16
Repeated measures ANOVA

Contents

16.1 Chapter overview

This chapter introduces repeated measures and mixed measures ANOVA as methods for dealing with correlated measures arising from paired, repeated or matched designs. Advantages and disadvantages of repeated measures models are considered, with focus on the increased statistical power arising when individual differences are prevalent, and on the assumptions of sphericity and multisample sphericity. Later sections provide a brief overview of several related models including MANOVA, repeated measures ANCOVA and analysis of gain scores.

16.2 Modeling correlated or repeated measures

A typical least squares regression model assumes measures are independent. For measures to be independent, each observation in the analysis should carry no information about the value of any other observation. Specifically, it will provide no additional information over and above that already accounted for by the structure of the model (e.g., factor or covariate values). If people are measured only once and randomly allocated to the conditions of an experiment this assumption is likely to be reasonable.[1] It is often plausible even if random assignment is not possible. On the other hand, if people are measured more than once, the responses for a given person are almost certainly correlated. This gives rise to what is termed a 'repeated measures' or 'within-subjects' design.

Correlated measures can also arise by design or because of the way the world is structured (e.g., if people are tested in pairs, or participants sampled from pre-existing groups). For field research with human participants, this kind of clustering is probably the norm rather than the exception. People have a tendency to create and associate within groups in many aspects of their lives: families, classrooms, schools or teams being good examples.

Correlated measures have historically been regarded as nuisance variables to be eliminated or controlled by careful experimental design (not least because independent measures are easier to work with). Correlated measures can also be viewed as an asset. A repeated measures or matched design might be selected because the resulting model has greater statistical power or precision. It is also increasingly common, as statistical models for correlated measures have become more sophisticated, that the clusters or correlated measures themselves are a focus of research. A researcher may be interested in the performance of individual children in a school, but they may also be interested in the performance of different schools. The main focus of this chapter is on ANOVA models with repeated measures on one or more factors. Multilevel regression models (that attempt to model the clustered structure of the data directly) are considered in a subsequent chapter.

16.3 ANOVA with repeated measures

Correlated measures present an obvious problem for a least squares regression model that assumes residuals are sampled from an independent, normal population of errors. In the simplest repeated measures models there are only two measurement points: the familiar paired design. By using the difference between pairs as the outcome Y it is possible to treat the differences as independent observations (e.g., in a one sample t test). This won't work if there

more than two levels on a factor or for a factorial design, but a similarly creative solution is feasible.

In a one-way independent measures design, the ANOVA parameterization of the model (with the usual sum-to-zero constraint for τ_j) is:

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

What if you represent the deviations of each participant from the grand mean in a similar way? This gives a repeated measures model of the form:

$$y_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij} \qquad \text{Equation 16.1}$$

In this model both τ_j and π_i must each sum to zero. Figure 16.1 shows the deviations of participant and level means for a trivial repeated measures ANOVA model with three levels and only two participants. The π_i term represents the deviation of each person's own mean (over all J levels of the factor) from the grand mean (after stripping out the average effect of the factor τ).

At first glance this looks almost exactly like a two-way independent measures ANOVA model incorporating only main effects (and no interaction). An important difference is that whereas the treatment τ is a fixed factor, π is a random factor.[2,3] Participants are viewed as randomly sampled from an infinite population, whereas the levels of the treatment are considered completely representative of the population of interest. The variance accounted for by the π_i term is due to systematic differences between participants: individual differences in Y (often termed *within-subject variance*). The shared subscript i of the participant and error variance indicates what is happening. In one-way independent measures ANOVA the variance not captured by

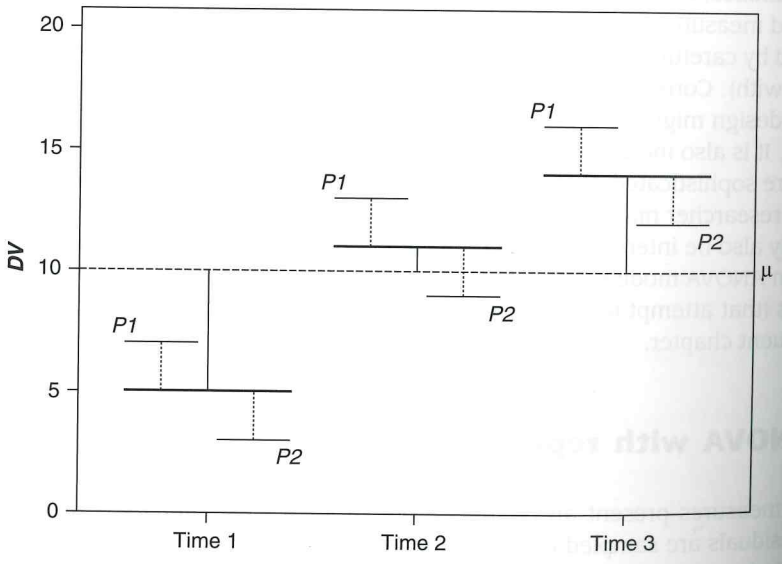

Figure 16.1 Deviations from the grand mean for a one-way repeated measures ANOVA, with three time points and two participants

the factor is subsumed within the ε_i term, whereas in the repeated measures design it is split between ε_i and π_i.

This model is implausible because it assumes that the effect of individual differences is exactly identical for all levels of the treatment. The model can be improved by permitting individual differences to vary across levels of τ. Allowing individual differences to vary in this way requires adding an interaction term:

$$y_{ij} = \mu + \pi_i + \tau_j + \tau\pi_{ij} + \varepsilon_{ij} \qquad \text{Equation 16.2}$$

This type of model is the one implemented in repeated measures ANOVA analysis, and if a subject-by-treatment interaction is present it will, in theory, produce more accurate tests of the factor (see Howell, 2002).[4]

Estimating this model in a repeated measures design is challenging because a typical application has exactly one measurement per participant in each combination of the levels of the fixed factor.[5] With only one observation per cell for the $\pi \times \tau$ interaction, the ε_i and $\tau\pi_{ij}$ terms cannot be estimated separately (and are said to be completely 'aliased'). This is not a problem for the test of the fixed factor (because its error term is formed from estimates of both ε_i and $\tau\pi_{ij}$ terms). Problems do arise for the test of the π_i subject term. Although a large F ratio provides evidence of systematic individual differences, a small F ratio is ambiguous (Kirk, 1995; Howell, 2002). It could occur because π_i is small relative to ε_i or because $\tau\pi_{ij}$ is large relative to π_i (e.g., if $\pi_i \approx \tau\pi_{ij}$).

It is possible to test whether the additive or non-additive interaction model is appropriate (see Kirk, 1995), but the choice of model ought to be decided *a priori* (Howell, 2002). The assumption of zero subject-by-treatment interaction is probably unjustified in research with human participants. To understand why, imagine a memory experiment comparing recall of own-race and other-race faces. If the subject by face interaction term were excluded from the model, this would be equivalent to assuming that individual differences in memory for faces are unaffected by the race of the face. This is highly implausible (e.g., it is contradicted by research that suggests lifetime exposure to other-race face influences ability to remember them). It is hard to come up with examples where one would be confident that individual differences are constant over all levels of a fixed factor unless individual differences or factor effects are themselves negligible. If $\tau\pi_{ij}$ is zero there will be no appreciable loss of power through adopting the non-additive model. What about the test of individual differences? The mere fact that you have decided to adopt a repeated measures model implies that you believe *a priori* that individual differences are present. A null hypothesis significance test (NHST) – particularly one with low statistical power – should not reverse that belief. Therefore a test of the π_i subject effect is only rarely of interest to researchers (and rarely reported or interpreted).

16.3.1 Advantages of repeated measures designs

A repeated measures design has one principal advantage over an independent measures design. In situations where both designs are feasible, it produces more accurate estimates of the fixed factor F ratio. This follows from the way in which the error term is constructed (and the logic of expected mean squares explained in Box 16.1). The error term in an independent measures design is influenced by at least two different sources of variation in the population.

These are experimental error and systematic between-subject variation (i.e., individual differences):

$$\sigma_E^2 = \sigma_{error}^2 + \sigma_{subjects}^2 \qquad \text{Equation 16.3}$$

In a repeated measures design, the individual differences can be separated out from the error term to get a purer estimate of error (i.e., $\hat{\sigma}_E^2 = \hat{\sigma}_{error}^2$). This is possible because multiple observations for each person allow $\sigma_{subjects}^2$ to be estimated from the variation between them. Having one observation per person per level of a factor (the state of affairs in an independent measures design) makes it impossible to untangle systematic individual differences from other sources of error. You can't tell if someone has scored high through chance or because their true score in the population is high.

There may be other advantages to repeated measures designs. One is practical: fewer participants need to be recruited to a study to obtain N data points. This is sometimes helpful if participants are scarce or difficult to recruit (because it allows a researcher to maximize the value of each participant's contribution). A further advantage of the repeated measures model is that the same equations also apply to matched or stratified designs. The calculations do not require that the same person be measured several times, just that observations are correlated with a particular structure. Equivalent correlations between measures arise for observations matched at the individual level (e.g., if each person is matched with a similar person in one or more comparisons on control conditions).

The main difference between matched and repeated measures designs is in their interpretation. A repeated measures design (assuming random or otherwise representative sampling) generalizes to all members of the population sampled, while a matched design generalizes only to the population of matched sets of individuals (Kirk, 1995). The matched design and repeated measures design thus reach similar conclusions only to the extent that matching is representative of the wider population. The difficulty of finding matches for some people might introduce bias into the sampling. Kirk also discusses the relative merits of matching (stratification) and ANCOVA for increasing statistical power. ANOVA with matched observations tends to be superior for low correlations between the confounding variable and the response, while ANCOVA is superior for high correlations. In practice, the difficulty of obtaining good matches makes ANCOVA more popular. In addition, ANCOVA usually assumes a linear effect of the covariate. Matching is a better strategy if the effect is curvilinear (Maris, 1998).

These characteristics imply certain disadvantages of repeated measures designs. The principal disadvantage is that, because measurements are typically spread over time, they may be influenced by the order in which they are obtained. If the order of measurement is controlled (e.g., by randomization or by counterbalancing) order effects can be reduced and perhaps eliminated, though this is only possible if the fixed factors can be manipulated by the experimenter. If order effects are controlled the study is a true experiment. Randomizing or counterbalancing the order of repeated measurements is essential if you want to infer that an experimental manipulation is causing differences in the outcome.

Order effects come in many different flavors (e.g., practice effects, fatigue effects or carryover effects) and are a particularly dangerous source of confounding. If the experimenter is not able to control the order of measurement it may be possible to reduce the influence of order effects (e.g., building in breaks to reduce fatigue effects) but not eliminate them. These additional controls are also useful for counterbalanced or randomized orders of testing because they should reduce error variance. It is also useful to model order of testing as a factor or

covariate. This strategy may increase statistical power in the same way that a covariate can in ANCOVA.

A final disadvantage of repeated measures designs is the increased complexity of the analysis and the assumptions of the model. Repeated measures ANOVA models can be inflexible. Analysis of an unbalanced design (in which some repeated measures are missing) is a good illustration. Multiple imputation provides a potential solution, if data are missing at random (MAR) or missing completely at random (MCAR).[6] A more flexible alternative for missing outcome data (one that overcomes these and other ANOVA limitations) is to use a multilevel model.

Box 16.1 Expected mean squares in one-way ANOVA

The smaller error term for a repeated measures design leads to tests with greater statistical power. Consider a pair of one-way designs with exactly the same sampling strategy and experimental manipulations: one with an independent measures design and one with a repeated measures design. All other things being equal, the test of the treatment (fixed factor) in each experiment would address the same null hypothesis: H_0: $\mu_1 = \mu_2 = \ldots = \mu_j$ (i.e., $\sigma^2_{factor} = 0$). The F ratio for the independent measures design is an estimate of the following ratio in the population:

$$F_{independent} = \frac{\sigma^2_{factor} + \sigma^2_{error} + \sigma^2_{subjects}}{\sigma^2_{error} + \sigma^2_{subjects}}$$

The true population variance associated with systematic differences between levels of the fixed factor σ^2_{factor}, systematic individual differences $\sigma^2_{subjects}$, and experimental error σ^2_{error} will be unknown, but imagine that they are 15, 10 and 5 respectively. The expected value of the F ratio for the values stated earlier will be:

$$E(F_{independent}) = \frac{15 + 10 + 5}{10 + 5} = \frac{30}{15} = 2$$

Doing the same arithmetic for the repeated measures design gives:

$$E(F_{repeated}) = \frac{\sigma^2_{factor} + \sigma^2_{error}}{\sigma^2_{error}} = \frac{15 + 5}{5} = \frac{20}{5} = 4$$

This illustrates why a repeated measures design has greater statistical power, and results in a more sensitive test of H_0.

A crucial feature of the expected F ratio is that, regardless of the design, $F = 1$ if the population effect of the fixed factor is zero. The numerator and denominator of the statistics are identical when $\sigma^2_{factor} = 0$. The increased power of the repeated measures therefore depends on the presence of systematic individual differences between the participants. The larger the population variance attributable to individual differences, the more sensitive the repeated measures design is (relative to an independent measures design). As the increased sensitivity to the effects of the factor is a consequence of the way that error variance is estimated, the statistical power advantage also applies to any inferential tool where the precision of estimation and hence estimation of error variance is fundamental (e.g., confidence intervals, as well as likelihood and Bayesian inference).

Up to this point, a number of technical points about expected F ratios have been glossed over (Howell, 2002; Kirk, 1995). A more detailed presentation of expected mean squares for one-way

independent measures ANOVA takes the form:

$$E(MS_{factor}) = \sigma_\varepsilon^2 + n\sigma_\tau^2$$
$$E(MS_{error}) = \sigma_\varepsilon^2$$

Greek letter subscripts are used to differentiate these quantities from the 'rough and ready' versions used earlier. An expected mean square is a long-run average (as if calculated over many samples). The constant n is the sample size per group (assuming a balanced design) and reflects the increase in observed treatment variance expected when larger samples are taken. The expected value of the F ratio thus depends on σ_τ^2 (the treatment variance, accounted for by a factor in the population). This expectation is one if $\sigma_\tau^2 = 0$ and greater than one if $\sigma_\tau^2 > 0$.

The expected mean squares for one-way repeated measures ANOVA depend on the presence of subject-by-treatment interactions. As the complete absence of subject-by-treatment variance is somewhat implausible, it is usual to adopt a structural model with interactions included. The mean squares for this structural model are:

$$E(MS_{subjects}) = \sigma_\varepsilon^2 + k\sigma_\pi^2$$
$$E(MS_{factor}) = \sigma_\varepsilon^2 + n\sigma_\tau^2 + \sigma_{\tau\pi}^2$$
$$E(MS_{error}) = \sigma_\varepsilon^2 + \sigma_{\tau\pi}^2$$

This leads to an F ratio for the factor that uses MS_{error} as the denominator (so that, on average, the $\sigma_{\tau\pi}^2$ terms cancel out). The F ratio for the subjects in this structural model is:

$$E(F_{subjects}) = \frac{\sigma_\varepsilon^2 + k\sigma_\pi^2}{\sigma_\varepsilon^2 + \sigma_{\tau\pi}^2}$$

This quantity is biased and can't be relied on as a test of systematic individual differences (at least if $\sigma_{\tau\pi}^2$ is not zero). In factorial repeated measures ANOVA, assuming that subject-by-treatment interactions are present leads to the aliasing of the error term with subject-by-treatment interactions. For this reason it is not wise to pool the error terms in factorial repeated measures designs. However, by using the treatment-by-subject interaction terms as error terms for treatment effects it is possible to construct true F ratios with an expected value of one when H_0 is true (Howell, 2002; Kirk, 1995).

16.3.2 ANOVA with repeated measures on all factors

The calculations for repeated measures ANOVA can be conducted using a similar approach to that of independent measures ANOVA. Hand calculation typically involves setting out *SS*, *df*, *MS* and *F* in an ANOVA table, though most computer software employs a general linear model solution. The standard sigma-restricted ANOVA parameterization treats the participants like an effect coded categorical predictor (using $n - 1$ indicator variables to code the n participants). Rutherford (2001) illustrates this by demonstrating how to run repeated measures ANOVA as a multiple regression with effect coding.

Table 16.1 One-way ANOVA with repeated measures

Source	*df*	*SS*	*MS*	*F*
Subjects	$n-1$	$SS_{subjects}$		
Factor A	$a-1$	SS_A	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{A\times subjects}}$
Error (A × subjects)	$(n-1)(a-1)$	$SS_{A\times subjects}$	$\frac{SS_{A\times subjects}}{df_{A\times subjects}}$	
Total	$N-1$			

One-way repeated measures ANOVA can be set out in a single table as illustrated in Table 16.1. In some software packages, the between-subjects and within-subjects components are split across separate tables. The *MS* and *F* ratio for the subjects is not included in Table 16.1. These can be derived in the usual manner if required (though the *F* ratio is not a pure measure of the ratio of effect variance to error variance).

The one-way table may be extended to incorporate additional factors. The extension to a two-way table is shown in Table 16.2. Several properties need to be emphasized. While the one-way design has only a single error term, the two-way design (and more generally any *k*-way design) has one error term for each fixed effect being estimated (i.e., for all effects other than the subjects and by-subjects interactions).

The logic for this is identical to that for use of A × subjects as the error term in a one-way design. With only one observation per cell, the residual term and by-subjects interaction terms are aliased. In addition, there is no longer an obvious route for testing the effect of subjects

Table 16.2 Two-way ANOVA with repeated measures

Source	*df*	*SS*	*MS*	*F*
Subjects	$n-1$	$SS_{subjects}$		
Factor A	$a-1$	SS_A	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{A\times subjects}}$
Error (A × subjects)	$(n-1)(a-1)$	$SS_{A\times subjects}$	$\frac{SS_{A\times subjects}}{df_{A\times subjects}}$	
Factor B	$b-1$	SS_B	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_{B\times subjects}}$
Error (B × subjects)	$(n-1)(b-1)$	$SS_{B\times subjects}$	$\frac{SS_{B\times subjects}}{df_{B\times subjects}}$	
A × B	$(a-1)(b-1)$	$SS_{A\times B}$	$\frac{SS_{A\times B}}{df_{A\times B}}$	$\frac{MS_{A\times B}}{MS_{A\times B\times subjects}}$
Error (A × B × subjects)	$(n-1)(a-1)(b-1)$	$SS_{A\times B\times subjects}$	$\frac{SS_{A\times B\times subjects}}{df_{A\times B\times subjects}}$	
Total	$N-1$			

(though you would not, in general, want to). In spite of these differences, the basic interpretation of the test of the main effects of A and B, or of the A × B interaction, are unchanged from a two-way independent measures design. The familiar tools of inspecting the level means and cell means or inspecting interaction plots can all be employed.

Formulas for *SS* and *MS* of both random and fixed effects factors are identical for the independent and repeated measures models. The main effects calculations use the level or participant means (as appropriate) after averaging over the other factors. Calculating the interaction *SS* or *MS* might seem difficult, but in a balanced design (which is common for repeated measures) this is simply done by calculating the cell means for a two-way table for A × B, A × subjects or B × subjects. Armed with the *SS* for A, B and subjects it is possible to calculate each interaction *SS* by subtraction from total *SS* for the two-way table (or from the residuals of the double-centered table just as for the interaction in a two-way independent measures design). The A × B × subjects interaction can in turn be calculated by subtraction from SS_{total} for the analysis.

16.3.3 Assumptions in repeated measures ANOVA

Repeated measures ANOVA is a form of general linear model and inherits those assumptions common to all least squares designs. However, having correlated measures violates one of the standard ANOVA assumptions (independence of observations). As repeated measures are correlated, it is necessary to fall back on the pure regression form of the assumption: that residuals are sampled from an independent, normal population of errors with constant variance.

A new complication is that the distribution of residuals – in particular its variance – depends both on the variances of the repeated measurements and the covariances between them. Stripping out the subjects and by-subjects variation does not necessarily guarantee independence and homogeneity of variance of the residuals (Kirk, 1995). Only under a restrictive condition known as *sphericity* will the patterns of variance and covariances between repeated measures meet the standard regression or general linear model assumptions.

For repeated measures, the specific distributional assumptions of the ANOVA model can therefore be set out as:

i) *correlated measures,*
ii) *normal distribution of errors* and
iii) *sphericity*.

The correlated measures assumption is justified *a priori* where matched or repeated measurements are employed. Unless there is a compelling argument that the correlation between these measures is exactly zero, a repeated measures model is required. Researchers are occasionally tempted to test the correlations for significance before using a repeated measures analysis for repeated or matched measurements. This is dangerous, and it is possible to show (e.g., via simulation) that correlations too small to be detected by a preliminary test can nevertheless have a material impact on the analysis.

The normal distribution of error assumption is identical to that for independent measures designs or for multiple regression. The residuals (of which there are several for each individual being measured) are assumed to have been sampled from a normal population of errors. While this assumption is identical to that for independent measures designs, it is important to note

that plotting the outcome Y by condition will no longer be an adequate procedure for checking the assumption. In independent measures ANOVA the outcome is merely a constant (the cell mean) plus the residual; residuals and Y therefore always show the same pattern within each group. In a repeated measures model this is no longer the case. In a one-way design, the outcome Y is a function of π, τ and ε. The distribution of Y for a given value of τ_j could be completely different from the distribution of ε. Although τ_j is constant within a group, π_j will vary. The outcome Y is therefore a joint function of both π and ε (i.e., the distribution of outcome depends on both individual differences and error).

The third assumption listed, that of sphericity, is the hardest to explain. It is considered in more detail in the next section. It can be summarized as entailing that the variances of the differences between pairs of repeated measures are equal. As with other distributional assumptions, it applies to the populations being sampled. Owing to sampling error, the assumption will rarely be met perfectly in a sample, whether or not it is true in the population (except where there are only two samples – as in a paired or matched pairs design).

16.3.4 Sphericity

Paired data provide a simple case in which to introduce the concept of sphericity. For a one-way repeated measures ANOVA design with two levels – equivalent to a paired t test – the population covariance matrix (see Key Concept 12.2) of the repeated measurements has this structure:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{bmatrix}$$

The main diagonal contains the population variances sampled by the first and second paired measurements. The covariance between the measurements is given by $\sigma_{1,2}$ or by $\sigma_{2,1}$. These quantities are necessarily identical, demonstrating the mirroring of covariances above and below the main diagonal of the matrix. For a two-group independent measures design the population covariance matrix is assumed to have the following structure:

$$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Even though the sample covariance matrix may not have zeroes in the off-diagonal cells, an independent measures ANOVA assumes that the true population covariance is zero. Viewed in this light, both designs make assumptions about the form of the population covariance matrix. The independent measures design assumes that covariances are precisely zero and that the variances σ_1^2 to σ_k^2 are equal in the population (respectively the assumptions of independent measures and homogeneity of variance). The repeated measures design assumes correlated measures (i.e., that the population covariances are non-zero) and that the covariance matrix is 'spherical'.

Sphericity was defined earlier as the assumption that the variances of differences between repeated measurements are equal in the population. The variance of a difference is obtained by subtracting one repeated sample from another to create a new variable (their difference). Its variance is the variance of a difference. Repeating this procedure produces variances of differences for every possible pair of the repeated measures samples. For paired data there

are two repeated measurements and therefore only one difference variable. The variance of the difference is therefore always consistent with the sphericity assumption for paired data; a single variance can't differ from itself. Violations of sphericity are only possible when a repeated measures factor has more than two levels.

The same logic applies to interaction terms. Sphericity is only a problem for effects with multiple degrees of freedom (*df*), and therefore coded by more than one indicator variable or contrast in a regression model. A factor with two levels or an interaction involving factors with more than two levels requires more than one predictor or contrast to code (regardless of whether dummy or effect coding is used). Sphericity is always true for effects with only 1 *df*.

Sphericity is a serious problem for multiple *df* effects in repeated measures ANOVA designs. For instance, Keselman *et al.* (2001) point out that violations of sphericity are likely to be common in longitudinal designs. This is because a commonly observed pattern is that of decreasing correlations over time (a pattern not consistent with sphericity). If sphericity is violated, the F ratio for an effect will not follow an F distribution with *df* defined by the numerator and denominator of the ratio. As a consequence, the p values for the usual test will be inaccurate. Violations of sphericity always lead to tests that are more liberal than suggested by the observed test statistic. The degree of inaccuracy is often substantial (e.g., the observed p value could be much smaller than a p value based on the true distribution of the test statistic). The degree of bias for the test increases with the severity of the violation. This, in turn, tends to increase as the number of repeated measurements rises (in essence because there is more room for severe departures to occur). If there are many repeated measures, serious violations of sphericity are likely to be common, leading to liberal inferences (e.g., increased Type I error rates).

It is therefore essential to check the sphericity assumption for all repeated measures ANOVA analyses (except those with two levels on all factors). An obvious starting point is the population covariance matrix for a one-way repeated measures design:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,3} & \dots & \sigma_{2,k} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \dots & \sigma_{3,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_{k,1} & \sigma_{k,2} & \sigma_{k,3} & \dots & \sigma_k^2 \end{bmatrix}$$

The pattern of covariances corresponding to sphericity is difficult to spot, because the variances of differences are not displayed directly. It is possible to calculate them using the variance-sum law (see Key Concept 3.2). For instance, the variance of the differences between the first two repeated measurements would be $\sigma_1^2 + \sigma_2^2 - 2\sigma_{1,2}$. This becomes tedious for large matrices. Many textbooks, particularly older ones, recommend looking for a related (but more restrictive) pattern known as *compound symmetry*. Compound symmetry entails both that the population variances are equal ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$) and that the population covariances are equal ($\sigma_{1,2} = \sigma_{1,3} = \ldots = \sigma_{k-1,k}$). To detect this pattern in the sample covariance matrix, requires that the values on the main diagonal are similar to each other ($\hat{\sigma}_1^2 \approx \hat{\sigma}_2^2 \approx \ldots \approx \hat{\sigma}_k^2$) and that the off-diagonal cells are also similar in value ($\hat{\sigma}_{1,2} \approx \hat{\sigma}_{1,3} \ldots \approx \hat{\sigma}_{k-1,k}$). If compound symmetry holds, then sphericity is always true (e.g., it holds when homogeneity of variance and independence are both true). Additionally, if approximate compound symmetry is present, sphericity is unlikely to be severely violated (though this can be hard to judge).

The trouble is that, while compound symmetry always implies sphericity, sphericity could be true even if compound symmetry is not. This is obvious from the covariance matrix for a paired

design, where the sphericity assumption is always met, but compound symmetry might not be true (if the variances are unequal).

A better approach to detecting violations of sphericity is to look at the epsilon estimate ($\hat{\varepsilon}$). Box (1954a, 1954b) showed that if the sphericity assumption is false the F statistic in one-way repeated measures ANOVA with $(J-1)$ and $(n-1)(J-1)$ df has an approximate F distribution with $\varepsilon(J-1)$ and $\varepsilon(n-1)(J-1)$ df. The ε parameter therefore indicates the degree to which the population departs from sphericity (with $\varepsilon = 1$ indicating that sphericity holds). As the departure becomes more extreme, ε approaches its lower bound. This lower bound depends on the design (starting at .5 if $J = 3$ and approaching zero as J becomes large). The sample estimate of ε is therefore an excellent descriptive statistic for assessing the degree to which sphericity is violated.

The lower bound of epsilon is

$$\hat{\varepsilon}_{lb} = \frac{1}{J-1}, \qquad \text{Equation 16.4}$$

where J is the number of levels of a repeated measures factor (and more generally the effective number of means be compared). Box (1954a, 1954b) provided an estimate of epsilon that was later explored by Geisser and Greenhouse (Geisser and Greenhouse, 1958; Greenhouse and Geisser, 1959). This estimate is now widely known as Greenhouse-Geisser epsilon ($\hat{\varepsilon}_{gg}$).[7] Huynh and Feldt (1976) proposed an alternative estimate $\hat{\varepsilon}_{hf}$.

There are several simulation studies comparing the performance of $\hat{\varepsilon}_{gg}$ and $\hat{\varepsilon}_{hf}$ under a range of different conditions. It can be established that $\hat{\varepsilon}_{gg} \leq \hat{\varepsilon}_{hf}$ and that $\hat{\varepsilon}_{gg}$ tends to be conservative, under-estimating ε (notably when departures from sphericity are modest). In contrast, $\hat{\varepsilon}_{hf}$ is liberal (overestimating ε and occasionally exceeding one). These results are relevant to repeated measures ANOVA because the estimates can be used to correct the effect and error df of the F ratio. The $\hat{\varepsilon}_{lb}$ statistic is extremely conservative and no longer of interest as a correction factor. It had previously been used as a first step in significance testing, because statistically significance with $\hat{\varepsilon}_{lb}$ eliminates the need for the more cumbersome calculation of $\hat{\varepsilon}_{gg}$ or $\hat{\varepsilon}_{hf}$.

The main choice is therefore between the conservative Greenhouse-Geisser estimate and the liberal Huynh-Feldt estimate. Huynh and Feldt (*ibid.*) proposed that the greater statistical power of the correction using $\hat{\varepsilon}_{hf}$ warranted its use when $\varepsilon \geq .75$. They also argued that it keeps the Type I error rate closer to nominal α than $\hat{\varepsilon}_{gg}$ does. Subsequent authors have tended to stick to this view, advocating $\hat{\varepsilon}_{gg}$ when ε is thought to be close to its lower bound and $\hat{\varepsilon}_{hf}$ when ε is thought to be high (e.g., greater than .75 or .80).

An important issue is how to decide whether sphericity is violated in the first place. One option is to form a test of the null hypothesis that $\varepsilon = 1$. The best-known such test (implemented in a number of software packages) is *Mauchly's sphericity test*. This kind of test of assumptions addresses a largely irrelevant hypothesis. What matters is the degree of violation rather than its presence. Furthermore, the Mauchly test is neither robust to violations of normality nor high in statistical power. It is therefore wise to ignore the procedure entirely. Instead, focus on $\hat{\varepsilon}_{gg}$ and $\hat{\varepsilon}_{hf}$. The average of these conservative and liberal estimates provides a reasonable guide to the extent of any sphericity violation. If this average is close to its lower bound then the Greenhouse-Geisser correction is a safe choice. If the average is around .75 or better $\hat{\varepsilon}_{hf}$ should offer greater power (with only modest risk of Type I error inflation). If both estimates are close to 1 (e.g., .95 or above) the uncorrected F test is defensible (with at most modest Type I error inflation). These decisions should be tempered by the relative cost of Type I or Type II errors. If Type I errors are considered the more costly of the two then $\hat{\varepsilon}_{gg}$ is a good choice. If statistical

power is the main concern then $\hat{\varepsilon}_{hf}$ should be preferred (unless the degree of sphericity violation is severe).

There is a surprising amount published on the relative merits of different sphericity corrections. Recent work has focused more on the relative power of different procedures (see Kirk, 1995). Two further strategies for dealing with sphericity ought to be considered. Probably the best strategy is to avoid tests of multiple *df* effects altogether. Repeated measures contrasts have 1 *df* and therefore make it possible to avoid sphericity concerns altogether. A second strategy is to use a type of *multivariate analysis of variance* (MANOVA) known as *profile analysis*.[8] This MANOVA approach will, under some conditions, provide more powerful tests than those of epsilon-corrected *F* ratios (Kirk, 1995). The precise conditions under which this occurs depend, in a somewhat unpredictable way, on the form of the population covariance matrix. A crude survey of the literature finds that MANOVA tends to provide greater power (relative to corrections using ε_{gg}) when n and the number of repeated measures is large or if ε is close to its lower bound. With small n and few repeated measures experience suggests that MANOVA only rarely reports statistical significance when the Greenhouse-Geisser test does not.[9]

Example 16.1 Uppal (2006) investigated young children's ability to recognize different emotions. Particular interest focused on the ability to recognize and discriminate pride from other emotions (e.g., happiness or surprise). She showed 90 children (aged between seven and nine years) several sets of pictures, each showing actors expressing either pride, happiness or surprise. For each set the children were asked to point to the picture that expressed one of the three emotions (as cued by the experimenter). Subsequent examples will refer to this data set as the pride data. Mean accuracy for the three picture types (averaging over other conditions of the experiment not considered here) was 68.1% for pride, 71.1% for happiness and 78.9% for surprise. Table 16.3 summarizes the output for one-way repeated measures ANOVA with emotion as the fixed factor. The main effect of emotion is not statistically significant, $F_{2,178} = 1.82$, $MS_e = 1,544$ $p = .165$. No differences are detected in average accuracy of recognizing the three emotions.

Table 16.3 Table for one-way ANOVA on the pride data

Source	*df*	*SS*	*MS*	*F*
Subjects	89	241,262	2,711	
Emotion	2	5,616	2,808	1.82
Error (emotion × subjects)	178	274,801	1,544	
Total	269			

In this case there is little reason to worry about sphericity. Even though there are more than two levels on the repeated measures factor, sphericity would only decrease statistical significance (i.e., make the non-significant p value less significant). For completeness, $\hat{\varepsilon}_{gg} = .9879$ and $\hat{\varepsilon}_{hf} = 1.0102$. Both suggest no violation of sphericity, and the upward bias of $\hat{\varepsilon}_{hf}$ is evident (as it exceeds the maximum value of the parameter it is estimating). The Greenhouse-Geisser correction (if implemented) would produce an almost identical main effect: $F_{1.98,175.84} = 1.82$, $p = .166$ (although no correction is warranted when both ε statistics are so close to the upper bound).

The main effect of emotion could also be tested via likelihood or Bayesian inference. AIC for the main effect model is 2808.2, while AIC for the intercept-only model (modeling accuracy of each emotion by the grand mean) is 2807.9. $\Delta AIC = 0.4$, equivalent to $LR_{AIC} = 1.2$ in favor of the grand mean model.

16.3.5 Confidence intervals for repeated measures ANOVA

Calculating confidence intervals (CIs) for a repeated measures ANOVA presents additional difficulties that do not arise in an independent measures design. The standard approach in an independent measures design is to plot error bars around each mean using a pooled standard error. This is useful for depicting the precision with which each mean is measured. These can be adjusted to support inference about differences between means (and for other factors such as multiple testing). The same approach is problematic for repeated measures designs because CIs based on individual means (with or without a pooled error term) may appear to be excessively wide, as they incorporate variance due to systematic individual differences. This variance is eliminated from tests of differences between means in repeated measures designs. Loftus and Masson (1994) argue that a better approach, for ANOVA, is to plot CIs that similarly exclude individual differences. These intervals tend to be much narrower (at least when individual differences are substantial) and therefore give a clearer indication of the presence of systematic differences between means.

Loftus and Masson (*ibid.*) proposed methods for calculating CIs for repeated measures designs using the error term from the corresponding ANOVA. Being based on the same error term, the width of a CI for an individual mean is related to the CI of a difference between means by the familiar factor of $\sqrt{2}$. Although these interval estimates are fairly widely employed, the Loftus-Masson approach has a number of drawbacks (Cousineau, 2005; Morey, 2008). One is that for effects with multiple *df* the pooled error term will be inappropriate if sphericity (equality of variances of differences between means) is not met. A further problem is that they can be awkward to calculate.

Cousineau (2005) proposed a simple alternative that is equivalent to the Loftus-Masson method when paired designs are employed. Morey (2008) showed that the intervals proposed by Cousineau tend to be too narrow (and explained how to correct for this problem). Like the Loftus-Masson approach, the Cousineau-Morey method attempts to strip out individual differences from the calculation of the interval estimate. Where the methods differ is that the Cousineau-Morey interval removes individual differences directly from the data prior to analysis. This is achieved by *participant mean centering* (subtracting the mean of each participant from their observed scores). While this strips out the individual differences, it also alters the mean score per condition. The remedy for this problem is to add the grand mean back on to each score. This process of participant mean centering followed by addition of the grand mean is termed *normalizing* (Loftus and Masson, 1994; Masson and Loftus, 2003).[10] Normalizing relocates all condition effects relative to the grand mean rather than participant means (and therefore relative to an average participant).

The discussion below assumes a one-way repeated measures ANOVA design with J levels (representing J different experimental conditions). If y_{ij} is the score of the i^{th} participant

in condition j, and $\hat{\mu}_i$ is the mean of participant i across all J levels, normalized scores can be expressed as:

$$y'_{ij} = y_{ij} - \hat{\mu}_i + \hat{\mu}_{grand} \qquad \text{Equation 16.5}$$

Why does computing a CI based on the normalized scores lead to intervals that are too narrow? It happens because adding the grand mean to all values induces a positive correlation between the levels that wasn't there before (Morey, 2008). The degree of positive correlation is related to the number of levels J (being largest when $J = 2$ and decreasing as J rises). As the normalized scores are positively correlated, the estimate of error variance computed from them is lower by a factor of $(J-1)/J$ than it would be for the original scores (and hence the CI is too narrow). This factor can be used to correct the CI computed from normalized scores. Thus Morey (2008) suggests computing a CI of the form

$$\hat{\mu}_j \pm t_{df_j,1-\alpha/2}\sqrt{\frac{J}{J-1}}\hat{\sigma}'_{\mu_j} \qquad \text{Equation 16.6}$$

where $\hat{\sigma}'_{\mu_j}$ is the standard error of the mean computed from the normalized scores of the j^{th} level. For factorial designs, Morey indicates that J should be replaced by the total number of conditions across all repeated measures fixed factors (i.e., excluding the subject's random factor). If the design also incorporates independent measures, the intervals can be computed separately for each of the groups defined by the independent measures factors. The intervals themselves have the same expected width as the Loftus-Masson CIs, but do not assume sphericity. Except when $J=2$, their width varies as a function of the variances and covariances of the repeated measures samples.

One criticism of this general approach is that it treats participants as a fixed effect rather than a random effect (Blouin and Riopelle, 2005). This does not matter if the focus is comparison between means, but it does overestimate the precision with which sample means are measured (because it neglects sampling error between participants). Blouin and Riopelle's solution is to obtain CIs for sample means from a multilevel model. A more pragmatic approach is to select the interval estimate to match the purpose of the plot. Cousineau-Morey intervals will be appropriate if means are being plotted to emphasize systematic differences between conditions. However, as the width of a Cousineau-Morey or Loftus-Masson interval is larger than that for a difference in means (by a factor of $\sqrt{2}$), plots of these intervals are easy to misinterpret. Overlap of 95% CIs around individual means is often misinterpreted as indicating that a 95% CI for the difference in means would include zero.

The same problem arises in independent designs and the solution proposed in Chapter 3 was to adjust the width so that absence of overlap corresponds to the required CI for the difference. For large numbers of means (or other statistics) an approach based on multipliers for standard errors can be used (Goldstein and Healy, 1995; Afshartous and Preston, 2010). For the small numbers of means encountered in repeated measures ANOVA analyses, Baguley (2011) proposes the following adjustment to the Cousineau-Morey interval:

$$\hat{\mu}_j \pm \frac{\sqrt{2}}{2}\left(t_{n-1,1-\alpha/2}\sqrt{\frac{J}{J-1}}\hat{\sigma}'_{\hat{\mu}_j}\right) \qquad \text{Equation 16.7}$$

This ensures that the joint width of the difference-adjusted Cousineau-Morey CI is $\sqrt{2}/2$ plus $\sqrt{2}/2 = \sqrt{2}$ times larger than their individual width (matching that for a CI of their difference).

Plotting these adjusted Cousineau-Morey intervals solves the problem of creating interval estimates that are informative about the pattern of differences between means. It is also possible to adjust the interval estimates in other ways (e.g., by incorporating corrections for multiple testing). If both the differences in means and their precision are of interest Baguley (2011) proposes using two-tiered error bars similar to those in Figure 3.8. For instance, the inner tier error bars could be used to display difference-adjusted Cousineau-Morey CIs, while the outer tier could display a 95% CI from a multilevel model (Blouin and Riopelle, 2005; Baguley, 2011). In small samples, the variances and covariances of the repeated measures may be poorly estimated and it may be better to plot CIs (e.g., Loftus-Masson intervals) that rely on a pooled estimate of the covariances between the repeated measures (Baguley, 2011).

Example 16.2 Table 16.4 reports the width and limits of 95% confidence intervals computed using the Loftus-Masson and Cousineau-Morey methods for the one-way ANOVA on the pride data.

Table 16.4 Comparison of Loftus-Masson and Cousineau-Morey 95% CIs, for accuracy of identifying emotions (from the one-way ANOVA of the pride data)

95% CI	Emotion		
	Happiness	**Pride**	**Surprise**
Loftus-Masson			
Lower	62.9	59.9	70.7
Upper	79.3	76.2	87.1
(Width)	(16.3)	(16.3)	(16.3)
Cousineau-Morey			
Lower	63.0	59.4	71.0
Upper	79.2	76.7	86.8
(Width)	(16.3)	(17.3)	(15.8)

The width of the Loftus-Masson CIs is constant, whereas the Cousineau-Morey intervals vary. They do not use a pooled error term (and assume neither homogeneity of variance nor homogeneity of covariance). Their expected value is equal to that of the Loftus-Masson intervals and is identical when $J = 2$.

When plotting the intervals, the main aims are to reveal the pattern of differences between means or the precision with which means are measured. Two-tiered error bar plots make it possible to display both in the same plot (Baguley, 2011). Figure 16.2 plots difference-adjusted Cousineau-Morey intervals constructed from Equation 16.7 as inner tier error bars, and CIs from a multilevel model as the outer tier.

The multilevel model for the outer tier CIs is fitted with an unstructured covariance matrix and, like the Cousineau-Morey interval, does not assume sphericity. The substantial overlap of the inner tiers of the error bars provides little indication of differences between means (consistent with the non-significant main effect of emotion). It is plausible that average accuracy to detect the different

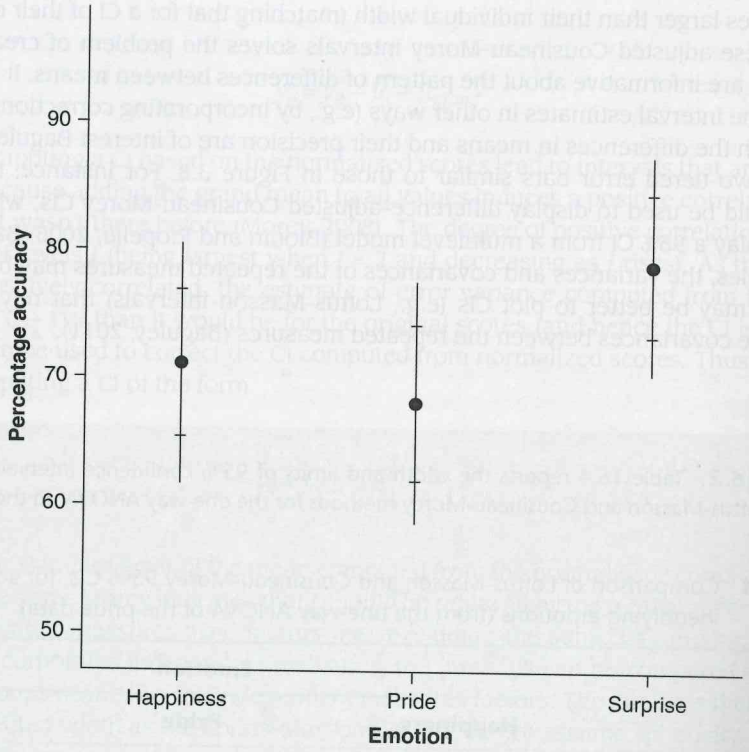

Figure 16.2 Two-tiered error bar plot for the one-way ANOVA of the pride data. The inner tier is a difference-adjusted 95% CI (so that overlapping CIs correspond to a 95% CI in the difference in means that includes zero). The outer tiers display 95% CIs for the individual means.

emotions is equal or similar in accuracy. The outer tier is also informative. Although the 95% CIs are rather wide, all comfortably exclude accuracy levels expected by random guessing (around 33%). There is evidence that children in this age range (seven to nine years) can recognize all three emotions.

16.4 Combining independent and repeated measures: mixed ANOVA designs

A design that includes both independent and repeated measures factors is termed a *mixed* or *mixed measures* design (though other labels such as *split-plot design* are also applied). A mixed ANOVA design allows a researcher to have some of the advantages of a repeated measures design (e.g., to account for individual differences) in the presence of one or more independent measures factors (e.g., individual difference factors). A common example is a pre-post design where different groups are compared at two or more time points (e.g., before and after an intervention).

Adopting a mixed design presents a number of challenges. These arise because individual differences can only be estimated between levels of repeated measures factors in a mixed design.[11] This carves up the design into repeated measures and independent measures components (and some software separates out these components into distinct ANOVA tables). The basic mixed design is a two-way design with a levels on the independent measures factor and b levels on the repeated measures factor. The ANOVA model for this design can be written as:

$$y_{iab} = \mu + \alpha_a + \pi_{i(a)} + \beta_b + \alpha\beta_{ab} + \beta\pi_{ib(a)} + \varepsilon_{iab} \qquad \text{Equation 16.8}$$

As with fully repeated measures designs, the error term is aliased with the $\beta\pi_{ib(a)}$ term (because there is only one observation per repeated measures condition).

Some explanation of the change in subscript notation is required: $\pi_{i(a)}$ represents the individual differences within the independent measures factor and is typically written as *subjects (A)* or *subjects within A* (which is also how it is spoken out loud). This acts as the error term for the independent measures (between subjects) factor. The $\beta\pi_{ib(a)}$ term is the interaction of the repeated measures factor B and subjects within A. This is used as the error term for both the effect of factor B and the A × B interaction. The ANOVA table for a two-way mixed design is shown in Table 16.5.

Mixed designs can be extended so that more than one repeated or independent measures factor is included (see Kirk, 1995; Howell, 2002). The general format remains the same. Independent measures effects are tested using a pooled error term (the subject variance within all independent measures factors) and repeated measures factors or interactions with independent measures factors are tested with the appropriate by-subjects error term.

A mixed design inherits characteristics from both repeated and independent measures designs. For a mixed design to be a true experiment, participants should be randomly assigned to independent measures conditions, and order effects should be controlled by randomization or counterbalancing for repeated measures conditions. It is also assumed that residuals are sampled from independent, normal populations with constant variance. As the residuals are represented by more than one error term, these assumptions should hold separately for both $\pi_{i(a)}$ and $\beta\pi_{ib(a)}$. For all repeated measures effects (e.g., B and A × B in the two-way mixed design) sphericity must be true for the residuals to be independent and have constant variance.

Table 16.5 Two-way ANOVA with mixed measures

Source	*df*	*SS*	*MS*	*F*
Factor A	$a-1$	SS_A	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{subjects(A)}}$
Subjects (A)	$a(n-1)$	$SS_{subjects(A)}$	$\frac{SS_{subjects(A)}}{df_{subjects(A)}}$	
Factor B	$b-1$	SS_B	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_{B\times subjects(A)}}$
A × B	$(a-1)(b-1)$	$SS_{A\times B}$	$\frac{SS_{A\times B}}{df_{A\times B}}$	$\frac{MS_{A\times B}}{MS_{B\times subjects(A)}}$
B × subjects (A)	$a(n-1)(b-1)$	$SS_{B\times subjects(A)}$	$\frac{SS_{B\times subjects(A)}}{df_{B\times subjects(A)}}$	
Total	$N-1$			

This leads to one further difficulty. It is possible to represent each effect by a single population covariance matrix in a repeated measures design. Independent measures designs have a separate covariance matrix for each independent group in the population. It is necessary not only for sphericity to hold for the population each group is sampled from, but also for the covariance matrices of the groups to be equal in the population for the repeated measures test statistics to follow an F distribution. This assumption of mixed ANOVA designs is known as *multisample sphericity*. It is often difficult to satisfy in practice.

Keselman *et al.* (2001) review approaches to dealing with multisample sphericity. In balanced designs – those with equal cell sizes for the independent measures factors – the Greenhouse-Geisser and Huynh-Feldt adjusted tests are known to be robust. MANOVA approaches are also robust in balanced designs for main effects, but may not always be robust to sphericity for interaction tests, even in balanced designs (Keselman *et al.*, 2001). For this reason it is probably safest to employ mixed ANOVA designs with equal or near equal cell sizes wherever possible. It also makes sense to use epsilon-corrected tests in balanced mixed designs (which necessarily always include an interaction test).[12] Keselman *et al.* consider a number of alternative approaches for unbalanced ANOVA designs, but several are difficult to implement without specialist software. Among the approaches they propose are procedures equivalent to those in multilevel regression models. For this reason, switching from mixed ANOVA (or MANOVA) approaches to multilevel modeling is recommended if the design is unbalanced unless sphericity is not an issue.

However, if all repeated measures factors have only two levels (and each effect has 1 df) neither sphericity nor multisample sphericity can be violated. Moreover, with only a few repeated measures in each factor, departures from multisample sphericity may be minor (e.g., if all epsilon estimates are close to one). In these cases ANOVA will probably be preferable to either MANOVA or a multilevel model approach (though there are other reasons, such as missing outcome data, that favor multilevel regression models).

Example 16.3 In Example 16.1, accuracy differences in the pride data were analyzed, ignoring an independent measures factor. Uppal (2006) showed children pictures depicting emotion with face only, torso only, or both face and torso visible. Exactly 30 participants took part in each of the three experimental conditions and it is therefore a 3 × 3 mixed measures ANOVA design. Mean percentage accuracy for the nine conditions (all possible combinations of the levels of each factor) are set out in Table 16.6.

The cell means suggest a more subtle pattern than evident from the earlier one-way analysis. Table 16.7 shows the output of a 3 × 3 mixed ANOVA for these data. There is no indication that sphericity is violated. $\hat{\varepsilon}_{gg} = .9874$ and $\hat{\varepsilon}_{hf} = 1.0102$ for both the emotion and emotion × condition effect.

Table 16.6 Cell means by condition and emotion for the pride data

	Both	Face	Torso
Pride	74.2%	50.0%	80.0%
Happiness	90.0%	83.3%	40.0%
Surprise	93.3%	76.7%	66.7%

Table 16.7 Two-way ANOVA with mixed measures for the pride data

Source	*df*	*SS*	*MS*	*F*	*p*
Condition	2	26,060	13,030	5.27	.007
Subjects (condition)	87	215,201	2,474		
Emotion	2	5,616	2,808	2.12	.123
Condition × emotion	4	44,231	11,058	8.34	< .001
Emotion × subjects (condition)	174	230,569	1,325		
Total	269				

In this analysis, effects of both condition and condition × emotion have been detected. Participants appear to be more accurate on average when both face and torso are visible ($M = 85.8\%$) than for either face ($M = 70.0\%$) or torso ($M = 62.2\%$). The mean accuracy by emotion (ignoring condition) analyzed earlier conceals a more complex pattern. Recognition of emotion appears to be facilitated by different information depending on the type of emotion. Figure 16.3 shows an

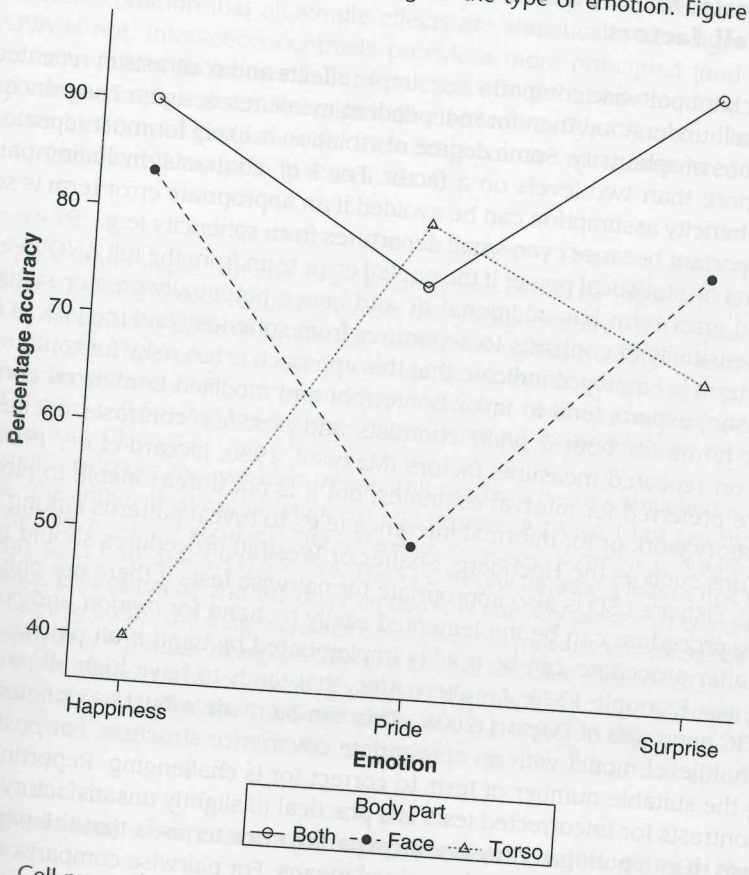

Figure 16.3 Cell means by condition and emotion for the pride data

interaction plot for the condition × emotion effect. The interaction plot suggests that having both face and torso visible produces consistently high performance (though there is a hint that pride is harder to identify than happiness).

Otherwise, it looks as though pride is very difficult to identify from facial expression alone, whereas as happiness is very hard to identify from body posture on its own. It seems likely that the interaction is largely down to this 'crossover' in performance between pride and happiness when only face or only torso is presented. Further exploration of the interaction is warranted to confirm this interpretation.

16.5 Comparisons, contrasts and simple effects with repeated measures

16.5.1 Comparisons, contrasts and simple effects with repeated measures on all factors

The best approach for pairwise comparisons, simple effects and contrasts in repeated measures ANOVA is less well understood than for independent measures designs. The principal problem is that of violations of sphericity. Some degree of violation is likely for most repeated measures analyses with more than two levels on a factor. For 1 *df* contrasts, including pairwise comparisons, the sphericity assumption can be avoided if an appropriate error term is selected. The error term is important because even small departures from sphericity (e.g., $.97 < \varepsilon < 1$) can have a material impact on statistical power if the pooled error term from the full ANOVA is used (Boik, 1981). A pooled error term has additional *df* and hence potentially greater statistical power. However, the sensitivity of contrasts to departures from sphericity and the lack of robustness if a pooled error term is employed indicate that this approach is too risky for routine application.

For this reason, experts tend to favor Bonferroni and modified Bonferroni corrections with separate error terms for both *a priori* contrasts and *post hoc* contrasts (including pairwise comparisons) on repeated measures factors (Maxwell, 1980; Jaccard *et al.*, 1984). Bonferroni corrections are preferred for interval estimates, but it is not unreasonable to plot uncorrected CIs for exploratory work or for informal inference (e.g., to reveal patterns among means). More powerful options such as the Hochberg, Shaffer or Westfall procedures should be adopted for formal testing. Fisher's LSD is also appropriate for pairwise tests if there are only three means. The Hochberg procedure can be implemented easily by hand for *a priori* and *post hoc* testing, while the Shaffer procedure can be readily implemented by hand if all pairwise comparisons are required (see Example 13.5). An alternative, that tends to have high all pairs power, is to adapt the PCIC approach of Dayton (2003). This can be made robust to violations of sphericity by fitting a multilevel model with an appropriate covariance structure. For *post hoc* contrasts, determining the suitable number of tests to correct for is challenging. Reporting the tolerance to implicit contrasts for uncorrected tests is a practical (if slightly unsatisfactory) solution.

One benefit of computing contrasts with separate error terms is that the procedure reduces to the calculation of a paired *t* test of weighted means. For pairwise comparisons this is merely the familiar paired *t* test of the differences between the level means you are interested in. For a contrast involving more than two means it may be necessary to calculate the contrast score

in the usual manner. For this reason it will often be convenient to choose contrast weights with absolute values that sum to two. One method is to compute the weighted positive and negative scores for each participant and use a paired t test. Alternatively, you could calculate the contrast score C for each participant and compute a one sample t test of the null hypothesis that $C = 0$ in the population. Frequentist, likelihood or Bayesian inferences can be derived from the resulting t statistic.

Simple main effects follow the same pattern. Howell (2002) suggests calculating a one-way ANOVA for the effect of one repeated measures factor at each level of the second factor. The advantage of this procedure is it uses a separate error term for each simple effect. If the simple effect has only 1 df it is equivalent to the paired t test approach already described. Furthermore, the simple effect can adopt an epsilon correction for sphericity if more than two means are involved. The same rationale applies to simple interaction effects within pure repeated measures designs. A simple interaction effect for a three-way interaction can be tested via two-way repeated measures ANOVA at each level of the third factor. As with simple main effects and simple interactions in independent designs, care must be taken with these analyses. Simple effects are tests of aggregate effects (decomposing sums of squares from, for example, a main effect and a two-way interaction). The strategy can have very low power to detect patterns of interest. It is quite common that all simple effects are statistically non-significant, even if the overall ANOVA is not. Interaction contrasts provide a more principled (and potentially more powerful) route to follow up an omnibus interaction effect that has multiple df. Testing many simple effects in a complex factorial design will also lead to Type I error inflation – an almost paradoxical combination of low power for the effects you are most interested in and high Type I error rates for the overall analysis (Maxwell, 2004).

Example 16.4 In Example 16.1 the mean accuracy was 68.1% for pride, 71.1% for happiness and 78.9% for surprise. Previous research suggests that young children, and to some extent adults, have difficulty recognizing pride (e.g., it is not classed among basic emotions such as happiness, surprise, or disgust). This hypothesis could be tested by a contrast comparing accuracy for pride versus the average accuracy for surprise and happiness. This contrast could use integer weights $\{-2, +1, +1\}$ or $\{-1, +0.5, +0.5\}$ to keep the contrast score on the original percentage scale. The latter is suitable for a CI.

The first step is to create the weighted means. This can be done by computing a new variable (*non-pride*) as the arithmetic mean of happiness and surprise. A paired t test can then be run to compare *pride* and *non-pride* accuracy. The difference in weighted means is 6.94%, 95% CI of $[-3.65, 17.54]$, $t_{89} = 1.30$, $SE = 5.33$, $p = .196$. This CI is helpful because it tells us that the estimate of the difference between pride and the other emotions is not very precise (it could plausibly be zero, or approaching 20%). The study has insufficient power to measure this difference precisely.

The t statistic makes it possible to apply a Bayesian t test or calculate a likelihood interval. Using a JZS prior, the Bayes factor in favor of the null hypothesis that pride is not harder to recognize than the other emotions is 5.25, while the 1/8 likelihood interval is $[-3.99, 17.88]$.

16.5.2 Comparisons, contrasts and simple effects with mixed measures

Calculating contrasts and comparisons for mixed designs depends on the source of the effect being investigated. For a pure independent measures effect (a factor or interaction with no

repeated measures), it is possible to treat the analysis exactly as for an independent measures design. For instance, with *a priori* contrasts and comparisons a powerful modified Bonferroni procedure such as the Westfall procedure could be employed. For repeated measures effects sphericity is a serious concern, and the approach set out for pure repeated measures designs is recommended.

The repeated measures approach of separate one-way ANOVAs (or factorial ANOVA for three-way and higher-order interactions) also works for repeated measures simple effects. For instance, if three types of problem (low, medium or high difficulty) were presented either to novice or expert problem solvers this could be analyzed with a 2×3 mixed design (with two levels on the independent factor and three levels on the repeated factor). The two simple main effects for the repeated measures factor test the effects of problem difficulty within the novice group and within the expert group. Each of these could be analyzed with a one-way ANOVA incorporating only one of the groups (and therefore not pooling error terms). Having three levels, there is a risk that sphericity is violated. This can be assessed by looking at the epsilon estimates and, if necessary, employing a correction.

Calculating a simple main effect for the independent factor in a mixed design is more treacherous. These effects are differences between groups (e.g., novices and experts) that are analyzed at just one measurement level of a repeated measures factor (e.g., only for low difficulty problems). The obstacle for the independent measures tests in this situation is that a simple main effect is not a method of decomposing an interaction. It decomposes *SS* attributable to both the main effect and the interaction. Hence the error variance for the independent factor simple main effect is distributed across the two error terms: subjects (A) and B $\times$ Subjects (A). The most accurate test will be obtained by pooling these two sources of error. The *SS* for the error term is the sum of the $SS_{\text{subjects}(A)}$ and $SS_{B\times\text{subjects}(A)}$, while the *df* is sum of $df_{\text{subjects}(A)}$ and $df_{B\times\text{subjects}(A)}$. The pooled error term is therefore:

$$MS_{error} = \frac{SS_{subjects(A)} + SS_{B\times subjects(A)}}{df_{subjects(A)} + df_{B\times subjects(A)}} \qquad \text{Equation 16.9}$$

The trouble is that this pooled error term combines estimates of very unequal population variances and the resulting test statistic will not follow the usual *F* distribution. This is a variant of the Behrens-Fisher problem encountered in the context of the independent *t* test (see Howell, 2002). It can be dealt with by applying a similar correction to the *df*. Corrected df_{error} for the simple main effect are computed as:

$$\nu' = \frac{\left(SS_{subjects(A)} + SS_{B\times subjects(A)}\right)^2}{\frac{\left(SS_{subjects(A)}\right)^2}{df_{subjects(A)}} + \frac{\left(SS_{B\times subjects(A)}\right)^2}{df_{B\times subjects(A)}}} \qquad \text{Equation 16.10}$$

For multiple *df* interactions, interaction contrasts can also be calculated. These can be generated along the same lines as those for an interaction contrast in an independent measures design. The contrast weights for the interaction effect define patterns of means that differ between groups (e.g., differential linear trends). The difficulty is deciding on what error term to use. If sphericity is not violated, the error term for the interaction in the mixed ANOVA should be employed. If sphericity is violated, but the design balanced, the error *df* could be adjusted using $\hat{\varepsilon}_{gg}$ or $\hat{\varepsilon}_{hf}$. If sphericity is violated and the design unbalanced, the contrast could be run via a multilevel model.

Example 16.5 Figure 16.3 shows a complex interaction among nine cell means. The interaction effect has 4 *df* and could be decomposed into one or more interaction contrasts. A hypothesis of particular interest in this case is how well the interaction is explained by a 2×2 interaction in which pride is better recognized from torso alone while happiness is better recognized by face alone. A reasonable choice of contrast weights for this hypothesis would be:

	Both	*Face*	*Torso*	Σ
Pride	0	−1	+1	*0*
Happiness	0	+1	−1	*0*
Surprise	0	0	0	*0*
Σ	*0*	*0*	*0*	*0*

Note that having absolute weights summing to two would have retained the original percentage scale.

The next step in the analysis would normally be to create a double-centered table of contrast weights (see Key Concept 15.1). By happenstance, the initial weights are already double-centered (with both row and column marginals summing to zero).

The contrast weights can then be multiplied by the cell means from Table 16.6 to give the contrast score:

	Both	*Face*	*Torso*	Σ
Pride	0	-1×50	$+1 \times 80$	*30*
Happiness	0	$+1 \times 83.3$	-1×40	*43.3*
Surprise	0	0	0	*0*
Σ	*0*	*33.3*	*40*	*73.3*

$SS_{contrast}$ is derived from the contrast score of 73.3, and is:

$$SS_{contrast} = \frac{C^2}{\frac{1}{n}\left(\sum_{j=1}^{J} w_j^2\right)} = \frac{73.3^2}{\frac{1}{30}(1+1+1+1)} = 40,296.7$$

The sphericity assumption was not violated in the earlier analysis. For this reason, it is reasonable to use the emotion × subjects (condition) MS_{error} term from the original ANOVA. This is 1,325, and so $F = 40,296.7/1,325$, and the contrast could be reported as: $F_{1,174} = 30.4$, $MS_e = 1,325$, $p < .01$.

As is often the case for a contrast, the explanatory power of the interaction contrast is of greater interest than the test. *SS* for the interaction is 44,231 and therefore the interaction contrast explains about 91% of the interaction effect ($r^2_{alerting} = .91$). This is equivalent to a correlation of .95 between contrast weights and the residualized cell means ($r_{alerting} = .95$). Although there is quite a bit going on in Figure 16.3, much of the variation is down to main effects. Of the variation that remains, the vast majority can be explained in terms of pride being harder to identify from facial expression than from body posture, and happiness being harder to identify from body posture than facial expression.

16.5.3 Effect size

Repeated measures and mixed designs are especially difficult to obtain appropriate standardized effect size metrics for. Many commonly calculated quantities (e.g., η_p^2 or g) are not comparable to similar metrics obtained from independent measures designs. It is a good idea to compare effects from different designs using unstandardized measures (e.g., simple mean differences) as a first – and possibly only – step. Standardized effect size metrics need to take into account factors that may distort the standardizer (the variance or standard deviation used to scale the effect). An important contributor to the standardizer in an independent measures design is individual differences (which in a repeated measures design are estimated separately from other sources of variance).

The generalized effect size measures of Olejnik and Algina (2003) provide a starting point for 'design neutral' standardized effect size metrics. Their approach is to calculate generalized statistics that treat repeated measures equivalently to independent measures designs. One proviso is that, for statistical power or sample size estimation, the appropriate metric is one that matches the design of the study being planned. In theory η_g^2 can be calculated with the formulas for other factorial designs by treating subjects as an additional measured factor. In practice, ANOVA software rarely provides the *SS* for such a calculation in a convenient format. Following Olejnik and Algina, calculating the required denominator for η_g^2 by subtraction is suggested. The goal is to exclude all manipulated factors or interactions with only manipulated factors (except the one under consideration). An indicator variable I is used to designate whether the effect under consideration is a manipulated factor ($I = 1$) or a measured factor ($I = 0$).

$$\eta_g^2 = \frac{SS_{effect}}{SS_{total} - \sum_{manip} SS_{manip} + I \times SS_{effect}} \qquad \text{Equation 16.11}$$

This formula excludes manipulated factors from the denominator (adding the focal effect back in only if the focal effect is a manipulated factor). Interactions with measured factors are considered measured factors. Repeated measures fixed factors tend to be manipulated factors, though it may be reasonable to treat them as measured factors in some situations.

Olejnik and Algina (2003) also extend ω_g^2 (generalized omega-squared) to designs with repeated measures factors. The correct formulas can become rather complex and the simplest solution is to refer to Tables 2, 3 and 4 of their paper.

Example 16.6 In the two-way mixed ANOVA for the pride data, the two factors are the emotion to be recognized and the experimental condition (whether the pictures showed face, torso or both face and torso). The experimental condition is a canonical example of a manipulated variable. Emotion is manipulated by the experimenter, and for comparisons with other experiments might be considered as such. For other purposes – for example to gauge impact on everyday performance – it may be considered a measured variable (because expressions of happiness, pride and surprise are a routine part of everyday experience).

Treating both variables as manipulated factors, η_g^2 for the interaction is:

$$\eta_g^2 = \frac{SS_{effect}}{SS_{total} + 1 \times SS_{effect} - \sum_{manip} SS_{manip}} = \frac{44231}{521678.2 + 1 \times 44231 - (44231 + 5616 + 26060)} = .090$$

The interaction would account for around 9% of the total sample variance in an equivalent independent measures design. Replacing SS_{effect} with $SS_{contrast}$ allows versions of η^2 to be calculated for a contrast. Thus η_g^2 for the interaction contrast is:

$$\eta_g^2 = \frac{SS_{effect}}{SS_{total} + I \times SS_{effect} - \sum_{manip} SS_{manip}}$$

$$= \frac{40296.7}{521678.2 + 1 \times 40296.7 - (40296.7 + 5616 + 26060)} = .083$$

In practice it is usually better to evaluate contrasts relative to the effect they are decomposed from, rather than in terms of total variance. The contrast explains about 91% of the interaction effect and about 8% of sample variance in an equivalent independent measures design.

16.6 MANOVA

Multivariate analysis of variance (MANOVA) is technique of potential interest whenever correlated measurements are obtained. MANOVA is designed for applications in which several correlated outcome variables or *DV*s (dependent variables) are measured.

An important application for repeated measures designs is a form of MANOVA called profile analysis. This does not assume sphericity and sometimes has greater statistical power than an epsilon-corrected ANOVA. In this application, already considered in passing, the repeated measures are treated as correlated outcome variables (which, in a sense, they are). Knowing in advance whether the MANOVA tests are more powerful than the corrected ANOVA analysis is far from simple. It would be possible to estimate the statistical power of each technique if the population covariance matrix Σ were known. Obtaining a good enough estimate of Σ (e.g., in a pilot study) to determine the relative power of the two approaches is likely to be difficult (but see Miles, 2003). Estimates of variances and covariances from small samples tend to be very imprecise. The MANOVA approach will tend to have greater power when many repeated measures are taken, when the degree of sphericity violation is large and when samples sizes are large (but there are departures from this general trend).

For designed experiments with small n and few repeated measurements, and certainly for designs with two levels on the repeated measures factors, ANOVA should probably be preferred. In unbalanced mixed designs MANOVA tests are not robust to violations of multisample sphericity. They may also be problematic for interaction tests in balanced designs (Keselman *et al.*, 2001; Olson, 1974). Keselman *et al.* (2001) also argue against combining MANOVA with epsilon corrections as this can produce unpredictable results. For mixed designs with imbalance, a multilevel model is recommended.

Two further applications of MANOVA, both controversial, need to be addressed. The first is the use of MANOVA to screen for effects prior to ANOVA. This is sometimes employed when a researcher has collected a number of different outcome measures. The second is to increase the statistical power to detect effects when correlated outcome measures are collected. Both practices should generally be avoided. Arguments against the MANOVA approach are part of a broader aversion both to screening tests and tests of multiple *df* effects.

Using MANOVA to screen for effects is analogous to using omnibus F tests prior to *post hoc* tests of all pairwise comparisons. If the omnibus H_0 is true (i.e., there are no differences between means for any of the outcome variables in the population) then the MANOVA test of an effect (e.g., a main effect of factor A) protects subsequent ANOVA tests on the separate outcome variables. This assumes that a researcher does not perform any further tests of differences, a rule that is not always adhered to (Huberty and Morris, 1989). Only very rarely should a researcher adopt this practice. Most research is conducted on the premise that some effects are likely to exist. Only rarely is the omnibus null hypothesis plausible.[13] It is more likely that a partial null hypothesis is true; that there are non-zero effects for some outcomes and zero or negligible effects for others (Huberty and Morris, 1989; Jaccard and Guillamo-Ramos, 2002).

Jaccard and Guillamo-Ramos (2002) provide a very clear illustration of the problem. Imagine a clinical study with one main outcome variable (Y_1 = depression) and four secondary outcomes (Y_2 to Y_5 measuring anxiety, self-confidence and so forth). There may be a substantial effect for Y_1 but negligible or zero effects for Y_2 to Y_5. If the Y_1 effect is large enough, the overall MANOVA main effect might be statistically significant. Subsequent ANOVA tests on the secondary outcomes would then be unprotected with respect to familywise error (considering tests of the same treatment effect on different outcomes as the family). Jaccard and Guillamo-Ramos argue that modified Bonferroni corrections to separate univariate tests of secondary outcomes are a better solution (though the primary outcome should rarely if ever be corrected in this way). An alternative strategy is to report tests of all outcomes without correction. This may be sensible if the outcomes are correlated (though a powerful correction such as the Westfall procedure may be preferred). Reporting unmodified effects may be reasonable if the correlations between predictors are positive and substantial, provided care is taken not to conceal the true extent of testing when communicating the results.

Using MANOVA to screen for effects prior to ANOVA is generally a very bad idea. It will usually lead either to inadequate Type I error protection or to decreased statistical power. The latter occurs when the partial null is true, but the omnibus null is not rejected. This is a consequence of the screening test itself lacking statistical power (see Zimmerman, 2004). The power issue is subtle, and will be addressed shortly. The main issue is that the test of the omnibus null hypothesis lacks focus relative to the tests of individual outcomes such as Y_1 or Y_2.

It has already been hinted that the MANOVA omnibus tests will lack statistical power, but this isn't quite true. Think about the rationale for using MANOVA to provide more powerful tests. In Jaccard and Guillamo-Ramos's example there were five outcome measures all likely to be correlated with successful treatment for depression. In a small study, none of the individual outcomes might reach statistical significance, but all might show an effect in the right direction. Could you not use MANOVA to analyze the whole set of outcome variables for a more sensible test? This seems like a good idea, but MANOVA will not always provide a more powerful test. The power of the omnibus test in MANOVA depends on sample sizes, effect sizes and the pattern of correlations between the outcomes (Cole *et al.*, 1994). Interestingly, Cole *et al.* show that if the outcome variables have high positive correlations (a situation quite likely where the outcomes are repeated measures) MANOVA will not always have high power (depending on the mix of effect sizes). With the right mix of effect sizes and correlations, MANOVA can have greater power than univariate ANOVA. However, other approaches may also have good power. In fact, the most obvious choice of outcome in some studies is probably just to average the variables (e.g., taking their mean or the mean of their z scores depending on whether they have the same or different scales). This approach seems particularly attractive when there are positive correlations between outcome measures and fairly consistent effects across those measures.

Many misapplications of MANOVA stem from a lack of understanding of how MANOVA works. There are many good introductions to MANOVA (e.g., Field, 2009), but most focus on calculation and basic interpretation. It will help to explore the simple case of a two-group design with y outcome variables Y_1 to Y_y. A fundamental (and potentially surprising) characteristic of MANOVA is that it is not an analysis of the separate outcome variables at all. It is an analysis of a linear transformation or a combination of the outcomes that we will refer to as Y_C. Technical details of the mathematics of the combination are given by Grayson (2004), but the basic form of Y_C is similar to a contrast:

$$Y_C = c_1 Y_1 + \ldots + c_y Y_y \qquad \text{Equation 16.12}$$

The weights or coefficients (c_1 to c_y) for the combination are chosen to maximize the value of a one-way ANOVA F statistic for the differences between the means Y_1 to Y_y.[14] It is important to remember that this maximizes the F ratio of the combination for the effect being considered. Different effects in the same design nearly always end up with different weights. MANOVA therefore is really a 'disguised' ANOVA with a transformed Y variable. Grayson points out that Type I protection (for the omnibus H_0) is obtained because an ANOVA test of an individual outcome such as Y_1 is also a linear combination of the full set of Y variables (one in which the weights are one for Y_1 and zero for all other outcomes):

$$Y_1 = 1 \times Y_1 + 0 \times Y_2 \ldots + 0 \times Y_y$$

If the omnibus null were true and this 'combination' was statistically significant by chance alone (the definition of a Type I error), then the combination that maximizes the F statistic, Y_C, would also be statistically significant.

What gives cause for caution about MANOVA (at least as it is routinely applied) is that the linear combination that maximizes an F ratio is an inherently atheoretical approach. There is no guarantee that this linear combination (the *discriminant function*) is interpretable. Grayson (*ibid.*) provides several plausible examples of simple data sets that do not have an interpretable structure (or at least not one that makes any kind of theoretical sense).

If a researcher has a hypothesis about the differences in a particular outcome measure, it does not seem like a good idea to test the hypothesis using a different outcome measure (whether that outcome measure is theoretically interpretable or not). A number of experts advise against MANOVA if you are really are interested in the differences between the means of individual outcome measures (Huberty and Morris, 1989; Jaccard and Guillamo-Ramos, 2002; Grayson, 2004). If you have an *a priori* reason to think that some linear combination of several outcomes is meaningful, then a better approach might be to combine the outcomes yourself (e.g., using an average or weighted average).

Huberty and Morris (1989) discuss legitimate research questions for MANOVA. These rest on whether the inter-relationships between outcome measures themselves are of interest to a researcher. In particular, MANOVA will sometimes throw up theoretically meaningful linear combinations. Although this is true, other approaches to this problem should be considered. If the primary interest is in combinations of predictors that best discriminate different outcomes then *discriminant analysis* may be appropriate. Multilevel regression models can also be extended to deal with multiple outcome variables in what is termed a multivariate multilevel model (Hox, 2010). The multilevel approach is attractive because it may permit explicit modeling of the correlations between outcome variables and can handle missing outcomes.

MANOVA provides two additional causes for concern, one widely known and the other less so. The widely known issue is that MANOVA produces several different, rival test statistics. With only two outcome variables they all reduce to the same statistic, *Hotelling's* T^2.[15] With more than two outcomes *Wilk's* Λ (lambda),[16] *Pillai's trace*, *Hotelling's trace* and *Roy's largest root* are usually provided by MANOVA software. Olson (1976) recommends *Pillai's trace* as the most robust to violations of MANOVA assumptions, but *Wilk's* Λ is also popular (see Field, 2009). The final cause for concern is in relation to effect size. A number of MANOVA effect size metrics have been developed, but none seem particularly useful. For instance, eta-squared variants can be derived from *Wilk's* Λ (a measure of unexplained sample variance for Y_C), but have unattractive properties. Because Y_C is maximized separately for each test, the total proportion of variance explained by all effects can exceed one, and will not strictly be comparable even within an analysis. As different studies will capitalize on sampling variability to determine Y_C, MANOVA effect sizes are also not strictly comparable between studies (and this is true also for unstandardized differences in Y_C).

In summary, MANOVA has a potential application in pure repeated measures analyses (the focus here). It sometimes provides more powerful tests than epsilon-corrected tests when sphericity is violated. It is less useful for mixed designs, but may be appropriate for tests on main effects in balanced designs. However, the power advantage of MANOVA is not consistent; switching to a multilevel model is recommended. The multilevel model approach has the flexibility to mimic both ANOVA and MANOVA analyses and to relax constraints inherent in both models (e.g., sphericity and multisample sphericity). The common strategy of MANOVA followed by univariate ANOVA is inappropriate for testing multivariate hypotheses and should be replaced by genuinely multivariate approaches (Huberty and Morris, 1989; Enders, 2003).

16.7 ANCOVA with repeated measures

ANCOVA with repeated measures expands the familiar repeated measures designs to include a covariate. This approach is most appropriate for randomized experimental designs with continuous confounding variable, but can also be applied to non-experimental designs (provided the usual cautions about using the covariate as a 'statistical control' are borne in mind). Adding a covariate appears to be a relatively harmless process, but can end up being rather messy. A one-way design with single covariate would take the form:

$$y_{ij} = \mu + b(C_i - \mu_C) + \pi_i + \tau_j + \tau\pi_{ij} + \varepsilon_{ij} \qquad \text{Equation 16.13}$$

The covariate C in Equation 16.13 is centered (by subtracting its mean μ_C) and takes the same value for each of the $i = 1$ to n observations. Both these points turn out to be very important.

As it turns out, this form of pure repeated measures ANCOVA design may be uninteresting. In Equation 16.13, the covariate is measured between participants; there is one covariate score for each of the participants. Variation in the covariate equates to individual differences between participants. If the covariate was not present, this variation would get absorbed by the subjects term of the repeated measures analysis. If the covariate varied across repeated measures (often termed a *time-varying covariate*) then the model would also be inappropriate, because it would fail to capture important variation across observations indexed by i and j. Time-varying covariates can be dealt with in a number of ways, but multilevel regression models provide an

elegant solution. The main motivation to add a time-stable covariate to a pure repeated measures model therefore differs from an independent measures design. The reason for including the covariate should be because its effect is of substantive interest, or because you are interested in testing covariate-by-treatment interactions.

Time-stable covariates are more interesting in mixed designs, because they potentially increase sensitivity to independent measures effects. Even so, there may still be advantages to separating out the independent measures analysis from the repeated measures analysis. In the basic one-way design ANCOVA above, it would perhaps be easier to run a one-way repeated measures ANOVA and a separate regression (or correlation) between outcome and covariate. One reason for this is that the way repeated measures ANCOVA is implemented in some software is problematic. A stronger reason to use a repeated measures ANCOVA model, in both mixed and pure repeated measures designs, is to include covariate-by-treatment interactions for the repeated measures factor. In some studies this will not be imperative, but in studies with measured factor by manipulated factor interactions it is strongly advisable (Yzerbyt *et al.*, 2004).

In the earlier presentation of factorial ANCOVA, centering of the covariate was a convenience. This is not so for tests of main effects of repeated measures factors in ANCOVA. For repeated measures analyses a number of packages use difference coding (as in MANOVA profile analyses). Difference coding strips out individual differences from the *SS* calculation on the repeated measures factor (because the means of differences between levels rather than the means of the levels are being compared). Delaney and Maxwell (1981) explain how adjusting the difference scores using a covariate messes up this calculation. One way to understand what happens is to realize that the difference scores have already eliminated the average effect of the covariate from the *SS*. A further adjustment for the covariate mean of each group confuses matters. It would, in effect, adjust the *SS* for the repeated measures main effect for the difference between the mean of the covariate and zero. This is not something you would usually want to do (*ibid.*).

Once a covariate is added, any tests of repeated measures main effects are no longer interpretable in isolation (though the overall model, including the prediction equation, is not compromised). This is a little like the problem with product terms between uncentered main effects in moderated multiple regression. The solution is the same: center the covariates prior to adding them to the ANOVA. For the same reasons that apply in moderated multiple regression, interactions between covariates and other predictors are unaffected. Adding any new predictor or set of predictors to a model allows you to test the effect of the predictor using standard model comparison approaches (e.g., *F* tests or ΔAIC).

As a general approach, repeated measures ANCOVA can be considered a mixed ANOVA in which the independent measures effect is a continuous covariate with a single *df*. This produces an ANOVA table resembling Table 16.5. Additional covariates and product terms to test covariate by factor interactions (moderator effects) can be added to the model. The product terms should be computed using the centered covariates and it is best to compute and add the centered covariate and product terms to the model yourself (unless you are certain that your software handles them correctly).

Two broad strategies for repeated measures ANCOVA are recommended. One is the full repeated measures ANCOVA with all repeated measures factors, all independent measures and all covariates of interest. If you adopt this strategy it is advisable that you use effect coding (or equivalent ANOVA parameterization) and center all covariates. If covariates are not centered, the repeated measures main effects may be uninterpretable. An alternative strategy is to conduct two separate analyses (see Rutherford, 2001). In the first analysis, no covariates are

included but all factors of interest (independent or repeated measures) and their interactions are present. This first analysis is used only to assess interactions between and main effects of repeated measures factors. The second analysis adds all covariates of interest and any covariate by factor interactions. The second model can be used to assess the remaining effects (i.e., any that are not pure repeated measures effects).

The former 'global' strategy is probably the safest method, provided all covariates are centered. The alternative 'two phase' method is useful if you want to estimate effects at a value of a covariate other than its mean. This is useful, for example, in developmental trajectory analysis (Thomas *et al.*, 2009). It can also be also useful for obtaining contrasts and simple effects for repeated measures factors or interactions between repeated measures factors that do not need covariate adjustment.

16.7.1 ANCOVA, pre-post designs and gain scores

A *pre-post design* is a repeated measures design in which measurements of an outcome variable are taken before an intervention or experimental manipulation (the pre-test or baseline) and again afterwards (the post-test). When paired measures from a pre-post design are analyzed, a popular strategy is to simplify the analysis using *gain scores*. A gain score is the change in outcome between the pre-test and post-test (i.e., post-test score minus pre-test score). For a repeated measures design with more than two time points it is possible to generalize gain scores as *change relative to baseline*. Here the baseline score is subtracted from all repeated measures prior to ANOVA analysis (and is equivalent to calculating separate gain scores for each post-test measurement).

Analysis of gain scores isn't the only option. Think about a two independent group design in which the aim of the study is to determine whether the change between pre-test and post-test scores differs between the groups. Two other alternatives could be selected, one more interesting than the other. First, one could use mixed measures ANOVA with pre-test and post-test scores as levels of the repeated measures factor. This is a fairly uninteresting alternative, because the F ratio from the ANOVA interaction is equivalent to the independent t test of the difference in gain scores between groups ($t^2 = F$). Analysis of gain scores (or analysis of change relative to baseline) and ANOVA of the raw scores are equivalent with respect to the test of the differential change in outcome.

The second alternative is to use the pre-test or baseline score as a covariate in the analysis. This models the change between the pre-test and post-test outcome in a very different way. To illustrate what is going on, we'll adapt the approach of Wright (2006) and present the equations for ANOVA and ANCOVA as regression models where group is a dummy coded predictor (X). In the gain score model (equivalent to ANOVA) the regression model is:

$$gain_i = post_i - pre_i = b_0 + b_1 x_i + e_i \qquad \text{Equation 16.14}$$

The corresponding ANCOVA model is:

$$post_i = b_0 + b_2 pre_i + b_1 x_i + e_i \qquad \text{Equation 16.15}$$

Comparing the two will be easier if the gain score model is rearranged in terms of the post-test scores (by adding the pre-test scores to both sides). This produces:

$$post_i = b_0 + pre_i + b_1 x_i + e_i \qquad \text{Equation 16.16}$$

The crucial difference between Equation 16.15 and Equation 16.16 is that the ANCOVA model estimates an additional parameter: the slope of the pre-test scores b_2. In the gain score model, the pre-test scores are a constant and therefore the value of b_2 is implicitly assumed to be one.

Having presented the models in this way, it should be clear that the two models can lead to different conclusions, because the tests of the differences between groups (the test b_1) are not identical. The most famous demonstration of this is *Lord's paradox* (Lord, 1967). Lord showed that two groups could show no difference when compared using gain scores, but a large difference when comparing the means adjusted for their pre-test score. Lord used an artificial example, but the 'paradox' can also be found in real data sets (Wainer and Brown, 2004; Wright and London, 2009). In the Wright and London example (using data from London *et al.*, 2009), children's recall for an event (a magic show) was measured two weeks after the event and again ten months later. The age of the children at the start of the study varied from about five to nine years. Aside from the predictor X being a continuous covariate (the age in months of a child at the start of the study), this is identical to the earlier example. One of the research questions was whether the change in recall differed for younger and older children. The analysis of gain scores leads to the following prediction equation:

$$post = 3.061 + pre - 0.085\,age$$

The ANCOVA analysis gives a very different outcome:

$$post = -1.729 + 0.105\,pre + 0.044\,age$$

The effects are not only different, but they lie in opposite directions. Which model is correct? The literature on Lord's paradox is substantial and not easy to summarize. It turns out that there is no definitive answer. Because the models differ, each of them addresses a different hypothesis. The correct answer depends on the context of the study and its objective. That stated, there are ways to approach an answer. A number of commentators have pointed out that if you don't care what causes the difference in post-test scores (as may occur in some applied work) it may be reasonable just to focus on the observed differences using gain scores (Wright, 2006). More likely, a researcher will care about the relationships between predictors and the outcome measure. Wainer's (1991) position is that the choice of model depends on untestable assumptions about the relationship between pre-test and post-test scores (the parameter estimated by b_2). If $b_2 = 1$ in the population, the gain score approach will be correct, but if $b_2 \neq 1$ then ANCOVA is appropriate. If $b_1 = 1$ then you are assuming that, on average, the mean of the post-test and pre-test scores would be the same in the population if the effects of the other predictors were not present. It is the counterfactual nature of the assumption that makes it untestable; it depends on information not present in the sample. As Wainer (*ibid.*, p. 149) puts it: 'The very nature of untestable assumptions means that there is no statistical procedure that can be counted on to settle this issue. The answer to this question must come from other sources.'

Returning to Wright and London's example, would you expect the post-test recall to be on average the same (all other things equal) as pre-test recall? The simple answer is no. After ten

months you'd expect all the children to have forgotten some of the information. This suggests the ANCOVA model is preferable. It is the more plausible of the two models, as it predicts older children remember more than younger children at the post-test (conditional on the initial level of recall). However, this is not necessarily the 'correct' answer. Even in this apparently clear-cut example Wright and London (2009) contrive a scenario in which the gain score model may be appropriate. What if an investigator has to select either an older or younger child to interview immediately after a crime (with the other child being interviewed much later)? The gain score model is useful here because it suggests that the older child should be interviewed first, because a delay would result in a larger absolute reduction in recall relative to the younger child.[17] Although the scenario is contrived it illustrates how great care should be taken in selecting the correct model. In addition, the usual ANCOVA concerns about the linearity of covariate effects and the absence of interactions apply (*ibid.*). In this case the linearity assumption is implausible and it might be sensible to assume an approximate power law relationship, though the precise nature of the forgetting curve is a matter of considerable debate (e.g., see Lansdale and Baguley, 2008).

Even though specifying the correct model can be hard, some guidelines are available. If the predictor of interest (e.g., group) is manipulated by the experimenter and therefore can be assigned at random, then all approaches produce unbiased tests, though ANCOVA tends to have greater statistical power (Wright, 2006). If pre-test score is confounded with the predictor of interest then ANCOVA can produce unbiased estimates of the de-confounded effects (though this is not necessarily the question of interest). ANCOVA is usually preferred in simple experimental and quasi-experimental studies looking to understand the effects of individual predictors. Maris (1998) also argues that ANCOVA is preferable if the pre-test is used to determine the assignment to the groups being compared. If the mechanism of assignment to groups is often unknown (e.g., because pre-existing 'groups' are measured) this presents a problem. These are the cases where it is necessary not only to know the precise hypothesis you wish to test, but also to try and work out which of the untestable assumptions about the model is most plausible. This issue is important because selecting the wrong model for the research question may introduce bias.

Some broad conclusions are possible. For randomized designs and situations where the pre-test determines assignment to groups ANCOVA has greater statistical power and is unbiased (Maris, 1998; Jamieson, 2004; Van Breukelen, 2006). For non-randomized studies where assignment to groups is not based on pre-test scores, it is argued that ANCOVA may have greater bias (Jamieson, 2004; Van Breukelen, 2006). However, recent work comparing randomized and non-randomized studies suggests that this is not inevitable (Cook *et al.*, 2008; Shadish *et al.*, 2008).

It appears that the key factor is being able to include predictors likely to have caused differential baseline performance, rather than just the usual range predictors used for matching or statistical control (e.g., easy to measure demographic factors such as age or sex). The argument here is closely related to that for dealing with missing data (e.g., drop out) by including predictors of missingness. This recent work is consistent with that of Senn (2006) who demonstrated that in situations where ANCOVA is biased it is difficult to design studies to detect a causal effect of a treatment in which change scores would not also be biased. It is sensible to take steps to remove as many sources of bias as possible (e.g., at the design stage or by inclusion of covariates). The question of which statistical model is best has no simple answer and will depend on the design and context of the study, with work on how best to select an analysis still ongoing (e.g., Dinh and Yang, 2011; Cousens *et al.*, 2011).

Box 16.2 Structuring repeated measures data: long form versus broad form

To run any repeated measures ANOVA it is necessary to structure the data in a way that preserves the relationship between repeated observations and the units being observed. To keep the explanation that follows manageable, assume that the repeated measures are on human participants and that the outcome is measured at two time points (*Time 1* and *Time 2*). Repeated measures analyses are tricky to deal with because different software requires data structured in different ways.

Independent measures data for regression and related analyses are usually represented in a data file or spreadsheet as follows:

Participant	*Predictor*	*Outcome*
P1	8.3	29
P2	6.1	12
P3	8.5	23
...	...	...

Each variable (a covariate or grouping variable) is a distinct column and each participant a separate row. These rows are often termed 'cases'.

The default repeated measures data structure in many packages is to represent data in what is sometimes called *broad form*:

Participant	*Predictor*	*Time 1*	*Time 2*
P1	8.3	29	33
P2	6.1	12	19
P3	8.5	23	20
...	...	...	...

This arrangement preserves the one case (row) per participant property of the independent measures data structure. It differs in the important respect that each repeated measures outcome is defined as a separate variable. Software that uses the broad form therefore has to have a method of linking the outcome variables in some way (e.g., requiring the user to define the appropriate columns as a repeated measures or within subjects factor).

An obvious alternative structure is to represent the data in *long form*:

Participant	*Predictor*	*Time*	*Outcome*
P1	8.3	1	29
P1	8.3	2	33
P2	6.1	1	12
P2	6.1	2	19
P3	8.5	1	23
P3	8.5	2	20
...	...	...	...

The long form violates the property that the data from a single participant is restricted on a single case, but retains the property of the standard independent measures structure that the outcome is described by observations in a single column. The broad form is popular because it provides a more efficient summary of the data (i.e., one with less repetition). However, for some regression models – especially multilevel models – it is easier to work with data in long form.

For multilevel data structures, the long form extends very easily to represent three, four or more levels. An important advantage is that it allows you to represent both cross-classified and nested data structures. In a nested design, observations are clustered uniquely within other measurement units. Here participants are nested within groups (with no participant in more than one group):

Group	*Participant*	*Outcome*
G1	P1	17
G1	P1	9
G1	P2	8
G1	P2	5
G2	P3	11
G2	P3	12
...	...	...

In a cross-classified design, lower-level observations are not clustered uniquely within high-level units. A fully crossed repeated measures structure is an extreme version where both measurement units are clustered within each other. Here the same four items in the experiment (I1 to I4) are clustered with every participant, though it would be just as valid to state that all *n* participants are clustered within every item:

Participant	*Item*	*Outcome*
P1	I1	10
P1	I2	11
P1	I3	7
P1	I4	9
P2	I1	15
P2	I2	8
...	...	...

Switching between the two formats can be time-consuming and is prone to error. Both SPSS and R have commands or functions for switching between the two data structures. For small data sets you may prefer to rearrange data in a spreadsheet such as Excel (because it is easy to check the results visually). In larger data sets it is better to use software to rearrange the data, but it is vital to check descriptive and other statistics to make sure the data are structured correctly.

16.8 R code for Chapter 16

16.8.1 One-way ANOVA with repeated measures (Example 16.1)

There are a number of different approaches to running repeated measures ANOVA models in R. There are pros and cons to each approach. Repeated measures ANOVA models with balanced data is one area in which the power and flexibility of R can be annoying; there are several very different methods to fit the same model (each with its own quirks).

For a basic model it is possible to use `aov()` by using the `Error()` function to specify the correct error terms to use in the analysis. It is also necessary to code a participant or subject variable as a factor for use within the `Error()` function. This approach uses the long form of the data rather than the broad form (used in most other software). The long form of the pride data

from Example 16.1 is in a file named `pride_longS.csv` while the broad form is in `prideS.csv`. To run a one-way repeated measures ANOVA using `aov()` try the following code:

```
pride.long <- read.csv('pride_longS.csv')
pride.long$participant <- as.factor(pride.long$participant)

pride.anov <- aov(accuracy ~ emotion +
  Error(participant/emotion), pride.long)
summary(pride.anov)
```

The participant identifier can also be defined as a factor within the command:

```
pride.anov <- aov(accuracy ~ emotion +
  Error(factor(participant)/emotion), pride.long)
summary(pride.anov)
```

The `Error()` function is required to get R to recognize the repeated measures structure and select the correct error terms. The `participant/emotion` argument indicates that the `emotion` factor is fully nested within participants.

Had the original data used number and letter combinations (e.g., P1, P2 etc.), R would have converted `participant` to a factor by default. There are several ways to convert between the long and broad form (see Box 16.1). The following code uses the `stack()` and `unstack()` functions, which work well for simple cases. More complex data sets should use `reshape()`. The next example turns the long form into the broad form and then turns the broad form back. The default column names for `stack()` output are 'values' and 'ind', so the last command just updates these. Note that the column orders in the `pride.long2` are different from `pride.long`. The broad form is handy for getting the means or *SD*s of the repeated measures conditions.

```
pride.broad <- unstack(pride.long, accuracy ~ emotion)
mean(pride.broad)
sd(pride.broad)

pride.long2 <- stack(pride.broad)
names(pride.long2) <- c('accuracy', 'emotion')
```

The `car` package provides a range of powerful functions for running repeated measures ANOVAs and related analyses (and includes Greenhouse-Geisser and Huynh-Feldt corrections). The `ez` package provides a more user-friendly ('SPSS-like') way to access some `car` functions. `ez` will automatically load `car` and several other packages that it needs. Here is how to run a one-way repeated measures ANOVA using the `ezANOVA()` function:

```
library(ez)

ezANOVA(data=pride.long, dv=.(accuracy), wid =.(participant),
  within =.(emotion))
```

`ezANOVA()` has several useful features. As well as providing sphericity corrections it converts the subject term defined by `wid` to a factor automatically. In addition, `car` and therefore `ez` defaults

to using hierarchical (Type II) sums of squares, whereas `aov()` defaults to sequential (Type I) sums of squares. `ez` also attempts to calculate η_G^2 (`ges` in the output) by treating subjects *SS* as a measured factor (other measured factors can be named using the `observed` argument). For one-way repeated measures designs this produces sensible output (equivalent to η^2).

Two additional methods could be mentioned at this point. One is to fit a MANOVA model using the `car` package, though this offers no advantage over the `ez` approach at this stage. The final method is to use a multilevel model (sometimes called a linear mixed model). This approach is more versatile than repeated measures ANOVA, but will produce equivalent results for a balanced design if the model is set up in the appropriate way. The example here uses the `lme()` function in the `nlme` package (part of the base distribution for R), although `lmer()` from `lme4` could also be used (but it has slightly different syntax).

```
library(nlme)

lme.fit <- lme(accuracy ~ emotion, random = ~1|participant,
  pride.long)
anova(lme.fit)
```

Further explanation of this and related functions will be provided in Chapter 18. One advantage of this approach is that if they are fitted by maximum likelihood, R will calculate AIC and related statistics for repeated measures models using `lme()` or `lmer()`. These are unavailable from `aov()` objects fitted including `Error()`. The default fitting method for `lme()` is *restricted maximum likelihood* (RML). This produces inferences that are identical to repeated measures ANOVA in a completely balanced design with a fully nested data structure. To compare AIC (or AIC_C or BIC) for fixed effects it is necessary to switch to maximum likelihood methods (so that the log-likelihood is estimated correctly). The following commands compare the one-way model with emotion as a factor against the intercept-only model (in which all emotions have the same mean level of accuracy):

```
io.ml <- lme(accuracy ~ 1, random = ~1|participant,
  pride.long, method='ML')

ow.ml <- lme(accuracy ~ emotion, random = ~1|participant,
  pride.long, method='ML')

AIC(io.ml, ow.ml)
delta.aic <- AIC(ow.ml) - AIC(io.ml)
exp(delta.aic/2)
```

16.8.2 Plotting repeated measures CIs (Example 16.2)

The Loftus-Masson CIs in Table 16.4 can be obtained using functions provided by Baguley (2011). Similar code is available from Wright (2007), who also provides a bootstrap version. The functions in Baguley (2011) support plotting of difference-adjusted Cousineau-Morey CIs and CIs from multilevel models with different covariance structures (for both one-way repeated measures and two-way mixed designs). Afshartous and Preston (2010) provide R code for extending Goldstein-Healy intervals (Goldstein and Healy, 1995) to dependent designs.

The examples here use the Baguley (2011) functions. The functions for one-way repeated measures analyses take data in broad form (e.g., `pride.broad` from the preceding section). The first function `lm.ci()` gives an unadjusted Loftus-Masson function:

```
lm.ci(pride.broad)
```

The second function `cm.ci()` gives a Cousineau-Morey interval. Its default is a difference-adjusted CI (in which overlapping CIs correspond to a CI for a difference in means that includes zero), using the formula in Equation 16.7. To get an unadjusted CI use the call:

```
cm.ci(pride.broad, difference = FALSE)
```

The plot in Figure 16.2 uses the `two.tiered.ci()` function. The default is to plot the inner tier as a difference-adjusted Cousineau-Morey CI and the outer tier as 95% CIs for individual means with a covariance matrix estimate that does not assume sphericity:

```
two.tiered.ci(pride.broad, ylab = 'Percentage accuracy', xlab
  = 'Emotion', grid=TRUE)
```

16.8.3 ANOVA with mixed measures (Example 16.3)

In Example 16.1, an independent measures factor with three levels was ignored for the pride data. This manipulated whether children saw pictures of expression with face, torso or both face and torso visible. Including the additional grouping factor presents no difficulty once the full data set is loaded.

```
pride.long <- read.csv('pride_long.csv')
pride.mixed <- aov(accuracy ~ emotion*condition +
  Error(factor(participant)/emotion), pride.long)
summary(pride.mixed)
```

To obtain the cell means and other summary statistics for data in long form the `ezStats()` function from the `ez` package can be used. This takes more or less the same format as `ezANOVA()`.

```
ezStats(data=pride.long, dv=.(accuracy), wid =.(participant),
  within =.(emotion), between =.(condition))
```

The output includes *n*, mean and *SD* per cell as well as Fisher's least significant difference (though this is not suitable for designs with repeated measures factors). To obtain sphericity corrections you can again use `ezANOVA()`.

```
ezANOVA(data=pride.long, dv=.(accuracy), wid =.(participant),
  within =.(emotion), between =.(condition))
```

The `lme()` function also works, though the `car` package output reported via `ez` is more informative.

```
lme.fit <- lme(accuracy ~ emotion*condition, random =
  ~1|participant, pride.long)
anova(lme.fit)
```

To plot something similar to Figure 16.3, `interaction.plot()` could be used:

```
interaction.plot(pride.long$emotion, pride.long$condition,
  pride.long$accuracy, xlab='Emotion', ylab= 'Percentage
  accuracy', legend = FALSE)
```

A prettier, color plot could be specified using the `ez` package `ezPlot()` function. This function returns an object that can be edited further using the `ggplot2` package.

```
ezPlot(data=pride.long, wid = .(participant), dv=.(accuracy),
  within = .(emotion), between = .(condition), x = .(emotion),
  split = .(condition), do_bars=FALSE)
```

The *x* argument defines which of the repeated (within) or independent (between) factors is on the *x*-axis and `split` determines whether to split by levels of another factor (as separate lines). Setting `do_bars=TRUE` adds error bars based on Fisher's LSD (not desirable in this case). Baguley (2011) includes a two-tiered error bar function for mixed designs based on `two.tiered.ci()`. Below, `reshape()` is used to get the full pride data set into broad form and the participant ID stripped out so that the grouping variable is the first column in the new data frame `pride.mixed`. Shown here is a basic plot that can be edited or relabeled:

```
pride.broad2 <- reshape(pride.long, idvar = 'participant',
  direction = 'wide', timevar = 'emotion', v.names =
  'accuracy')[2:5]

two.tiered.mixed(pride.broad2, group.var='first', lines=TRUE)
```

16.8.4 Contrasts on a repeated measures factor (Example 16.4)

A contrast is a form of weighted comparison of means. With pure repeated measures designs, contrasts can be run as paired *t* tests between weighted means. To illustrate this Example 16.4 runs a contrast for the pride data set. Working with the broad form of the data set, first extract two vectors, one for the accuracy of the pride emotion and one for the mean happiness and surprise. The *t* test can then be run comparing the means of these vectors.

```
pride.mean <- pride.broad$pride
nonpride.mean <- (pride.broad$happiness +
  pride.broad$surprise)/2

t.test(nonpride.mean, pride.mean, paired=TRUE)
```

As an alternative to the t test or 95% CI the t statistic or mean and SE can be used calculate Bayes factors or likelihood intervals if desired (using functions from Chapter 11).

```
JZS.prior.Bf.1s(1.30, 90)
t.lik.int(6.94444, 5.33, 90)
```

There is nothing special about repeated measures contrasts – any ANOVA contrast that outputs a t statistic can be used to calculate a Bayes factor or a likelihood interval.

16.8.5 Interaction contrasts with repeated measures (Example 16.5)

Interaction contrasts for pure repeated measures effects can be calculated using the t test approach described earlier. This avoids sphericity problems (possibly with some loss of power), but does not extend to mixed measures interaction terms. One way to deal with mixed effects is to adopt the approach described in Example 16.5. To calculate an interaction contrast it is necessary to decide on the interaction weights and it is usually best to work with them in matrix form. For instance, the matrix for the contrast in Example 16.5 would be:

```
cont <- matrix(c(0,0,0,1,-1,0,-1,1,0), 3, 3)
```

The interaction residuals (the cell means after sweeping out the main effects) could be obtained in several ways. However, R can provide them directly using the `model.tables()` function. Using the interaction model from the `aov()` command fitted earlier (`pride.mixed`) the residuals are given in the table for the `emotion:condition` interaction.

```
model.tables(pride.mixed)
int.resids <-
  model.tables(pride.mixed)$tables$'emotion:condition'
```

The contrast score is the summed product of the interaction residuals multiplied by the contrast weights, while $SS_{contrast}$ is obtained from contrast score and n per cell.

```
c.score <- sum(cont*int.resids)
c.score

n <- 30
ss.contrast <- c.score^2/(sum(cont^2)/30)
ss.contrast
```

$SS_{contrast} = 40333.33$ and is slightly larger than in the worked example (because of rounding error). F, p, $r_{alerting}$ and $r^2_{alerting}$ are:

```
r.alerting <- cor(as.vector(resids),as.vector(cont))
r2.alerting <- r.alerting^2
c(r.alerting, r2.alerting)

ms.error <- 1325
F <- ss.contrast/ms.error
F ; pf(F, 1, 174, lower.tail = FALSE)
```

As the F ratio is very large the p value is tiny, though it is more interesting to focus on $r^2_{alerting}$. This suggests that the interaction contrast accounts for most of the variance of the interaction effect.

It is also possible to adapt the cell means approach described in Chapter 15 to mixed and repeated measures models. These analyses are run as multilevel models. First, the cell means interaction model needs to be fitted using `lme()`.

```
library(nlme)
pride.cmm <- lme(accuracy ~ 0 + emotion:condition, random =
  ~1|participant, pride.long)
pride.cmm
```

The `glht()` function isn't the most flexible option for running the contrast. Instead, we'll use `estimable()` from the package `gmodels`. This is used for calculating 'estimable functions' (linear functions of model parameters) which contrasts are a special case of. The function is more flexible than `glht()`, but doesn't directly support corrections for multiple testing. The key advantage of `estimable()` is that it takes input from a wide range of model objects plus a contrast matrix. To reduce confusion, the contrast matrix works with named parameters (with unnamed parameters set to zero). The following example grabs the names required for the contrast from the model object and adds them to the contrast matrix. Because a cell means model has been fitted, the names are those of the nine cell means (e.g., `emotionhappiness : conditionboth` is the cell mean for the happiness emotion with both torso and face presented).

```
library(gmodels)

labels <- names(coef(pride.cmm))[1:9]

contr <- matrix(c(0,0,0,1,-1,0,-1,1,0),1,9,
  dimnames=list('contrast',labels))

estimable(pride.cmm, contr)
```

The reported t statistic is 5.517036. Its squared value is 30.43769, identical, allowing for rounding error, to the 30.44025 obtained for F by the other method. The function provides a 95% CI if a `conf.int=.95` argument is added, but these are on the incorrect scale. The following code will rescale the contrast score, *SE* and CI without also (incorrectly) rescaling the df:

```
estimable(pride.cmm, contr, conf.int=.95)[1:2]/2
estimable(pride.cmm, contr, conf.int=.95)[6:7]/2
```

The multilevel approach has several advantages. It can be extended to deal with violations of sphericity and multisample sphericity or to more complex models (e.g., with unbalanced designs or time-varying covariates). The resulting models will not always produce statistics with an exact t or F distribution (though in this case, the statistics are known to be exact if the assumptions are met).

16.8.6 Generalized η^2 measures for repeated measures (Example 16.6)

Calculators for generalized η^2 are not easy to automate, because the decision over whether to consider variables as manipulated or measured is slightly subjective (and may vary with context). Using R can make the calculations easier – particularly for the total *SS*.

```
N <- 30*3*3
ss.tot <- var(pride.long$accuracy)*(N-1)
```

To get the generalized measures then involves manually adjusting the denominator using information from the ANOVA output:

```
ss.effect <- 44231
ges <- ss.effect / (ss.tot + ss.effect - sum(ss.effect + 5616
  + 26060))
ges.contrast <- ss.contrast / (ss.tot + ss.contrast -
  sum(ss.effect + 5616 + 26060))
c(ges, ges.contrast)
```

Note that `ezANOVA()` also provides η_g^2 output for the interaction. Its output of 0.09026790 matches that calculated in Example 16.6 – indicating that it is treating the repeated measures factor and the independent measures factor as manipulated factors. This is debatable. Emotions vary naturally in our environment so the repeated measures factor could be considered measured. The emotion factor should therefore be considered measured. If so, η_g^2 for the interaction should be:

```
ges <- ss.effect/(ss.tot - 26060)
ges
```

This can be checked with `ezANOVA()` where the observed argument allows you to override the defaults and treat a fixed factor as measured rather than manipulated.

```
ezANOVA(data=pride.long, dv=.(accuracy), wid =.(participant),
  within =.(emotion), between = .(condition), observed=
  .(emotion), detailed=TRUE)
```

16.8.7 R packages

Bates, D. M., Maechler, M., and Bolker, B. M. (2011) *lme4*: Linear mixed-effects models using S4 classes. R package version 0.999375-39.
Fox, J., and Weisberg, S. (2010) *An R Companion to Applied Regression*, (2nd edn). Thousand Oaks CA: Sage.
Lawrence, M. A. (2011) *ez*: Easy analysis and visualization of factorial experiments. R package version 3.0-0.
Pinheiro, J., Bates, D. M., DebRoy, S., Sarkar, D, and the R Core team (2011) *nlme*: Linear and Nonlinear Mixed Effects Models. R package version 3.1–98.
Warnes, G. R., *et al.* (2011) *gmodels*: Various R programming tools for model fitting. R package version 2.15.1.

16.9 Notes on SPSS syntax for Chapter 16

16.9.1 ANOVA with repeated measures (Examples 16.1 and 16.3)

Repeated measures ANOVA analyses in SPSS use the general linear model `GLM` command (which can also run `UNIANOVA` commands using the same syntax).

SPSS data file: `pride_rm.sav`

```
GLM pride happiness surprise
  /WSFACTOR=emotion 3
  /METHOD=SSTYPE(2)
  /PRINT=DESCRIPTIVE
  /WSDESIGN=emotion.
```

The statement `pride happiness surprise` defines the variables making up the within subjects (repeated measures factors) and `/WSFACTOR=emotion 3` tells SPSS that there is one repeated measures factor with three levels. The `/PRINT` subcommand requests descriptive statistics. Fitting a mixed ANOVA is straightforward with the same command:

```
GLM pride happiness surprise BY condition
  /WSFACTOR=emotion 3
  /METHOD=SSTYPE(2)
  /PRINT=DESCRIPTIVE
  /WSDESIGN=emotion.
```

A covariate could be added with a `WITH` statement. Additional repeated measures or independent measures can be added. For instance, a four-way mixed ANOVA with two repeated measures factors and two independent measures factors (`A` and `B`) could be run as:

```
GLM C1D1 C1D2 C2D1 C2D2 BY A B
  /WSFACTOR=factor1 2 Polynomial factor2 2 Polynomial
  /METHOD=SSTYPE(2)
  /WSDESIGN=factor1 factor2 factor1*factor2
  /DESIGN=A B A*B.
```

The repeated measures factors each have two levels and are defined across the variables `C1D1 C1D2 C2D1 C2D2`. The independent measures factors are defined using `BY A B` and the model is specified in the `/WSDESIGN` and `/DESIGN` subcommands. The `Polynomial` statement in the `/WSFACTOR` definition tells SPSS to run polynomial contrasts on the repeated measures factors and their interactions.

These models can also be fitted using the SPSS multilevel modeling `MIXED` command. As well as being able to relax the sphericity assumption for these models, it is also possible to obtain AIC, AIC_C and BIC.

16.9.2 Repeated measures CIs (Example 16.2)

Wright (2007) provides SPSS syntax for running Loftus-Masson CIs, while Cousineau provides SPSS syntax for the uncorrected CI using normalized data. To obtain the corrected Cousineau-Morey intervals for the pride data you can also adjust nominal confidence (in this case to 98.2%) to obtain the correct width for a 95% CI when $n = 30$ and $J = 3$ (see Baguley, 2011). The following syntax obtains the normalized data from the participant means and the grand mean (72.6851851851852). The latter is calculated from the participant means using `DESCRIPTIVES`.

SPSS data file: `pride_rm.sav`

```
COMPUTE pmeans=(pride+happiness+surprise)/3.
DESCRIPTIVES VARIABLES=pmeans
  /STATISTICS=MEAN.

COMPUTE n_pride = pride - pmeans + 72.6851851851852.
COMPUTE n_happiness = happiness - pmeans + 72.6851851851852.
COMPUTE n_surprise = pride - pmeans +72.6851851851852.
EXECUTE.

GRAPH
  /ERRORBAR(CI 98.2) = n_pride n_happiness n_surprise.
```

16.9.3 Contrasts for repeated measures designs (Examples 16.4 and 16.5)

Many repeated measures contrasts can be run using the paired *t* test commands. The following commands compute a weighted contrast for the pride versus other emotion contrast in Example 16.4.

SPSS data file: `pride_rm.sav`

```
COMPUTE pride_v_other=pride-(happiness+surprise)/2.
EXECUTE.

T-TEST
  /TESTVAL=0
  /VARIABLES=pride_v_other
  /CRITERIA=CI(.95).
```

This is a very versatile method for running contrasts on repeated measures. It avoids a pooled error term, but may sacrifice power if sphericity is true. The contrast can also be run using GLM, by changing the default within-subject contrasts:

```
GLM pride happiness surprise
/WSFACT = emotion(3)
special
(1 1 1
-1 .5 .5
0 -1 1).
```

This gives an *F* test of the contrast by default. With three levels there are two orthogonal contrasts by default (usually linear and quadratic polynomials specified by `Polynomial`). The contrasts are set out as a 'matrix' over several lines to reveal the structure (though it is not strictly necessary). The first row `1 1 1` defines the intercept, the second the contrast of interest and the third is an arbitrary contrast orthogonal to the first (which happens to compare the happiness and surprise means).

The default repeated and mixed measures ANOVA output (see the mixed ANOVA syntax above) includes polynomial contrasts for the repeated measures factors and interaction contrasts for effects involving repeated measures factors. Unfortunately, these are only really interpretable if the repeated measures factor is ordered (e.g., time points), which it is not for the pride data. In principle, it is possible to get SPSS to run custom interaction contrasts for mixed designs, but the difficulty of setting up the contrast coefficients often makes it easier to carry out the analysis by another route (e.g., by hand).

16.10 Bibliography and further reading

Keselman, H. J., Algina, J., and Kowalchuk, R. K. (2001) The Analysis of Repeated Measures Designs: A Review. *British Journal of Mathematical and Statistical Psychology*, 54, 1–20.

Kirk, R. E. (1995) *Experimental Design* (3rd edn) Belmont, CA: Brooks/Cole.