Copyright 1993 by Marcel Dekker, Inc. Reprinted by Permission of Marcel Dekker, Inc.
13
Hierarchical Linear Models and Experimental Design

STEPHEN W. RAUDENBUSH Michigan State University, East Lansing, Michigan

13.1 INTRODUCTION
13.1.1 Purpose of the Chapter

This chapter demonstrates how data analysis based on recently developed hierarchical linear models (1) duplicates the results of standard ANOVA models for an important class of experimental designs and (2) extends the study of fixed and random effects to include unbalanced data, predictors that are either continuous or discrete, and random effects that covary. We illustrate these principles by reanalyzing data from Kirk's (1982) widely used text and data from 103 primary schools in Thailand. We consider one-way designs, two-factor nested and crossed designs, and randomized block (repeated measures) designs. The hierarchical linear model, like the ANOVA and standard regression models, is itself a special case of a more general mixed linear model. In this chapter, we briefly consider how the general mixed linear model can be tailored to analyze data for which hierarchical linear models do not apply.

13.1.2 Mixed Model Analysis of Variance

Of all the statistical approaches taught in graduate schools of education and psychology, probably none receives more attention than does the analysis of variance (ANOVA). The ANOVA is elegant in its computational simplicity, but, more important, it facilitates estimation in models having fixed effects, random effects, or both. This flexibility makes it appropriate for a tremendous variety of designs, including nested designs and repeated measures designs.
Nested designs are common in experimental behavioral and social research. In education, a common design involves students nested within classrooms and classrooms nested within methods of instruction. In counseling psychology, a standard design assigns subjects to therapists and therapists to treatments. However, nested designs also arise routinely in nonexperimental research such as multistage social surveys using census tracts, blocks, or schools as clusters. Similarly, repeated measures designs are common in experiments such as clinical trials where patients are administered different drugs on different occasions; but such designs also arise frequently in nonexperimental settings including longitudinal surveys and program evaluations using follow-up interviews.

Each of the above settings usually calls for a "mixed" statistical model, that is, a model including both fixed and random effects. In a study of students nested within classrooms and classrooms nested within methods of instruction, classroom effects will logically be specified as random and method effects as fixed. Repeated measures designs call for estimation of random effects of persons and fixed effects of treatments or programs. Multisite program evaluations also yield data appropriate for mixed models. For example, Raffe's (1991) study of a new vocational education initiative in Britain involved a treatment-control comparison at each of 19 sites. This is really a randomized block design in which the sites are the blocks, viewed as random, and the two programs are fixed. The students nested within site-by-program cells are "replications" (see Kirk, 1982, p. 293).

13.1.3 Limitations of ANOVA

Although the ANOVA methods commonly taught in experimental design courses work well for balanced designs having discrete independent variables, they are not widely applicable when the data are unbalanced and some predictors are continuous. Balanced designs with discrete independent variables arise primarily in carefully designed, small-scale experiments. However, in field experiments, quasi-experiments, and surveys, unbalanced data and a mix of discrete and continuous predictors will be the rule rather than the exception. Because of this, researchers often turn away from the mixed-model ANOVA to embrace multiple regression as an alternative.

13.1.4 Applicability of Multiple Regression

For fixed-effects designs, multiple regression is a more general analytic technique than ANOVA. Any fixed-effects analysis of variance can be reformulated as a regression problem, and the regression analysis will exactly reproduce the ANOVA results. In addition, regression allows a
mix of discrete and continuous predictors. In contrast, the ANOVA cannot generally reproduce the regression results because the ANOVA cannot incorporate continuously measured predictor variables.

Of course, both ANOVA and regression are special cases of a “general linear model” (see e.g., Kirk, 1982, Chapter 5). This general linear model offers several approaches to estimation of deficient rank ANOVA models including specification of constraints on the parameters, use of the generalized inverse, and full-rank reparameterization via regression. For many researchers, the regression approach has become the method of choice. Not only does regression allow a single, simple framework for handling both continuous and discrete predictors, it also provides a natural and sensible way to handle unequal cell sizes.

Unfortunately, the benefits of standard regression analysis are available only in fixed-effects models. Standard computing packages for regression, based on ordinary least squares estimation, are inappropriate when some factors are viewed as random. Maximum likelihood estimation via SAS and BMDP allows estimation only for some special cases of a mixed model (Kang, 1991, Chapter IV).

The inappropriate use of multiple regression for data yielded by nested designs has a long and disreputable history in education, sociology, and psychology (cf. Robinson, 1950; Cronbach and Webb, 1975; Burstein, 1980). Yet this use has continued. Torn between the limitations of the mixed-model ANOVA and the fixed-effects regression, researchers have often clung to regression, especially in nonexperimental studies where it is necessary to control for multiple continuously measured covariates in order to avoid misspecifying estimates of theoretically interesting predictors.

In repeated measures analysis, multivariate methods have offered a reasonable alternative to the univariate mixed-model ANOVA. Within-subjects contrasts are represented as multiple dependent variables, allowing the investigator to make less stringent assumptions about the covariance structure of the repeated measures than are possible using the univariate mixed-model ANOVA (Bock, 1975). However, the standard multivariate approach does not accommodate missing data, uneven spacing of time series observations, or time-varying covariates (Ware, 1985).

The difficulty in extending general linear model analysis to designs having fixed and random effects has been technical. When designs are completely balanced, it is possible to estimate efficiently the fixed effects (for example, regression coefficients) without information about the variance components. Once the fixed effects are estimated, it is not difficult to estimate the variance components. However, in unbalanced designs, efficient estimates of the fixed effects and the variance components are mu-
tually dependent, and some sort of iterative procedure is needed to arrive at a unique and efficient set of estimates.

The more general analytic methods described in this chapter are based on such iterative methods, which have become available as a result of advances in statistical theory and computation. Statisticians are now making rapid progress in algorithmic development, essentially enabling researchers to "catch up" with advances in computing. The result is an array of new analytic approaches rapidly becoming accessible to researchers.

13.1.5 Hierarchical Linear Models

This chapter considers hierarchical linear models, a class of models that combine the advantages of the mixed-model ANOVA, with its flexible modeling of fixed and random effects, and regression, with its advantages in dealing with unbalanced data and predictors that are discrete or continuous. Results based on hierarchical linear models duplicate the results of many classical ANOVA models (Kirk, 1982; Winer, 1971) and expand possibilities for data analysis.

Hierarchical linear models have been referred to as random coefficient models (Rosenberg, 1973), multilevel linear models (Mason, Wong, and Entwisle, 1983; Goldstein, 1987), covariance components models (Dempster, Rubin, and Tsutakawa, 1981), and unbalanced models with nested random effects (Longford, 1987). However, the term hierarchical linear models captures two defining features of the models. First, the data appropriate for such models are hierarchically structured, with first-level units nested within second-level units, second-level units nested within third-level units, and so on. Second, the parameters of such models may be viewed as having a hierarchical linear structure. The investigator may specify a level-one model, the parameters of which characterize linear relationships occurring between level-one units. These parameters are then viewed as varying across level-two units as a function of level-two characteristics. Higher levels may be added, in principle, without limit, although to date no published applications involve more than three levels.

The connection between hierarchical linear models and classical experimental design models has not been well understood in the past. Some assume that hierarchical models apply only to designs having nested factors and not to designs having crossed factors, but this intuition is false. What distinguishes hierarchical linear models is that the random factors are nested and never crossed. However, fixed factors can be crossed with random factors (or with each other) and random factors may be nested within fixed factors. The data can be unbalanced at any level, and continuous predictors can be defined at any level. Both discrete and continuous predictors can
be specified as having random effects, and these random effects are allowed to covary. We shall return later to a broader class of models under the rubric of the general mixed linear model, which subsumes hierarchical linear models but also includes models with crossed random factors.

### 13.1.6 Organization of the Chapter

The next section formalizes the two-level hierarchical linear model. Subsequent sections apply this model to one-way designs, two-factor nested designs, two-factor crossed designs, and randomized blocks (repeated measures) designs. In the case of balanced data, hierarchical model results are shown to duplicate classical results, as demonstrated by reanalysis of data from Kirk (1982). In each case, the capacity of the hierarchical analysis to generalize application of the classical model is demonstrated on large-scale field data that would pose serious problems for a classical approach. In the final section, we consider more complex designs and indicate how a general mixed model fills in gaps left even by the hierarchical model.

### 13.2 THE TWO-LEVEL HIERARCHICAL LINEAR MODEL

In this section we formulate a general two-level hierarchical linear model. For concreteness, let us assume that the design involves students nested within schools. The level-one model specifies how student-level predictors relate to the student-level outcome. At level two, each of the regression coefficients defined in the level-one model, including the intercept, may be predicted by school-level predictors, and each may, in addition, have a random component of variation. The combined level-one and level-two models constitute a mixed linear model with fixed and random regression coefficients. A number of classical ANOVA and regression models may be shown to represent simplifications of this mixed model, as illustrated in Sections 13.3–13.6.

#### 13.2.1 The Level-One Model

At level one, the student level, the outcome $y_{ij}$ for student $i$ in school $j$ ($i = 1, \ldots, n_j; j = 1, \ldots, J$), varies as a function of student characteristics, $X_{qij}$, $q = 1, \ldots, Q$, and a random error $r_{ij}$ according to the linear regression model

$$y_{ij} = \beta_0 + \sum \beta_q X_{qij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2),$$

where $\beta_0$ is the intercept and each $\beta_q, q = 1, \ldots, Q$, is a regression coefficient indicating the strength of association between each $X_{qij}$ and the
outcome within school $j$. Note that the intercept and the regression slope are each subscripted by $j$, allowing them to vary from school to school. The error of prediction of $y_{ij}$ by the $X$'s is $r_{ij}$, which is assumed normally distributed and, for simplicity, homoscedastic.

### 13.2.2 The Level-Two Model

At level two (the school level), each regression coefficient, $\beta_{qj}$, $q = 0, 1, \ldots, Q$, defined by the level-one model, becomes an outcome variable to be predicted by school-level characteristics, $W_{sj}$, $s = 1, \ldots, S$, according to the regression model

$$\beta_{qj} = \Theta_{q0} + \Sigma Q_s \Theta_{qs} W_{sj} + u_{qj},$$

(13.2)

where $\Theta_{q0}$ is an intercept; each $\Theta_{qs}$, $s = 1, \ldots, S$, is a regression slope specifying the strength of association between each $W_{sj}$ and the outcome $\beta_{qj}$; and the random effects are assumed sampled from a $(Q + 1)$-variate normal distribution, where each $u_{qj}$, $q = 1, 2, \ldots, Q$, has a mean of 0 and a variance $\tau_{qq}$, and the covariance between $u_{qj}$ and $u_{q'j}$ is $\tau_{qq'}$.

The user is faced with a considerable number of options in modeling each $\beta_{qj}$, $q = 0, 1, \ldots, Q$, in Eq. (13.2). If every $W_{sj}$ is assumed to have no effect, the regression coefficients $\Theta_{qs}$, $s = 1, \ldots, S$, are set to zero. If the random effect $u_{qj}$ is also constrained to zero, then $\beta_{qj} = \Theta_{q0}$, i.e., $\beta_{qj}$ is fixed across all schools. This option is useful when the user wishes to constrain a regression to be homogeneous. If the user believes that one or more of the $W$'s does predict $\beta_{qj}$, but wishes to constrain $u_{qj}$ to zero, then $\beta_{qj}$ does indeed vary, but strictly as a function of the $W$'s: its variation is nonrandom. Of course, $\beta_{qj}$ could also could have a random component (not constrained to zero) but might not be predicted by the $W$'s. Its variation is strictly random, with no predictable component. Clearly, Eq. (13.2) can be simplified in a large number of ways, which enables the user to tailor the model to conform to a variety of conventional linear models, as we shall demonstrate in the next four sections.

### 13.2.3 Research Applications

Applications of the two-level model in cross-sectional school effects research are presented by Aitkin and Longford (1986), deLeeuw and Kreft (1986), Raudenbush and Bryk (1986), and Goldstein (1987). We consider applications of this type in Sections 13.3 and 13.4.

In research on growth, Eq. (13.1) may be viewed as a within-subject model relating the outcome $Y_{ij}$ of person $j$ at time $i$ to age or time where the $X$'s might be polynomial trend coefficients. In that case the $\beta$'s are the growth parameters of subject $j$, to be predicted in Eq. (13.2) by subject
Hierarchical Linear Models and Experimental Design

background characteristics or between-subjects treatments (the W's). Applications in growth research (Laird and Ware, 1982; Bryk and Raudenbush, 1987) will be considered in Section 13.6.

In research synthesis, Eq. (13.1) is a within-study model, perhaps relating treatment contrasts (the X's) to subject outcomes (Y). The treatment effects defined in that equation (the β's) then become the outcomes in the between-study model [Eq. (13.2)] where the W's are study characteristics that predict different effects. Raudenbush and Bryk (1985) and Becker (1988) used two-level models in this way.

A book-length treatment of the use of these models in studying organizational effects, growth, and research synthesis is given in Bryk and Raudenbush (1992). It also provides a detailed account of the statistical theory underlying estimation. The present chapter is novel in clarifying linkages between hierarchical linear models and classical models for experimental design.

13.3 THE ONE-WAY ANALYSIS OF VARIANCE

13.3.1 Classical Approach

13.3.1.1 The One-Way ANOVA Model

The usual statistical model for the one-way classification may be written as

\[ y_{ij} = \mu + \alpha_j + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2), \]  

(13.3)

where \( y_{ij} \) is the observation for subject \( i \) assigned to level \( j \) of the independent variable \( (i = 1, \ldots, n_i; j = 1, \ldots, J) \); \( \mu \) is the grand mean; \( \alpha_j \) is the effect associated with level \( j \); and \( r_{ij} \) is assumed normally distributed with a mean of 0 and homogeneous variance, \( \sigma^2 \). In the fixed-effects case, the effect \( \alpha_j \) is a fixed constant and the constraint \( \sum n_i \alpha_j = 0 \) is introduced. In the random-effects approach, each \( \alpha_j \) is typically assumed independently, normally distributed, i.e., \( \alpha_j \sim N(0, \tau^2) \).

13.3.1.2 Data and Results

The data in Table 13.1 were collected by Nelson (1976) and reprinted by Kirk (1982, p. 168). Fifty children were randomly assigned to five experimental methods of training in discrimination among blocks; the outcome was the number of blocks correctly identified. Thus, \( n_i = n = 10 \) for all \( j \), \( J = 5 \), and the total sample size is \( nJ = 50 \).

We consider the results under the random-effects assumption simply because it offers an interesting comparison to the hierarchical linear model.
Table 13.1 One-Way ANOVA Data and Source Table

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
<th>Method 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

ANOVA Source Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>( F(\text{mean square}) )</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>4</td>
<td>21.88</td>
<td>5.47</td>
<td>( n\tau^2 + \sigma^2 )</td>
<td>5.37</td>
</tr>
<tr>
<td>Within groups</td>
<td>45</td>
<td>45.80</td>
<td>1.02</td>
<td>( \sigma^2 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>67.68</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

results. The grand mean estimate and its estimated standard error are

\[
\hat{\mu} = \Sigma \Sigma y_{ij}/50 = 2.08,
\]

\[
s(\hat{\mu}) = \sqrt{(MS_p/50)} = .331.
\]

The ANOVA source table shows evidence of significant variation among groups, \( F(4, 45) = 5.37, p < .001 \). Under the random-effects model, this suggests rejection of the null hypothesis \( H_0: \tau^2 = 0 \). To estimate the variance components, the observed mean squares are set equal to their expected values given in Table 13.1:

\[
MS_w = \sigma^2,
\]

\[
MS_b = n\tau^2 + \sigma^2,
\]

leading to the solutions

\[
\hat{\sigma}^2 = MS_w = 1.02,
\]

\[
\hat{\tau}^2 = (MS_b - MS_w)/n = .445,
\]

where \( MS_b \) and \( MS_w \) refer to mean squares between and within, respectively. This result implies that the proportion of variation between groups is about
.445/(.445 + 1.02) = .30 or 30%. This proportion is also the estimated intraclass correlation coefficient, that is, the estimated correlation between pairs of observations sharing membership in the same group, j.

13.3.2 Analysis via the Hierarchical Linear Model

13.3.2.1 The Model

Let us first set every regression coefficient in Eq. (13.1), save the intercept, to zero, so that the level-one model becomes

\[ y_j = \beta_{0j} + r_j, \quad r_j \sim N(0, \sigma^2). \tag{13.4} \]

According to this model, the outcome for subject i in group j is predicted only by the intercept, \( \beta_{0j} \), which is the group mean, so that the variance \( \sigma^2 \) is the within-group variance. Only one parameter, \( \beta_{0j} \), is to be predicted at level two [Eq. (13.2)], and the model for that parameter is similarly simplified so that all regression coefficients except the intercept are set to zero:

\[ \beta_{0j} = \Theta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau^2). \tag{13.5} \]

Here \( \Theta_{00} \) is the grand mean and \( u_{0j} \) is the effect associated with level j. In the random-effects model, that effect is typically assumed normally distributed with a mean of zero. In terms of Eq. (13.2), this variance might be denoted by \( \tau_{00} \), but we use the notation \( \tau^2 \) for conformity to standard ANOVA usage. In the fixed-effects models, each \( u_{0j} \) is a fixed constant. Combining Eqs. (13.4) and (13.5) yields the single model

\[ y_j = \Theta_{00} + u_{0j} + r_j, \tag{13.6} \]

with \( r_j \sim N(0, \sigma^2) \) and \( u_{0j} \sim N(0, \tau^2) \). This is clearly the one-way random effects ANOVA, identical to Eq. (13.3) with \( \Theta_{00} = \mu \), and \( \alpha_j = u_{0j} \).

13.3.2.2 Estimation

For all hierarchical analyses in this chapter we employ the computer program of Bryk, Raudenbush, Seltzer, and Congdon (1988), which computes restricted maximum likelihood (ML) estimates by means of an iterative approach known as the EM algorithm (Dempster, Laird, and Rubin, 1977). (We discuss the logic of this approach in some detail here; readers uninterested in the logic of estimation and hypothesis testing may skip to the results in Section 13.3.2.4.)

The restricted likelihood is the full likelihood with the fixed effects integrated out, leaving the likelihood of the data strictly as a function of the variance-covariance components (see Dempster et al., 1981). Estimates of variance-covariance components based on the restricted likelihood have
the distinct advantage of taking into account uncertainty about the effects. In balanced designs, this means that variance estimates are corrected for loss of degrees of freedom due to estimation of fixed effects (e.g., to estimate a residual variance in regression, the residual sum of squares is divided by \( n - p \) rather than \( n \), where \( n \) is the sample size and \( p \) is the number of regression coefficients including the intercept). The reader is warned that computing programs based on the full rather than the restricted likelihood will not duplicate exactly the standard ANOVA results for balanced designs.

In the case of the one-way analysis of variance with random effects, the ML grand mean estimate is a precision-weighted average of the group means. Let \( \bar{y}_j = \frac{\sum y_j}{n_j} \) denote the sample mean for group \( j \). Then the variance of \( \bar{y}_j \) is

\[
V_j = \text{Var}(\bar{y}_j) = \tau^2 + \sigma^2/n_j.
\]  
(13.7)

The precision of \( \bar{y}_j \) is then defined as the inverse of its variance

\[
P_j = V_j^{-1},
\]

and, with \( \sigma^2 \) and \( \tau^2 \) known, the ML estimate and also the unique, minimum variance unbiased estimate of the grand mean is the precision-weighted average

\[
\hat{\theta}_{\infty} = \frac{\sum P_j \bar{y}_j}{\sum P_j},
\]  
(13.8)

Note that the best estimate is not the grand mean in the sample. Nor is the best estimate the arithmetic mean of the sample means. Raudenbush (1984) showed that Eq. (13.8) is a compromise between the grand mean \( \frac{\sum y_j}{\sum n_j} \) and the mean of the means \( \sum \bar{y}_j/J \). Equation (13.8) approaches the grand mean as \( \tau^2 \) approaches zero (holding \( n_j \) constant) and approaches the mean of the means as \( \tau^2 \) increases. Also, as \( n_j \) increases (with \( \tau^2 \) held constant), Eq. (13.8) approaches the mean of the means. When the data are balanced, the two estimates are identical.

When \( \sigma^2 \) and \( \tau^2 \) are not known, the ML estimate of the grand mean is Eq. (13.8) with ML estimates substituted for \( \sigma^2 \) and \( \tau^2 \). However, when the data are balanced (\( n_j = n \) for all \( j \)), \( P_j \) is constant for every group; that is, every group mean has the same precision and Eq. (13.8) becomes the grand mean

\[
\hat{\theta}_{\infty} = \bar{y}_\cdot = \frac{\sum \bar{y}_j}{J}.
\]  
(13.9)

As Eq. (13.9) shows, in the case of balanced data, knowledge about the variance components \( \sigma^2 \) and \( \tau^2 \) is not needed to obtain the ML estimate of the grand mean. This principle is important and generalizes to more complex designs. ML estimates of fixed effects do not depend on variance
Hierarchical Linear Models and Experimental Design

components when the data are balanced, so that, in the balanced case, the
standard ANOVA estimators and the ML estimators of fixed effects converge.

With \( \sigma^2 \) and \( \tau^2 \) known, the precision of the ML estimate is the sum of
the respective precisions of the group means

\[
\text{Precision}(\hat{\theta}_{oo}) = \Sigma P_j
\]

so that the sampling variance of \( \hat{\theta}_{oo} \) is the inverse of its precision

\[
\text{Var}(\hat{\theta}_{oo}) = \frac{1}{\Sigma P_j}
\]

However, when the data are balanced, the precisions are equal and

\[
\text{Var}(\hat{\theta}_{oo}) = \frac{1}{J_P} = \frac{\tau^2 + \sigma^2/n}{J} = \frac{nr^2 + \sigma^2}{Jn}.
\]

(13.10)

The reader will notice that the numerator in Eq. (13.10) is equivalent to
the expected \( MS_g \) (see Table 13.1) so that the restricted ML estimate is

\[
\text{Estimated Var}(\hat{\theta}_{oo}) = \frac{MS_g}{Jn},
\]

(13.11)

and the estimated standard error of \( \hat{\theta}_{oo} \) is the square root of Eq. (13.11).
We shall illustrate the principle emerging from this simple example in
subsequent sections. In balanced designs, standard errors for fixed-effects
estimates will be simple functions of sufficient statistics such as mean squares.

Standard texts assert without qualification that the best estimate of a
group mean in the one-way ANOVA is simply the group's sample mean,
\( \bar{y}_j \). However, as Lindley and Smith (1972) pointed out, this point estimate
can be improved upon whenever the design includes more than two groups.
An alternative point estimator of the group mean, \( \beta_{0j} \), is the weighted average

\[
\beta_{0j}^* = \pi_j \bar{y}_j + (1 - \pi_j) \hat{\theta}_{oo},
\]

(13.12)

where \( \pi_j = \tau^2(\tau^2 + a^2/n_j) \). Note that \( \beta_{0j}^* \) is a weighted average of group
\( j \)'s sample mean, \( \bar{y}_j \), and the grand mean, \( \hat{\theta}_{oo} \). The weight accorded \( \bar{y}_j \) is
\( \pi_j \), which, in fact, is the reliability of \( \bar{y}_j \) as an estimate of \( \beta_{0j} \), where reliability is measured by the ratio of the "true-score" variance, \( \text{Var}(\beta_{0j}) = \tau^2 \), to the "observed-score" variance, \( \text{Var}(\bar{y}_j) = \tau^2 + \sigma^2/n_j \). Note that the observed-score variance includes both the true-score variance, \( \tau^2 \), and the error variance, \( \sigma^2/n_j \), of \( \bar{y}_j \) as an estimate of \( \beta_{0j} \). Whenever the error variance is large relative to \( \tau^2 \), \( \pi_j \) will be small and \( \beta_{0j}^* \) will rely heavily on the
grand mean. However, when the error variance is small relative to \( \tau^2 \),
\( \pi_j \) will be large and \( \beta_{0j}^* \) will rely heavily on the group's sample mean. The
weighted average, \( \beta_{0j}^* \), has a smaller expected mean squared error of estimation than does the sample mean, \( \bar{y}_j \) (Efron and Morris, 1975).
Note that Eq. (13.12) requires knowledge of the variances \( \sigma^2 \) and \( \delta^2 \). However, when these are unknown, results of the hierarchical linear model analysis yield maximum likelihood estimates of these variances, even when the data are unbalanced. These estimates can be substituted into Eq. (13.12) to obtain what are sometimes called “empirical Bayes’” estimates (Morris, 1983).

We caution that these empirical Bayes estimates are justifiable only if the investigator has no a priori hypotheses regarding the expected magnitude of the group means. For example, if the treatments represent duration of exposure so that larger durations are expected to lead to greater effect sizes, these empirical Bayes estimates should not be used. In this case, “conditional” empirical Bayes estimators are available that represent a compromise between the sample mean and its predicted value based on a particular estimated a priori contrast (see Raudenbush, 1988).

13.3.2.3 Hypothesis Testing
A simple test of the null hypothesis of no group effects, i.e.,

\[ H_0: \tau^2 = 0 \]

is given by the statistic

\[ H = \sum \hat{\delta} (\bar{y}_j - \bar{y})^2, \tag{13.13} \]

where \( \hat{\delta} = n_i/\delta^2 \). The statistic \( H \) has a large sample chi-square distribution with \( J - 1 \) degrees of freedom under the null hypothesis. In the case of balanced data, this sum of precision-weighted squared differences reduces to

\[ H = \hat{\delta} \sum (\bar{y}_j - \bar{y})^2 = (J - 1)MS_j/MS_\nu, \tag{13.14} \]

revealing clearly that \( H/(J - 1) \) is the usual \( F \) statistic for testing group differences in ANOVA.

13.3.2.4 Results
The results of the hierarchical analysis (Table 13.2) are mathematically identical to the results based on the usual ANOVA, although they are presented in a different form. The point estimates for all parameters are the same as those in Table 13.1. Note that the chi-square test for group differences is \( H = 21.50 \) and that \( H/(J - 1) = 21.50/4 = 5.37 \), the \( F \) statistic in the ANOVA table (Table 13.1).

13.3.2.5 Estimation of Group Effects
Table 13.3 lists the standard group effect estimators, \( \bar{y}_j \), and the empirical Bayes estimators, \( \hat{\beta}_j \), for the data in Table 13.1. Note that the empirical
Table 13.2 One-Way ANOVA Results Based on the Hierarchical Linear Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, $\theta_{00}$</td>
<td>2.08</td>
<td>.331</td>
</tr>
</tbody>
</table>

Random effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^2$</td>
<td>.445</td>
<td>21.50</td>
<td>4</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bayes estimators are closer to the grand mean of 2.08 than are the five sample means.

In this case, the differences between the standard and empirical Bayes estimators are small because the reliability of the sample means is reasonably high at .81. Hence, the group mean $\bar{y}_j$ is weighted heavily relative to the grand mean $\bar{y}_{..}$ in composing the empirical Bayes estimator, $\beta_j^e$ [Eq. (13.12)].

13.3.3 Generalization

To illustrate application of the one-way random effects ANOVA in large-scale survey data, we consider the results of Raudenbush, Eamsukkawat, Di-Ibor, Kamali, and Taoklam (1991), who analyzed data from 103 small primary schools in rural Thailand. Each school had just one sixth-grade classroom, and the outcome variable was total academic achievement as measured by a test covering the five major areas of the Thai curriculum.

Table 13.3 Standard and Empirical Bayes Estimates of Group Effects in the One-Way ANOVA

<table>
<thead>
<tr>
<th>Group</th>
<th>$\bar{y}_j$</th>
<th>$\beta_j^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.50</td>
<td>1.55</td>
</tr>
<tr>
<td>2</td>
<td>2.40</td>
<td>2.34</td>
</tr>
<tr>
<td>3</td>
<td>2.50</td>
<td>2.42</td>
</tr>
<tr>
<td>4</td>
<td>2.90</td>
<td>2.75</td>
</tr>
<tr>
<td>5</td>
<td>1.20</td>
<td>1.37</td>
</tr>
</tbody>
</table>
Table 13.4 One-Way ANOVA Results of
Thailand Classroom Data Based on the Hierarchical
Linear Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, $\theta_{00}$</td>
<td>-24.45</td>
<td>6.25</td>
</tr>
</tbody>
</table>

Random effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^2$</td>
<td>3785</td>
<td>1947.8</td>
<td>102</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>4099</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The outcome variable was standardized nationally to have a mean of 0 and a standard deviation of 100. Classrooms averaged 19.2 students with a minimum of 9 and a maximum of 35. The one-way ANOVA provided a number of useful preliminary results in their analysis (see Table 13.4).

The results show, first, that average achievement in these small rural schools is significantly less than the national average of 0 ($\hat{\theta}_{00} = -24.45$, se = 6.25). Second, the between-classroom variance component is large. The proportion of variance between classrooms (equivalent to the intraclass correlation) is estimated to be $3785/(3785 + 4099) = .48$ or 48%. This proportion of variation between classrooms is much larger than that typically reported in U.S. primary schools (cf. Bryk and Raudenbush, 1988) and, not surprisingly, is highly statistically significant, with an approximate chi-square of 1947.8, in this case equal to $F(102, 1875) = 1947.8/102 = 19.1$, $p < .001$. Note that the denominator degrees of freedom are equal to the total sample size of 1978 minus 103, the number of classrooms.

Because of the large variation among classrooms (i.e., $\tau^2$ is large), a typical classroom's sample mean has a high reliability of about .94 as an estimate of its unknown true mean. Hence, the empirical Bayes estimators of the classroom means will be very similar to the classroom sample means as shown in Figure 13.1. In contrast, in studies where $\tau^2$ is small, little of the variation among the classroom sample means reflects variation in the true means. In such studies, the empirical Bayes distribution looks much less dispersed than does the distribution of these classroom means (see Willms and Raudenbush, 1989, for an example).

The example illustrates a generalization of the standard one-way random effects ANOVA in that maximum likelihood estimates of the grand mean
and the variance components are available even though the data are unbalanced. These estimated variance components, besides enabling an estimate of the reliability of the group means, facilitate computation and interpretation of the empirical Bayes estimators.

13.4 THE TWO-FACTOR NESTED DESIGN

As mentioned in the introduction, two-factor nested designs are common in education (e.g., students within classrooms, classrooms within instructional methods), counseling psychology (e.g., patients within counselors, counselors within therapies), and sociology, for example, cross-national studies of demography in which women are nested within countries and countries within types of family-planning efforts (Mason et al., 1983). In these designs, classrooms, counselors, or countries are typically viewed as random factors because only a sample of the interesting levels is available and one wishes to make generalizations of such levels to the universe of levels. Noniterative maximum likelihood estimates are available in the balanced case treated in standard experimental design texts. Hierarchical linear models generalize maximum likelihood estimation to the case of unbalanced data, and covariates measured at each level can be discrete or continuous and can have either fixed or random effects. In this section, we (1) consider the classical analytic approach, (2) show how the hierarchical linear model can be formulated to duplicate the classical results in the balanced case, and (3) illustrate how the hierarchical analysis gener-
alizes to cases involving unbalanced data and incorporation of discrete and continuous covariates.

13.4.1 Classical Approach

13.4.1.1 The Model

The standard model for the two-factor nested design may be written

\[ y_{ijk} = \mu + \alpha_i + \pi_{i(k)} + r_{ijk}, \quad \pi_{i(k)} \sim N(0, \tau^2) \]
\[ r_{ijk} \sim N(0, \sigma^2) \]  

(13.15)

where \( y_{ijk} \) is the outcome for subject \( i \) nested within level \( j \) of the random factor, which is, in turn, nested within level \( k \) of the fixed factor \( (i = 1, \ldots, n_j; j = 1, \ldots, J; k = 1, \ldots, K) \); \( \mu \) is the grand mean; \( \alpha_i \) is the effect associated with the \( k \)th level of the fixed factor; \( \pi_{i(k)} \) is the effect associated with the \( j \)th level of the random factor within the \( k \)th level of the fixed factor; and \( r_{ijk} \) is the random (within-cell) error. In the case of balanced data \( (n_{jk} = n \) for every level of the random factor), the standard analysis of variance method and the method of restricted maximum likelihood coincide.

13.4.1.2 Example and Results

Table 13.5 lists artificial data provided in Kirk (p. 460). For ease of understanding we shall refer to the design as involving classes nested within methods of instruction with the outcome being the number of correct responses on a posttreatment cognitive test. There are four classes within each of two methods.

The estimated grand mean and its estimated standard error are given by

\[ \hat{\mu} = \bar{y} = 5.375, \]
\[ \text{se}(\hat{\mu}) = \sqrt{\text{MS}_{\text{class}}/32} = .738. \]

The estimated contrast between methods 1 and 2 is

\[ \hat{\alpha}_1 - \hat{\alpha}_2 = 3.75, \]

and its significance is tested by the ratio of mean squares (methods to classes) yielding \( F(1, 6) = 6.46, p < .05 \). Variance estimates are given by equating the expected mean squares to their observed values and solving as in the one-way ANOVA, yielding

\[ \hat{\sigma}^2 = \text{MS}_{\text{within}} = 0.77, \]
\[ \hat{\tau}^2 = (\text{MS}_{\text{classes}} - \text{MS}_{\text{within}})/n = 4.16. \]
Table 13.5 Nested Two-Factor Data and Source Table

<table>
<thead>
<tr>
<th></th>
<th>Method 1</th>
<th></th>
<th>Method 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Class 2</td>
<td>Class 3</td>
<td>Class 4</td>
<td>Class 5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

ANOVA source table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean square</th>
<th>$\bar{E}(\text{mean square})$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>1</td>
<td>112.50</td>
<td>112.50</td>
<td>$\sigma^2 + n\tau^2 + nJ \sum \alpha_j^2 (K - 1)$</td>
<td>6.46</td>
</tr>
<tr>
<td>Classes</td>
<td>6</td>
<td>104.50</td>
<td>17.42</td>
<td>$\sigma^2 + n\tau^2$</td>
<td>22.59</td>
</tr>
<tr>
<td>Within cell</td>
<td>24</td>
<td>18.50</td>
<td>0.77</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>235.50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As in the one-way ANOVA case, negative values of $\tau^2$ are set to 0. (An advantage of estimation for the hierarchical model based on the EM algorithm is that estimates will be nonnegative.) Note that $\tau^2$ is the residual between-classes variance after removing the effects of the method factor, so it will differ from the estimate resulting from a one-way ANOVA. A test of the null hypothesis of no class-level variance, i.e., $H_0: \tau^2 = 0$, is the $F$ statistic computed as the ratio of mean squares (classes to within-cell), in this case yielding $F(6, 24) = 22.59, p < .01$.

13.4.2 Analysis by Means of a Hierarchical Linear Model

13.4.2.1 The Model

As in the case of the one-way ANOVA, we first set every regression coefficient in Eq. (13.1), save the intercept to zero, so that the level-one model is

$$y_{ij} = \beta_0 + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2), \quad (13.16)$$

where $y_{ij}$ is the outcome for subject $i$ within class $j$ ($i = 1, \ldots, n_i; j =$
1, ..., J), \( \beta_{0j} \) is the mean of class \( j \), and \( r_{ij} \) is the residual assumed normally distributed with a mean of zero and within-class variance \( \sigma^2 \).

The level-two (between-class) model is a regression model in which the class mean \( \beta_{0j} \) is the outcome and the predictor is a contrast between the two treatments (methods of instruction). This level-two model,

\[
\beta_{0j} = \Theta_{00} + \Theta_{01}W_j + u_{0j}, \quad u_{0j} \sim N(0, \tau^2), \tag{13.17}
\]

has \( W_j = 1/2 \) for classes experiencing instructional method 1 and \( W_j = -1/2 \) for classes experiencing instructional method 2. Hence, the correspondences between the hierarchical model and the ANOVA model are

\[ \Theta_{00} = \mu, \quad \Theta_{01} = \alpha_1 - \alpha_1, \quad u_{0j} = \pi_{j(k)} . \]

Combining Eqs. (13.16) and (13.17) yields the single model

\[
y_{ij} = \Theta_{00} + \Theta_{01}W_j + u_{0j} + r_{ij}. \tag{13.18}
\]

Note that \( K - 1 \) contrasts must be included to represent the variation among the \( K \) methods in order to duplicate the ANOVA results.

13.4.2.2 Results

The results of the analysis via the hierarchical model using Table 13.6 are again mathematically identical to the results based on the usual ANOVA, although they are presented in a different form. The reader will again notice that squaring the \( t \)-statistic associated with the method contrast will yield the corresponding \( F \) of the ANOVA in Table 13.5, and dividing the chi-square for the class variance by its degrees of freedom will yield the corresponding \( F \) in the ANOVA table.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, ( \Theta_{00} )</td>
<td>5.375</td>
<td>.738</td>
<td>—</td>
</tr>
<tr>
<td>Method contrast, ( \Theta_{01} )</td>
<td>3.750</td>
<td>1.48</td>
<td>2.54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance components</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau^2 )</td>
<td>4.16</td>
<td>135.6</td>
<td>6</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
13.4.3 Generalization

We again turn to the analysis of the primary school data from Thailand. Our goal is to examine the difference in achievement between classrooms taught by teachers with a bachelor's degree and classes taught by teachers with less than a bachelor's degree. Complicating the analysis are the following facts: (1) the data are unbalanced; and (2) because classes were not assigned at random to levels of teacher education, it will be important to control for potentially confounding variables. These covariates may exist at both the student and the class levels and they may be either continuously or discretely measured. Hence, although the design is indeed a nested two-factor design, the balanced two-factor nested ANOVA will be inappropriate.

Raudenbush, Kidchanapanish, and Kang (1991) addressed these issues with the following level-one model:

\[ y_{ij} = \beta_{0j} + \beta_{1j}(GPA)_{ij} + \beta_{2j}(Repetition)_{ij} + \beta_{3j}(SES)_{ij} \\
+ \beta_{4j}(Time)_{ij} + \beta_{5j}(Dialect)_{ij} + \beta_{6j}(Breakfast)_{ij} + r_{ij}, \]  

(13.19)

where GPA is fifth-grade point average, Repetition is an indicator for having ever repeated a grade, SES is a measure of socioeconomic status, Time is the time needed to travel to school, Sex is an indicator for males, Dialect is an indicator for speaking Central Thai dialect, and Breakfast is an indicator for eating breakfast daily.

At level two, only the intercept was allowed to be random. Other predictors could have been specified to have random coefficients, but they were not. In addition, only the intercept was specified as being predicted by school characteristics, so that the level-two model was

\[ \beta_{0j} = \Theta_{00} + \Theta_{01}(Infrastructure)_j + \Theta_{02}(Remoteness)_j \\
+ \Theta_{03}(Supervision)_j + \Theta_{04}(Bachelor’s)_j \\
+ u_{0j}, \quad \beta_{qj} = \Theta_{q0} \text{ for } q > 0, \]  

(13.20)

where Infrastructure is a scale measuring the modernity of the community in which the school is located; Remoteness is a composite measure of the school's distance from the district capital, the highway, and the market; Supervision is the intensity of the supervision provided the teacher; and Bachelor's is an indicator (1 = bachelor's degree; 0 = less than bachelor's).

The results (Table 13.7) indicate, after controlling for covariates, a mean difference of 32.0 points in favor of classes of teachers with a bachelor's degree. The among-classroom standard deviation (Table 13.4) is \( \sqrt{3785} = 61.5 \), indicating that the standardized effect size associated with a bachelor's degree is about 32/61.5 = .52 in units of the classroom
Table 13.7  Two-Factor Nested Results for Thailand Classroom Data Based on the Hierarchical Linear Model

<table>
<thead>
<tr>
<th>Predictors</th>
<th>One-way ANOVA</th>
<th></th>
<th>Full model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>se</td>
<td>t</td>
<td>Coeff</td>
</tr>
<tr>
<td>School/classroom level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-24.45</td>
<td>6.25</td>
<td>-3.92</td>
<td>-50.29</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>37.25</td>
<td>15.80</td>
<td>2.36</td>
<td></td>
</tr>
<tr>
<td>Remoteness</td>
<td>6.53</td>
<td>2.76</td>
<td>2.37</td>
<td></td>
</tr>
<tr>
<td>Bachelor's</td>
<td>31.99</td>
<td>12.27</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td>Internal supervision</td>
<td>10.25</td>
<td>4.71</td>
<td>2.18</td>
<td></td>
</tr>
<tr>
<td>Student level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>25.67</td>
<td>1.15</td>
<td>22.29</td>
<td></td>
</tr>
<tr>
<td>Repetition</td>
<td>-22.86</td>
<td>3.77</td>
<td>-6.06</td>
<td></td>
</tr>
<tr>
<td>SES</td>
<td>4.86</td>
<td>3.22</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>Time to school</td>
<td>-4.60</td>
<td>2.32</td>
<td>-1.98</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>-2.09</td>
<td>2.51</td>
<td>-0.83</td>
<td></td>
</tr>
<tr>
<td>Dialect</td>
<td>19.01</td>
<td>8.33</td>
<td>2.82</td>
<td></td>
</tr>
<tr>
<td>Breakfast</td>
<td>12.37</td>
<td>4.49</td>
<td>2.75</td>
<td></td>
</tr>
</tbody>
</table>

### Variance components

<table>
<thead>
<tr>
<th>Parameter</th>
<th>One-way ANOVA</th>
<th>Full model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between class, $\delta^2$</td>
<td>3785</td>
<td>2693</td>
</tr>
<tr>
<td>% Explained</td>
<td>0.0</td>
<td>28.8</td>
</tr>
<tr>
<td>Within class, $\delta^2$</td>
<td>4099</td>
<td>3072</td>
</tr>
<tr>
<td>% Explained</td>
<td>0.0</td>
<td>25.0</td>
</tr>
</tbody>
</table>

The effect size is about .36 in units of the overall standard deviation (note the overall standard deviation of the outcome shown in Table 13.4 is just less than 90). Notice also, that the estimated residual variances ($\delta^2$ at the student level and $\delta^2$ at the classroom level) are reduced by 28.8% and 25.0%, respectively, revealing how the explanatory power of the model may be monitored in this type of analysis. The analysis has allowed for control of continuous and discrete covariates at two levels of aggregation and produces ML variance components estimates even though the data are unbalanced. These variance components estimates not only are interesting in themselves but also are required to obtain estimates of the regression coefficients, estimated via a generalized least squares approach that weights the data from each school in proportion to its precision (Aitkin and Longford, 1986; Raudenbush, 1988).
13.5 THE TWO-FACTOR CROSSED DESIGN (WITH REPlications WITHIN CELLS)

Experimental design texts distinguish among three types of models for two-factor designs: Model I (both factors fixed), Model II (both factors random), and Model III (the "mixed" case with one factor fixed and the other random). In Section 13.4 on nested factors, we considered only the mixed case, the case that arises most naturally in field research. In the present section on crossed two-factor designs, we again focus on the mixed case (Model III).

It is well known that the fixed-effects case (Model I) can readily be represented within the framework of standard multiple regression, an especially useful approach when the data are unbalanced. Multiple regression offers a flexible approach for incorporating effects of covariates measured either continuously or discretely. Available texts (for example, Cohen and Cohen, 1975) consider such applications in detail. Here we will simply indicate that standard multiple regression represents a special case of the hierarchical linear model [Eqs. (13.1) and (13.2)] in which all variance components except $\sigma^2$ have been set to zero.

On the other hand, the case in which both factors are random (Model III) cannot be represented within the hierarchical model; and generalizing Model III to include unbalanced data and covariates requires a more general mixed model (see Dempster, Rubin, and Tsutakawa, 1981). Goldstein (1987) has considered estimation for such models, and Kang (1991) has developed a workable computational approach based on the EM algorithm that allows fixed and random effects of rows, columns, and interactions where predictors may be either continuous or discrete and the data may be unbalanced.

Crossed two-factor designs with one factor fixed and one factor random arise commonly in social and behavioral research. For example, Raffe (1991) considers the case of a treatment-control contrast estimated within each of 19 sites in the context of evaluating Britain's innovative technical education initiative. In this context, sites and treatments are crossed with site being the random factor, treatment the fixed factor, and the site-by-treatment interaction also a random factor. The advantages of such a design over the two-factor nested design (sites nested within treatments) are well known. In the crossed design, site-by-treatment interactions are estimable, and, indeed, in Raffe's analysis, the varying effect of treatment across sites constituted the key focus of the inquiry. The main effect of treatment is estimated against within-site variation, making the analysis more powerful than that of a nested design when among-site variation is large. We note that meta-analyses of treatment-control studies are crossed designs of this
type. A treatment-control contrast is estimated in each study, and the variation in these "effect sizes" across studies is the focus of inquiry.

As in the two-factor nested design, the two-factor crossed analysis of variance represents a workable approach when the data are balanced and no covariates are employed. We first illustrate this approach on artificial data. However, even in this simple case, the hierarchical analysis offers extra information such as the correlation between the random effect of site and the site-by-treatment interaction. After replicating the results of the ANOVA via the hierarchical approach, we consider a generalization to include unbalanced data and multiple within-site covariates, both continuous and discrete.

13.5.1 Classical Approach

13.5.1.1 The Model

The mixed model for the two-factor crossed design may be written as

$$y_{ijk} = \mu + \alpha_i + \pi_j + (\alpha \pi)_{jk} + r_{ijk}, \quad \pi_j \sim N(0, \tau^2),
(\alpha \pi)_{jk} \sim N(0, \delta^2), \quad r_{ijk} \sim N(0, \sigma^2),$$

(13.21)

where $y_{ijk}$ is the outcome for subject $i$ nested within cell $jk$; $\mu$ is the grand mean; $\alpha_i$ is the effect associated with level $i$ of the fixed effect; $\pi_j$ is the effect associated with level $j$ of the random effect; $(\alpha \pi)_{jk}$ is the interaction effect associated with cell $jk$; and $r_{ijk}$ is the within-cell error ($i = 1, \ldots, n_j; j = 1, \ldots, J; k = 1, \ldots, K$). The random terms, $\pi_j$, $(\alpha \pi)_{jk}$, and $r_{ijk}$, are assumed mutually independent, each with a mean of zero and variances $\tau^2$, $\delta^2$, and $\sigma^2$, respectively. As before, in the case of balanced data ($n_{jk} = n$ for every level of the random factor), the standard analysis of variance method and restricted maximum likelihood coincide except when the ML estimate of either $\tau^2$ or $\delta^2$ is at or very near zero (Hartley and Rao 1967).

13.5.1.2 Example and Results

Table 13.8 lists artificial data created by the author. Let us assume that these data resulted from an experiment in which 40 subjects were assigned at random to one of 10 tutors and that, within tutors, subjects were then assigned at random to instruction without or with practice on a cognitive task. The outcome is the score on a test measuring proficiency on the task. The result is a 2-by-10 cross classification with 2 levels of the fixed effect of practice ($K = 2$) and 10 levels of tutor ($J = 10$). The interesting question is not so much whether there is a main effect of practice (the effect is huge) but whether the magnitude of that effect varies across tutors. Table 13.8 provides the source table for the ANOVA.
Table 13.8 Crossed Two-Factor Data and Source Table

<table>
<thead>
<tr>
<th>Tutor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = No</td>
<td>65</td>
<td>70</td>
<td>62</td>
<td>56</td>
<td>62</td>
<td>45</td>
<td>56</td>
<td>82</td>
<td>53</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>78</td>
<td>66</td>
<td>64</td>
<td>70</td>
<td>48</td>
<td>69</td>
<td>86</td>
<td>54</td>
<td>88</td>
</tr>
<tr>
<td>2 = Yes</td>
<td>140</td>
<td>159</td>
<td>163</td>
<td>139</td>
<td>127</td>
<td>141</td>
<td>130</td>
<td>139</td>
<td>128</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>163</td>
<td>181</td>
<td>142</td>
<td>138</td>
<td>146</td>
<td>138</td>
<td>144</td>
<td>130</td>
<td>165</td>
</tr>
</tbody>
</table>

ANOVA source table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>E(mean square)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice</td>
<td>1</td>
<td>63840.1</td>
<td>63840.1</td>
<td>$\sigma^2 + n\delta^2 + nJ\Sigma\alpha_i/(K - 1)$</td>
<td>298.74</td>
</tr>
<tr>
<td>Tutors</td>
<td>9</td>
<td>4325.0</td>
<td>480.6</td>
<td>$\sigma^2 + nK\tau^2$</td>
<td>14.14</td>
</tr>
<tr>
<td>Practice × tutors</td>
<td>9</td>
<td>1923.4</td>
<td>213.7</td>
<td>$\sigma^2 + n\delta^2$</td>
<td>6.28</td>
</tr>
<tr>
<td>Within cell</td>
<td>20</td>
<td>679.0</td>
<td>34.0</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>70767.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimated contrast between methods 1 and 2 is

$\hat{\alpha}_1 - \hat{\alpha}_2 = 53.12,$

and its significance is tested by the ratio of mean squares (practice to practice-by-tutors interaction), yielding $F(1, 9) = 298.74, p < .05.$ Variance estimates are given by equating the expected mean squares to their observed values as in the one-way ANOVA, yielding

$\sigma^2 = MS(\text{within-cell}) = 34.0,$

$\delta^2 = [MS(\text{practice} \times \text{tutors}) - MS(\text{within-cell})]/n = 89.6,$

$\tau^2 = [MS(\text{tutors}) - MS(\text{within-cell})]/(nK) = 111.65.$

Negative estimates of $\delta^2$ and $\tau^2$ would be set to 0. A test of the null hypothesis of no practice-by-tutors interaction, i.e., $H_0: \delta^2 = 0,$ is the $F$ statistic computed as the ratio of mean squares (practice-by-tutors interaction to within-cell), yielding, in this case, $F(9, 20) = 6.28, p < .01.$ A test of the null hypothesis of no tutor effect, i.e., $H_0: \tau^2 = 0,$ is given by the ratio of mean squares (tutors to within-cell), yielding $F(9, 20) = 14.14, p < .01.$
13.5.2 Analysis by Means of a Hierarchical Linear Model

13.5.2.1 The Model

Recall that, in the case of the two-factor (mixed) nested model, the fixed-effects contrasts were specified in the level-two model. In the two-factor (mixed) crossed model, the fixed factor is specified in the level-one (within-tutor) model, which, in general, is Eq. (13.1) with \((K - 1) X\)'s. In our case, with \(K = 2\), only one \(X\) is specified, so the level-one model becomes

\[
y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2),
\]  

(13.22)

where \(y_{ij}\) is the outcome for subject \(i\) having tutor \(j\), \(\beta_{0j}\) is the mean for the \(j\)th tutor, \(\beta_{1j}\) is the contrast between the practice and no-practice conditions within tutor \(j\), \(X_{1ij} = 1\) for subjects of tutor \(j\) having practice and \(-1\) for those having no practice, and \(r_{ij}\) is the within-cell error.

Notice that the level-one model defines two parameters that are allowed to vary across tutors: \(\beta_{0j}\), the tutor mean, and \(\beta_{1j}\), the treatment contrast. To replicate the results of the ANOVA, the level-two model is formulated to allow these to vary randomly across tutors:

\[
\beta_{0j} = \Theta_{00} + u_{0j} \quad \text{and} \quad \beta_{1j} = \Theta_{10} + u_{1j},
\]  

(13.23)

where \(\Theta_{00}\) is the grand mean, \(u_{0j}\) is the unique effect of tutor \(j\) on the mean level of the outcome, \(\Theta_{10}\) is the average value of the treatment contrast, and \(u_{1j}\) is the unique effect of tutor \(j\) on that contrast. The random effects \(u_{0j}\) and \(u_{1j}\) are assumed multivariate normal with variances \(\tau_{00}\) and \(\tau_{11}\), respectively, and covariance \(\tau_{01}\). Here the hierarchical model departs from the ANOVA model because the latter assumes that the tutor main effect and the tutor-by-practice interaction are independent. When the contrast coefficients sum to zero and the data are balanced, this covariance does not affect estimation or hypothesis testing regarding the fixed effects and variance components.

The correspondences between the hierarchical model and the ANOVA model are

\[
\begin{align*}
\Theta_{00} &= \mu, \quad \Theta_{10} = \frac{\alpha_2 - \alpha_1}{2}, \quad u_{0j} = \tau_{0j}, \\
u_{1j} &= \frac{(\alpha \beta)_2 - (\alpha \beta)_1}{2}, \quad \tau_{00} = \tau^2, \quad \tau_{11} = \delta^2/2.
\end{align*}
\]

Combining Eqs. (13.22) and (13.23) yields the single model

\[
y_{ij} = \Theta_{00} + \Theta_{10}X_{1ij} + u_{0j} + u_{1j}X_{1ij} + r_{ij}.
\]  

(13.24)
13.5.2.2 Results

The results of the analysis (Table 13.9) are again mathematically identical to the results based on the usual ANOVA. The reader will again notice that squaring the t statistic associated with the practice contrast will yield the corresponding F of the ANOVA in Table 13.8; and dividing each chi-square (for the tutor-effect variance and the tutor-by-practice variance) by its degrees of freedom will yield the corresponding F in the ANOVA table.

13.5.3 Generalization

We again turn to the analysis of the Thai primary school data, and our goal now is to examine the relationship between pre-primary school experience and total achievement. Within at least some of our 103 classrooms, there will be some children who have experienced pre-primary education and some children who have not. The proportion having pre-primary experience in the sample is .38. We may view pre-primary experience (yes, no) as crossed with classrooms, meaning that the design is a 103-by-2 crossed design. Of course, the proportion having pre-primary education will vary from class to class so that the design is unbalanced. Indeed, in quite a few classrooms there is no variation in pre-primary experience. To be meaningful, the estimate of the effect of pre-primary experience must be adjusted for other background variables. For example, past research has shown that Thai children of high social class are substantially more likely than Thai children of low social class to have the benefit of pre-primary experience (Raudenbush, Kidchanapanish, and Kang, 1991). Given

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, θ₀₀</td>
<td>106.25</td>
<td>3.47</td>
<td>—</td>
</tr>
<tr>
<td>Practice contrast, θ₁₀</td>
<td>39.95</td>
<td>2.31</td>
<td>17.28</td>
</tr>
</tbody>
</table>

**Variance components**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ₀₀</td>
<td>111.65</td>
<td>127.39</td>
<td>9</td>
</tr>
<tr>
<td>τ₁₁</td>
<td>44.94</td>
<td>56.65</td>
<td>9</td>
</tr>
<tr>
<td>σ²</td>
<td>33.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the unbalance of the design and the need to control for covariates measured both continuously and discretely, use of the mixed two-way ANOVA is inappropriate, and we have utilized a hierarchical linear model instead.

The level-one (within-classroom) model relates the outcome, $y_{ij}$, to the effect of pre-primary experience, adjusting for several covariates:

$$
y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + \beta_{2j}(\text{SES})_j + \beta_{3j}(\text{Time})_j + \beta_{4j}(\text{Sex})_j + \beta_{5j}(\text{Dialect})_j + \beta_{6j}(\text{Breakfast})_j + \epsilon_{ij},
$$

(13.25)

where the covariates SES, Time, Sex, Dialect, and Breakfast were already defined in Section 13.4, and the $X_{ij}$ is an indicator taking on a value of 1 if child $i$ in classroom $j$ has had pre-primary experience and 0 if not. Note that the key independent variable is $X_{ij} - \bar{X}_j$, representing the contrast between those having and those not having pre-primary experience within classroom, $j$. By centering the indicator around its classroom mean, we guarantee that student-level predictor is orthogonal to all school-level variables and therefore represents the within-classroom contrast.

At level two (between classrooms), only the intercept and the effect of pre-primary education were allowed to be random. Other covariates could have been specified to have random coefficients, but they were not; and classroom-level predictors could have been utilized in this analysis, but they were not. Thus, the level-two model is

$$
\begin{align*}
\beta_{0j} &= \Theta_{00} + u_{0j}, \\
\beta_{1j} &= \Theta_{10} + u_{1j}, \\
\beta_{qj} &= \Theta_{q0} \quad \text{for } q > 1,
\end{align*}
$$

(13.26)

where the random effects $u_{0j}$ and $u_{1j}$ are assumed multivariate normal as above.

The results (Table 13.10) indicate that, controlling for the covariates, children attending pre-primary school scored, on average, 10.4 points higher than did their classmates who did not attend pre-primary school, $t = 2.02$, $p < .05$. This represents a small effect of about 10% of a standard deviation overall and 16% of the within-classroom standard deviation. Results show no indication that the pre-primary education effect varies across classrooms, as indicated by a chi-square of 57.4 (to be compared to the percentiles of a central chi-square with 58 degrees of freedom) equivalent to an $F$ of 57.4/58 < 1.00. The results do suggest that significant mean differences remain across classrooms, even after controlling for the covariates, as indicated by a chi-square of 1206.8 with 58 degrees of freedom, $p < .001$. Note that the chi-squares are based on only 59 of 103 schools because only 59 schools had sufficient variation in pre-primary experience to esti-
Table 13.10 Two-Factor Nested Results for Thailand Classroom Data Based on the Hierarchical Linear Model

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Coeff</th>
<th>se</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>School/classroom level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-36.00</td>
<td>8.33</td>
<td>-</td>
</tr>
<tr>
<td>Student level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-primary education</td>
<td>10.38</td>
<td>5.12</td>
<td>2.02</td>
</tr>
<tr>
<td>SES</td>
<td>11.28</td>
<td>3.97</td>
<td>2.84</td>
</tr>
<tr>
<td>Time to school</td>
<td>-6.62</td>
<td>2.66</td>
<td>-2.48</td>
</tr>
<tr>
<td>Sex</td>
<td>-10.44</td>
<td>2.85</td>
<td>-3.66</td>
</tr>
<tr>
<td>Dialect</td>
<td>27.18</td>
<td>9.08</td>
<td>2.99</td>
</tr>
<tr>
<td>Breakfast</td>
<td>11.85</td>
<td>5.12</td>
<td>2.32</td>
</tr>
</tbody>
</table>

Variance components

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>Degrees of freedom</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\tau}_{00}$</td>
<td>3520</td>
<td>1206.8</td>
<td>58</td>
<td>.000</td>
</tr>
<tr>
<td>$\hat{\tau}_{11}$</td>
<td>236</td>
<td>57.4</td>
<td>58</td>
<td>.500</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>3997</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To estimate an ordinary least squares regression. In this situation, a better test of the significance for the interaction between classroom and pre-primary education is a likelihood ratio test.

To conduct this test, the analyst reruns the model with the variance of the pre-primary effect constrained to zero and then compares the goodness-of-fit statistics for the two models. Goodness of fit is measured by the deviance, that is, $-2$ times the log of likelihood function evaluated at the maximum. We did this, and the results were as follows. First, in the "unconstrained" model, the covariance component estimates were

$\hat{\tau}_{00} = 3520$,  
$\hat{\tau}_{11} = 236$,  
$\hat{\theta}_{01} = 9$,  
$\hat{\sigma}^2 = 3997$,  


yielding a deviance of 2278.2. Under the constrained model, two parameters, $\hat{r}_{11}$ and $\hat{r}_{01}$ were set to zero with the remainder of the model unchanged; and the variance components estimates were

$$\hat{r}_{00} = 3517,$$
$$\hat{\sigma}^2 = 4017,$$

yielding a deviance of 2782.4, meaning that the increase in deviance was only .2 when the model was simplified. This increase in deviance is approximately distributed as chi-square with degrees of freedom equal to the difference in the number of parameters estimated by the two models under the null hypothesis that the constrained model is adequate. Clearly, this difference between deviances is nonsignificant. Under the simpler model, the pre-primary contrast was unchanged at 10 points, but the $t$-test was a bit larger ($t = 2.33$ as opposed to 2.02), reflecting the greater precision of the simpler model.

13.6 RANDOMIZED BLOCK (AND REPEATED MEASURES) DESIGNS

Randomized block designs will typically involve mixed models having both fixed and random effects. Blocks (often subjects) will typically be viewed as having random levels, and, within blocks, there will commonly be a fixed-effects design. The fixed effects may represent experimental treatment levels, or, in longitudinal studies, they may involve polynomial trends. When the within-blocks design is identical for every block, and there are no missing data or within-block covariates, classical ANOVA procedures will often be appropriate. Under the same circumstances, multivariate ANOVA will allow more flexible assumptions regarding the variances and covariances of the within-block observations. However, when the within-blocks design varies, as (1) when the number or spacing of time series observations differs in a panel study, (2) when some blocks have missing data, or (3) when within-block covariates are present, these classical approaches are problematic. Under these circumstances, analysis by means of a hierarchical linear model offers a more flexible approach.

We first consider a simple randomized block design from the standpoint of the ANOVA and show how to formulate the hierarchical model to duplicate the results. We will then consider a generalization in which simplification of the model enables one to disentangle blocks-by-treatments variance from within-cell variance.
13.6.1 Randomized Block Design: Classical ANOVA Approach

13.6.1.1 The Model
We consider a randomized block design in which there are no between-blocks factors. This is a two-way crossed design (blocks crossed with a fixed-effects factor) identical to that studied in Section 13.5 except that there are no replications within cells. Each block is observed once and only once under each treatment or on each occasion.

The standard ANOVA model for the randomized block design may be written as

\[ y_{ij} = \mu + \alpha_i + \pi_j + e_{ij}, \]  

(13.27)

where \( y_{ij} \) is the outcome for block \( j \) under treatment \( i \) \((i = 1, \ldots, p; j = 1, \ldots, f)\), \( \alpha_i \) is the effect of treatment \( i \), \( \pi_j \) is the effect of block \( j \), and the error, \( e_{ij} \), has two components: \((\alpha \pi)_{ij}\) and \( r_{ij} \). That is, because there is only one observation per cell, the treatment-by-blocks interaction effect, \((\alpha \pi)_{ij}\), is confounded with the within-cell error, \( r_{ij} \). Typical assumptions of the model are that the random components are mutually independent and normally distributed:

\[ \pi_j \sim N(0, \tau^2), \quad e_{ij} \sim N(0, \sigma^2), \]

although, of course, the error \( \sigma^2 \) confounds variance attributable to block-by-treatment interaction and variance attributable to within-cell error. We initially assume that the blocks-by-treatment interaction effects are null.

13.6.1.2 Example and Results
The data in Table 13.11 are from Kirk (p. 244) and involve eight blocks, each observed under four treatment conditions. For ease of understanding, let us assume that the treatment is the duration of instruction prior to a cognitive test and that the outcome is the number correct on the test. We assume that 32 subjects have been classified into eight blocks of four each based on a pretest of cognitive ability. Within each block, subjects are then assigned at random to have 1, 2, 3, or 4 minutes of instruction prior to the test. We view blocks as random and treatments as fixed. The table shows the results, with the variation attributable to treatment effects explained by linear, quadratic, and cubic polynomial trend components.

The results indicate a highly statistically significant effect of duration, most of which is accounted for by the linear trend. Variance estimates are given by equating the expected mean squares to their observed values as
Table 13.11 Repeated Measures Data and Source Table (Additive Effects Model)

<table>
<thead>
<tr>
<th>Block</th>
<th>Treatment (duration)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

ANOVA source table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>$E(\text{mean square})$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>3</td>
<td>194.5</td>
<td>64.8</td>
<td>$\sigma^2 + n \sum \alpha_i^2/(p - 1)$</td>
<td>47.78</td>
</tr>
<tr>
<td>Linear</td>
<td>1</td>
<td>184.9</td>
<td>184.9</td>
<td></td>
<td>136.26</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1</td>
<td>8.0</td>
<td>8.0</td>
<td></td>
<td>5.90</td>
</tr>
<tr>
<td>Cubic</td>
<td>1</td>
<td>1.6</td>
<td>1.6</td>
<td></td>
<td>1.18</td>
</tr>
<tr>
<td>Blocks</td>
<td>7</td>
<td>12.5</td>
<td>1.8</td>
<td>$\sigma^2 + pr^2$</td>
<td>1.32</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>28.5</td>
<td>1.4</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>235.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

in the one-way ANOVA, yielding

\[
\delta^2 = MS(\text{error}) = 1.36,
\]

\[
\gamma^2 = [MS(\text{blocks}) - MS(\text{error})]/p = .11.
\]

A test of the null hypothesis of no block effects, i.e., $H_0: \gamma^2 = 0$, is given by the ratio of mean squares (blocks to error) yielding $F(7, 21) = 1.32$, implying that little evidence of block effects exists. This inference, of course, is fragile in light of the assumption needed to justify it—that no treatment-by-blocks interaction exists.

13.6.2 Analysis by Means of a Hierarchical Linear Model

13.6.2.1 The Model

Specification of the hierarchical linear model for the randomized block design is similar to specification for the two-factor crossed design discussed
in Section 13.5. The difference is that in the case of the randomized block design, there is no replication within cells. Hence, the model must be simplified.

According to the level-one (within-block) model, the outcome depends on polynomial trend components plus error:

\[ y_{ij} = \beta_0 + \beta_1 \text{(LIN)}_{ij} + \beta_2 \text{(QUAD)}_{ij} + \beta_3 \text{(CUBE)}_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2), \]

where \( y_{ij} \) is the outcome for subject \( i \) in block \( j \); \( \beta_0 \) is the mean for block \( j \); \( \text{(LIN)}_{ij} \) assigns the linear contrast values \((-1.5, -0.5, 0.5, 1.5)\) to durations \((1, 2, 3, 4)\), respectively; \( \text{(QUAD)}_{ij} \) assigns the quadratic contrast values \((-1.5, -0.5, -0.5, 0.5)\); \( \text{(CUBE)}_{ij} \) assigns the cubic contrast values \((-0.5, 1.5, -1.5, -0.5)\); \( \beta_{1j}, \beta_{2j}, \) and \( \beta_{3j} \) are the linear, quadratic, and cubic regression parameters, respectively; and \( r_{ij} \) is the within-cell error.

Notice that with only four observations per block and four regression coefficients \((\beta's)\) in the level-one model, no degrees of freedom remain to estimate within-cell error. If we assume, however, that the contrast values do not vary across blocks \(\) (no blocks-by-duration interactions), we can treat the trend parameters as fixed, yielding the level-two (between-blocks) model

\[ \beta_{0j} = \Theta_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau^2), \]

\[ \beta_{1j} = \Theta_{10}, \]

\[ \beta_{2j} = \Theta_{20}, \]

\[ \beta_{3j} = \Theta_{30}, \]

where \( \Theta_{00} \) is the grand mean and \( u_{0j} \) is the unique effect of block \( j \) assumed normally distributed with a mean of zero and a variance of \( \tau^2 \). The coefficients, \( \beta_{1j}, \beta_{2j}, \) and \( \beta_{3j} \), are constrained to be invariant across blocks.

The correspondences between the hierarchical model and the ANOVA model are

\[ \Theta_{00} = \mu, \]

\[ \Theta_{10} = (-1.5\alpha_1 - 0.5\alpha_2 + 0.5\alpha_3 + 1.5\alpha_4), \]

\[ \Theta_{20} = (0.5\alpha_1 - 0.5\alpha_2 - 0.5\alpha_3 + 0.5\alpha_4), \]

\[ \Theta_{30} = (-0.5\alpha_1 + 1.5\alpha_2 - 1.5\alpha_3 + 0.5\alpha_4), \]

\[ u_{0j} = \pi_j. \]

Combining Eqs. (13.28) and (13.29) yields the single model

\[ y_{ij} = \Theta_{00} + \Theta_{10} \text{(LIN)}_{ij} + \Theta_{20} \text{(QUAD)}_{ij} + \Theta_{30} \text{(CUBE)}_{ij} + u_{0j} + r_{ij}. \]
This model, like the ANOVA model above, assumes that the variance-covariance matrix of the repeated measures is compound symmetric: \( \text{Var}(Y_{ij}) = \tau^2 + \sigma^2; \text{Cov}(Y_{ij}, Y_{ik}) = \tau^2. \)

### 13.6.2.2 Results

The results of the analysis via the hierarchical model (Table 13.12) are again mathematically identical to the results based on the usual ANOVA. The HLM computer program allows the investigator to specify an omnibus test for the effect of duration which is the combined null hypothesis

\[ H_0: \Theta_{10} = \Theta_{20} = \Theta_{30} = 0. \]

The result is a large-sample chi-square test with 3 df, which, in this case, takes on a value of 143.33. Note that this value divided by 3 yields the omnibus \( F(3, 21) = 47.78 \) of the ANOVA. The reader will again notice that squaring the \( t \) statistic associated with each of the trend components yields the corresponding \( F \) in the ANOVA table (Table 13.11). Based on the hierarchical model, the \( t(\text{LIN}) = 11.67 \), when squared, yields the ANOVA \( F(\text{LIN}) = 136.26 \), and similarly, the \( t \) values for the quadratic and cubic trends can be converted into \( F \) values. The estimates of \( \sigma^2 \) and \( \tau^2 \) are also identical. Finally, the null hypothesis that \( \tau^2 = 0 \) is tested in the hierarchical analysis by means of the chi-square of 9.21, which, when divided by its degrees of freedom, yields the ANOVA \( F(3, 21) \) of 1.32.

### Table 13.12 Randomized Block Results Based on the Hierarchical Linear Model (Additive Effects Model)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, ( \Theta_{00} )</td>
<td>5.38</td>
<td>.236</td>
<td>—</td>
</tr>
<tr>
<td>Linear effect, ( \Theta_{10} )</td>
<td>2.15</td>
<td>.184</td>
<td>11.67</td>
</tr>
<tr>
<td>Quadratic effect, ( \Theta_{20} )</td>
<td>1.00</td>
<td>.412</td>
<td>2.43</td>
</tr>
<tr>
<td>Cubic effect, ( \Theta_{30} )</td>
<td>-0.20</td>
<td>.184</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau^2 )</td>
<td>.107</td>
<td>9.21</td>
<td>7</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>1.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
13.6.3 Generalization: A More Credible Model

The hierarchical results duplicate the ANOVA results because they are based on the same statistical assumptions, the most fragile of which is that block-by-treatment interaction is null. This assumption does not endanger the statistical basis of the tests of the fixed effects (the effects of duration and the associated trends) (Kirk, p. 249). However, if block-by-treatment interactions are large and falsely assumed zero, two consequences would be apparent. First, the generality of the inferences about duration would be endangered in that, if the effect of treatment varies across blocks, the effect is less general than if the effect is invariant across blocks. Second, the estimate of the variance of the block effects and the associated test could be wrong, perhaps misleading future researchers into believing that the basis upon which the blocks were formed was irrelevant to the outcome.

One feature of the results offers an opportunity to escape this impasse. We see in Table 13.12 that there is no evidence of a cubic trend in the data. If the cubic trend is dropped from the model, both the linear and quadratic trends can be allowed to vary randomly across blocks while still leaving a degree of freedom within each block to estimate the within-cell variance. In fact, the results of such an analysis (not reported here) indicated significant main effects of both the linear and quadratic trends. Although there was quite strong evidence that the linear trend varied significantly across blocks, there was no evidence that the quadratic trend varied across blocks. These results laid the basis for estimation of a more parsimonious and informative model, the results of which are displayed in Table 13.13. In this model, there is a random linear effect and a fixed quadratic effect, while the cubic effect is dropped entirely. Using this model, one can simultaneously account for between-block variance, block-by-treatment variance, and within-cell variance.

The level-one (within-blocks) model becomes

\[
y_{ij} = \beta_0 + \beta_y(LIN)_{ij} + \beta_y(QUAD)_{ij} + r_{ij},
\]

\[
r_{ij} \sim N(0, \sigma^2),
\]

which is the same model as Eq. (13.28) except that the cubic trend has been dropped. The level-two (between-blocks) model is now

\[
\beta_0 = \Theta_0 + \mu_0,
\]

\[
\beta_y = \Theta_y + \mu_y,
\]

and

\[
\beta_{2y} = \Theta_{2y}.
\]

Here the base and linear trend vary randomly around their average effects while the quadratic effect is constrained to be constant across blocks. Var-
Table 13.13  Randomized Block Results Based on the Hierarchical Linear Model (Linear-by-Blocks Interaction Model)

Fixed effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean, $\Theta_{00}$</td>
<td>5.38</td>
<td>.236</td>
<td></td>
</tr>
<tr>
<td>Linear effect, $\Theta_{10}$</td>
<td>2.15</td>
<td>.263</td>
<td>8.19</td>
</tr>
<tr>
<td>Quadratic effect, $\Theta_{20}$</td>
<td>1.00</td>
<td>.300</td>
<td>3.34</td>
</tr>
</tbody>
</table>

Variance-covariance components

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Chi-square</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{00}$</td>
<td>.268</td>
<td>17.41</td>
<td>7</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>.408</td>
<td>26.88</td>
<td>7</td>
</tr>
<tr>
<td>$\tau_{01}$</td>
<td>-.320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>.718</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

iances to be estimated include $\tau_{00}$, the variance of the block effects; $\tau_{11}$, the variance of the linear trends (block-by-treatments variance); and $\sigma^2$, the within-cell variance. We also estimate $\tau_{01}$, the covariance between $\mu_{0b}$, the block effect on the base, and $\mu_{1b}$, the block effect on the linear trend.

The results in Table 13.13 have several noteworthy features. First, we see a very strong negative covariance between the block mean and the linear effect: blocks with high means have small linear trends. The covariance $\tau_{01} = -.32$ is equivalent to a correlation of $-.97$, which could result from a ceiling effect on the outcome. Given this relationship, it is clear that the linear trend must be modeled as varying across blocks; indeed, a likelihood ratio test, comparing a model with a fixed linear trend to a model with a random linear trend, produced clear evidence in favor of the random linear trend.

Second, with the model respecified, the between-blocks variance is significantly greater than zero (note the chi-square of 17.4 with 7 degrees of freedom). In contrast, the compound-symmetry model (Table 13.12) seemed to indicate no significant between-block variation. However, that model was based on the assumption of no blocks-by-treatment interaction, an assumption strongly undermined by our reanalysis.

Third, the level-one random error variance, $\sigma^2$, is considerably smaller in the reanalysis than it had been in Table 13.12. Recall that this variance confounds within-cell variance and block-by-treatment variance. The removal of the block-by-linear variance allowed isolation of the within-cell variance.
Hierarchical Linear Models and Experimental Design

Other error structures may also be evaluated by means of hierarchical models. For example, Goldstein (1987) shows how to formulate and estimate a model in which the within-cell error is dropped and the linear, quadratic, and cubic effects along with the mean are assumed to vary randomly across blocks. This is really a fully multivariate model formulated via a hierarchical linear modeling approach.

13.7 A SCHEME FOR CLASSIFYING DESIGNS

This chapter has illustrated how two-level hierarchical linear models generalize analytic approaches to a number of common experimental designs. The hierarchical linear model itself is a specific case of a more general mixed linear model in that the hierarchical model applies only to random effects that are nested. Lindley and Smith (1972), Dempster, Rubin, and Tsutakawa (1981), and Goldstein (1987) have shown how a general class of crossed random effects models can be formulated and estimated. Kang (1991) has developed a general computing algorithm for such crossed random effects models. As these models become accessible, it becomes possible to develop a fairly complete taxonomy for generalizing analysis for experimental designs.

Table 13.14 Common Experimental Designs and Analytic Approaches That Generalize Their Applicability

<table>
<thead>
<tr>
<th>Experimental design</th>
<th>Generalized analytic approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way ANOVA</td>
<td></td>
</tr>
<tr>
<td>Fixed-effects model</td>
<td>ANOVA or multiple regression</td>
</tr>
<tr>
<td>Random-effects model</td>
<td>Two-level hierarchical model</td>
</tr>
<tr>
<td>Two-way ANOVA (within cell replications)</td>
<td></td>
</tr>
<tr>
<td>Both effects fixed</td>
<td>Multiple regression</td>
</tr>
<tr>
<td>One fixed, one random</td>
<td>Two-level hierarchical model</td>
</tr>
<tr>
<td>Both random</td>
<td>Crossed random effects model</td>
</tr>
<tr>
<td>Randomized block model (no within-cell replications)</td>
<td></td>
</tr>
<tr>
<td>Blocks random, treatments fixed</td>
<td>Two-level hierarchical model</td>
</tr>
<tr>
<td>Split-plot design</td>
<td></td>
</tr>
<tr>
<td>Blocks random, within- and between-block factors fixed</td>
<td>Two-level hierarchical model</td>
</tr>
<tr>
<td>Blocks random, within-block factor fixed,</td>
<td>Three-level hierarchical model</td>
</tr>
<tr>
<td>between-block factor random</td>
<td></td>
</tr>
<tr>
<td>Blocks random, within-block factor random,</td>
<td></td>
</tr>
<tr>
<td>random, between-block factor fixed</td>
<td>Crossed random effects model</td>
</tr>
</tbody>
</table>
Table 13.14 lists common experimental designs and, for each design, it identifies the statistical model that generalizes analytic possibilities. That is, the recommended approach (1) captures the relevant sources of variation and covariation, (2) allows for unbalanced data at any level, and (3) allows specification of both continuous and discrete predictors (including the possibility of both fixed and random regression coefficients for those predictors).

At present, reasonably user-friendly software is available for all the designs listed in Table 13.14 except for the crossed random effects models. Undoubtedly, that gap will soon be filled. The implication of these advances seems to be that the teaching of research designs linked to appropriate analytic models should soon be revolutionized. The conceptual strengths of current experimental and quasi-experimental design courses will be matched by the data analytic flexibility of a broad family of approaches integrated under the rubric of a general mixed linear model with crossed or nested random effects.

ACKNOWLEDGMENT

Research reported here was supported by grants from the John D. and Catherine T. MacArthur Foundation and the National Institute of Justice to the Program on Human Development and Criminal Behavior.

REFERENCES


Hierarchical Linear Models and Experimental Design


