

Factorial ANOVA with unbalanced data: A fresh look at the types of sums of squares

Carrie E. Smith^{1*} and Robert Cribbie¹

¹*Department of Psychology, York University*

Abstract: In this paper we endeavour to provide a largely non-technical description of the issues surrounding unbalanced factorial ANOVA and review the arguments made for and against the use of Type I, Type II and Type III sums of squares. Though the issue of which is the 'best' approach has been debated in the literature for decades, to date confusion remains around how the procedures differ and which is most appropriate. We ultimately recommend use of the Type II sums of squares for analysis of main effects because when no interaction is present it tests meaningful hypotheses and is the most statistically powerful alternative.

Key words: sums of squares, unbalanced factorial ANOVA.

1. Introduction

The fixed effects analysis of variance (ANOVA) procedure has been a staple of introductory statistics courses in the behavioural sciences since it was introduced by R.A. Fisher. Fisher developed factorial ANOVA for use with data sets with equal numbers of observations across the levels of each experimental factor, termed 'balanced' data. It quickly became clear that troublesome results can follow when this condition is not met. In 1934 Yates published his landmark paper in which he proposed all the methods used today in the analysis of main effects with unbalanced designs. Three of Yates' procedures, the unadjusted method of fitting constants, the adjusted method of fitting constants and the weighted squares of means, have become the most commonly employed methods for partitioning the sums of squares. Researchers in psychology are likely more familiar with the labelling introduced by the software package SAS as Type I, Type II and Type III sums of squares, respectively.

Many introductory statistics texts avoid the types of sums of squares controversy by only presenting factorial ANOVA in the context of equal sample sizes. Researchers are often made aware of the issue when they discover that different options are available for analysis of factorial designs in their statistical software.

Though the issue of which is the 'best' approach has been debated in the literature for decades, the discussion can be quite difficult to follow as such a wide variety of terminology

* Corresponding author.

has been introduced (a summary of the language employed has been assembled in Appendix A). Frustrated researchers continue to post questions to message boards with answers to these questions rarely providing any detailed recommendations. The purpose of this paper is to provide researchers with a largely non-technical description of the issues surrounding unbalanced factorial ANOVA, review the arguments made for and against each of the standard methods, and provide simple recommendations.

2. Example

To frame the discussion we will use an example inspired by the research design of Weiss and Loubier (2008), but using simulated data. In this study a survey was conducted using random sampling in which respondents were asked about their gambling behaviours. The participants were categorized as being former athletes, current athletes, or non-athletes, with the working hypothesis that former athletes would demonstrate stronger gambling behaviours than current and non-athletes. There was also a hypothesized gender effect. The study is therefore a classical a 3 x 2 factorial ANOVA. The data was simulated with some imbalance, as we might expect to see when random assignment to condition is not possible. The fictitious data is presented in Table 1 and the cell means are plotted in Figure 1. The question of interest is whether there is an effect of the variables, athletic status or gender, on the gambling behaviour scores.

To express the problem in the typical ANOVA language, we are looking for evidence of mean differences in the outcome (gambling) across the levels of athletic status and gender. To assess this statistically, the ANOVA methodology involves breaking down the observed variance in gambling into parts: variance that is ‘explained’ by the two factors and their interaction, and ‘unexplained’ variance (otherwise referred to as ‘residual’ or ‘error’ variance). If the variation attributed to a factor is large relative the residual variance, then that variable is considered to have a significant effect on the dependent variable. We then conclude that the means are not all equal across the levels of that factor.

We might begin to appreciate the difficulty in assessing mean differences with unbalanced factorial designs by considering that the mean gambling behaviour score for, say, the former athletes might be computed in any number of different ways. We could take the average of the mean gambling score for female former athletes ($\bar{Y} = 3.93$) and male former athletes ($\bar{Y} = 4.97$), arriving at a mean of means for former athletes [$\bar{Y} = \frac{1}{2}(3.93 + 4.97) = 4.45$]. This is otherwise known as the equally weighted (or ‘unweighted’) mean and is implicit in the Type III method of sums of squares (SS). Alternatively, we might disregard gender and compute a mean time over all the former athlete observations ($\bar{Y} = 4.62$). This method is a form of weighted mean because this value is more heavily influenced by the cell with the largest sample, and is implicit in the first factor in a Type I SS analysis. Other possibilities for mean weighting exist, for example the Type II Sums of Squares use yet another form of weighted means. When there are an equal number of observations in each cell, the unweighted and weighted means will be equivalent, but when there are unequal frequencies they generally will not.

Table 1: Simulated gambling scores

		Athletic Status					
		Current Athlete		Former Athlete		Non-Athlete	
Gender	Male	3.0	3.0	5.1	5.2	2.1	1.9
	Female	2.8		4.7	4.9	2.0	1.8
Gender	Female	2.3	2.4	3.9	4.1	1.2	1.3
	Male	2.1		3.8		1.1	1.1
							1.0

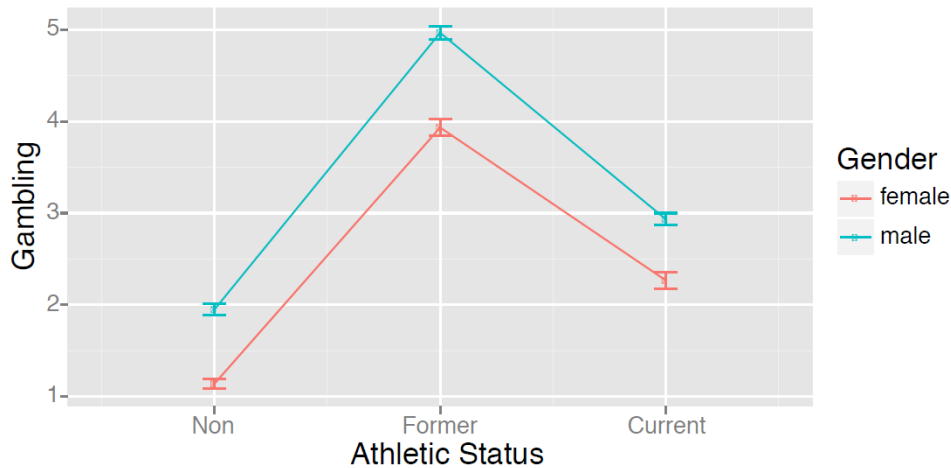


Figure 1: Graph of the cell means for the example data

Difficulties also arise in partitioning the explained variance with unbalanced data. To understand this more clearly it is important to understand that ANOVA is nothing more than a special case of linear regression (see Fox & Weisberg, 2011). We can conduct ANOVA using least squares regression and this is precisely what your statistical software is doing ‘behind the scenes’. To fit this model using least squares linear regression, the categorical variables athletic status and gender should be coded using deviation or ‘sum-to-zero’ regressors. This is similar to the more familiar ‘dummy coding’, but, as the name implies, these variables sum to zero. Sum to zero coding ensures that important constraints are always maintained. To represent all the levels of a factor, we require the number of levels less one deviation regressors. For the gender factor, which has only two levels, we require only one deviation regressor which we will call D_{Gen} . The factor athletic status has three levels, and so we need two deviation regressors $D1_{AthStat}$ and $D2_{AthStat}$. Since we rarely interpret (or even see!) the regression coefficients in ANOVA, the category set to -1 is completely arbitrary and will not impact our analysis. Applying this coding scheme to our data, we arrive at a data set as shown in Table 3

and we can fit a regression of the form Equation 1 where \hat{Y} represents the predicted gambling score.

$$\hat{Y} = c_0 + c_1 D_{Gen} + c_2 D1_{AthStat} + c_3 D2_{AthStat} + c_4 D_{Gen} \times D1_{AthStat} + c_5 D_{Gen} \times D2_{AthStat} \tag{1}$$

Table 2: Sum-to-zero coded regressors representing Gender (D_{Gen}) and Athletic Status ($D1_{AthStat}$ and $D2_{AthStat}$)

	D_{Gen}		$D1_{AthStat}$	$D2_{AthStat}$
Male	-1	Current Athlete	-1	-1
Female	1	Former Athlete	0	1
		Non-Athlete	1	0

Table 3: Simulated data recoded with sum-to-zero deviation regressors

Gambling	Gender	Athletic Status	D_{Gen}	$D1_{AthStat}$	$D2_{AthStat}$	$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
3.0	Male	Current Athlete	-1	-1	-1	1	1
3.0	Male	Current Athlete	-1	-1	-1	1	1
2.8	Male	Current Athlete	-1	-1	-1	1	1
5.1	Male	Former Athlete	-1	0	1	0	0
5.2	Male	Former Athlete	-1	0	1	0	0
4.7	Male	Former Athlete	-1	0	1	0	0
4.9	Male	Former Athlete	-1	0	1	0	0
4.9	Male	Former Athlete	-1	0	1	0	0
5.0	Male	Former Athlete	-1	0	1	0	0
2.1	Male	Non-Athlete	-1	1	0	-1	0
.
.
.
1.1	Female	Non-Athlete	1	1	0	1	0
1.0	Female	Non-Athlete	1	1	0	1	0

We are not, for better or worse, particularly interested in the coefficients of this regression in the ANOVA framework. Instead, we are interested in the proportion of variability in the outcome gambling (Y) that can be accounted for by gender (D_{Gen}), athletic status ($D1_{AthStat}$ and $D2_{AthStat}$ combined) and the interaction ($D_{Gen} \times D1_{AthStat}$ and $D_{Gen} \times D2_{AthStat}$ combined). The total variability (SS_{total}) can always be partitioned into explained variance (the sum of squares of the regression, SS_{reg} , Equation 2) and unexplained variance (SS_{error} , Equation 3) as in Equation SS_{sum} . What we would like to do is further decompose the explained variance (SS_{reg}) into parts ‘due to’ the factors and interaction.

$$SS_{reg} = \sum (\hat{Y}_i - \bar{Y})^2 \tag{2}$$

$$SS_{error} = \sum(Y_i - \hat{Y}_i)^2 \quad (3)$$

$$SS_{total} = SS_{reg} + SS_{error} = \sum(Y_i - \bar{Y})^2 \quad (4)$$

In balanced designs the correlations between regressors associated with separate factors are zero (Table 4). When this is true, SS_{reg} can be separated into unique portions associated with the factors and the interaction, as per the top left Venn diagram in Figure 2a. Furthermore, the sums of squares associated with each of the factors and the interaction can be simplified and we can compute the sums of squares directly using equations commonly presented in introductory statistics texts. As outlined earlier, with balanced data it makes no difference which method is used to compute the sums of squares, Type I, II and III will all yield identical results.

Table 4: If the data were balanced the regressors associated with one factor would be uncorrelated with the regressors of the other

	D_{Gen}	$D1_{AthStat}$	$D2_{AthStat}$	$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
D_{Gen}	1	-	-	-	-
$D1_{AthStat}$	0	1	-	-	-
$D2_{AthStat}$	0	0.5	1	-	-
$D_{Gen} \times D1_{AthStat}$	0	0	0	1	-
$D_{Gen} \times D2_{AthStat}$	0	0	0	0.5	1

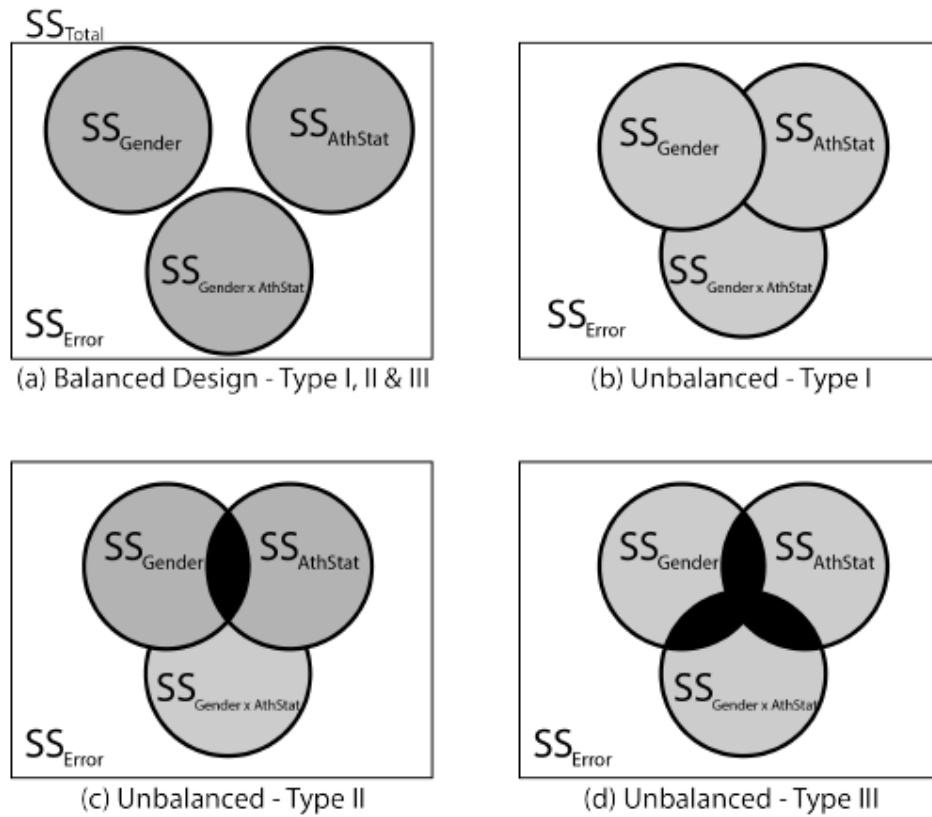


Figure 2: Venn Diagrams

Table 5: Correlated regressors with unbalanced data

	D_{Gen}	$D1_{AthStat}$	$D2_{AthStat}$	$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
D_{Gen}	1	-	-	-	-
$D1_{AthStat}$	0.122	1	-	-	-
$D2_{AthStat}$	-0.171	0.406	1	-	-
$D_{Gen} \times D1_{AthStat}$	0.122	0.011	0.077	1	-
$D_{Gen} \times D2_{AthStat}$	0.052	0.077	-0.198	0.406	1

The unequal frequencies in our simulated study produce an unfortunate side-effect, as correlations are induced between our regressors D_{Gen} , $D1_{AthStat}$, $D2_{AthStat}$, $D_{Gen} \times D1_{AthStat}$ and $D_{Gen} \times D2_{AthStat}$ (Table 5). In multiple regression terminology, this is known as collinearity. Said differently, our ‘predictors’ are correlated, since there is a small relationship between Gender and Athletic Status. As a consequence of this collinearity, the sums of squares associated with the factors and interaction overlap or underlap to the extent that they are correlated (Figure 2 b, c and d). The Venn diagram analogy is not perfect, since ‘underlapping’ variance cannot be appropriately represented in this format, but it does make clear that we now

have a problem in how to divvy up the shared variation, and it is precisely this that is handled differently in the Type I, Type II and Type III methods.

There is not a general set of equations to compute the sums of squares under these conditions, instead the process of arriving at the Type I, Type II and Type III methods of sums of squares is one of computing differences between regression sums of squares of appropriate nested sub-models. This idea will be explored in more detail in following sections, but the regression sums of squares required for these computations are summarized in Table 6.

3. Error and Interaction

The sums of squares for two entries in the ANOVA table will be the same regardless of the method selected to compute the sums of squares: the error or ‘residual’ SS and the interaction sum of squares.

The error term is simply the residual variability having fit the full model (i.e. all factors and interactions included) as per Equation 3. Since we are interested in computing the proportion of variability explained by the factors in the model, we will compute F-ratios whereby the denominator is the mean square error ($MS_{error} = SS_{error}/df_{error}$).

The interaction term in a 2-factor ANOVA is also computed in the same way in the Type I, II and III methods of sums of squares. To arrive at $SS_{Gen \times AthStat}$ we must subtract regression sum of squares for the model including gender and athletic status (Table 6 model 2) from SS_{reg} for the full model (Table 6 model 1). The computation of the incremental sum of squares by comparing nested sub-models is depicted in Figure 3. We then say that the interaction term $SS_{Gen \times AthStat}$ is the effect of the gender x athletic status interaction ‘controlling’ for the main effects of gender and athletic status, for which we will use the shorthand $SS(Gender \times AthleticStatus|Gender, AthleticStatus)$.

Table 6: Regression sums of squares and associated deviation regressors for the full regression model and nested sub-models

Model	Regression Sum of Squares	Model Regressors				
		Gender	Athletic Status		Interactions	
1	$SS_{reg}(Gen, AthStat, Gen \times AthStat)$	D_{Gen}	$D1_{AthStat}$	$D2_{AthStat}$	$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
2	$SS_{reg}(Gen, AthStat)$	D_{Gen}	$D1_{AthStat}$	$D2_{AthStat}$		
3	$SS_{reg}(Gen, Gen \times AthStat)$	D_{Gen}			$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
4	$SS_{reg}(AthStat, Gen \times AthStat)$		$D1_{AthStat}$	$D2_{AthStat}$	$D_{Gen} \times D1_{AthStat}$	$D_{Gen} \times D2_{AthStat}$
5	$SS_{reg}(Gen)$	D_{Gen}				
6	$SS_{reg}(AthStat)$		$D1_{AthStat}$	$D2_{AthStat}$		

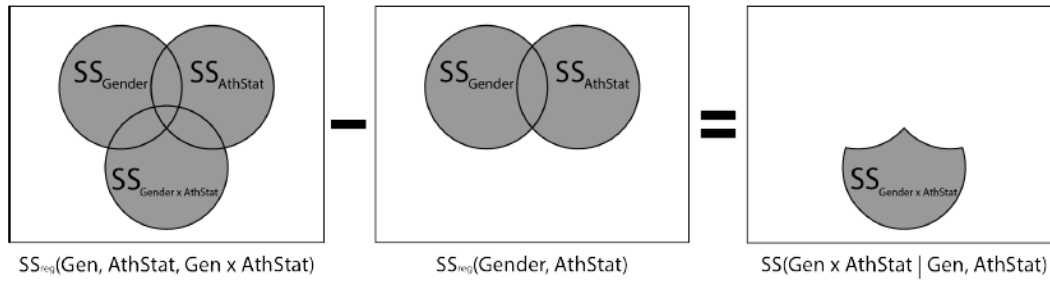


Figure 3: Computing the sum of squares for the gender x athletic status interaction by incremental regression sum of squares

We can now begin to fill in the standard ANOVA table (Table 7). The interaction effect is non-significant at $\alpha = 0.05$, with $F_{2,18} = 2.906$ and $p = 0.081$. We strongly recommend that researchers supplement the standard hypothesis test with visual inspection of the data and measures of effect size such as η^2 (Equation 5) or ω^2 (Equation 6). For the example at hand, we saw little evidence of an interaction in the profile plot of the means (Figure 1), and the effect of the interaction is negligible with $\eta^2 = 0.0024$ and $\omega^2 = 0.0016$.

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \quad (5)$$

$$\omega^2 = \frac{SS_{effect} - df_{effect}MS_{error}}{SS_{total} + MS_{error}} \quad (6)$$

Table 7: The ANOVA summary table entries for the interaction and error terms will be the same regardless of the method of SS employed

Source	SS	df	MS	F	Sig.
.
.
Gender x Athletic Status	0.121	2	0.061	2.906	0.081
Error	0.375	18	0.021		

Since both the interaction sum of squares and the error term are the same regardless of the method (Type I, II, III) selected for analysis, the test of statistical significance (F-ratio and p-value) and effect sizes for the interaction term of gender x athletic status will be equivalent across all methods. In our example, there is little evidence of an interaction, thus we can proceed with analysis of the main effects.

Had the interaction term been significant, this would indicate that the effect of one factor is not consistent across the levels of the other. Main effects in this situation are generally not meaningful, so in the presence of a significant interaction the method of sums of squares becomes irrelevant. Rather, we would proceed by interpreting the interaction directly using an appropriate follow-up analysis, such as interaction contrasts. Interaction contrasts break down larger interactions into all potential 2x2 combinations, which facilitates an understanding of the root of the interaction (Abelson & Prentice, 1997). Interaction contrasts are generally preferred

over simple effects tests because they provide a statistical test of the component interactions, rather than simply providing separate statistical tests at each level of one predictor (which often provides unclear or even confusing interpretations of the interaction).

In general this statement holds for the highest-order interaction term. That is to say, in a 3-factor ANOVA the 3-way interaction term is the same regardless of the method of analysis selected, but the 2-way interactions will differ.

4. Main Effects

The issues pertaining to analysis of main effects are split, roughly, into three domains: underlying regression models, the null hypotheses associated with the methods, and finally statistical power. The reader should keep in mind that these topics are intertwined, and dividing the arguments presented in the literature along these lines is rather artificial. We will generally discuss the methods in reverse order, starting with the Type III procedure, for ease of discussion.

4.1 Regression Models

4.1.1 Type III Sum of Squares

The main effect for gender by the Type III method can be interpreted as the main effect of gender controlling or adjusting for athletic status and the interaction gender x athletic status ($SS[Gender|AthleticStatus, Gender \times AthleticStatus]$). A major criticism of the Type III methodology is that the model does not respect marginality, that it is generally wrong to interpret main effects in the presence of an interaction (Cramer & Appelbaum, 1980; Nelder 1994; Nelder & Lane, 1995; Langsrud, 2003). If a significant interaction is detected, this result should be pursued. If it is not, and we haven't a particularly strong reason to believe in its existence, it should be removed from the model so that main effects can be interpreted alone.

The Type III sum of squares for the main effect of gender is computed by subtracting the regression sum of squares for a model with athletic status and gender x athletic status (Table 6 model 4) from the full model (Table 6 model 1) as shown in Figure 4. Notice that this involves performing a very unusual regression including a single main effect and an interaction term. The main effect of athletic status is computed in an analogous way, taking the difference in SS_{reg} between models 1 and 3. The results for the example are summarized in Table 8.

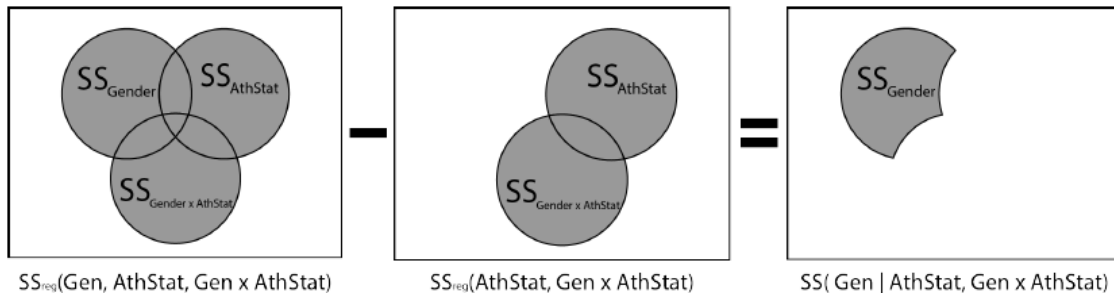


Figure 4: Computing the Type III sum of squares for gender by incremental regression sum of squares.

Table 8: ANOVA summary table using Type III SS

Source	SS (Type III)	df	MS	F	Sig.
Gender	3.897	1	3.897	186.888	<.0001
Athletic Status	35.989	2	17.995	862.971	<.0001
Gender x Athletic Status	0.121	2	0.061	2.906	0.081
Error	0.375	18	0.021		

4.1.2 Type II Sum of Squares

Using the Type II Method, the sums of squares for main effects are computed adjusting for other main effects in the model, but omitting higher-order terms. The main effect for gender ($SS[Gender|AthleticStatus]$) is therefore computed by taking the regression sum of squares for a two-factor model (without interaction) and subtracting the sum of squares for a model with athletic status alone (Figure 5). Referring back to Table 6, this is the difference between models 2 and 6. The incremental sum of squares for athletic status is computed by taking the difference between models 2 and 5. The results of the Type II method are summarized in Table 9.

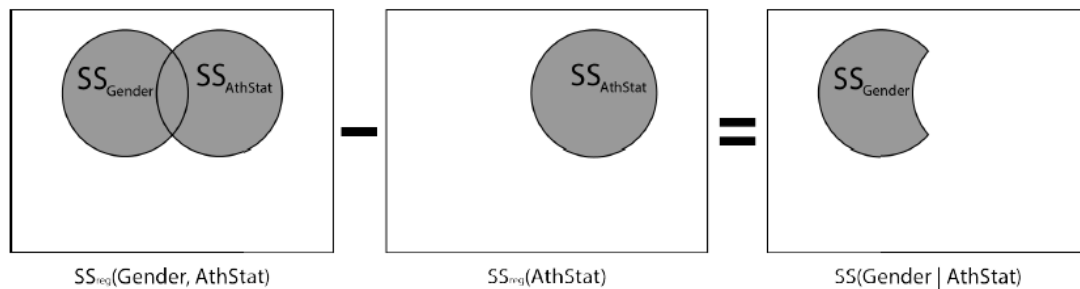


Figure 5: Computing the Type II sum of squares for Shape by incremental regression sum of squares.

Table 9: ANOVA summary table using Type II SS

Source	SS (Type II)	df	MS	F	Sig.
Gender	4.139	1	4.139	198.497	<.0001
Athletic Status	37.94	2	18.97	909.757	<.0001
Gender x Athletic Status	0.121	2	0.061	2.906	0.081
Error	0.375	18	0.021		

Unlike the Type III method, the Type II Sums of Squares for main effects do not violate marginality and is therefore logically sound in terms of model specification. The main effects are adjusted for each other, but not for potential higher-order interaction terms. Cramer1980 argue that the ANOVA method should be viewed as a logical sequence of model comparisons, beginning with interaction terms and proceeding to analysis of main effects. In regression analysis this is common practice, models are tested and reduced seeking the most parsimonious model.

4.1.3 Type I Sum of Squares

The Type I Sums of Squares are known as 'sequential' sums of squares, because effects are adjusted only for the terms that appear 'above' them in the ANOVA table. That is to say, the first factor entered into the analysis is not controlled or adjusted for any other variables, so $SS(\text{Gender})$ is simply equal to the SS_{reg} for a model including only the variable gender (though the statistical test is computed using the error term for the full model). The second factor, athletic status, is computed adjusting for the first ($SS[\text{Athletic}|\text{Gender}]$), which is equivalent to the Type II Sum of Squares for this factor. The interaction term remains the same, gender x athletic status adjusted for the two main effects. Thus the sum of squares for gender is equal to SS_{reg} for Table 6 model 5, and $SS(\text{AthleticStatus}|\text{Gender})$ is the difference in SS_{reg} between models 2 and 5. The results for the example data are summarized in Table 10.

Table 10: ANOVA summary table using Type I SS

Source	SS (Type I)	df	MS	F	Sig.
Gender	11.023	1	11.023	528.625	<.0001
Athletic Status	37.94	2	18.97	909.757	<.0001
Gender x Athletic Status	0.121	2	0.061	2.906	0.081
Error	0.375	18	0.021		

Unlike the other two, the Type I method will produce different values for the sums of squares if we swap the ordering of the factors and compute athletic status before gender. This property of the Type I method is generally unappealing and many authors recommend that the Type I method be used with caution and in the unlikely scenario that a researcher has a valid a priori reason for the ordering (Langsrud, 2003; Hector, von Felten, & Schmid, 2010). By rotating the order in which factors are entered into the model one can compute all the sums of squares associated with the Type II and Type III methods. If the design involves only a few

variables, Hector *et al.* (2010) suggest that this additional effort may be worthwhile since exploring the relationships between factors in this way may provide a better feel for the data. Furthermore, the Type I method does conform to marginality, and is therefore considered a viable option by those authors who consider adherence to marginality constraints to be necessary (Appelbaum & Cramer, 1974; Cramer & Appelbaum, 1980; Nelder, 1994; Nelder & Lane, 1995; Langsrud, 2003).

The Type I method yields sums of squares for the factors, interactions and error that add to the sums of squares total (See Figure 2b), while both the Type II and Type III methods typically result in double-counted or missed variance. This quality of the Type I method is not a particularly good reason to choose it over either of the other two, but it is mentioned in passing for the interested reader.

4.2 Null Hypotheses

4.2.1 Type III

In its favour, the null hypotheses associated with the Type III method can be interpreted as equivalence of unweighted means (Table 11) without making further assumptions regarding the presence of interactions (which we will see is not the case for the Type II method). The Type III method is the most common default method for factorial ANOVA in statistical software (Table 12) and this reasoning is frequently cited in help files (SAS 9.2, SPSS 19 and Statistica). Many authors have also recommended the Type III method because of the form of the null hypotheses (Carlson & Timm, 1974; Blair & Higgins, 1978; Kutner, 1974, 1975; Howell & Mcconaughey, 1982; Searle, 1995)

Table 11: Null hypotheses associated with Type III SS (Speed, 1978).

Type III Null Hypotheses
$H_{0A} : \bar{\mu}_{i\bullet} = \bar{\mu}_{i'\bullet} \text{ for all } i \neq i'$
$H_{0B} : \bar{\mu}_{\bullet j} = \bar{\mu}_{\bullet j'} \text{ for all } j \neq j'$
$H_{0A \times B} : \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'} \text{ for all } i \neq i', j \neq j'$

Table 12: Default and available methods in popular statistical software

Program and Version	Type I	Type II	Type III
R 2.13.0 <i>stats::anova()</i>	D		
R 2.13.0 <i>car::Anova()</i> V2.0-11		D	A
SAS 9.2 proc anova	D	A	D
SPSS 19	A	A	D
Stata 12	A		D
Statistica	A	A	D
Minitab 16.1.1	A		D
Matlab 2011b <i>anovan()</i>	A	A	D

4.2.2 Type II

One of the arguments commonly levied against the use of the SAS Type II method is that the null hypotheses for the full model (including interactions) are not easily interpretable (Table 13, left). It is for this reason that some authors roundly reject the Type II method (Carlson & Timm, 1974; Howell & McConaughy, 1982), and others gently recommend against it in favour of Type III (Overall & Spiegel, 1969; Searle, 1995; Shaw & Mitchell-Olds, 1993). However, the Type II hypotheses simplify to the test equality of equally weighted means if the interaction term is assumed to be zero (Speed, Hocking, & Hackney, 1978) (Table 13, right). In other words, if we are willing to assume that there is no interaction, the hypotheses for the Type II method become equivalent to those associated with Type III.

Table 13: Null hypotheses associated with Type II SS (Speed, 1978). If no inter-action is assumed, the null hypotheses for the main effects simplify to equality of equally weighted means (right) which are equivalent to the Type III null hypotheses.

Type II - Full Model	Type II - No Interaction
$H_{0A} : \sum_j n_{ij} \mu_{ij} = \sum_{i'} \sum_j \frac{n_{ij} n_{i'j} \mu_{i'j}}{n_{\bullet j}}$ for all i	$H_{0A} : \bar{\mu}_{i\bullet} = \bar{\mu}_{i'\bullet}$ for all $i \neq i'$
$H_{0B} : \sum_j n_{ij} \mu_{ij} = \sum_i \sum_{j'} \frac{n_{ij} n_{ij'} \mu_{ij'}}{n_{i\bullet}}$ for all j	$H_{0B} : \bar{\mu}_{\bullet j} = \bar{\mu}_{\bullet j'}$ for all $j \neq j'$
$H_{0A \times B} : \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$ for all $i \neq i', j \neq j'$	

Howell and McConaughy (1982) argue that researchers are rarely in a position to make such a strong claim about the absence of an interaction in the population, and making this decision on the basis of a non-significant test amounts to ‘proving the null hypothesis’. For this reason, they say that the Type III method is a ‘safer’ alternative to the Type II because if an interaction exists (whether the test has the power to show it or not) then the tests for main effects are made against ‘meaningful’ hypotheses.

While it must be conceded that we can never speak with certainty as to the true nature of the interaction in the population, Cramer and Appelbaum (1980) argue that we must be willing to allow our statistical tests to guide our model selection process. As Nelder and Lane (1995)

and Cramer and Appelbaum point out, if an interaction is present, no form of mean weighting (including the equally weighted Type III mean) is particularly meaningful or interesting. If we have decided to proceed with the interpretation of main effects, we must then be willing to assume no interaction exists and that we are testing the meaningful Type II hypotheses (Table 13, right).

4.2.3 Type I Sum of Squares

The Type I method tests for equivalence of fully weighted means for the first variable entered into the model, and the null hypothesis for the second factor is the same as the Type II method (Table 14, left). The same criticisms that are levied against the Type II method are often raised with regards to the Type I method, that the null hypotheses are a function of the sample size. If the interaction is assumed to be zero, the second null hypothesis simplifies to an acceptable form. The first hypothesis, however, is only really appropriate if the frequencies are related to population frequencies and one wished to make conclusions respecting the demographics (Carlson & Timm, 1974; Hector *et al.*, 2010). Further, it is generally not desirable to have different forms of null hypotheses for each main effect.

Table 14: Null hypotheses associated with Type I SS (Speed, 1978). If no interaction is assumed, the null hypotheses for the main effects simplify to equality of equally weighted means (right) which are equivalent to the Type III null hypotheses.

Type I - Full Model	Type I - No Interaction
$H_{0A} : \sum_j \frac{n_{ij}\mu_{ij}}{n_{i\bullet}} = \sum_j \frac{n_{i'j}\mu_{i'j}}{n_{i'\bullet}}$ for all $i \neq i'$	$H_{0A} : \sum_j \frac{n_{ij}\mu_{ij}}{n_{i\bullet}} = \sum_j \frac{n_{i'j}\mu_{i'j}}{n_{i'\bullet}}$
$H_{0B} : \sum_j n_{ij}\mu_{ij} = \sum_{j'} \sum_i \frac{n_{ij}n_{i'j'}\mu_{ij'}}{n_{i\bullet}}$ for all j	$H_{0B} : \bar{\mu}_{\bullet j} = \bar{\mu}_{\bullet j'}$ for all $j \neq j'$
$H_{0A \times B} : \mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$ for all $i \neq i', j \neq j'$	

4.3 Statistical Power

The last consideration is statistical power. It can be shown analytically that if there is no interaction present in the population, the Type II method will necessarily be maximally powerful (Monette & Fox, 2009). In other words, when no interactions exist the Type II method is always more powerful than the Type III method for unbalanced ANOVA. A discussion of power for Type I sums of squares is complicated by the fact that calculations depend on the order of entry of the variables. For the first variable entered using Type I sums of squares is equivalent to a one-way ANOVA, and hence direct comparisons to Type II and Type III sums of squares are not meaningful. For the second variable entered using Type I sums of squares the computations are equivalent to the Type II method, and hence power will be the same.

Otherwise, the power depends on the structure of the imbalance, and it cannot be guaranteed that Type II is the most powerful in all cases (Shaw & Mitchell-Olds, 1993). Lewsey and Gardiner (2001) conducted some limited simulations comparing the behaviour of Type II and Type III methods when small, non-significant interactions exist in a 2×3 factorial

design. Their main finding was that, while the Type II method was on average more powerful for most unbalanced structures investigated, it was also more strongly influenced by cell patterning. They were therefore unable to provide strong conclusions as to which particular patterns of imbalance result in a power advantage for the Type II sums of squares.

Research from pharmaceutical statistics offers one specific scenario in which the Type II method has substantially higher power than Type III. Consider a clinical trial conducted at several medical facilities of differing capacity. In the analysis one would like to assess the effectiveness of the medication while controlling for location. Gallo (2000) demonstrates that, provided the within-centre treatment allocation is well balanced, that the SAS Type II method maintains high power. Power for detecting treatment effects deteriorates rapidly using the Type III method when center sample sizes vary. Gallo presents an example in which one medical center is 1/3 as large as the other three in a hypothetical study. For this case the Type III method had higher power if the small center was omitted rather than kept in the analysis, clearly signalling a problem with the Type III method.

5. Conclusions and Recommendations

The purpose of this paper was to explore the different methods available for computing sums of squares in unbalanced factorial ANOVA. It is clear from the above discussion that the crux of the issue is whether or not there is an interaction effect. When interactions are hypothesized, it is of the utmost importance that studies have adequate power to detect them, which requires careful consideration of sample size requirements. If a significant interaction is detected, then the analysis should proceed directly to appropriate follow up tests (e.g., interaction contrasts), and the issue of which method of sums of squares to employ is not relevant. If a non-significant interaction is obtained, but the effect size (e.g., ω^2 , η^2) is not-negligible, say exceeding Cohen's (1998) recommendation of 0.01 for a small effect and certainly 0.06 for a medium effect, researchers should strongly consider collecting more data in order to improve statistical power for detecting the interaction. Since an interaction is suggested it may be preferable to proceed with interaction contrasts rather than interpreting main effects that are possibly contaminated by the interaction.

If there is no evidence of an interaction, either by way of significant hypothesis tests or effect sizes, we agree with the assessment of Cramer and Appelbaum (1980) that one of three eventualities has unfolded: (1) no interaction was detected because none exists in the population in question. In this circumstance the Type II method is definitively more powerful and we will necessarily lose power by electing to use the Type III method instead. (2) A very small interaction exists in the population, in which case it is not definitive which method will provide for the best statistical power for main effects. (3) A large interaction exists in the population but we have been extremely unfortunate in selecting a sample that does not evidence it. In this case the Type III method may hold better statistical power, but in this unfortunate situation the main effects will be of dubious value anyway. As Stewart-Oaten (1995, p.2007) quipped "the Type III SS is 'obviously' best for main effects only when it makes little sense to test main effects at all". Further, we do not recommend the Type I sums of square since we

believe it will be rare to find a situation when researchers would be interested in testing distinct hypotheses for the two predictors and that the cell sizes represent the population proportions.

Since the likelihood of a large interaction being present in the population but not detected in the sample is small (case 3), it is recommended that the decision between Type II and Type III SS be based on case 1 and case 2. For case 1 Type II sums of squares will always be more powerful (Monette & Fox, 2009), whereas for case 2 the Type II method is likely more powerful particularly for light cell imbalance (Lewsey & Gardiner, 2001). Therefore, since it is rare to find cases in which we would recommend Type I or TYPE III sums of squares, we suggest that researchers routinely use the Type II sum of squares to analyze between group factorial ANOVA designs. Currently, this method is the default for only the "Anova" function (from the 'car' package) in R, but hopefully this paper will help convince other software developers to make Type II sums of squares the default for their programs.

Appendix A: Summary of terminology employed in the literature

Table 15

	Separate One-Way ANOVAs	Sequential SS	Adjusted SS	
SAS Institute	One-way Fact A One-way Fact B	Type I	Type II	Type III partial sums of squares
Yates (1934)	Method of fitting constants (unadjusted) Special case: Method for proportional cell sizes		Method of fitting constants (adjusted)	Weighted squares of means Approximation: Method of unweighted means
Henderson (1953)	Method 1	Method 3		
Overall & Spiegel (1969)		Method 1 A priori ordering	Method 2 Experimental design method	Method 3 Complete least squares
Searle (1971)	$SS_A : R(\alpha \mu)$ $SS_B : R(\beta \mu)$ $SS_{A \times B} : R(\gamma \mu, \alpha, \beta)$	$SS_A : R(\alpha \mu)$ $SS_A : R(\beta \mu, \alpha)$ $SS_{A \times B} : R(\gamma \mu, \alpha, \beta)$	$SS_A : R(\alpha \mu, \beta)$ $SS_A : R(\beta \mu, \alpha)$ $SS_{A \times B} : R(\gamma \mu, \alpha, \beta)$	$SS_A : R(\alpha \mu, \beta, \gamma)$ $SS_A : R(\beta \mu, \alpha, \gamma)$ $SS_{A \times B} : R(\gamma \mu, \alpha, \beta)$
Howell & McCon (1982)	Method W Weighted means???		"Peculiarly" weighted means	Method U - Unweighted means
Herr (1986)	Weighted Means (WM)	Hierarchical, rows first then columns (HAB)	Each adjusted for other (EAD)	Standard parametric (STP)
Macnaughton (1998)			Higher-level terms omitted (HTO)	Higher-level terms included (HTI)

References

- [1] Abelson, R., & Prentice, D. (1997). Contrast tests of interaction hypothesis. *Psychological Methods*, 2 (4), 315–328. Retrieved from <http://psycnet.apa.org/journals/met/2/4/315/>
- [2] Appelbaum, M. I., & Cramer, E. M. (1974). Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin*, 81 (6), 335–343. Retrieved from <http://content.apa.org/journals/bul/81/6/335> doi: 10.1037/h0036315

-
- [3] Blair, R. C., & Higgins, J. (1978). Tests of Hypotheses for Unbalanced Factorial Designs Under Various Regression/Coding Method Combinations. *Educational and Psychological Measurement*, **38** (3), 621–631. Retrieved from <http://epm.sagepub.com/cgi/doi/10.1177/001316447803800303>
doi:10.1177/001316447803800303
- [4] Carlson, J. E., & Timm, N. H. (1974). Analysis of nonorthogonal fixed-effects designs. *Psychological Bulletin*, **81** (9), 563–570.
- [5] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition ed.). Routledge Academic.
- [6] Cramer, E. M., & Appelbaum, M. I. (1980). Nonorthogonal analysis of variance—once again. *Psychological Bulletin*, **87** (1), 51–57. Retrieved from <http://content.apa.org/journals/bul/87/1/51> doi: 10.1037//0033-2909.87.1.51
- [7] Fox, J., & Weisberg, S. (2011). *An r companion to applied regression*. SAGE Publications.
- [8] Gallo, P. P. (2000). Center-weighting issues in multicenter clinical trials. *Journal of biopharmaceutical statistics*, **10** (2), 145–63. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10803722> doi: 10.1081/BIP-100101019
- [9] Hector, A., von Felten, S., & Schmid, B. (2010). Analysis of variance with unbalanced data: an update for ecology & evolution. *The Journal of animal ecology*, **79** (2), 308–16. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20002862>
doi: 10.1111/j.1365-2656.2009.01634.x
- [10] Howell, D. C., & McConaughy, S. H. (1982). Nonorthogonal Analysis of Variance: Putting the Question before the Answer. *Educational and Psychological Measurement*, **42** (1), 9–24. Retrieved from <http://epm.sagepub.com/cgi/doi/10.1177/0013164482421002>
doi: 10.1177/0013164482421002
- [11] Kutner, M. H. (1974). Hypothesis testing in linear models (Eisenhart Model I). *The American Statistician*, **28** (3), 98–100.

- [12] Kutner, M. H. (1975). Hypothesis tests in linear models. *The American Statistician*, **29** (3), 133–134. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22044322> doi: 10.2460/javma.239.10.1288
- [13] Langsrud, O. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, **13** (2), 163–167. Retrieved from <http://www.springerlink.com/index/U671865627245351.pdf>
- [14] Lewsey, J., & Gardiner, W. (2001). A study of type II and type III power for testing hypotheses from unbalanced factorial designs. *Communications in Statistics - Simulation and Computation*, **30** (3), 597–609. Retrieved from <http://cat.inist.fr/?aModele=afficheN&cpsidt=14165716>
- [15] Monette, G., & Fox, J. (2009). A Framework for Hypothesis Tests in Statistical Models With Linear Predictors. In *user*. Rennes. Retrieved from <http://www.r-project.org/conferences/useR-2009/slides/Monette+Fox.pdf>
- [16] Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statistics and Computing*, **4** (4), 221–234. Retrieved from <http://www.springerlink.com/index/10.1007/BF00156745> doi: 10.1007/BF00156745
- [17] Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: in memoriam-Frank Yates. *The American Statistician*, **49** (4), 382–385. Retrieved from <http://www.jstor.org/stable/2684580>
- [18] Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, **72** (5), 311–322. Retrieved from <http://content.apa.org/journals/bul/72/5/311> doi: 10.1037/h0028109
- [19] Searle, S. R. (1995). Comments on J. A. Nelder. 'The statistics of linear models: back to basics' *. *Statistics and Computing*, **5**, 103–107.
- [20] Shaw, R., & Mitchell-Olds, T. (1993). Anova for unbalanced data: An overview. *Ecology*, **74** (6), 1638–1645. Retrieved from <http://www.jstor.org/stable/1939922>
- [21] Speed, F. M., Hocking, R. R., & Hackney, O. P. (1978). Methods of Analysis of Linear Models with Unbalanced Data. *Journal of the American Statistical Association*, **73** (361), 105. Retrieved from <http://www.jstor.org/stable/2286530?origin=crossref> doi: 10.2307/2286530

-
- [22] Stewart-Oaten, A. (1995). Rules and Judgments in Statistics: Three Examples. *Ecology*, **76** (6), 2001–2009. Retrieved from <http://www.jstor.org/stable/1940736>
- [23] Weiss, S., & Loubier, S. (2008). Gambling behaviors of former athletes: The delayed competitive effect. *UNLV Gaming Research & Review Journal*, **12** (1), 53–61. Retrieved from <http://www.cabdirect.org/abstracts/20093028438.html>
- [24] Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, **29** (185), 51

Received March 15, 2013; accepted November 10, 2013.

Carrie E. Smith
Department of Psychology York University
Toronto, ON M3J 1P3, Canada
smithce@yorku.ca

Robert Cribbie
Department of Psychology York University
Toronto, ON M3J 1P3, Canada
cribbie@yorku.ca

