More about Partial Effect Sizes

In Section C, I explained why the ordinary formula for eta squared (SS_effect / SS_total) produces a misleading result when both factors in a two-way ANOVA involve experimental manipulations, and why this is not the case when one of the factors is based on pre-existing individual differences that you would expect to find in the samples of most experiments (because they exist in the general population of interest). Partial η², as described by J. Cohen (1973), was recommended when both factors are manipulated. However, if one of your experimental (i.e., manipulated) factors has only two levels, and you want to express its effect size with a distance measure, such as Cohen's d, a recent article by Gillett (2003) defines a partial version of d in a manner quite analogous to the defining of partial eta squared. He uses “d” (without bolding) to represent the sample estimate of a partial d. Expressed in terms of the symbols I have been using, Gillett's (2003) formula for estimated partial d is:

\[ d_A = \frac{\bar{X}_{A_1} - \bar{X}_{A_2}}{s_{\text{cell}}} (1 - \frac{3}{4 \cdot df_{W} - 1}) \]  

Formula 14.10

This formula is a simple extension to the two-factor case of the sample measure I have been calling “g”; note that it includes the same bias correction factor as Formula 12.27 (Chapter 12, section D). The two means being subtracted are the marginal means for the two levels of the A factor (averaged across all of the levels of some B factor), and \( s_{\text{cell}} \) is just the square root of the ordinary MS_within-cell error term of the two-way ANOVA. Imagine that your original experiment was designed to use separate groups in order to compare recall under two different sets of instruction (i.e., two levels of Factor A), involving either subvocal repetition or the formation of visual images. Then you decide to divide each of your groups in half (randomly, of course) so that you can introduce a second factor (Factor B) – perhaps type of material to memorize (e.g., half of each instruction group gets abstract words and the other half gets concrete words). Given the relation between d and t and F, and dropping the bias correction factor (which is certainly not needed for conceptual understanding or crude power approximations), Formula 14.10 can be rewritten more simply as:

\[ d_A = \sqrt{\frac{2 F_A}{bn}} \]  

Formula 14.11

where b is the number of levels of the B factor (2 in our example) and n is the number of participants per cell, so that bn is the number of participants being averaged together to create each of the marginal means for Factor A (as in the numerator of the first part of Formula 14.10). Because Factor B also has only two levels, it is just as reasonable to find dB using Formula 14.11, changing FA to FB and bn to an.

The Variance-Reduction Model

The scenario I have just described, adding a second experimental factor to an experimental one-way ANOVA, falls under what Gillett (2003) labeled the “Variance-Preservation Model”, in that there is no reason to expect that the error term of the two-way ANOVA (MS_within-cell) will be any smaller or larger than the error term of the one-way ANOVA (MSW) before the second factor was added – hence, within-group variance is “preserved.” The contrasting model, the “Variance-Reduction Model”, applies
to the design that I used to introduce the two-way ANOVA in Section A of this chapter; that is, a “grouping” factor is added to an experimental factor. In terms of the example in this section, imagine that again you are starting with a one-way ANOVA studying two types of instruction, but this time the second factor involves dividing each of the groups by a median split on their visual imagery vividness (VIV; as measured by a questionnaire given previously in an unrelated setting), rather than adding any new manipulation. As I have pointed out in earlier sections of this chapter, this is the situation in which you can expect MS_{within-cell} to be smaller than MS_W to the extent that VIV has some effect on memory ability, which is certainly plausible. This is the same situation I described in Section C as not requiring partial eta squared; ordinary eta squared should not be misleading, because SS_{total} is not being inflated by a second experimental factor. In fact, in this case it is partial \( \eta^2 \) that would usually be misleading, because its denominator is likely to be artificially decreased by assessing variability separately for high and low VIV participants. Partial \( \eta^2 \) from the experiment just described may be a considerable overestimate of the instruction effect in a future experiment, in which participants are not divided by VIV. Of course, ordinary \( \eta^2 \) is easy to find for the instruction (i.e., experimental) effect; it is just the SS of the effect divided by SS_{total} (on the other hand, a partial \( \eta^2 \) would be appropriate for assessing the grouping factor, but that is less likely to be of interest).

**Estimating \( d \) for an Experimental Factor that is Crossed with a Grouping Factor**

Ironically, \( d \), which is estimated very simply as a partial \( d \) by Formula 14.10 or 14.11 when dealing with a two-way ANOVA with variance preservation, requires a more complicated estimating formula under the Variance-Reduction model, though the concept is straightforward. The estimate of \( d \) is labeled “\( g \)” in this model by Gillett (2003) to distinguish it from the estimate of partial \( d \) discussed above. The simple version of the formula for \( g \) follows:

\[
g_A = \frac{\bar{X}_{A_i} - \bar{X}_{A_t}}{s_{level}}
\]

where \( s_{level} \) is the square root of MS_W from the original one-way ANOVA (i.e., before dividing participants into subgroups based on VIV). In other (more complicated) words, \( s_{level} \) is based on a pooling of the variances of the participants at each level of \( A \), ignoring the grouping factor. [By the way, Gillett (2003) uses the term stratified factor to refer to the same kind of IV that I have been calling a grouping factor. Unfortunately, this label is not standard. For example, Olejnik and Algina (2003) use the terms “measured” factor or “blocking” factor for the same purpose. All of these terms (and there are others) refer to a variable that is based on pre-existing, individual differences rather than on any experimental manipulation.]

If you have already divided SS_{total} into all of the components necessary for a two-way ANOVA, then \( s_{level} \) can be found by adding back to \( s_{cell} \) the SS’s for the main effect of \( B \) and the \( A \times B \) interaction, then dividing by the corresponding sum of df’s, and finally taking the square root. After a good deal of algebraic manipulation, the formula for \( g_A \) can be written in terms of Formula 14.11 multiplied by a correction factor that compensates for the fact that the F ratios are based on \( s_{cell} \), rather than (the usually larger) \( s_{level} \). Once again leaving out the bias correction factor shown in Formula 14.10, the formula for \( g_A \) based on the already calculated F’s for the two-way ANOVA is given by Gillett (2003) as:

\[
g_A = \sqrt{\frac{2F_A}{bn} \left( \frac{df_{cell} + df_B + df_{A \times B}}{df_{cell} + df_B F_B + df_{A \times B} F_{A \times B}} \right)}
\]

Formula 14.12
You can see that if the F’s were zero for both the B factor and the interaction, the correction factor would actually be a bit larger than 1.0. In that case, there is no variance reduction at all, just some “extra” df’s in the two-way ANOVA, and as n gets larger in comparison to the number of levels of the factors, the correction factor gets closer to 1.0. If the two other F’s are both 1.0 (the more typical type of result when there is no population effect of Factor B or the interaction), the correction factor equals 1.0 – no correction is needed in that case. As the F’s for the other two effects get larger, the correction factor becomes a smaller fraction, but a large n will lessen the impact of the other effects, because the same F associated with a larger n implies a smaller effect size.

Bear in mind that Formula 14.12 would not be appropriate for assessing the effect of the grouping factor in the two-way ANOVA; for that purpose you would want to use a partial η² or estimate of partial d, to avoid using an error term that is inflated due to the variance added by the experimental factor. Gillett (2003) also includes more complex versions of Formula 14.12 that are appropriate for three-way ANOVA’s with one or two "stratified" factors. If all three factors involve experimental manipulations, then Formula 14.10 is all you need (if you want to use Formula 14.11 instead, the term “bn” becomes “bcn”, where c is the number of levels of the third experimental factor).

Estimating Omega Squared under the Variance-Reduction Model

Gillett’s (2003) formulas are specifically designed for the case in which the factor of interest (usually an experimental factor) has only two levels (the other factor can be either an experimental or grouping factor with any number of levels), and you would like to use a distance measure of effect size. If the factor for which you would like to estimate effect size has more than two levels, you can use an estimate of Cohen’s f, but although this measure may be useful for power analysis, it has not become popular as a descriptive measure of effect size (in part, perhaps, because very different patterns of three or more population means can yield exactly the same value for f). For multi-level factors, you can use the ordinary or partial η² for descriptive purposes, but these measures are often quite biased, in that they can overestimate the population effect size (ω²) considerably. Formula 12.14 shows you how to correct the bias in the one-way ANOVA case; this formula also applies to estimating ω² for an experimental factor with any number of levels, when it is crossed with a grouping factor with any number of levels. However, if the second factor also involves an experimental manipulation, and you want to correct the bias inherent in using partial η², you can use an estimate of partial omega squared, as given by Olejnik and Algina (2003):

\[
est.\omega^2 = \frac{SS_A - df_A MS_W}{SS_A + (N - df_A) MS_W} \text{ Formula 14.13}
\]

where MS_W is the usual within-cell error term of a factorial ANOVA, and N is the total number of participants, summed across all groups. Olejnik and Algina (2003) present formulas for estimating ω² in three-way factorial ANOVA designs with various combinations of experimental and blocking factors, including factors involving repeated measures. I will discuss these considerably more complicated effect size estimates in Chapter 16, section D.

The Power of Complex Interaction Contrasts

As mentioned in Section B, when two experimental factors are combined in a factorial ANOVA, but they do not interact with each other, a useful (if not terribly interesting) advantage of the two-way ANOVA is its economy. Of course, this economy is essentially lost if there is a strong (especially, if disordinal) interaction between the two factors. However, a significant interaction often provides
information that is more interesting (or, at least, more specific) than the main effects. For example, suppose that there are three major methods for teaching reading (traditional, phonics, visual/holistic), and that for each method there are a number of primers recommended for use in conjunction with that method (there may be few, if any, texts that are considered suitable for all three methods). If you want to confirm that the match of text with method is important to optimize performance, you might select the most highly recommended text for each method, and then cross the three texts with the three methods in a 3 X 3 factorial design, as shown in Table 14.11 (each cell mean represents the average reading score of 10 sixth-grade pupils). A sizeable main effect of text would suggest that there are overall differences in the effectiveness of one text compared to another, and a large main effect of method would suggest that there are method differences regardless of the text used. However, if the matching hypothesis is strongly supported by your data, the interaction will be large – quite possibly large enough to drown out the main effects and/or render them difficult to interpret.

Table 14.11

<table>
<thead>
<tr>
<th>Method</th>
<th>Trad Text</th>
<th>Visual Text</th>
<th>Phonics Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>6.1</td>
<td>5.6</td>
<td>6.0</td>
</tr>
<tr>
<td>Visual</td>
<td>5.9</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Phonics</td>
<td>6.3</td>
<td>5.7</td>
<td>6.6</td>
</tr>
</tbody>
</table>

You can (and probably should) verify for yourself (using Formula 14.3) that for the cell means in Table 14.11, SSText is 1.867, and SSMethod is 1.4. By subtraction, SSInteraction is found to be 4.333, so MSInteraction equals 4.333 / 4 = 1.0833. If averaging the cell variances (not shown) were to yield MSW = .5, then FInteraction would equal 1.0833 / .5 = 2.167, which falls short of significance [F.05 (4, 81) = 2.49]. However, according to Abelson and Prentice (1997) pointed out, an appropriately designed (complex) interaction contrast can capture a good deal of the interaction with only one degree of freedom (providing much more power for your test), if the cell means conform fairly well to the pattern you predicted. The coefficients in Table 14.12 represent the “matching” prediction by giving extra weight to the cell means for which the match of method and text are expected to give a boost to reading performance, and lesser weight to all of the others (note that the coefficients add up to zero, making this a legitimate linear contrast).

Table 14.12

<table>
<thead>
<tr>
<th>Method</th>
<th>Trad Text</th>
<th>Visual Text</th>
<th>Phonics Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>+2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>Visual</td>
<td>-1</td>
<td>+2</td>
<td>-1</td>
</tr>
<tr>
<td>Phonics</td>
<td>-1</td>
<td>-1</td>
<td>+2</td>
</tr>
</tbody>
</table>

As with any linear contrast, the coefficients can be used to combine the cell means into a single “difference” score, which I have been calling L. For this example, L equals 2*6.1 + 2*6.2 + 2*6.6 - 5.6 - 6.0 - 5.9 - 6.3 - 5.7 = 37.8 - 35.4 = 2.4. According to Formula 13.12, $SS_{contrast} = 10*2.4^2 / \Sigma c^2 = 57.6 / 18 = 3.2$. Note that more than two-thirds of SSInteraction is captured by this single-df contrast, so whereas MSInteraction is only 1.0833, MSContrast is 3.2, and therefore FContrast for the matching hypothesis is
3.2 / .5 = 6.4, which would be well above its critical value \([F_{.05}(1, 81) = 3.96]\) had this contrast been planned (of course, given that the omnibus interaction was not significant, post hoc contrasts would not be justified – nor could any of them reach significance with Scheffé's test).

Abelson and Prentice (1997) emphasize the importance of testing the residual SS for a complex contrast (such as the one just performed) to ensure that your underlying hypothesis serves to explain your data well, all by itself. For the example above, \(SS_{\text{residual}} = SS_{\text{interaction}} - SS_{\text{contrast}} = 4.333 - 3.2 = 1.133\), and \(df_{\text{residual}} = df_{\text{interaction}} - df_{\text{contrast}} = 4 - 1 = 3\), so \(MS_{\text{residual}} = 1.133 / 3 = .378\). Therefore, \(F_{\text{residual}} = .378 / .5 = .756\), which is obviously not significant. The matching hypothesis gives a good accounting of the data in this example. Finally, Abelson and Prentice (1977) also point out that it can be informative to estimate the effect size associated with your complex contrast. If you have already calculated \(F_{\text{contrast}}\), you can estimate the contrast effect size in the population by using Formula 14.11, but with only “n” (i.e., the cell size) rather than “bn” in the denominator. However, in the single-df case (e.g., pairwise comparison, complex linear contrast), you have the option of calculating \(t\) instead of \(F\) (\(t_{\text{contrast}} = \sqrt{F_{\text{contrast}}}\)), so est. \(d_c\) can be written as:

\[
est. d_c = t_{\text{contrast}} \sqrt{\frac{2}{n}} \quad \text{Formula 14.14}
\]

Notice the close resemblance to Formula 8.5 for “g”. For this example, \(t_{\text{contrast}} = \sqrt{F_{\text{contrast}}} = \sqrt{6.4} = 2.53\), so est. \(d_c = 2.53 \sqrt{2 / 10} = 2.53 \times .447 = 1.13\), a rather large effect size. It looks like matching the text to the method plays an important role in teaching children to read, in this hypothetical study.

The Two-Factor Nested Design

The design of the experiment just described rests on the assumption that there is a most popular text associated with each method. However, it is also quite possible that there are numerous texts written for each method, with no obvious choice in each case, and that the texts are written so specifically for one method or another that it really doesn’t make sense, for example, to teach the phonics method with a “visual” text. If your chief purpose were to compare the effectiveness of the three methods, you could just pick one appropriate text randomly for each method, and then perform an ordinary one-way ANOVA on the group means. But in that case, there is no way to test whether some texts designed for a particular method are better than other texts designed for that same method. Moreover, if the choice of text within each method were to make a difference, the validity of the one-way study would depend heavily on the typicality of the text selected for each method. Given that the reading primers are written to be used for one particular method, and not another, a factorial design that crosses text with method would not make sense. However, you can still use a two-factor design to test whether text, as well as method, makes a difference in how quickly reading is learned – in just one study. In this case, you need what is called a nested design.

If there were perhaps just three or four texts available for each method, you might want to look at the effects of those texts in particular; in that case, “text” would be considered a fixed-effects factor, just like “method.” However, if many texts are available for each method, and we are more interested in whether the choice of text makes any difference at all than how particular texts compare to each other, it makes sense to choose texts randomly from those available for each method. Text would then be viewed as a random-effects factor. Because a different set of texts would then be associated with each method, we would say that the text factor is nested in the methods factor (this is illustrated in Table 14.13). As an example of the shorthand notation that is used to describe such experimental designs, I
will use “T” to represent the text factor, and “M” to represent the method factor. The sum of squares associated with the text factor would then be represented by \( \text{SS}_{T|M} \); the vertical bar in the subscript can be read as “nested in.” The SS for method would be written as \( \text{SS}_{M} \), because method is not nested in some other factor. The other SS components for this design will be explained in the context of a numerical example below.

When two fixed-effects factors are either crossed (as in the earlier examples of this chapter), or nested one inside the other, the proper analysis follows what is called the fixed-effects model. Similarly, the crossing or nesting of two random factors leads to the random-effects model. However, for reasons that I will discuss later, I do not think you will ever see a nested design that has only two fixed factors, or only two random factors. When one random- and one fixed-effects factor are combined, the appropriate analysis follows what is often called a mixed-model design (psychologists tend not to like this term because it is too easily confused with the mixed-design ANOVA, which will be covered in a later chapter of this text). The mixed-model can be applied to a factorial (i.e., crossed) design (as I will show later in this section), as well as a nested design. However, whereas it is fairly common to nest a random-effects factor in a fixed-effects factor, I cannot imagine any situation in which you would want to use the reverse pattern. Therefore, “random nested in fixed” is the only nested design I will describe in detail.

The Distinction between Fixed and Random Effects

It is not the variables themselves that are fixed or random, it is a matter of how we select levels of the variables for our particular study, and what kind of inferences we are trying to draw. For instance, if we want to compare cognitive, psychodynamic, and humanistic therapy for the treatment of some emotional disorder, these three therapy levels are considered fixed; anyone claiming to replicate this study would have to use these three types of therapy. However, if our point is to show that all forms of talking therapy are (or are not) equally effective, we might pick three types of talking therapies from the many choices available, in which case therapy types would be considered levels of a random-effects factor. Someone could randomly select three other types of talking therapy and claim to be replicating our experiment. In this case, we are not interested in drawing conclusions about the three particular forms of therapy used in our study; we want to draw a conclusion about talking therapy, in general.

Are you a bit uncomfortable with talking therapy as a random-effects factor? Fortunately, most cases are more clear-cut (e.g., comparing a few very popular texts, in order to decide which of them to use vs. selecting a few texts at random from a large pool of choices to see if text choice makes any difference), but there will always be ambiguous cases. One way to understand the fixed versus random distinction is that it is for fixed effects that the one-way ANOVA gives very little useful information (this is ironic in that one-way ANOVA’s with fixed effects are far more common than those with random effects). When dealing with a fixed-effects ANOVA with more than two groups, a significant F ratio is merely a signal that if you look further some more specific (and hopefully meaningful) comparison will be significant. In fact, the ANOVA is not at all necessary (even for Type I error control) if you plan to use Tukey’s HSD to compare pairs of means, or you plan to test a set of orthogonal contrasts, though there is a very strong tradition of reporting the ANOVA before presenting more specific results. Indeed, when dealing with fixed effects, it is very hard to imagine that someone would stop his/her data analysis after obtaining a significant F ratio, and not conduct additional tests. On the other hand, it is for a random-effects ANOVA that a significant F ratio may be sufficient for drawing one’s conclusion (e.g., type of therapy does make a difference). Most often, a random-effects factor is considered a nuisance factor (e.g., the different levels are different research assistants, or experimental rooms) that is included only to estimate properly the effects of another factor. A significant F for such a factor is not likely to be followed by more specific comparisons (e.g., Did the
participants run by Jane perform better than those run by Richard?). In such cases, we usually prefer that the main effect of the random factor not be significant, but unless we select a large number of levels, we should not be quick to conclude that the random factor has no effect in the population just because our results fell a bit short of significance. In such a case, we might want use an unusually large alpha, as I will soon discuss.

The Ordinary One-Way ANOVA as a Nested Design

It may be helpful to understand that an ordinary one-way ANOVA with independent groups is actually a rather clear example of a nested design. First, you need to think of Subjects as a random-effects variable with very many possible levels from which to choose. For instance, suppose you want to compare the effects of two different sets of instructions (imagery vs. subvocalization) on the recall of a list of words. If you assign your subjects at random to one condition or the other, the larger the variability in recall due to individual differences among your subjects within each condition, the more this variability will tend to increase the difference between the means of the two conditions; that is why we need to perform a t test or one-way ANOVA on the data (the more that subjects vary from each other, the easier it is to accidentally assign much better subjects to one condition rather than the other). In this case, you can think of the Subject factor as involving general differences in memory ability; each subject will be at a different level of the Subject factor in terms of having a better or worse overall ability to recall words, apart from the effect of whatever instructions s/he receives. Because a different set of subjects is selected for each condition, we can say that the subject levels are nested in the instructions factor. If we use a generic A to represent the experimental (fixed-effects) factor in a one-way independent-groups ANOVA, the sum of squares associated with the variability of subjects within each group – what I have been calling SS_W – can be symbolized as SS_S | A. This notation is preferred by some text authors, who deal with statistics in the context of complex experimental designs. It is also possible to cross the Subject factor with (rather than nest it in) an experimental factor (each subject participates at all levels of the IV); this creates a repeated-measures design, which will be the main topic of the next chapter.

Calculating an Example of a Nested Design

The table below presents hypothetical results from a text-nested-in-method experiment, in which the reading levels of sixth graders are measured after a semester of advanced reading instruction (this table is adapted from Table 7.3 in Cohen and Lea, 2003).

<table>
<thead>
<tr>
<th>Method:</th>
<th>Traditional</th>
<th>Visual</th>
<th>Phonics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text:</td>
<td>A  B  C  D</td>
<td>E  F  G  H</td>
<td>I  J  K  L</td>
</tr>
<tr>
<td></td>
<td>5.6  6.0  5.5  5.8</td>
<td>5.7  5.8  6.4  5.9</td>
<td>6.0  6.5  6.2  6.1</td>
</tr>
<tr>
<td>Method Mean:</td>
<td>5.725</td>
<td>5.95</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Under each of the 12 texts is a number that represents the mean reading score for the set of 10 children randomly assigned to that text (i.e., N_T = 120). Of course, assignment to a text automatically entails assignment to one of the three methods, as well. Because the design is balanced, the mean for each method is just the ordinary average of the means of the four texts assigned to that method (the grand mean comes out to 5.96). The appropriate analysis yields two separate one-way ANOVA’s, one nested inside the other, as you will soon see.
To test the main effect of the method factor you can treat the text means in Table 14.13 as though they were individual scores, ignoring that they are actually means of 10 scores each. As with any one-way ANOVA, you can begin by calculating SS_total. Applying Formula 14.3 to the text means, we find that SS_total (which can be called SS_text in this context) equals \( N_T \sigma^2 \text{ (text means)} = 120 * .08576 = 10.292 \). The next step is to calculate SS_between, which in this case is SS_method: \( 120 * \sigma^2 (5.725, 5.95, 6.2) = 120 * .0376 = 4.517 \). Then, SS_within can be found by subtracting SS_between (i.e., SS_method) from SS_total (i.e., SS_text). Recall that SS_within stands for SS_within, which in this example is the SS for the variability of texts within each method. Because the texts are nested within the methods, SS_within can be written as SS_T|M, which equals 10.292 - 4.517 = 5.775. The next step is to divide SS_method and SS_T|M by df_method and df_T|M, respectively. Remember, this is just an ordinary one-way ANOVA, in which the texts are the "subjects", so df_method = \# of methods - 1 = 3 - 1 = 2, and df_T|M corresponds to df_within = \# of texts - \# of methods = 12 - 3 = 9. Therefore, MS_method equals 2.26, and MS_T|M equals .6417. Finally, the F ratio for testing the method effect is: \( F_{\text{method}} = \frac{MS_{\text{method}}}{MS_{T|M}} = 2.26 / .6417 = 3.52 \). The F calculated for method is not larger than F.05 (2, 9) = 4.26, so the null hypothesis (that the population means corresponding to the three methods are the same) cannot be rejected.

The other effect that can be tested in this design is the main effect of text (because text is nested in method, there is no interaction to test). Fortunately, we have already calculated the numerator for the F ratio we will need. The numerator for F_text is MS_T|M. Note that it is not MS_text (which I didn't bother to calculate), because SS_text is based on all of the variability from text to text, which includes variability produced by any differences among the three methods. On the other hand, MS_T|M, by measuring the variability of texts within each method, is not inflated by any differences among the methods. But what would be the proper error term? The way Table 14.13 is presented you may have forgotten that each of the 12 text values is a mean of ten children. I didn't include the SD's for each text, but if I had, you could square them to get the unbiased variances, and then, assuming homogeneity of variance and equal sample sizes, you could simply average them to obtain the error term to put under MS_T|M. Normally, I would call that error term MS_within, but in a two-factor nested design, both error terms have equal claims to that title. In this particular example, the two error terms could be referred to as MS_within-Methods and MS_within-Texts, but for a general notation system, it is more meaningful to label these two error terms as MS_T|M and MS_S|T|M, respectively. In the latter term, S represents subjects; the notation indicates that subjects are nested in texts (a different group of subjects is randomly assigned to each of the 12 texts), which in turn are nested in methods (a different set of texts are selected randomly for each method). MS_S|T|M can be calculated from the SD's, but it can also be found by first obtaining SS_S|T|M by subtraction, and then dividing by df_S|T|M, which equals the total number of subjects minus the total number of different texts (120 - 12 = 108, in this case). If SS_S|T|M were to equal 54.0, then MS_S|T|M = 54.0 / 108 = .50, and FT|M, which equals MS_T|M / MS_S|T|M would be .6417 / .50 = 1.28. Because this F is smaller than F.05 (9, 108) = 1.97, the null hypothesis (that all of the text population means are the same within each method) cannot be rejected.

The Lack of Power in Nested Designs

The results of this hypothetical study seem quite disappointing, especially given how large the F ratio was for testing method (3.52). It doesn't help that the df for the error term of this test was so small (df_T|M = 9). In fact, if df_T|M had been 20 or more (with no change in F), the main effect of method would have been significant at the .05 level. This is a common problem with nested designs. The fixed-effects factor is usually the one in which the experimenter is more interested, whereas the random-effects factor is more likely to be included only to increase the generality of the results (if just one text were used for each method – three different texts in all – how would you know that method differences
were not due entirely to differences among the particular texts used?). However, as in our example, the df are often much larger for the random- than the fixed-effects error term. Wampold and Serlin (2000) demonstrated that, given a limited number of participants available for the type of study in our example, you would gain more power for the fixed-effects factor by using more levels of the text factor, with consequently fewer subjects at each level (e.g., had we used 8 instead of 4 different texts for each method, we would have had only 5 instead of 10 subjects per text, but \( df_T | M \) would have been 21, and the same \( F_{\text{method}} \) that failed to reach significance above would be larger than the new, reduced critical \( F \) for method).

Fortunately, if it is clear that the random-effects factor is not having an effect, you are allowed to ignore it entirely, which can greatly increase your power, as I will soon demonstrate. However, if the text effect had just barely failed to reach significance in our example, that could hardly be taken as evidence that choice of text has no effect. Although it is virtually impossible to prove the null hypothesis, various statisticians recommend the use of a very liberal alpha when testing the random-effects factor; then, if significance is still not obtained, they recommend pooling the two error terms and using the pooled error term to test the fixed-effects factor (I will show that this is the same as simply ignoring the random-effects factor entirely). In the preceding example, the \( p \) value associated with \( F_{T | M} \) (1.28) was .26. The most liberal alpha normally recommended for testing the random factor is .25, so, because in this case \( p > .25 \), we can justify pooling the error terms (text does not seem to be having much, if any, effect).

**Pooling the Error Terms in a Nested Design**

Pooling error terms is not new to you; it is the same as pooling the two sample variances in a t test. The easy way is to add the two SS’s first, and then divide by the sum of the two df’s. In our example, the two sources of error are text-to-text variability (\( SS_T | M \)) and subject-to-subject variability (\( SS_S | T | M \)), so \( SS_{\text{pooled}} = 5.775 + 54.0 = 59.775 \), \( df_{\text{pooled}} = 9 + 108 = 117 \), and \( MS_{\text{pooled}} = 59.775 \times 117 = .51 \). Testing \( MS_{\text{method}} \) with \( MS_{\text{pooled}} \), we get \( F_{\text{method}} = 2.26 \times .51 = 4.43 \). Not only has \( F_{\text{method}} \) increased (from 3.52), but its critical value has been lowered considerably, as well: \( F_{.05} (2, 117) = 3.08 \). The new \( F_{\text{method}} \) is easily significant at the .05 level. But are you comfortable with this result? As with pooling variances in a t test, there is a judgment call involved. Even with an alpha of .25, the test of the random factor won’t have much power if your samples are quite small. And, it is well-known that if the random-effects factor does have some effect in the population, you will be inflating your Type I error rate (above whatever alpha you are using to find your critical values) if you ignore its existence. As Wampold and Serlin (2000) point out, random factors are often ignored, resulting in exaggerated estimates of the effect size of the fixed-effects factor.

For instance, to compare three teaching methods it would not be unusual to have, perhaps, a total of 12 teachers, four trained in each of the methods, each teaching a different, randomly selected set of children (I’m assuming individual instruction in all cases; teaching the children in groups threatens the independence of your observations). If you think of the teachers within each method as interchangeable (having all been trained alike), you are likely to ignore the teacher factor, and run the risk of inflating your F ratio for testing the methods. I will return to the issue of why it can be important to include random factors as part of your data analysis shortly, but first I want to show you how pooling the error terms in a nested design is the same as ignoring the random factor.

I began the analysis of the method/text example by referring to the text-to-text SS as \( SS_{\text{total}} \), but by now you should realize that that was just to keep things simple. For this example, the real \( SS_{\text{total}} \) is more than that; it is equal to \( SS_{\text{text}} \) plus the subject-to-subject SS within each text group, \( SS_S | T | M \). So,
Explaining Psychological Statistics  Chapter 14 (Section D)  B. H. Cohen

SS\textsubscript{total} = 10.292 + 54.0 = 64.292 (recall that the 54.0 doesn’t come from Table 14.13, it comes from the SD’s that I left out of the table). If you ignore the text factor completely, you need subtract only SS\textsubscript{method} from SS\textsubscript{total}, and then call the remainder SS\textsubscript{W}, in which case SS\textsubscript{W} = 64.292 - 4.517 = 59.775. Similarly, df\textsubscript{W} = df\textsubscript{total} - df\textsubscript{method} = 119 - 2 = 117. Thus, ignoring the text factor, the new error term, MS\textsubscript{W}, is 59.775 / 117 = .51. You should notice two things: one, all I just did was a simple one-way ANOVA using method as the factor; and, two, MS\textsubscript{W} is exactly the same as MS\textsubscript{pooled}. Put another way, pooling the error terms is the same as not separating them in the first place.

**Expected Mean-Squares for the Nested Design**

One way to understand just what can happen if you ignore a random factor that is present in your study is to look at the expected MS for each part of the analysis. E(MSeffect) is the value you would get for a particular MS (i.e., MSeffect) if you were measuring the entire population. Assuming that the way you calculate MSeffect produces an unbiased estimate of its value in the population, E(MSeffect) is also what the average of your MSeffect’s would approach if you calculated MSeffect for very many replications of the same study. As I mentioned in Chapter 12, section C, the expected value for subject-to-subject variability within groups can be written simply as σ\textsuperscript{2}; this is the variability you can expect within a group in which all of the participants are treated exactly alike (or as closely as you can control your experimental conditions). Because this variability occurs in spite of your efforts to treat all participants within a group identically, it is usually written as σ\textsuperscript{2}\textsubscript{error} [this is, for instance, E(MS\textsubscript{W}) in an ordinary one-way ANOVA]. In the nested design I have been describing, E(MSS | T | M) = σ\textsuperscript{2}\textsubscript{error}. MST | M is affected by subject-to-subject variability (even if text has no effect, the means of different text groups will vary as much as any random samples of a given size would), but any population differences among the texts will contribute, as well (differences among the methods will not contribute to this MS, because the SS of text means is calculated separately for each method before being summed). Remember that any MS\textsubscript{between} is calculated as n s\textsuperscript{2}, so E(MST | M) = n σ\textsuperscript{2}T + σ\textsuperscript{2}\textsubscript{error}, where n is the number of participants in each text group and σ\textsuperscript{2}T is the variance of the population means of all possible texts that could be chosen for each method. Notice that the ratio MST | M / MSS | T | M can be expected to follow the F distribution, and is an appropriate way to test for text effects, because although this ratio averages to (n σ\textsuperscript{2}T + σ\textsuperscript{2}\textsubscript{error}) / σ\textsuperscript{2}\textsubscript{error} when there are text effects, when the null hypothesis concerning texts is true (σ\textsuperscript{2}T = 0), F\textsubscript{text} reduces to a ratio of two estimates of subject-to-subject population variance, as in the ordinary one-way ANOVA.

The E(MS) for the method effect is more complex still. It is affected by the variability of the texts (e.g., the texts randomly chosen for one method could be accidentally much better than the texts selected for the other methods), which in turn is also affected by subject variability (even if the texts don’t differ, the subjects assigned to a particular text could accidentally be much above average), so E(MS\textsubscript{method}) includes both n σ\textsuperscript{2}T and σ\textsuperscript{2}\textsubscript{error}. However, any difference in the population means of the different methods will also produce an increase in E(MS\textsubscript{method}). Because method is treated as having fixed effects in our study, the variance of the method means in the population is just the variance of the three population means that correspond to the three particular methods used in our example. A different symbol is used to distinguish the variance of a few particular population means (usually, θ\textsuperscript{2}) from the variance of an entire, virtually infinite, distribution of population means (e.g., σ\textsuperscript{2}T). Therefore, the contribution of method-to-method variability to E(MS\textsubscript{method}) can be symbolized as tnθ\textsubscript{M}\textsuperscript{2}, where t is the number of texts used in each method, and n is therefore the number of participants assigned to each of the methods. Adding these three sources of variance, we can see that E(MS\textsubscript{method}) = tnθ\textsubscript{M}\textsuperscript{2} + nσ\textsuperscript{2}T + σ\textsuperscript{2}\textsubscript{error}.
Explaining Psychological Statistics

Chapter 14 (Section D) B. H. Cohen

11

The Danger of Ignoring Random Factors

If you try to test \( MS_{\text{method}} \) by just dividing it by the subject-to-subject variability within each text group (i.e., \( MS_{S|T|M} \)), your expected values will form the following ratio: \( \frac{(t\theta_M^2 + n\sigma_T^2 + \sigma_{\text{error}}^2)}{\sigma_{\text{error}}^2} \). The problem with this approach should be easy to see. What if the null is true for method (i.e., \( \theta_M^2 = 0 \)), but not for text? In that case, \( F_{\text{method}} \) would become a test of the text effect because text effects are contributing to the numerator of \( F_{\text{method}} \); this situation can greatly increase the Type I error rate for testing the method effect, and is therefore quite obviously unacceptable. That is why the error term for \( F_{\text{method}} \) is not \( MS_{S|T|M} \), but rather \( MS_{M|T} \), yielding the following ratio of \( E(\text{MS}) = \frac{(t\theta_M^2 + n\sigma_T^2 + \sigma_{\text{error}}^2)}{(n\sigma_T^2 + \sigma_{\text{error}}^2)} \). Note that when the null is true for method, the numerator and denominator are now properly estimating the same amount of variance.

Even when the null is not true for method, any text effect that exists will contribute to the numerator of \( F_{\text{method}} \), making that \( F \) larger than it would otherwise be, unless you use the proper error term. For the same reason, if you ignore the random factor when estimating the effect size for method, you run the risk of greatly overestimating that effect size. The following formula provides a virtually unbiased estimate of the proportion of variance accounted for by the method factor in the population. I have adapted the formula from Siemer and Joormann (2003), changing the notation to agree with mine.

\[
est. \omega^2 = \frac{SS_M - df_M \times MS_{T|M}}{SS_M - df_M \times MS_{T|M} + t_m \times (MS_{T|M} - MS_{S|T|M}) + N_T \times MS_{S|T|M}}
\]

Note that \( N_T \) is the total number of participants, which could also be written as \( t_m n \). The numerator of this formula is conceptually the same as in Formula 12.14, in that it consists of the SS for the effect whose size we want to know minus a term that is the \( df \) for that effect multiplied by the error term used to test that effect. The denominator is more complicated than in Formula 12.14, because it contains unbiased estimates for the three sources of variability in the data: method (which also appears in the numerator); texts within each method; and subjects within each text (which are, in turn, nested in the methods).

Would calculating partial omega squared to estimate the method effect make sense for this example? The answer is not immediately obvious. We are not classifying students in terms of the texts they happen to be already using; rather, we are randomly assigning students to texts. You might think that adding text as an experimental factor in the design could increase the total amount of variance, thus making the contribution of the method factor a smaller proportion of variance accounted for than it should be. However, if many different texts are presently being used in the population, our study may be just mirroring the amount of variance that is already in the population. Given that our goal is to generalize our results, we have to decide whether it is more useful to estimate the method effect as though for each method everyone is using the most typical text for that method (partial omega squared), or to estimate the method effect in the context of people using the full range of available texts for each method (Formula 14.15). The latter approach seems more appropriate for most purposes. Calculating a partial omega squared for method would involve using Formula 14.15, after deleting the middle of the three terms in the denominator (the variance due to using different texts). Depending on the true size of the text effect, partial omega squared can greatly overestimate the effect of method in the real-world situation, in which not all students being taught by the same method are using the same text.

A similar problem would occur if you were to use Formula 12.14 (instead of Formula 14.15) to calculate omega squared for a nested design; you would be estimating \( \omega^2 \) for a one-way ANOVA, completely ignoring the random factor. Wampold and Serlin (2000) demonstrated just how greatly you could be overestimating the effect size of the fixed-effects factor in that case. On the other hand, Siemer and Joormann (2003) argue that treating the second factor as based on random-effects, when it could be
more realistically viewed as a fixed-effects factor, greatly underestimates the effect-size estimate of the first factor (method, in this example). I will return to this issue when I briefly discuss the nested design with two fixed factors.

**Estimating the Effect of the Random Factor**

Although in most cases you are hoping that the population effect size for your random factor is not large, it is certainly useful to estimate its size, especially for planning future studies (in some cases, you may direct your efforts at reducing the variance associated with the random factor by some form of standardization, and later you would want to assess your success). To make the following discussion more concrete, I will continue to discuss effect-size estimates in terms of the method/text experiment, and assume that for more general purposes you will substitute “fixed-effect” for method and “random-effect” for text. As I just explained, you probably would not want to use a partial omega squared estimate for the method effect, because it is unrealistic to estimate the method effect in a situation where there is no variability due to different texts. However, this reasoning is generally reversed when looking at the effect of text. If you want to know, proportionally, how much variability is added by the use of different texts to the usual variability due to individual differences in reading aptitude (and other uncontrollable error factors), you now want to use a partial omega squared estimate, because it is not useful to tie your estimate of text-to-text variability to a population situation in which, for instance, one third of the population is using each of the three methods in your experiment and no others.

An important assumption, rarely discussed explicitly for nested designs, is homogeneity of variance for the random factor across the different levels of the fixed-effects factor. In terms of our example, text variability is considered to be the same (in the population) for each method, and any random sample of texts is expected to produce about the same amount of variability as any other random sample. That is why it is useful to estimate the variability of texts holding method constant (as though everyone was being taught by the same method). The usual estimate for the partial ω² of a nested random-effects factor looks like this:

\[
\text{est. } \omega_{\text{partial}}^2 = \frac{MS_{T|M} - MS_{S|T|M}}{MS_{T|M} + (n - 1) MS_{S|T|M}}
\]

Formula 14.16

The proportion expressed in Formula 14.16 is often referred to as an intraclass correlation coefficient (there are others), and symbolized as \(\rho_I\) (the subscript stands for intraclass), although why this is not termed a squared correlation, or symbolized as \(\rho_I^2\) (as in \(\eta^2\)), is something I admit I have been unable to discern.

**The SS Breakdown for the Nested Design**

By way of summarizing (and hopefully, clarifying) the analysis of a two-factor mixed-model nested design, I will compare its division of SS components to the more familiar breakdown of the two-way factorial design. To do this I have forced the data of Table 14.13 into a factorial design, as shown below.

**Table 14.14**

<table>
<thead>
<tr>
<th></th>
<th>Text 1</th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>5.6</td>
<td>6.0</td>
<td>5.5</td>
<td>5.8</td>
<td>5.725</td>
</tr>
<tr>
<td>Visual</td>
<td>5.7</td>
<td>5.8</td>
<td>6.4</td>
<td>5.9</td>
<td>5.95</td>
</tr>
<tr>
<td>Phonics</td>
<td>6.0</td>
<td>6.5</td>
<td>6.2</td>
<td>6.1</td>
<td>6.2</td>
</tr>
<tr>
<td>Column Mean</td>
<td>5.77</td>
<td>6.1</td>
<td>6.03</td>
<td>5.93</td>
<td>5.96</td>
</tr>
</tbody>
</table>
Please note that Text 1 is not literally the same text for each method; it just arbitrarily denotes the first text in each set of texts (i.e., Text A for Traditional, Text E for Visual, and Text I for Phonics), and Text 2 denotes the second text in each set, and so on. Therefore, the column means are not really meaningful, but they can be used to show how the SS’s compare between the two designs (I will write the corresponding SS for the nested design in parentheses after each component I extract from the factorial design). For any factorial design, the analysis begins with dividing SS\text{total} (same as in the nested design) into SS\text{between-cell} (SS_{\text{S|T|M}}) and SS\text{within-cell} (SS_{\text{S|T|M}}). Then SS\text{between-cell} is further divided into SS\text{column} (which can be called SS_{\text{method}} in this example, and will have the same value, as well as the same name, as SS_{\text{method}} in the nested design), SS_{\text{Row}} and SS_{\text{interaction}}. The latter two components are not divided in the nested design; the sum of SS_{\text{Row}} and SS_{\text{interaction}} equals SS_{\text{S|T|M}}. If you choose to ignore the random (i.e., text) factor in the nested design, you will be dividing MS_{\text{method}} (same in both designs) by SS_{\text{total}} - SS_{\text{method}} (also equal to the sum of SS_{\text{Row}}, SS_{\text{interaction}} and SS_{\text{within-cell}}) divided by NT - # of methods. To clarify the breakdown of the degrees of freedom in the nested design, I will draw the appropriate df tree (compare to Figure 14.6). Note that in the figure below, n = the number of subjects in each text group (i.e., at each level of the random factor), t = the number of texts within each method (i.e., the number of levels of the random factor at each level of the fixed-effects factor), and m = the number of different methods (i.e., levels of the fixed factor). Therefore, NT = tmn = 4*3*10 = 120.

\[
\begin{align*}
\text{df}_{\text{total}} &\quad [NT - 1 = 119] \\
\text{df}_{\text{text}} &\quad [tm - 1 = 11] \\
\text{df}_{\text{S|T|M}} &\quad [tm (n-1) = 108] \\
\text{df}_{\text{method}} &\quad [m - 1 = 2] \\
\text{df}_{\text{T|M}} &\quad [m (t - 1) = 9]
\end{align*}
\]

Obtaining a Nested Design Analysis from SPSS

As of this writing, the latest version of SPSS for Windows (12.0) directs you to its Mixed Models ...Linear dialog box when you click on the Nested topic under Help. I have not been able to get the F ratios calculated above from the Mixed Models program, but fortunately you do not need this new SPSS module to perform a nested design. You will, however, need to use the Syntax window. The first step is to enter your data as you would for a factorial design. Table 14.15 displays how the data might look in an SPSS spreadsheet for the first and last two children in the reading study.

Table 14.15

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
<th>Readscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5.9</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6.1</td>
</tr>
<tr>
<td>.......</td>
<td>.....</td>
<td>......</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5.7</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Note that Text 1 for Method 1 is really Text A, whereas, for instance, Text 1 for Method 3 is Text I, and Text 4 for Method 3 is Text L. Although numbering the texts 1 to 4 within each method makes it look like you are using the same texts for each method, when this is not the case, you will have the opportunity to tell SPSS about the nesting in a syntax command. To minimize the amount of typing you
have to do (and possible typos you could make), I recommend that you use the Analyze menu, and click on General Linear Model... Univariate. Then move your Fixed and Random factors (Method and Text, respectively, in the example of this section) to the appropriate Factor boxes, and your DV (e.g., Readscor) to the Dependent Variable box. Choose whatever options, plots, or contrasts you want and then click on Paste, instead of OK. A syntax window should be created with a command file that starts like this:

```
UNIANOVA
   readscor  BY method text
   /RANDOM = text
```

and ends like this:

```
/DESIGN = method  text  method * text .
```

To get the nested design analysis I have been describing you need only change the last statement to the following:

```
/DESIGN = method text (method).
```

The Nested Design in which Both Factors have Fixed-Effects

As I mentioned earlier, it is rare to see a fixed-effects factor nested in another one (I have never seen it done deliberately), because interpreting the results of such a design would be so problematic. For one thing, whereas it is reasonable to expect a random-effects nuisance factor to exhibit the same population variability for each level of a fixed factor in which it is nested, it is not obvious that it is reasonable to assume that a nested fixed-effects factor would show the same homogeneity from level to level of another fixed factor. However, I also mentioned that the boundary between a fixed-effects and a random-effects factor can be fuzzy. In particular, Siemer and Joormann (2003) argued that when therapists are nested in particular types of psychotherapy, it is usually not reasonable to treat each set of therapists as a random sample of all possible therapists practicing a specific form of therapy. As an alternative, Siemer and Joormann (2003) advocate the use of the “fixed-nested-in-fixed” analysis on the data from such a study. This suggestion is controversial, for reasons I will mention soon, but the analysis is easy to describe in terms of the expected MS’s I have already defined.

Imagine that for each of the three methods in the reading study we include the four most popular texts. Given that text is now a fixed-effects factor, its expected MS would be expressed as $n\theta_T^2 + \sigma^2_{\text{error}}$ instead of $n\theta_T^2 + \sigma^2_{\text{error}}$. Using fixed levels of text implies that there is no chance that an exact replication will involve different sets of texts that favor one method over another in a way that is different from the original study. Therefore, the E(\text{MS}) for the method effect does not include any contribution from the variability of one text to another; $E(\text{MS}_{\text{method}})$ = $tn\theta_M^2 + \sigma^2_{\text{error}}$. Consequently, the F ratios for both the method and text effects can properly use an estimate of $\sigma^2_{\text{error}}$ alone (i.e., $\text{MSS}_{T|M}$) as their error terms. In fact, the only difference from the ordinary two-way factorial ANOVA is that there is no interaction to test in this design.

The big difference between the mixed-model and fixed-model analyses involves the error terms used to test the main effect of method; in the mixed-model the error term is $\text{MST}_{T|M}$, and in the fixed-model it is $\text{MSS}_{T|M}$. Not only is the latter error term generally smaller (it is not influenced by text to text variability), but it is usually associated with many more degrees of freedom. In our example, these two error terms are .6417 and .5, with df's of 9 and 108, respectively. $F_{\text{method}}$, which was not significant
in the mixed-model analysis, increases to 4.52 when the fixed-model is applied, and with 108 df for the error term, it is easily significant for this analysis. There is no question that the fixed-model usually has much more power to detect an effect in the factor of interest (e.g., teaching method), but there is some debate about when it is appropriate to apply each model.

Siemer and Joormann (2003) point out that in the case of comparing different types of psychotherapy with therapists nested in therapy type, the experimenter is in the position of having to select particular therapists that are skilled in each type of therapy, and it is virtually impossible to do this in a random way that insures that all possible practitioners of a particular therapy have an equal chance of being selected, as the mixed-model requires. The problem this creates for drawing conclusions is similar to the problem inherent in any study in which participants are selected from pre-existing categories (e.g., vegetarians and non-vegetarians) rather than being randomly assigned to conditions. You cannot, of course, make firm causal conclusions, and without random sampling, you cannot even be sure that your samples are representative of their respective populations (e.g., how can you know if you have selected a typical group of psychoanalysts, and a typical group of behavior therapists, etc., to administer the therapies you are comparing?). Siemer and Joormann (2003) argued that, given the lack of randomness in your “random” factor, you might as well gain the considerable extra power associated with the fixed-model analysis. Serlin, Wampold, and Levin (2003) replied that, whereas the fixed-model is undoubtedly more powerful than the mixed-model, that power comes at too high a price, because the use of the fixed model limits your conclusions to the particular therapists who participated in your study, and has an unacceptably high Type I error rate should you try to extrapolate your results beyond those particular therapists.

This debate makes it clear that neither statistical model can correct the fundamental flaw in the experimental design: that you are comparing the therapies by using therapists that already favor one therapy over the others (e.g., if the most empathic therapists are drawn to, say, Rogerian therapy more than behavioral therapy, your results may be misleading due to a confounding variable). Amid the debate of these two parties about the logic of statistical and scientific inference, however, some useful recommendations can be derived. When designing a study in which participants necessarily fall into subgroups, either because each subgroup is run in a different location, or by a different research assistant, or is treated differently in any way that could affect your dependent variable, you should use as many subgroups (i.e., levels of the nuisance factor) as possible for each level of the factor of interest (e.g., type of therapy; reading method), even if that means the subgroups will be fairly (but not ridiculously) small. Then, if a journal editor or reviewer demands that you use a mixed-model analysis to present your results, you will have as much power as possible. At the same time, it is also in your interest to take steps to minimize the variability from one subgroup to another (e.g., for each therapy type all of the therapists refer to the same manual), and therefore the size of your fixed-effects error term in the mixed-model – without completely eliminating the generality of your study (if your study is conducted under very narrowly constrained conditions, it may be difficult for others to replicate that degree of control, and the study conditions may have little in common with the way treatments are administered in the real world).

The presence of random factors seems especially likely in large medical/clinical studies in which patients are treated at several different hospitals, or large educational studies in which different schools are randomly chosen for various interventions. However, random factors can easily impact small laboratory studies when, for instance, different experimental conditions require different sets of stimuli (e.g., comparing visual to verbal memory). These situations usually involve nested factors within levels of a repeated-measures (RM) factor, which complicates the analysis a good deal. Random factors in RM designs will be dealt with in section D of Chapter 16.
The Mixed-Model Factorial Design

In the previous example, all of the possible texts were designed with one method or another in mind, so it would not have made much sense to select four texts at random and then use the same texts for all three methods. But suppose that the situation were such that none of the teaching methods studied required specially suited texts, and that there were no reasons to expect any particular text to work better in concert with one method than another. In that case, you might want to study the three methods using only the most popular text, or if there were, for instance, four leading texts in contention, you could cross the four texts with the three methods to gain the economy of a factorial design, as pointed out in Section B. But, what if there were hundreds of reasonable choices for reading primers, all seemingly interchangeable, with none much more popular than any of the others? Selecting just one text at random with which to compare the three methods would be risky. With only one text, how could you feel confident that any other text would be likely to produce a similar pattern of performance across the three methods? Moreover, with only one text in the study, you could not test whether some texts are generally better than others, regardless of the method used.

The situation I have been describing can be handled nicely by selecting perhaps as few as four texts at random from the pool of possibilities, and then crossing those four texts with the three teaching methods in a factorial design. A significant main effect of the text factor would suggest that some texts are generally better than others, and a significant text by method interaction would suggest that some texts are indeed better suited for use with some methods as compared to others. On the other hand, a significant main effect of method, together with only a small amount of interaction, would give you greater confidence in the generality of the method differences you found than if you had found those differences using only one randomly selected text. The factorial design just described could be analyzed with an ordinary two-way ANOVA, and this ANOVA would work well in screening out studies in which the null hypotheses were true for all three of the effects (two main effects, and the interaction). Unfortunately, because one of the factors (i.e., text) is a random-effects factor in this scenario, the ordinary two-way ANOVA will not work well if it turns out that the two main effects are truly null in the population, but the interaction of the two factors is not. Let's take a look at how such a situation could arise, and how the ordinary (fixed-effects) factorial ANOVA could lead to misleading results.

How a Random Factor Can Affect a Factorial Design

To explain how the interaction of a random with a fixed-effects factor can spill over into the main effect of the fixed factor, I will describe one of the simplest situations that can produce this confound. Imagine that among the thousands of texts that could be selected for our reading study, exactly half produce reading scores that are always .2 units higher when used with the traditional method as compared to the other two methods (I'll call it pattern 1), while the other half produce precisely the opposite pattern (i.e., .2 units lower for traditional than either the visual or phonics methods—let's call it pattern 2), thus creating a disordinal interaction in the population. Averaging across the entire pool of texts, there would be no main effect of method, because the .2 increases and decreases for the traditional method would balance out (I'm assuming that none of the texts creates any differences between the visual and phonics methods). The problem occurs when you pick a few texts at random from the many that could be selected. In our example, four texts are selected, and if two of them yield pattern 1 and the other two produce pattern 2, no main effect of method would be created. However, when selecting four texts from a pool that is half pattern 1 and half pattern 2, winding up with two of each pattern is actually somewhat less likely than drawing 3 from one pattern and 1 from the other.
Suppose our random selection consists of three texts that exhibit pattern 1 and one text with pattern 2. The traditional method will tend to come out .1 units higher \([3*.2 - 1*.2] / 4\] than the other methods in our data, even though the three methods have the same means in the population. Even when the method by text interaction balances out the method effects exactly across all possible texts, our small, very finite selection is not likely to retain that balance; the interaction will produce a spurious main effect of the fixed factor due to the sampling error involved in selecting levels of the random factor. Lest you think that this confounding is only problematic in the very restricted situation I just described (e.g., the null hypothesis is exactly true for method), I should point out that a large amount of interaction can noticeably inflate a relatively small main effect. In this latter case, you would not be making a Type I error by rejecting the null hypothesis for the fixed-effects factor, but you could be greatly overestimating the size of that effect. This is similar to the problem discussed with respect to nesting, in which the effect of the random factor increases the apparent size of the fixed effect in which it is nested. The solution to the problem is similar, as well, as I will show next.

Expected Mean-Squares for the Mixed-Model

The intrusion of the interaction into a main effect in the mixed-model design can be understood by looking at the expected MS for the main effect of the fixed factor. In our example, \(E(MS_{\text{method}}) = \frac{\theta_2 M^2 + \sigma^2_{TM} + \sigma^2_{\text{error}}}{n}\), where \(t\) is the number of texts and \(n\) is the number of participants in each cell of the factorial design. The first term is based on any actual differences that may exist among the population means for the different methods (as mentioned earlier, the symbol \(\theta_2^2\) represents the variance due to a finite number of fixed effects). The second term represents the intrusion of the interaction into this main effect. Note that the symbol \(\sigma^2\) tells us that, like the main effect of text, the interaction is also a random-effects factor (as would be any interaction that involves a random factor). The third term represents the usual error variance due to the differences of individuals within each group or cell. As you may have guessed, \(E(MS_{TM}) = n \sigma^2_{TM} + \sigma^2_{\text{error}}\), so, conveniently, the ratio of \(MS_{\text{method}}\) to \(MS_{TM}\) \([\frac{\theta_2 M^2 + \sigma^2_{TM} + \sigma^2_{\text{error}}}{n \sigma^2_{TM} + \sigma^2_{\text{error}}}\] can be expected to follow the F distribution, when the null is true for the main effect of method.

The within-cell error term can be represented as \(MS_{S|TM}\) (i.e., subjects are nested in the cells formed by crossing the T and M factors), so \(E(MS_{S|TM}) = \sigma^2_{\text{error}}\). Because text is a random factor, its expected MS is: \(E(MS_{\text{text}}) = m \sigma^2_{\text{text}} + \sigma^2_{\text{error}}\), where \(m\) is the number of methods. Note that the interaction of the two factors does not intrude upon the main effect of the random factor; if the interaction balances things out so that there is no effect of the random factor in the population, there is no random selection of fixed-effects levels to upset that balance. Therefore, the main effect of text can be tested by the ratio of \(MS_{\text{text}}\) to \(MS_{S|TM}\) (more simply known as \(MS_{W}\)); you can see that when the null is true for text, the expected values follow the F distribution: \(\frac{m \sigma^2_{\text{text}} + \sigma^2_{\text{error}}}{\sigma^2_{\text{error}}}\). I have already showed you that \(E(MS_{TM}) = n \sigma^2_{TM} + \sigma^2_{\text{error}}\), so it makes perfect sense to test the interaction effect with the same error term as for the main effect of text – i.e., \(MS_{S|TM}\).

The SS Breakdown for the Factorial Mixed Model

I have reproduced Table 14.14 below with one small change in labeling. In Table 14.16, the texts have been assigned letters to remind you that this time there are only four texts in the study, and each text is being combined with all three methods. The appropriate analysis begins by calculating exactly the same SS components as you would for any two-way ANOVA with two fixed factors, and dividing by the usual df's to produce the same MS's you would compute in the usual two-way ANOVA. The difference from the usual factorial analysis occurs when you form the F ratios to test the three possible effects.
Table 14.16

<table>
<thead>
<tr>
<th></th>
<th>Text A</th>
<th>Text B</th>
<th>Text C</th>
<th>Text D</th>
<th>Row Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>5.6</td>
<td>6.0</td>
<td>5.5</td>
<td>5.8</td>
<td>5.725</td>
</tr>
<tr>
<td>Visual</td>
<td>5.7</td>
<td>5.8</td>
<td>6.4</td>
<td>5.9</td>
<td>5.95</td>
</tr>
<tr>
<td>Phonics</td>
<td>6.0</td>
<td>6.5</td>
<td>6.2</td>
<td>6.1</td>
<td>6.2</td>
</tr>
<tr>
<td>Column Mean</td>
<td>5.77</td>
<td>6.1</td>
<td>6.03</td>
<td>5.93</td>
<td>5.96</td>
</tr>
</tbody>
</table>

Two of the MS’s we now need were previously calculated when the data in Table 14.16 were viewed in the context of a nested design. MS\text{method} equals 2.26, as before, and MS_{S | T | M} is the same as MS_{S | T | M} from the nested design, which is equal to .50. However, in the nested design, it did not make sense to calculate a component based on the column means, because the texts differed across the methods. In the factorial design it does make sense: SS_{text} equals \( N_T \sigma^2 (5.77, 6.1, 6.03, 5.93) = 120 \times .01537 = 1.84 \). It also did not make sense to calculate the interaction in the nested design, but SS\text{between-cell} (by another name) was the first component we calculated. I’ll repeat that calculation here, changing the labeling, as appropriate: SS\text{between-cell} equals \( N_T \sigma^2 \) (cell means) = 120 \times .08576 = 10.29. Now we can find SS\text{interaction} by subtracting both SS\text{method} and SS_{text} from SS\text{between-cell} : SS\text{interaction} equals 10.29 - 4.517 - 1.84 = 3.935. The next step is to divide SS_{text} and SS\text{interaction} by df_{text} and df_{interaction}, respectively, yielding MS_{text} = .613 and MS\text{interaction} = .656. Finally, we are ready to calculate the three F ratios. Although it may seem counterintuitive, it is the fixed-effects factor in a crossed mixed-model that involves the unusual error term (for reasons explained, I hope, above), whereas the random factor takes the usual, within-cell error term. So, F\text{method} = MS\text{method} / MS_{T X M} = 2.26 / .656 = 3.45; F\text{text} = MS_{text} / MS_{S | TM} = .613 / .50 = 1.23; and F\text{interaction} = MS\text{interaction} / MS_{S | TM} = .656 / .50 = 1.31.

Pooling the Error Terms in a Mixed-Model Factorial Design

As in the nested design, F\text{method} falls short of significance. It is of no help that the error df are even lower in the crossed than the nested design: F_{.05} (2, 6) = 5.14. The other two F ratios are obviously not significant, either. In fact, the F for interaction is so far from being significant (p = .27) that we can consider the same power-boosting trick we used for the nested design – pooling the error terms. One could argue that because p > .25 for the interaction, there is little evidence that the fixed and random factors interact in the population, and thus little reason to separate SS\text{interaction} from SS_{S | TM}. Summing both the SS's and the df's, we find that MS_{pooled-error} = (3.935 + 54) / (6 + 108) = 57.935 / 114 = .508. This smaller error term results in F_{method} increasing to 4.45, but even this larger F ratio would not be significant were it not for the increase in df; the new critical F for the main effect of method is F_{.05} (2, 114) = 3.08, so the null hypothesis can be rejected at the .05 level, if you are comfortable with pooling the error terms (as usual, this is a judgment call, even if p > .25 for the test of the interaction). F_{text} barely changes at all with the new pooled error term, and remains far from significant.

Connection to RM ANOVA

The use of MS\text{interaction} as the fixed-effects error term in the design just described may seem less strange if you consider that only a small amount of interaction will result if each randomly selected text exhibits roughly the same pattern across the method means (e.g., phonics is always about half a point higher and the visual method only a quarter of a point higher than the traditional method). It is not likely that the different texts will follow virtually parallel patterns across the methods just by accident, so a relatively small amount of interaction suggests that the method differences are due to some true population differences. Therefore it makes sense that as the interaction gets smaller, the F ratio for testing the main effect should get larger. I will be using this type of argument again in the next chapter, when I show that the RM ANOVA can be viewed as a special case of the two-way mixed-model ANOVA.
References


