in neuroimaging. Although a variety of specialized tools are now available (e.g., Chen et al., 2014; Guillaume et al., 2014; McFarquhar et al., 2016), it is surprising that to-date there has been no clear discussion of how researchers can analyse repeated measurements using the traditional voxel-wise

general linear model (GLM) approach, implemented in software such as SPM (http://www.fil.ion.ucl.ac.uk) and FSL (https://fsl.fmrib.ox.ac.uk)[1]. Despite the implication from some authors (e.g., McLaren et al., 2011; Chen et al., 2014), the analysis of complex repeated-measurement designs *is* possible within these software packages. However, because they have often not been designed with these analyses in mind, specification of the correct models can be difficult. For instance, the traditional modeling of repeated measurements requires the inclusion of subject and all possible interactions with subject as factors in the model, as well as specifying multiple models to force the use of appropriate error terms for the *F*-ratios. Ignorance of the correct way to model these effects would, at best, lead to an analysis that lacked sensitivity, but at worst would lead to an analysis with an increased Type I error rate and *F*-ratios that do not test the intended model effects.

For researchers who are less familiar with classical linear model theory, some of the requirements of repeated-measurement models can seem esoteric. However, these models are based on the key statistical concept of *expected mean squares* (EMS) which, once understood, should make the logic behind these methods clear. Given the importance of the EMS for understanding the logic of hypothesis tests in repeated measures models, this paper will begin with a detailed exposition of the concept. No claims of originality are made on this exposition as these are issues well-known in the statistical literature. However, the degree of confusion surrounding these models in the neuroimaging literature has prompted an explicit discussion of this core statistical concept. An ancillary issues, in the form of *estimable functions* in overparameterized ANOVA models, will also be discussed. Once these concepts are understood, it should become clear how these models need to be treated in neuroimaging software. To that end, a set of simple steps will be given alongside an example of how to specify a 3-way repeated-measures ANOVA model in SPM12. Some further issues with repeated measurement models, such as the inclusion of continuous covariates and the assumption of sphericity, will also be discussed. The hope is that these discussions will provide some clarity on this complex topic, giving researchers the knowledge and understanding needed to confidently use these models within familiar software packages.

2. EMS IN ANOVA MODELS

In order to begin understanding the requirements of repeated measurement models when implemented in the GLM, the concept of EMS in ANOVA models must be understood. A basic aim of any ANOVA model is to split the data into different sources of variation. These sources of variation are formalized in terms of the calculation of *sums-of-squares* for each model component, which are converted to *mean squares* using the degrees of freedom of the model terms. The *F*-ratios are then formed by dividing a suitable mean square for the effect of interest by a suitable mean square for the error. In order to understand the logic of these *F*-ratios, it is necessary to consider the expected value for the mean squares. The EMS represents the theoretical mean of the sampling distribution of each of the mean squares and take the form of the addition of several sources of variation. In order to test a specific effect, the *F*-ratio should be formed from two terms whose EMS differ only by the source of variation associated with that effect. Under the null hypothesis that the effect of interest is 0, the magnitudes of the two mean squares should be similar and the *F*-ratio should be close to 1. As such, the larger the discrepancy between the mean squares the larger the *F*-ratio and the greater the evidence against the null. This logic of constructing ANOVA tests is central to the ANOVA methodology, but is also one of the major sources of misunderstanding when attempting to construct tests of effects in repeated-measurement models. As such, this first section aims to describe the derivation of the EMS and their use in forming meaningful hypothesis tests.

2.1. EMS in Between-Subject Designs

To understand the use of EMS in deriving suitable *F*-ratios, consider a balanced two-way between-subjects ANOVA model with n observations per-cell

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{1}$$

where μ is the overall mean, α_i is the effect of the ith level of factor A ($i = 1, \ldots, a$), β_j is the jth level of factor B ($j = 1, \ldots, b$), $(\alpha\beta)_{ij}$ is the ij interaction effect and ϵ_{ijk} is random error ($k = 1, \ldots, n$) assumed $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

For this basic ANOVA design, the correct error term for the omnibus main effects and interaction is the model variance σ^2. To see why this is the case, we can calculate the EMS. Although possible to derive the EMS formally through the use of the expectation operator, a more practical approach involves following some basic rules. In this paper, the rules given by Kutner et al. (2004) are used. As an example, **Table 1** gives an outline of the arithmetic involved in constructing the EMS for the model in Equation (1). These tables are provided throughout this paper to give direct correspondence between the method of Kutner et al. (2004) and the eventual forms of the EMS used to derive appropriate *F*-ratios.

Using **Table 1**, the EMS for the terms in Equation (1) are given in Equation (2). Construction of an appropriate *F*-ratio involves using the EMS to identify two mean squares which differ only by the effect of interest. For instance, to test the effect of factor A we must identify which EMS in Equation (2) differs only from EMS_A by $bn\frac{\sum \alpha_i^2}{a-1}$. In this instance, the only choice is EMS_E. As such, a test for the effect of A can be constructed using the ratio of the mean square of A (MS_A) and the mean square of the errors (MS_E). Continuing in this fashion, suitable *F*-ratios for all the model effects can be derived, as given in **Table 2**. This confirms the initial statement that a suitable denominator for all tests from

[1]It is notable that AFNI (https://afni.nimh.nih.gov) has facilities included to accommodate random effects in the group-level ANOVA programs 3dANOVA2 and 3dANOVA3, as well as facilities for multivariate approaches to repeated-measurements in 3dMVM. As such, the discussions in the paper are less relevant for AFNI users.

the model in Equation (1) is the overall error term.

$$\begin{aligned} \text{EMS}_A &= \sigma^2 + bn\frac{\sum \alpha_i^2}{a-1} \\ \text{EMS}_B &= \sigma^2 + an\frac{\sum \beta_j^2}{b-1} \\ \text{EMS}_{AB} &= \sigma^2 + n\frac{\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)} \\ \text{EMS}_E &= \sigma^2 \end{aligned} \tag{2}$$

2.2. EMS in Mixed-Measures Designs

For between-subject designs containing only fixed-effects it is rarely necessary to calculate the EMS as a suitable denominator for the *F*-ratios is always given by the overall error term. The situation becomes quite different when considering ANOVA models containing random-effects. Although not usually applicable in neuroimaging for between-subjects designs, when within-subject and mixed within-subject and between-subjects (mixed-measures) designs are considered, it is usually desirable to include *subject*, as well as all possible interactions with *subject*, as random-effects. In doing so, the structure of the EMS changes and the derivation of a suitable error term for testing hypotheses about particular effects becomes more complex.

2.2.1. A Single Within-Subject Factor

To begin, consider a basic mixed-measures design containing a single within-subject factor and a single between-subjects factor. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + S_{k(j)} + \epsilon_{ijk} \tag{3}$$

where μ is the overall mean, α_i is the effect of the ith level of the within-subject factor ($i = 1, \ldots, a$), β_j is the effect of the jth level of the between-subjects factor ($j = 1, \ldots, b$) and $(\alpha\beta)_{ij}$ is the ij interaction effect. $S_{k(j)}$ is the random effect of the kth subject ($k = 1, \ldots, n$) assumed $S_{k(j)} \sim \mathcal{N}(0, \sigma_s^2)$ and ϵ_{ijk} is random error assumed $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. The notation $S_{k(j)}$ indicates that the kth subject is *nested* within group j. This conveys the fact that, for example, $k = 1$ refers to a different subject depending on the value of j (see Chapter 26 in Kutner et al., 2004).

One of the key differences between the model in Equation (1) and the model in Equation (3) is the inclusion of the random subject effects $S_{k(j)}$. Although possible to forego the subject effects and work with a pooled error term (Penny and Henson, 2007), doing so produces tests which are more conservative (see Casella, 2008, p. 85). As such, the inclusion of $S_{k(j)}$ allows one to *partition* the model errors in order produce more sensitive tests of the model effects. Some intuition can be gained here by re-writing Equation (3) as

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{k(j)}^{(1)} + \epsilon_{ijk}^{(2)} \tag{4}$$

where the splitting of the singular error term is now more explicit. The complication for the traditional neuroimaging GLM framework is that the error term used as the denominator for the test statistics is derived implicitly from the difference between the data and the model prediction. This means that for the model in Equation (4), only $\epsilon_{ijk}^{(2)}$ will be used in the formation of the test statistics. Furthermore, in order to correctly derive the final error term, $\epsilon_{k(j)}^{(1)}$ must also be included in the design matrix, despite its status as a random-effect.

As with before, the breakdown of the arithmetic in **Table 3** gives the calculation of the EMS for the mixed-measures ANOVA model given in Equation (3). The final EMS are given in Equation (5) with suitable ratios for testing the main effects and interactions given in **Table 4**. Of particular importance here is to recognize how both the A and A × B effects use the overall error term, but that a suitable *F*-ratio for the effect of B requires the use of MS_S as the denominator instead. As discussed above, only MS_E will be used as the denominator in neuroimaging software implementing the traditional GLM approach. Testing of the effect of B therefore requires specifying a separate model where the final error term is forced to become MS_S. This can be achieved by averaging the raw data over the levels of the within-subject factor, and will be discussed in more detail in the example analysis given at the end of this paper.

$$\begin{aligned} \text{EMS}_A &= \sigma^2 + bn\frac{\sum \alpha_i^2}{a-1} \\ \text{EMS}_B &= \sigma^2 + a\sigma_s^2 + an\frac{\sum \beta_j^2}{b-1} \\ \text{EMS}_{AB} &= \sigma^2 + n\frac{\sum (\alpha\beta)_{ij}^2}{(a-1)(b-1)} \\ \text{EMS}_S &= \sigma^2 + a\sigma_s^2 \end{aligned} \tag{5}$$

TABLE 1 | Arithmetic for the derivation of the EMS in a 2-way between-subjects ANOVA model, using the method of Kutner et al. (2004).

	i F *a*	*j* F *b*	*k* R *n*	Variance	EMS_A *i*	EMS_B *j*	EMS_{AB} *ij*	EMS_E *k(ij)*
α_i	0	b	n	σ_α^2	bn	0	0	0
β_j	a	0	n	σ_β^2	0	an	0	0
$(\alpha\beta)_{ij}$	0	0	n	$\sigma_{\alpha\beta}^2$	0	0	n	0
$\epsilon_{k(ij)}$	1	1	1	σ^2	1	1	1	1

TABLE 2 | The numerator and denominator mean squares from Equation (2) used to form appropriate *F*-tests for the model in Equation (1).

Effect	Test
Factor A	MS_A/MS_E
Factor B	MS_B/MS_E
A × B	$\text{MS}_{AB}/\text{MS}_E$

TABLE 3 | Arithmetic for the derivation of the EMS in a 2-way mixed ANOVA with a single within-subject and a single between-subjects factor.

	i **F** ***a***	***j*** **F** ***b***	***k*** **R** ***n***	**Variance**	$\mathbf{EMS}_A$ ***i***	$\mathbf{EMS}_B$ ***j***	$\mathbf{EMS}_{AB}$ ***ij***	$\mathbf{EMS}_S$ ***k(j)***	$\mathbf{EMS}_E$ ***k(ij)***
α_i	0	b	n	σ^2_α	bn	0	0	0	0
β_j	a	0	n	σ^2_β	0	an	0	0	0
$(\alpha\beta)_{ij}$	0	0	n	$\sigma^2_{\alpha\beta}$	0	0	n	0	0
$S_{k(j)}$	a	1	1	σ^2_s	0	a	0	a	0
$\epsilon_{k(ij)}$	1	1	1	σ^2	1	1	1	1	1

TABLE 4 | The EMS ratios used to form appropriate F-tests for the main effects and interactions in a 2-way mixed-measures ANOVA.

Effect	**Test**
A	MS_A /MS_E
B	MS_B /MS_S
A × B	MS_{AB} / MS_E

$$\text{EMS}_E = \sigma^2$$

2.2.2. Multiple Within-Subject Factors

The situation with multiple error terms becomes more complex as the number of within-subject factors increases. Consider adding another within-subject factor to the model in Equation (3). This produces a 3-way mixed-measures ANOVA model, which can be written as

$$\begin{aligned} y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ + (\alpha\beta\gamma)_{ijk} + S_{l(k)} + (S\alpha)_{il(k)} + (S\beta)_{jl(k)} + \epsilon_{ijkl} \end{aligned} \quad (6)$$

where β_j is now the effect of the jth level of the additional within-subject factor, γ_k is the effect of the kth level of the between-subjects factor ($k = 1, \ldots, c$) and the subject effects are indexed by $l = 1, \ldots, n$. In comparison to the model in Equation (3), inclusions of additional within-subject factors provides an opportunity for further interactions with the subject effects. These are given by the interaction with the first within-subject factor $(S\alpha)_{il(k)}$ and the interaction with the second within-subject factor $(S\beta)_{jl(k)}$. Because these are interactions with a random factor, these effects are also considered random-effects and thus represent a further partitioning of the error term. Although it may initially appear as though a 3-way interaction with subject could also be included, this is not possible as it would be perfectly collinear with the errors. This is a clue to the fact that the error term in this model *is* the 3-way interaction with the subject effects.

As this is a much larger model than the previous example, the derivation of the EMS is more lengthy process. As with before, the arithmetic is presented in **Table 5** and the EMS are given in Equation (7). We can see that there are now four possible error terms, given by MS_S, MS_{SA}, MS_{SB}, and MS_E, respectively. As indicated above, MS_E could equivalently be written as MS_{SAB} to denote the equivalence with the highest-order interaction between the subjects and within-subject factors. Suitable tests for the model effects are given in **Table 6**, presenting a much more complex arrangements where no more than two effects are tested using the same error term.

$$\begin{aligned}
\text{EMS}_A &= \sigma^2 + b\sigma^2_{s\alpha} + bcn\frac{\sum \alpha_i^2}{a-1} \\
\text{EMS}_B &= \sigma^2 + a\sigma^2_{s\beta} + acn\frac{\sum \beta_j^2}{b-1} \\
\text{EMS}_C &= \sigma^2 + ab\sigma^2_s + abn\frac{\sum \gamma_k^2}{c-1} \\
\text{EMS}_{AB} &= \sigma^2 + cn\frac{\sum (\alpha\beta)^2_{ij}}{(a-1)(b-1)} \\
\text{EMS}_{AC} &= \sigma^2 + b\sigma^2_{s\alpha} + bn\frac{\sum (\alpha\gamma)^2_{ik}}{(a-1)(c-1)} \\
\text{EMS}_{BC} &= \sigma^2 + a\sigma^2_{s\beta} + an\frac{\sum (\beta\gamma)^2_{jk}}{(b-1)(c-1)} \\
\text{EMS}_{ABC} &= \sigma^2 + n\frac{\sum (\alpha\beta\gamma)^2_{ijkl}}{(a-1)(b-1)(c-1)} \\
\text{EMS}_S &= \sigma^2 + ab\sigma^2_s \\
\text{EMS}_{SA} &= \sigma^2 + b\sigma^2_{s\alpha} \\
\text{EMS}_{SB} &= \sigma^2 + a\sigma^2_{s\beta} \\
\text{EMS}_E &= \sigma^2
\end{aligned} \quad (7)$$

2.3. Section Summary

EMS are a necessary concept in ANOVA models in order to define suitable tests for the model effects. In purely fixed-effects models it is rarely necessary to explicitly calculate the EMS as a suitable denominator for each test is always given by the overall error term. When random effects are included, such as in repeated-measurements models with partitioned errors, complications arise in the derivation of suitable tests. As a minimum, models with a single within-subject factor have a choice of two error terms to form tests, whereas those with multiple within-subject factors have multiple possibilities when forming tests. It is precisely this issue of specifying the correct error term that leads to problems when using neuroimaging software designed to only use a single error term. Unless the

TABLE 5 | Arithmetic for the derivation of the EMS in the 3-way mixed-measures ANOVA with two within-subject and one between-subjects factor.

	i	***j***	***k***	***l***											
	F	**F**	**F**	**R**		$\mathbf{EMS}_{A}$	$\mathbf{EMS}_{B}$	$\mathbf{EMS}_{C}$	$\mathbf{EMS}_{AB}$	$\mathbf{EMS}_{AC}$	$\mathbf{EMS}_{BC}$	$\mathbf{EMS}_{ABC}$	$\mathbf{EMS}_{S}$	$\mathbf{EMS}_{SA}$	$\mathbf{EMS}_{SB}$	$\mathbf{EMS}_{E}$
	a	***b***	***c***	***n***	**Variance**	***i***	***j***	***k***	***ij***	***ik***	***jk***	***ijk***	***l*(*k*)**	***il*(*k*)**	***jl*(*k*)**	***l*(*ijk*)**
α_i	0	*b*	*c*	*n*	σ^2_{α}	*bcn*	0	0	0	0	0	0	0	0	0	0
β_j	*a*	0	*c*	*n*	σ^2_{β}	0	*acn*	0	0	0	0	0	0	0	0	0
γ_k	*a*	*b*	0	*n*	σ^2_{γ}	0	0	*abn*	0	0	0	0	0	0	0	0
$(\alpha\beta)_{ij}$	0	0	*c*	*n*	$\sigma^2_{\alpha\beta}$	0	0	0	*cn*	0	0	0	0	0	0	0
$(\alpha\gamma)_{ik}$	0	*b*	0	*n*	$\sigma^2_{\alpha\gamma}$	0	0	0	0	*bn*	0	0	0	0	0	0
$(\beta\gamma)_{jk}$	*a*	0	0	*n*	$\sigma^2_{\beta\gamma}$	0	0	0	0	0	*an*	0	0	0	0	0
$(\alpha\beta\gamma)_{ijk}$	0	0	0	*n*	$\sigma^2_{\alpha\beta\gamma}$	0	0	0	0	0	0	*n*	0	0	0	0
$S_{l(k)}$	*a*	*b*	1	1	σ^2_{S}	0	0	*ab*	0	0	0	0	*ab*	0	0	0
$(S\alpha)_{il(k)}$	0	*b*	1	1	$\sigma^2_{S\alpha}$	*b*	0	0	0	*b*	0	0	0	*b*	0	0
$(S\beta)_{jl(k)}$	*a*	0	1	1	$\sigma^2_{S\beta}$	0	*a*	0	0	0	*a*	0	0	0	*a*	0
$\epsilon_{l(ijk)}$	1	1	1	1	σ^2	1	1	1	1	1	1	1	1	1	1	1

TABLE 6 | The EMS ratios used to form appropriate *F*-tests for the main effects and interactions in a 3-way mixed-measures ANOVA.

Effect	**Test**
A	MS_A/MS_{SA}
B	MS_B/MS_{SB}
C	MS_C/MS_S
A × B	MS_{AB}/MS_E
A × C	MS_{AC}/MS_{SA}
B × C	MS_{BC}/MS_{SB}
A × B × C	MS_{ABC}/MS_E

EMS are taken into consideration it is entirely possible to end up with *F*-ratios that do not actually test the intended model effects. For instance, testing of MS_B from Equation (3) using MS_E would not result in a test of the between-subject effect, but a test of the between-subject effect *plus* the between-subject error, leading to an artificial inflation of the *F*-statistic (as noted previously by McLaren et al., 2011). Considering that the between-subject results are often of great interest in clinical neuroimaging studies, the ramifications of inflating these effects could be dire. Furthermore, these issues are not constrained to just between-subject effects. Considering the breakdown of the EMS in Equation (6), it is clear that the use of MS_E as the denominator of the *F*-ratios could lead to over-inflated statistics for all but the A×B and A×B×C interaction effects. As such, it is vital that the correct error terms are derived and then enforced to make sure that the tests of the model effects are accurate.

3. CONTRAST WEIGHTS

In the previous section we saw the importance of using EMS to derive suitable error terms in ANOVA models. Although this represents the core issue at the heart of implementing repeated measurement models in the GLM, it is also worth considering the practical question of how questions can be asked of these models in the form of contrast weights. Although the contrast framework is a well-established aspect of hypothesis testing in the neuroimaging GLM (e.g., Poline et al., 2007), additional complications arise when implementing repeated measurements models. This is due to both the inclusion of the subject effects in the model and the necessity of using overparameterized designs in certain software packages (such as SPM). These complications have unfortunately lead to some dubious advice on forming contrast weights in these models, which shall be discussed below. In addition, the use of contrast weights provides further insights into the topic of the EMS and formation of *F*-ratios, and so provides a complimentary perspective on the issues discussed in the previous section.

3.1. Contrast Weights for Overparameterized Repeated-Measurement Models

Consider the overparameterized design matrix for a 2 × 2 mixed-measures ANOVA with $n = 2$, given in Equation (9). Each row represents the linear combination of parameters which form one of the model predictions. As an example, the first row is given by

$$\mathbf{X}_1^{(A)} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

which tells us the combination of parameters needed to calculate the prediction for the A1B1 cell for subject 1, defining an estimable function of the parameters (see McFarquhar, 2016). Because linear combinations of estimable functions are themselves estimable (see McCulloch et al., 2008, p. 122), the rows of the design matrix can be used as the building-blocks for deriving contrast weights, irrespective of the form that the design matrix takes. Furthermore, note that this prediction is given by

$$\begin{aligned} \mu_{111} &= \mathbf{X}_1^{(A)}\boldsymbol{\beta} \\ &= \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + S_{1(1)} \end{aligned} \tag{8}$$

which is a combination of both the fixed and random model effects. Although the subject effects are not of interest, their

inclusion in the design matrix means we cannot simply give them a weight of 0 when calculating cell or marginal means as this would define a non-estimable function. As such, calculation of the ANOVA effects from cell and marginal means will include the subject terms. For certain ANOVA effects, the subject terms will cancel in the numerator, whereas for others they will not. For those where they do not, an error term must be selected such that the subject effects are also present in the denominator. This is simply a re-statement of the general approach to constructing *F*-ratios using the EMS, but from the perspective of contrast weights.

$$\mathbf{X}^{(A)} = \begin{array}{c} \begin{array}{cccccccccccc} \mu & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & (\alpha\beta)_{11} & (\alpha\beta)_{21} & (\alpha\beta)_{12} & (\alpha\beta)_{22} & S_{1(1)} & S_{2(1)} & S_{1(2)} & S_{2(2)} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{9}$$

As an example, consider deriving the weights for testing the main effect of the within-subject factor A. To do so, we can first average the rows in Equation (9) which code the first level of factor A and then average the rows which code the second level of factor A[2]. This gives

$$\begin{aligned} \mathbf{G}_{A1}^{(A)} &= \begin{bmatrix} 1 & 1 & 0 & 1/2 & 1/2 & 1/2 & 0 & 1/2 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \\ \mathbf{G}_{A2}^{(A)} &= \begin{bmatrix} 1 & 0 & 1 & 1/2 & 1/2 & 0 & 1/2 & 0 & 1/2 & 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \end{aligned} \tag{10}$$

providing the weights for calculating the marginal means of factor A. Notice that these weights are non-zero for the subject effects. The weights for the main effect are then formed from the subtraction of the weights for the marginal means, giving

$$\mathbf{L}_A^{(A)} = \mathbf{G}_{A1}^{(A)} - \mathbf{G}_{A2}^{(A)} \tag{11}$$

$$= \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 1/2 & -1/2 & 1/2 & -1/2 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{12}$$

where we can see that the subject effects have canceled. Now consider deriving the weights for testing the main effect of the between-subject factor B. Taking a similar approach to above we find

$$\begin{aligned} \mathbf{G}_{B1}^{(A)} &= \begin{bmatrix} 1 & 1/2 & 1/2 & 1 & 0 & 1/2 & 1/2 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix} \\ \mathbf{G}_{B2}^{(A)} &= \begin{bmatrix} 1 & 1/2 & 1/2 & 0 & 1 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} \end{aligned} \tag{13}$$

which provide the weights for the marginal means of factor B which again contain non-zero weights for the subject effects. Subtracting these weights gives

$$\mathbf{L}_B^{(A)} = \mathbf{G}_{B1}^{(A)} - \mathbf{G}_{B2}^{(A)} \tag{14}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 1/2 & 1/2 & -1/2 & -1/2 & 1/2 & 1/2 & -1/2 & -1/2 \end{bmatrix} \tag{15}$$

which notably is still non-zero for the subject terms. This has direct correspondence with the definitions of the EMS from earlier where EMS_B in Equation (3) contains σ_s^2. In order to form a meaningful *F*-ratio using the weights given above, one would need to select an error term that also contained the subject effects. As this is not possible by default in most neuroimaging implementations of the GLM, the use of the above contrast would be inappropriate for testing the between-subject effect of factor B. This is because, as stated earlier, the magnitude of the test-statistics would be inflated by the inclusion of the subject effects in the numerator, but not in the denominator. This speaks to a general rule-of-thumb for implementing these models in neuroimaging software, namely that appropriate contrast weights should always contain zeros for all terms containing the subject effects. Notably, this goes against the methods given by Gläscher and Gitelman (2008), where contrasts containing weights for the subject-terms are given as means of testing all the ANOVA effects within the same model. Hopefully it is now clear why this advice is inappropriate.

3.2. Contrast Weights for Non-overparameterized Repeated-Measurement Models

To see how the discussions in the previous section are readily applicable to non-overparameterized models (such as those used in FSL FEAT), consider the design matrices given in Equation (16). These are both constrained versions of the matrix from Equation (9), with $\mathbf{X}^{(B)}$ using "treatment" coding and $\mathbf{X}^{(C)}$ using "sigma-restricted" coding (see McFarquhar, 2016). Of note is the fact that "cell means" coding could also be used to simplify Equation (9), but that the design would remain overparameterized (although the contrast weights would be simpler).

$$\mathbf{X}^{(B)} = \begin{array}{c} \begin{array}{cccccc} \mu & \alpha_1 & \beta_1 & (\alpha\beta)_{11} & S_{1(1)} & S_{1(2)} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \quad \mathbf{X}^{(C)} = \begin{array}{c} \begin{array}{cccccc} \mu & \alpha_1 & \beta_1 & (\alpha\beta)_{11} & S_{1(1)} & S_{1(2)} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 \end{pmatrix} \end{array} \tag{16}$$

Application of the earlier approach to deriving contrasts leads to the weights for the effect of within-subject factor A in model B ($\mathbf{L}_A^{(B)}$) and model C ($\mathbf{L}_A^{(C)}$), as given in Equation (17).

$$\begin{aligned} \mathbf{G}_{A1}^{(B)} &= \begin{bmatrix} 1 & 1 & 1/2 & 1/2 & 1/4 & 1/4 \end{bmatrix} \quad \mathbf{G}_{A1}^{(C)} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} \\ \mathbf{G}_{A2}^{(B)} &= \begin{bmatrix} 1 & 0 & 1/2 & 0 & 1/4 & 1/4 \end{bmatrix} \quad \mathbf{G}_{A2}^{(C)} = \begin{bmatrix} 1 & -1 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} \\ \mathbf{L}_A^{(B)} &= \begin{bmatrix} 0 & 1 & 0 & 1/2 & 0 & 0 \end{bmatrix} \quad \mathbf{L}_A^{(C)} = \begin{bmatrix} 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \tag{17}$$

[2]Be aware that this approach will not work with unbalanced design matrices. See section 4.2.4 for how this procedure can be adjusted for those cases.

which, as with before, do not contain weights for the subject effects. Similarly, the contrast for between-subjects factor B can be derived for both alternative codings as shown in Equation (18).

$$\begin{aligned}
\mathbf{G}_{B1}^{(B)} &= \begin{bmatrix} 1 & 1/2 & 1 & 1/2 & 1/2 & 0 \end{bmatrix} \quad \mathbf{G}_{B1}^{(C)} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1/2 & 0 \end{bmatrix} \\
\mathbf{G}_{B2}^{(B)} &= \begin{bmatrix} 1 & 1/2 & 0 & 0 & 0 & 1/2 \end{bmatrix} \quad \mathbf{G}_{B2}^{(C)} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 1/2 \end{bmatrix} \\
\mathbf{L}_{B}^{(B)} &= \begin{bmatrix} 0 & 0 & 1 & 1/2 & 1/2 & -1/2 \end{bmatrix} \quad \mathbf{L}_{B}^{(C)} = \begin{bmatrix} 0 & 0 & 2 & 0 & 1/2 & -1/2 \end{bmatrix}
\end{aligned} \tag{18}$$

which again contain weights for the subject effects and are therefore inappropriate when dealing with software that only implements a single error term.

3.3. Contrast Weights for Follow-Up Tests

Another aspect of hypothesis testing in ANOVA models is the use of *post-hoc* contrasts to follow-up omnibus main effects or interaction results. Although often dealt with using *t*-contrasts, most of the discussion in the preceding sections is equivalent for using either *t*- or *F*-contrasts. Indeed, given that $F = t^2$ when rank $(\mathbf{L}) = 1$, the discussions in this paper can be taken as equivalent for both the *t* and the *F*. In terms of the actual follow-up tests, the theory remains the same insofar as the *post-hoc* tests of the main effects can be conducted using the same error term as the omnibus test. For interactions, more care must be taken as the error term for the follow-up tests will not necessarily be the same as the error term used for the omnibus effect. As an example, consider following-up the A×B interaction from the model in Equation (9). Using the approach of simple main effects, we may wish to examine the effect of the within-subject factor A at the first level of the between-subjects factor B. The weights for this would be derived as follows

$$\begin{aligned}
\mathbf{G}_{A1B1}^{(A)} &= \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix} \\
\mathbf{G}_{A2B1}^{(A)} &= \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix} \\
\mathbf{L}_{A1-A2(B1)}^{(A)} &= \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned} \tag{19}$$

which contains no weights for the subject effects and so can be tested with the overall error term of the model. Alternatively, if we wanted to examine the effect of the between-subjects factor B at the first level of the within-subject factor A, the weights would be

$$\begin{aligned}
\mathbf{G}_{A1B1}^{(A)} &= \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \end{bmatrix} \\
\mathbf{G}_{A1B2}^{(A)} &= \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix} \\
\mathbf{L}_{B1-B2(A1)}^{(A)} &= \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 1 & 0 & -1 & 0 & 1/2 & 1/2 & -1/2 & -1/2 \end{bmatrix}
\end{aligned} \tag{20}$$

which does contains non-zero values for the subject effects. This is perhaps not surprising given that this simple main effect is a between-subject comparison, constrained to only use the estimates from the first level of factor A. Nevertheless, it demonstrates that the error term for the omnibus test may not always be appropriate for testing the simple effects. If one did wish to test this effect, another between-subjects model would need to be specified containing only the data from the first level of the within-subject factor, adding further complication to the approach necessitated by the implementation of the GLM in common neuroimaging packages.

3.4. Section Summary

Although contrast weights are a familiar concept for hypothesis testing in the GLM, the inclusion of the random subject effects can make their derivation more difficult depending on the design matrix coding options available. A general approach has been given whereby weights can always be reliably derived using the rows of the design matrix. In addition, this section has shown how contrast weights can provide a complimentary perspective on the issue of suitable error terms. In particular, weights that are derived correctly but contain non-zero values for the subject effects are not suitable for testing with the overall error-term of the model. This provides a useful rule-of-thumb for neuroimaging researchers, particularly when it comes to follow-up tests of interactions, where extra care must be taken given that a suitable error term is not necessarily the same as the error term used for the omnibus effect.

4. BUILDING REPEATED MEASURES MODELS IN NEUROIMAGING SOFTWARE

Now that the core theoretical concepts of repeated measurements models have been described, we turn to the practical aspect of specifying partitioned-error ANOVA models in neuroimaging software. Based on the discussions in the preceding sections, four generic steps for correctly specifying these models are:

1. Calculate the EMS for the complete model. The number of error terms corresponds to the number of separate models that need to be estimated.
2. For each model, identify which within-subject factors are *not* tested under that error term. Those factors that are missing must be averaged over.
3. Use contrasts at the 1st-level to average-over the various factors identified above.
4. Specify the 2nd-level models using the 1st-level contrasts created in the previous step and then derive the contrast weights using the design matrices.

These steps can be used with any software implementing the mass-univariate GLM approach to modeling group-level neuroimaging datasets. To make these steps clear, an example will now be provided of specifying a 3-way mixed-measures ANOVA using the Flexible Factorial module in SPM12.

4.1. Example Data Set

The example data set comes from a previously reported fMRI study by Trotter et al. (2016) investigating the role of serotonin in discriminatory and affective touch perception. Subjects were scanned whilst experiencing gentle stroking of either the arm or the fingers using brushes of different textures. Sixteen of the subjects were administered a tryptophan-depleting amino acid drink prior to the scan, with the remaining 14 subjects receiving a balanced (control) amino acid drink. The design

was therefore a 2 × 3 × 2 factorial design with a within-subject factor of *Location* (arm/fingers), a within-subject factor of *Texture* (soft/medium/coarse) and a between-subjects factor of *Drink* (balanced/tryptophan-depleting). It is worth noting that Trotter et al. (2016) conducted the analysis appropriately using the sandwich estimator (SwE) toolbox (Guillaume et al., 2014), however for the purpose of the current paper we shall explore how the data could have been modeled using SPM12 instead.

4.2. Example Analysis

The model we wish to fit is given in Equation (6) and is repeated below

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + S_{l(k)} + (S\alpha)_{il(k)} + (S\beta)_{jl(k)} + \epsilon_{ijkl}$$

In the context of the example dataset, α_i is the ith level of *Location* ($i = 1, 2$), β_j is the jth level of *Texture* ($j = 1, 2, 3$) and γ_k is the kth level of *Drink* ($k = 1, 2$). The indices for the subjects run from $l = 1, \ldots, n_k$ where $n_1 = 14$ and $n_2 = 16$. As discussed earlier, the *Subject* effects are nested within *Drink* and are considered random-effects.

4.2.1. Step 1: Calculate the EMS and Identify the Number of Models

The EMS for this design have already been derived using **Table 5** and are listed in Equation (7). Although the method of Kutner et al. (2004), used to calculate these expressions, is designed for balanced models, it can be applied to unbalanced designs for the purpose of deriving the necessary error terms for each *F*-ratio. The final ANOVA table for this dataset is given in **Table 7**. The degrees of freedom can be calculated using the rules given in Appendix D.2 of Kutner et al. (2004) and provide one of several ways of checking that each model has been specified correctly in the analysis software. It is also worth mentioning that the EMS can be calculated automatically using the algorithms available in software such as the `R` package `EMSaov` (Choe et al., 2017).

TABLE 7 | ANOVA table for the example 3-way mixed-measures model.

Effect	Df
Drink	1
Error: Subject(Drink)	28
Location	1
Location × Drink	1
Error: Subject(Drink) × Location	28
Texture	2
Texture × Drink	2
Error: Subject(Drink) × Texture	56
Texture × Location	2
Texture × Location × Drink	2
Error: Subject(Drink) × Location × Texture	56

4.2.2. Step 2: Identify the Factors to be Averaged-Over

Based on the tests given in **Table 7**, we can see that there are 4 models needed. The model with the *Subject(Drink)* × *Location* × *Texture* error term uses the original dataset. The model with the *Subject(Drink)* × *Texture* error term requires a data set with the *Location* factor averaged over. Similarly, the model with the *Subject(Drink)* × *Location* error term requires a data set with the *Texture* factor averaged over. Finally, the dataset with only the *Subject(Drink)* error terms requires both the *Location* and *Texture* factors averaged-over for each subject.

To understand why averaging over different effects produces the correct error term, consider the model where we wish to enforce *Subject(Drink)* × *Texture* as the error term. After averaging over the *Location* factor, we can specify the following model

$$y_{.jkl} = \mu + \beta_j + \gamma_k + (\beta\gamma)_{jk} + S_{l(k)} + \epsilon_{.jkl} \qquad (21)$$

Clearly we cannot include any of the terms containing α_i from the full model, but notice that we cannot include $(S\beta)_{jl(k)}$ either. This is because this term would now be *perfectly collinear* with the errors. Because of this, we know that the error term *is* $(S\beta)_{jl(k)}$. This connects directly with the EMS from Equation (7) where the averaged model has effectively removed the σ^2 term from EMS_B and EMS_{BC}, leaving only $\sigma^2_{s\beta}$ as the overall error. Thus the *F*-ratios from the averaged model allow for the effective isolation of the effects of interest.

4.2.3. Step 3: Create the 1st-Level Contrasts for Each of the 2nd-Level Models

The basic 1st-level models for this dataset contain boxcar regressors for each of the *Location* × *Texture* cells of the design. To create the raw data for the *Subject(Drink)* model a single contrast per-subject was specified to average across all the cells. For the *Subject(Drink)* × *Location* model 2 contrasts per-subject were specified, one for each level of *Location* averaged over the levels of *Texture*. For the *Subject(Drink)* × *Location* model 3 contrasts per-subject were specified, one for each level of *Texture* averaged over the level of *Location*. Finally, for the *Subject(Drink)* × *Location* × *Texture* model 6 contrasts per-subject were specified, one for each of the *Texture* × *Location* cells. Of note is the fact that all the condition effects in these contrasts were specified as subtractions from preceding rest periods. This was done to make all images taken to the group-level readily interpretable. This is an important point when implementing repeated measures models with neuroimaging data, given that generally the images taken to the group-level represent within-subject averages, rather than contrasts *per-se*.

4.2.4. Step 4: Specify the Models and Derive the Contrast Weights From the Design Matrices

Example design matrices for the different models, as specified in SPM12 using the Flexible Factorial module[3], are given in **Figure 1**. Once the models have been specified, contrasts for

[3]For the current release of SPM12, it is necessary to alter the spm_cfg_factorial_design.m file (inside /spm12/config) and change the line

the ANOVA effects can be derived from the design matrix by averaging over the rows after the removal of the subject blocks and reduction to unique-row form[4]. As an example, deriving the weights for testing the *Location* × *Drink* effect from the *Subject(Drink)* × *Location* error model can be achieved in MATLAB using

```
load('SPM.mat');

% Get design matrix, remove subjects and
  reduce to unique rows
X = SPM.xX.X;
X = X(:,1:8);
X = unique(X, 'rows', 'stable');

% Get weights for the cell means
A1B1 = mean(X(X(:,1) == 1 & X(:,3) == 1,:));
A2B1 = mean(X(X(:,2) == 1 & X(:,3) == 1,:));
A1B2 = mean(X(X(:,1) == 1 & X(:,4) == 1,:));
A2B2 = mean(X(X(:,2) == 1 & X(:,4) == 1,:));

% Create interaction weights
L = (A1B1 - A2B1) - (A1B2 - A2B2);
```

where columns 1 and 2 code the levels of *Location* and columns 3 and 4 code the levels of *Drink*.

4.2.5. Results

The results from this model for a single voxel are shown in both SPM and SPSS version 23 (http://www.ibm.com) in **Figure 2**. Note that to make this comparison valid, all non-sphericity corrections were switched off in SPM (see section 6 for more on this). On the left are the tables from SPSS under the assumption of sphericity and on the right are the test statistic values reported by SPM. The equivalency of the *F*-ratios and degrees of freedom demonstrate how the correct error terms have been selected in SPM and that the contrast weights derived from the design matrix have resulted in the calculation of the correct Type III test statistics.

It is worth noting that although the *F*-ratios are equivalent in **Figure 2**, the sums-of-squares derived from the models containing averaged datasets will not be the same as those calculated by other statistical software. For instance, consider the *Texture* × *Drink* effect from the *Subject(Drink)* × *Texture* model. In SPSS the sums-of-squares are given as 0.264 for the interaction effect and 4.522 for the error. In SPM these same sums-of-squares are given as 0.132 for the interaction effect and 2.261 for the error. Notice how these both differ by a factor of 2 due to averaging over the two levels of *Location* before fitting the model. Similarly, all the sums-of-squares for the *Subject(Drink)* × *Location* model will be out by a factor of 3, and for the *Subject(Drink)* model will be out by a factor of 2 × 3. Because these differences only amount to a constant in both the numerator and denominator, the *F*-ratios remain the same. As such, this discrepancy is of little concern.

5. USE OF CONTINUOUS COVARIATES IN REPEATED MEASUREMENT MODELS

In the previous sections we have seen how complex mixed-measures models can be specified using the GLM framework, as implemented in standard neuroimaging software packages. However, the discussion has so far neglected the formation of ANCOVA models by the inclusion of continuous covariates. Putting aside issues of whether it is meaningful to use certain covariates to "control" for concomitant factors in quasi-experimental situations (see Miller and Chapman, 2001, for discussion), the use of covariates to reduce error variance is an attractive and useful means of increasing the sensitivity of an analysis. This is particularly pertinent for neuroimaging given the general noisiness of the data. Despite this, the use of continuous covariates in classical mixed-measures designs had received limited attention in both the literature and statistical textbooks. Notable exceptions include Federer (1955), Winer (1962), Federer and Meredith (1992), and Federer and King (2007). In particular, the text by Federer and King (2007) contains extensive coverage of this issue and will be used as the basis for the discussion in this section.

5.1. The Mixed-Measures ANCOVA Model

The extension of the basic mixed-measures ANOVA model from Equation (3) to a mixed-measures ANCOVA model is given by Federer and King (2007) as

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \beta_1 \bar{x}_{.k(j)} + \beta_2 x_{ik(j)} + S_{k(j)} + \epsilon_{ijk} \quad (22)$$

where $x_{ik(j)}$ is the raw covariate value for repeated measurement i from subject k in group j and $\bar{x}_{.k(j)}$ is the average covariate value for subject k in group j. In this parameterization, β_1 gives the between-subject regression slope and β_2 gives the within-subject regression slope. The inclusion of both regression coefficients is in-line with the recommendations of Federer and King (2007) who state that "in an analysis of covariance...there are as many...regression coefficients as there are error terms in an analysis of variance..." (p. 240). Using this approach, the between-subjects effects are adjusted for β_1 and the within-subject effects are adjusted for β_2. The model in Equation (22) is therefore the basis for any mixed-measures model that contains a continuous covariate. Implementation will largely depend on whether the covariate in question is measured between-subjects or within-subject, as will be discussed below.

5.1.1. Between-Subjects Covariates

A between-subjects covariate (also known as *constant* or *time-invariant*) is defined based on having a single value per-subject that does not depend on the within-subject manipulation. The inclusion of a between-subjects covariate in Equation (22) renders the $x_{ik(j)}$ term redundant and the model can be simplified

`fnums.num = [2 1]` to `fnums.num = [Inf 1]` to allow for interactions higher than a two-way.

[4]Removal of the subject blocks allow for derivation of Type III sums-of-squares in unbalanced designs by reducing the design matrix to a balanced unique-row form. This is based on assuming that the effect in question is being testing within an appropriate model and thus the weights on the subject effects will be zero. If this is not the case, then a non-estimable contrast will be returned.

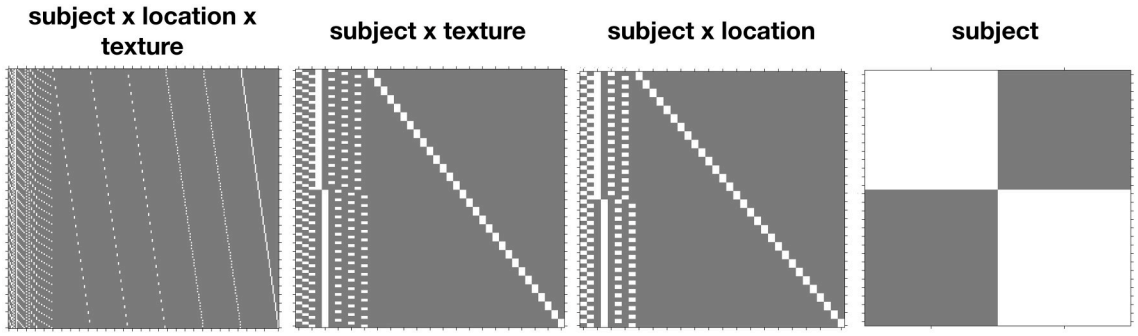

FIGURE 1 | Comparison of the design matrices produced by SPM12 for the different error terms.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
location * texture	Sphericity Assumed	1.459	2	.730	11.238	.000
location * texture * drink	Sphericity Assumed	.152	2	.076	1.167	.319
Error (location*texture)	Sphericity Assumed	3.636	56	.065		

co-ordinates x = -17.40 y = 24.80 z = 28.00 statistic 11.24
co-ordinates x = -17.40 y = 24.80 z = 28.000 statistic 1.17
SPM{F2,56}

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
texture	Sphericity Assumed	.111	2	.055	.685	.508
texture * drink	Sphericity Assumed	.264	2	.132	1.634	.204
Error(texture)	Sphericity Assumed	4.522	56	.081		

co-ordinates x = -17.40 y = 24.80 z = 28.000 statistic 0.69
co-ordinates x = -17.40 y = 24.80 z = 28.00 statistic 1.63
SPM{F2,56}

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
location	Sphericity Assumed	.003	1	.003	.038	.847
location * drink	Sphericity Assumed	.011	1	.011	.150	.702
Error(location)	Sphericity Assumed	2.031	28	.073		

co-ordinates x = -17.40 y = 24.80 z = 28.00 statistic 0.04
co-ordinates x = -17.40 y = 24.80 z = 28.00 statistic 0.15
SPM{F1,28}

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	.064	1	.064	.481	.494
drink	.026	1	.026	.193	.664
Error	3.741	28	.134		

co-ordinates x = -17.40 y = 24.80 z = 28.000 statistic 0.48
co-ordinates x = -17.40 y = 24.80 z = 28.000 statistic 0.19
SPM{F1,28}

FIGURE 2 | Comparison of the results produced by SPM and SPSS 23 for data from a single voxel. Equivalence of the *F*-statistics and the degrees of freedom confirms that the correct error terms have been selected and that correct Type III weights have been derived.

to only contain the $x_{.k(j)}$ term. Note that when using the multiple-model approach advocated in the previous section, the covariate must be tested within the same model as the other between-subject effects and interactions (i.e., the *Subject(Drink)* model from the previous example). All other models should contain the covariate (by replicating the per-subject values), but should not be used for testing the effect. The only exception is when an interaction between the covariate and a within-subject factor is included. In this instance, the interaction effect would be tested within the same model as the test of the interacting within-subject factor. Inclusion of these interactions is a modeling choice that provides equivalence with the repeated measurement models generated by software such as SPSS, as well as providing equivalence with the multivariate approach to repeated measures (see McFarquhar et al., 2016, for details).

5.1.2. Within-Subject Covariates

A within-subject covariate (also known as *time-varying* or *time-dependent*) is defined based on having multiple values per-subject that depend on the within-subject manipulation. Unlike a between-subject covariate, there are no redundancies in Equation (22) and both terms are therefore necessary. Furthermore, when implementing the multiple-model approach from the previous section, it is important to include as many of the covariates as possible in each model. For some of the models, averaging over certain factors will create redundancies across

the covariates that can be removed. For instance, the model used to test the between-subject effects will only contain $\bar{x}_{.k(j)}$, whereas the model used to test the within-subject main effect and interaction can contain both $\bar{x}_{.k(j)}$ and $x_{ik(j)}$. In a similar vein to testing the traditional ANOVA effects, the parameter associated with $\bar{x}_{.k(j)}$ should be tested using the between-subject error term and the parameter associated with $x_{ik(j)}$ should be tested using the within-subject error term.

As a more involved example, consider an ANCOVA for the complete $2 \times 3 \times 2$ mixed-measures design from section 4. Assuming there is a covariate value per-cell of the design, extension of the *split block* ANCOVA model presented in Federer and King (2007) provides the model form

$$\begin{aligned} y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \\ + \beta_1\bar{x}_{..l(k)} + \beta_2\bar{x}_{i.l(k)} + \beta_3\bar{x}_{.jl(k)} + \beta_4 x_{ijl(k)} \\ + S_{l(k)} + (S\alpha)_{il(k)} + (S\beta)_{jl(k)} + \epsilon_{ijkl} \end{aligned} \tag{23}$$

where $x_{ijl(k)}$ is the raw covariate value, $\bar{x}_{..l(k)}$ is the average covariate value for subject l from group k, $\bar{x}_{i.l(k)}$ is the average covariate value for subject l from group k from level i of the first within-subject factor and $\bar{x}_{.jl(k)}$ is the average covariate value for subject l from group k from level j of the second within-subject factor. The ANOVA table for this model, indicating the most suitable error terms for testing the covariates, is given in **Table 8**.

An additional complication arises when one of the covariates in Equation (24) is associated with one within-subject factor, but is constant over the other. In this situation there will be redundancies in the definitions of the four covariates. For instance, if one of the within-subject factors was time and a value was measured only once per-visit, the value would be constant across any other within-subject variables. Implementation of this design would then be similar in spirit to the use of between-subject covariates in Equation (22), insofar as all covariates terms would attempt to enter each model, but would then be dropped wherever redundancies are found.

TABLE 8 | ANOVA table for the example 3-way mixed-measures model including a within-subject covariate.

Effect	Df
Covariate ($\bar{x}_{..l(k)}$)	1
Drink	1
Error: Subject(Drink)	27
Covariate ($\bar{x}_{i.l(k)}$)	1
Location	1
Location × Drink	1
Error: Subject(Drink) × Location	27
Covariate ($\bar{x}_{.jl(k)}$)	1
Texture	2
Texture × Drink	2
Error: Subject(Drink) × Texture	54
Covariate ($x_{ijl(k)}$)	1
Texture × Location	2
Texture × Location × Drink	2
Error: Subject(Drink) × Location × Texture	54

6. THE ASSUMPTION OF SPHERICITY

One final important issue to discuss is the much-maligned sphericity assumption of the traditional repeated-measures ANOVA. In brief, the validity of the F-ratios, in terms of following an exact F-distribution under the null, is predicted on assuming a spherical structure to the variance-covariance matrix. This can be expressed as

$$\mathrm{Var}(y_i - y_j) = \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij} = \lambda \quad \forall i \neq j \tag{24}$$

which indicates that for all pairs of measurements the variance of their differences are identical. This is therefore a similar (but less restrictive) case of compound symmetry (Davis, 2002).

Traditionally, departures from sphericity are assessed using hypothesis tests, such as described by Mauchly (1940). If significant departures from sphericity are found then corrections to the degrees of freedom, such as those after Greenhouse and Geisser (1959) or Huynh and Feldt (1976), can be applied. Generally speaking, neuroimaging software does not implement such corrections. For instance, use of the OLS algorithm in FSL means assuming sphericity for the validity of the F-tests at every voxel. Use of permutation tests via Randomize (Winkler et al., 2014) do not have such restrictive assumptions, although the exchangeable structure of the data is more complex and must be accommodated for accurate derivation of the null distribution (see Winkler et al., 2015, for details). SPM, on the other-hand, implements a correction for departures from sphericity, but there are some caveats. Firstly, the covariance matrix used to derive the correction comes from a subset of pooled voxels, rather than being applied at each voxel individually. As discussed by McFarquhar et al. (2016), the method used to select voxels to enter this pool can have a dramatic effect on the number of voxels that subsequently survive correction. Secondly, the impact of this correction on the specification of repeated measurement models remains unclear. According to Glaser and Friston (2007), the non-sphericity correction employed by SPM is essentially a whitening procedure that renders the covariance structure a scalar multiple of an identity matrix. This is therefore equivalent to the method of generalized least-squares (e.g., Faraway, 2016), which can be used to specify a marginal model that accommodates a variety of covariance structures without the need for random effects (see Guillaume et al., 2014). The implication here would seem to be that when the non-sphericity procedure is employed the inclusion of the random subject blocks is unnecessary. However, the current default SPM implementations of both the paired t-test and one-way within-subject ANOVA make use of the non-sphericity correction with random subject blocks. At present it is unclear why this is the case and requires further clarification from the SPM authors.

7. CONCLUSIONS

This paper has discussed the modeling of group-level repeated measurements in neuroimaging, using the traditional GLM framework. The core statistical concept of the EMS has

been discussed and from these discussion a set of steps for implementing these forms of models in software have been given. Additional considerations, such as covariates and the assumption of sphericity, have also been discussed. The main conclusion from this paper is that if one wishes to use traditional neuroimaging analysis tools for this purpose, great care must be taken to correctly derive the tests from the EMS and then to carefully implement the multiple models necessitated by the GLM framework. In doing so, it is important to carefully consider the contrasts used and the error-terms employed, especially for follow-up tests from interaction effects. To that end, the oft-quoted advice given by Gläscher and Gitelman (2008) should no longer be relied-upon for deriving contrast weights as it may lead to inappropriate tests of the model effects. Furthermore, the example given in this paper has highlighted how tedious and complicated the implementation of these models can be in standard software. Ultimately, there are much better alternative tools available for this purpose. Examples include the author's own multivariate and repeated measures (MRM) toolbox (McFarquhar et al., 2016), the previously mentioned SwE toolbox (Guillaume et al., 2014), the AFNI tool 3dMVM (Chen et al., 2014), and the mixed-effects approaches discussed by Chen et al. (2013). There are also tools available which seek to simplify the specification of the traditional partitioned-error ANOVA, such as GLMFlex (McLaren et al., 2011), although the limitations of the traditional ANOVA framework should be enough for researchers to consider the more modern alternatives mentioned above. McFarquhar et al. (2016) provides comparison between several of these tools (including MRM, SwE, and GLMFlex) noting that largely their differences come down to assumptions about the covariance structure of the data, available methods for multiple-comparison correction, the ability to accommodate certain features of the data (e.g., missing data, within-subject covariates) and the user-friendliness of the implementation.

In terms of a more general conclusion from this paper, it is important for developers of neuroimaging analysis packages to recognize that the onus of correctly specifying these models should not be placed on the users. Indeed, considering the methods outlined in this paper it would seem wholly unfair to expect that users would know to perform the given steps without any clear guidance. Instead, software developers should strive for improved usability and clarity in their implemented methods. Although usability is not always considered as carefully compared with commercial software, it is hopefully clear that only by providing user-friendly and well-documented software can the neuroimaging community be confident in the accuracy of their methods. This is particularly true given recent concerns about the accuracy of common neuroimaging analysis approaches (Eklund et al., 2016) as well as more general concerns about the replicability of behavioral research (Open Science Collaboration, 2015). At present it is unclear how many published neuroimaging analyses were conducted using inappropriate methods of analysing repeated measurements. What is clear is that neuroimaging needs a renewed focus on the methods and software that are readily employed to analyse brain imaging data. This is the only way for the neuroimaging community to have faith that published analyses have been implemented correctly and provide results that can be interpreted accurately.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The author would like to thank Dr. Paula Trotter for kindly allowing the use of her dataset in this paper.

REFERENCES

Casella, G. (2008). *Statistical Design*. New York, NY: Springer.

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., and Cox, R. W. (2014). Applications of multivariate modeling to neuroimaging group analysis: a comprehensive alternative to univariate general linear model. *Neuroimage* 99, 571–588. doi: 10.1016/j.neuroimage.2014.06.027

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., and Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage* 73, 176–190. doi: 10.1016/j.neuroimage.2013.01.047

Choe, H.-M., Kim, M., and Lee, E.-K. (2017). Emsaov: An r package for the analysis of variance with the expected mean squares and its shiny application. *R J.* 9, 252–261. doi: 10.32614/RJ-2017-011

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York, NY: Springer.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why FMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Faraway, J. J. (2016). *Linear Models With R*. Boca Raton, FL: Chapman and Hall/CRC.

Federer, W. T. (1955). *Experimental Design*. London: Macmillan.

Federer, W. T., and King, F. (2007). *Variations on Split Plot and Split Block Experiment Designs*. Hoboken, NJ: John Wiley & Sons.

Federer, W. T., and Meredith, M. P. (1992). Covariance analysis for split-plot and split-block designs. *Am. Stat.* 46, 155–162.

Gläscher, J., and Gitelman, D. (2008). Contrast weights in flexible factorial design with multiple groups of subjects. Available online at: https://www.researchgate.net/publication/267779738

Glaser, D., and Friston, K. (2007). "Covariance components," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images, Vol. 1* (Academic Press), 126.

Greenhouse, S. W., and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24, 95–112. doi: 10.1007/BF02289823

Guillaume, B., Hua, X., Thompson, P. M., Waldorp, L., Nichols, T. E., Initiative, A. D. N., et al. (2014). Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. *NeuroImage* 94, 287–302. doi: 10.1016/j.neuroimage.2014.03.029

Huynh, H., and Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J. Educ. Stat.* 1, 69–82. doi: 10.3102/10769986001001069

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models, 5th Edn*. New York, NY: McGraw-Hill.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *Ann. Math. Stat.* 11, 204–209. doi: 10.1214/aoms/1177731915

McCulloch, C., Searle, S., and Neuhaus, J. (2008). *Generalized, Linear, and Mixed Models, 2nd Edn*. Hoboken, NJ: Wiley.

McFarquhar, M. (2016). Testable hypotheses for unbalanced neuroimaging data. *Front. Neurosci.* 10:270. doi: 10.3389/fnins.2016.00270

McFarquhar, M., McKie, S., Emsley, R., Suckling, J., Elliott, R., and Williams, S. (2016). Multivariate and repeated measures (mrm): a new toolbox for dependent and multimodal group-level neuroimaging data. *NeuroImage* 132, 373–389. doi: 10.1016/j.neuroimage.2016.02.053

McLaren, D., Schultz, A., Locascio, J., Sperling, R., and Atri, A. (2011). "Repeated-measures designs overestimate between-subject effects in fmri packages using one error term," in *17th Annual Meeting of Organization for Human Brain Mapping*, (Quebec City, QC), 26–30.

Miller, G. A., and Chapman, J. P. (2001). Misunderstanding analysis of covariance. *J. Abnorm. Psychol.* 110:40. doi: 10.1037/0021-843X.110.1.40

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Penny, W., and Henson, R. (2007). "Analysis of variance," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (London: Academic Press), 193–210.

Poline, J.-B., Kherif, F., Pallier, C., and Penny, W. (2007). "Contrasts and classical inference," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images, Vol. 1* (London: Academic Press), 126.

Trotter, P. D., McGlone, F., McKie, S., McFarquhar, M., Elliott, R., Walker, S. C., et al. (2016). Effects of acute tryptophan depletion on central processing of ct-targeted and discriminatory touch in humans. *Eur. J. Neurosci.* 44, 2072–2083. doi: 10.1111/ejn.13298

Winer, B. J. (1962). *Statistical Principles in Experimental Design.* is New York, NY: McGraw-Hill Book Company.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage* 92, 381–397. doi: 10.1016/j.neuroimage.2014.01.060

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., and Smith, S. M. (2015). Multi-level block permutation. *NeuroImage* 123, 253–268. doi: 10.1016/j.neuroimage.2015.05.092

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.