Using Multilevel Models to Analyze Single-Case Design Data

David M. Rindskopf and John M. Ferron

Much of the discussion of statistical models for single-case designs (SCDs) has focused on the application of single-level models (e.g., regression and time-series models) that have been developed to analyze the interrupted time-series data from a single case (e.g., Glass, Willson, & Gottman, 1975; Huitema & McKeen, 2000; Maggin et al., 2011). In many single-case studies, however, there are actually multiple cases (Shadish & Sullivan, 2011), such as with multiple-baseline, replicated ABAB, or replicated alternating treatment designs (see Chapter 1, this volume). When time-series data from multiple cases are available, a separate single-level analysis could be made for each case, but it is also possible to examine all cases simultaneously using a multilevel model, where part of the model describes the behavior of each case and another part of the model describes the commonalities and differences among cases (e.g., Nugent, 1996; Shadish & Rindskopf, 2007; Shadish, Rindskopf, and Hedges, 2008; Van den Noortgate & Onghena, 2003a, 2007).

In doing so, multilevel modeling allows researchers to answer a wider range of research questions than single-level models. In addition to addressing questions about how large the treatment effect is for a particular case and how the treatment effect changes over time for that case, multilevel models also allow researchers to address questions about the average treatment effect and how that average effect changes over time. Furthermore, questions can be addressed about the degree to which the treatment effect varies across cases and whether this variation can be explained by characteristics of the cases. The ability to address a variety of specific questions can also be seen as an advantage of multilevel models over randomization tests (Chapter 5, this volume). Single-case randomization tests lead to general inferences about whether there was an intervention effect, but not to more specific inferences.
about the size of the effect, how the effect changes over time, and how the effect varies across persons.

A limitation of multilevel models is that to obtain these more specific inferences, a set of distributional assumptions needs to be made. For example, (a) is the variance the same in baseline and treatment phases? (b) Is the variance the same across participants? (c) Can a normal distribution be assumed? (d) Can the residuals within a case be considered independent or is a dependent error structure more appropriate? Ideally, researchers know the answers to questions like these a priori, but if not, they may turn to their data for guidance. Unfortunately, in many situations the data available are not sufficient to provide definitive answers to these types of questions. In such circumstances, researchers may use sensitivity analyses to estimate the model multiple times under alternative plausible sets of assumptions. If the inferences of interest remain substantially the same across the different plausible sets of assumptions, the conclusions are strengthened.

Note that multilevel models not only require more assumptions than randomization tests, but also require more assumptions than the single-level models. By making additional assumptions across cases, multilevel models not only extend the range of research questions that can be addressed, but they capitalize on similarities among the cases, which can lead to better inferences about the specific cases, particularly when the additional assumptions are accurate. Also of note, multilevel models have the flexibility to accommodate a variety of modeling challenges that may arise in single-case studies, such as (a) the need to account for outcomes that are counts or proportions, (b) the need to consider potential dependencies among errors (e.g., autocorrelation), and (c) the need to model linear or nonlinear trends. The recognition of the flexibility of these models and their compatibility with the research goals of many single-case researchers motivated the writing of this chapter.

Our focus is on analyses of data from primary studies, as opposed to meta-analyses, and thus we focus on two-level models, where observations are nested within cases, such as what would result from a multiple-baseline design, a replicated ABAB design, or a replicated alternating treatments design. Those interested in meta-analytic multilevel models for single-case data are referred to extensions of the two-level models to three levels, where observations are nested within cases and cases are nested within studies (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013; Owens & Ferron, 2012; Van den Noortgate & Onghena, 2008), and to multilevel analyses where effect-size measures are computed for each case and these effect sizes are then modeled as nested within studies (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012; Van den Noortgate & Onghena, 2003b, 2008).

In the course of this discussion we frequently refer to the effect size of a treatment. This reference typically means the treatment effect in terms of raw score units, or rates, proportions, or odds. These units allow researchers to represent results in terms that are easily understandable. A different perspective is needed when the results of many studies are combined; then the effects need to be expressed in standardized units so that studies using different outcomes can be appropriately compared. One such effect size for combining studies is $d$, the effect size discussed in other chapters of this book.
We begin our treatment of multilevel models with a relatively simple two-level model, which could be used for analyzing data from a multiple-baseline design, where the outcome is continuous and there are no trends in either the baseline or treatment phases. We then move through a series of increasingly more complex applications where we illustrate the inclusion of trends, accommodations for designs with more than two phases, extensions to adapt the model for counts (using the Poisson distribution) and proportions (using the binomial distribution), a Bayesian model to solve some of the problems of small sample size, and a method for examining gradual change between phases using a non-linear model. By considering a wide range of SCD applications and various levels of complexity, we hope to highlight both the advantages and limitations of multilevel models and, in doing so, facilitate the consideration of these models by SCD researchers. Those who would like to see examples illustrating statistical software commands and the resulting output are referred to Nagler, Rindskopf, and Shadish (2008).

**Model With No Time Effect and Two Phases**

Let us consider a simple example of a multiple-baseline design with several participants. Each person has a baseline phase and a treatment phase, begun at different time points to satisfy the replication requirement of this design. Suppose that during each phase there is no trend up or down over time, but that there is a jump going from one phase to another. We treat first the case of a continuous outcome that is normally distributed within each phase. Data consistent with this general scenario are presented graphically in Figure 7.1.

We begin with the model for each individual's behavior; for person \( j \) we let \( y_{ij} \) be the response at time \( i \). Similarly for each time point we let \( x_{ij} \) be the phase of person \( j \) at time \( i \), which will be a value of 0 for the baseline phase points, and 1 for treatment phase points. In other words, \( x_{ij} \) is a dummy variable representing phase. We let \( r_{ij} \) be the residual. The level-1 model is written

\[
y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + r_{ij}. \tag{7.1}
\]

During baseline phase, \( x_{ij} = 0 \), and so the expected response for the \( j^{th} \) person during baseline is \( \beta_{0j} \). This is illustrated graphically in Figure 7.1, where the baseline level for the first participant is 70 (\( \beta_{01} = 70 \)), for the second participant 60 (\( \beta_{02} = 60 \)), and for the third participant 65 (\( \beta_{03} = 65 \)). During treatment phase, \( x_{ij} = 1 \), so the expected response for the \( j^{th} \) person is \( \beta_{0j} + \beta_{1j} \), which means that \( \beta_{1j} \) is the jump between phases for the \( j^{th} \) person. This jump can be thought of as the person-specific treatment effect. In Figure 7.1 we see the treatment effects for the three participants are -50, -40, and -55, respectively (i.e., \( \beta_{11} = -50 \), \( \beta_{12} = -40 \), and \( \beta_{13} = -55 \)).

The residual \( r_{ij} \) is the difference between the observed value at the \( i^{th} \) point in time for the \( j^{th} \) person and what would be expected given the model. These residuals can be seen visually in Figure 7.1 as the vertical gaps between the points in the graph (observed values) and the lines (expected values based on
the model), i.e., \( r_{ij} = y_{ij} - (\beta_{0j} + \beta_{1j}x_{ij}) \). For example, the first observation for the first participant in Figure 7.1 is 68, but the expected baseline level is 70, and thus the residual for this observation is \(-2\). With a continuous outcome variable, like we are currently considering, the residuals are typically assumed to be sampled independently from a normal distribution with variance \( \sigma^2 \). If so, there is a single variance parameter (i.e., \( \sigma^2 \)) to be estimated for the level-1 model.

There are situations, however, where a more complex level-1 error structure may be warranted. For example, the treatment may be expected not only to shift the level of responding, but also to reduce the variance in responding, in which case the treatment phase variance would be different from the baseline variance. In other contexts the variance may be assumed to differ across participants. Multilevel models that group the level-1 residuals and provide

\[ \beta_{01} \]
\[ \beta_{11} \]
\[ \beta_{02} \]
\[ \beta_{12} \]
\[ \beta_{03} \]
\[ \beta_{13} \]

**Figure 7.1.** Graph of data for a hypothetical multiple-baseline study illustrating parameters in the multilevel model.
separate variance estimates for the different groups have been proposed and illustrated (Baek & Ferron, 2013).

Another type of complexity is encountered when researchers assume non-independent, as opposed to independent, residuals. Researchers may think, for example, that for each participant the residuals that are closer together in time are more likely to be similar than the residuals that are further apart in time, suggesting a nonindependent (or autocorrelated) error structure. In such circumstances, researchers using multilevel models can choose from a variety of possible error structures, including first-order autoregressive, Toeplitz, and banded Toeplitz structures (Baek & Ferron, 2013; Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009).

Just because it is possible to model the error structure as autocorrelated doesn’t mean that researchers should. There is no agreement among methodologists regarding whether residuals from behavioral time-series should be, or are, autocorrelated (Ferron, 2002; Huitema & McKeen, 1998; Kratochwill et al., 1974; Matyas & Greenwood, 1997; Shadish & Sullivan, 2011), and thus no consensus about how the level-1 error structure should be modeled. Furthermore, residuals that are independent can appear to be autocorrelated as a result of misspecifying another part of the model (e.g., modeling a nonlinear trend as linear), and conversely, residuals that are autocorrelated may appear to be independent because of imprecision in the estimation. Our advice is to think carefully about the case being studied, the behavior being observed, and the spacing between the observations, and to ask whether there is reason to believe that residuals should be independent or autocorrelated, and then to model accordingly. If it is unclear, researchers could consider a sensitivity analysis to assess the degree to which the primary inferences of the study (e.g., inferences about the size of the treatment effect) are sensitive to the assumptions made about the level-1 error structure. (In a sensitivity analysis, one changes a peripheral aspect of the analysis and sees how much the main results are affected. For example, one could make different assumptions about the value of the autocorrelation and see whether the effect size or standard errors are greatly affected.)

Now that we have considered specification of the level-1 model, we turn to the level-2 model that allows us to model the variation between participants. The values of \( \beta_{0j} \) and \( \beta_{1j} \) may be the same across participants, or they may vary. A simple level-2 model that allows for differences among participants, but does not attempt to explain that variation, is used as a starting point. There is one equation for each \( \beta \):

\[
\beta_{0j} = \gamma_{00} + u_{0j} \quad (7.2)
\]

\[
\beta_{1j} = \gamma_{10} + u_{1j}. \quad (7.3)
\]

The \( \gamma \) parameters are the average values. More specifically, \( \gamma_{00} \) is the average baseline level (referring to Figure 7.1 it would be the average of \( \beta_{01} \), \( \beta_{02} \), and \( \beta_{03} \)), and \( \gamma_{10} \) is the average treatment effect (referring to Figure 7.1 it would be the average of \( \beta_{11} \), \( \beta_{12} \), and \( \beta_{13} \)). Thus, the estimate of \( \gamma_{10} \) could be used to address...
questions regarding the size of the average treatment effect. Each equation in this level-2 model also has a residual term, where \( u_{0j} \) is the difference between the \( j \)th person’s baseline level \( (\beta_{0j}) \) and the average baseline level \( (\gamma_{00}) \), and \( u_{1j} \) is the difference between the \( j \)th person’s treatment effect \( (\beta_{1j}) \) and the average treatment effect \( (\gamma_{10}) \), and therefore the \( u_{ij} \) show that an individual might vary above or below that average. It is typically assumed that pairs of level-2 residuals \( (u_{0j}, u_{1j}) \) are sampled independently from a multivariate normal distribution. The \( u_{0j} \) have variance \( \tau_{00} \), the \( u_{1j} \) have variance \( \tau_{11} \), and their covariance is \( \tau_{01} \).

To some degree we are limited by the small sample size in explaining differences across participants, and may not be able to explain variation among them, but if the sample size is large enough, we could include an explanatory variable (either categorical or continuous) to explain variation in the \( \beta_{0j} \) or \( \beta_{1j} \).

Suppose we think that age is a factor that explains variation, and that the average participant is 10 years old. We center age by defining \( \text{Age10} = \text{Age} - 10 \), and add \( \text{Age10} \) as a predictor:

\[
\begin{align*}
\beta_{0j} &= \gamma_{00} + \gamma_{01}\text{Age10}_j + u_{0j}; \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\text{Age10}_j + u_{1j}.
\end{align*}
\]

In these equations, \( \gamma_{01} \) is the amount of change in the baseline level for each year of age, and \( \gamma_{11} \) is the change in treatment effect for each year of age.

**Fixed Versus Random Effects**

Unexplained variation (residuals) among participants (subjects) is represented by a random effect \( u_{0j} \) (with variance \( \tau_{00} \)) for intercepts and \( u_{1j} \) (with variance \( \tau_{11} \)) for slopes. Problems with this assumption include (a) the residuals may not be normally distributed, as is often assumed; (b) they may be not be well-estimated if there are only a few participants (say two or three); and (c) it may be questionable whether we consider the participants to be a random sample from any population. An alternative is to consider subjects as fixed; instead of fitting a multilevel model, we would fit an ordinary regression model to all the observations, and have variables (either dummy coded or effects coded) representing subjects. Further, we could have interactions of treatment phase with subjects to test whether the effect of treatment is the same across all participants.

Normally, using variables for subjects in this way would waste degrees of freedom and would suggest that multilevel models be used; however, with two or three participants it makes little difference and does not require the assumptions of normality that multilevel models usually implement. We can make inferences about the participants in a study, and whether each is similar to the others in the study. We lose some accuracy in doing this, as Bayesian methods make use of all the data without restricting the estimate to be the same for each person.

The basic two-level model that we have presented here has been studied using Monte Carlo simulation methods for a variety of multiple-baseline study
conditions (e.g., four cases with 20 observations per case). The results of these simulation studies suggest that when the model is correctly specified, (a) the estimates of the average treatment effect (and other $\gamma$ coefficients) are unbiased and that the corresponding inferences are accurate when either the Satterthwaite or Kenward-Roger method (Kenward & Roger, 1997) is used to estimate degrees of freedom, but (b) the restricted maximum likelihood estimates of the variance parameters tend to be biased (Ferron et al., 2009). In addition, the empirical Bayes estimates of the individual effects ($\beta_{ij}$) are biased as expected, but the interval estimates for these individual effects are accurate, assuming that the model is correctly specified and the Kenward-Roger approach is used for making the inferences (Ferron, Farmer, & Owens, 2010).

**Adding a Time Trend**

Adding a time trend is straightforward (statistically) if it is a constant time trend, that is, it does not change across phases. As will be seen, there are conceptual difficulties with including trend, but in terms of the equations it just means adding a time trend to the equation for individuals:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \beta_{2j}t_{ij} + r_{ij}. \quad (7.6)$$

In this equation $t_{ij}$ represents the time (generally session number) for the observation at time $i$ for person $j$. Consequently, $\beta_{2j}$ represents the slope of the growth trajectory for person $j$, or the expected change in the outcome for each unit increase in time. The common slope across phases is illustrated in Figure 7.2, which provides a graphical display and visual representation of each of the regression coefficients in Equation 7.6. Because the baseline and treatment slopes are assumed to be equal, the vertical gap between the two parallel lines in Figure 7.2 stays constant over time. This gap represents the treatment effect, or shift in behavior that occurred with treatment, and is indexed by $\beta_{1j}$.

Finally, $\beta_{0j}$ is the expected response when both the treatment variable and the time variable are equal to zero. In Figure 7.2, time is coded such that 0 corresponds to the first baseline observation and the treatment variable is dummy coded so that baseline observations are coded 0. Consequently, $\beta_{0j}$ is the expected response at the start of the study. If the researcher wanted $\beta_{0j}$ to represent the baseline measure at the end of baseline period, we have to scale time separately for each individual by subtracting the time at which the phase changes. For example, if time is measured in number of sessions, and the first person has 12 baseline sessions before changing to treatment, then we must calculate $t_{12} = t - 12$, so that $t_{12}$ will be zero when $t = 12$. Of course, in a multiple-baseline design every person will have the phase change at a different session, and will need a different constant subtracted to produce the right transform of time.

Most SCDs will not include a constant time trend, because a baseline phase should continue until a trend is no longer present. However, it is possible for a flat baseline to be followed by a trend during treatment phase, or a trend in
baseline to be followed by a different (or no) trend in the treatment phase. Therefore, allowing for a change in trend would be necessary under those conditions. This assessment can be accomplished by adding an interaction term to the model that allows the time effect to depend on treatment. More specifically, we would create a new variable representing the product of the dummy-coded variable representing phase ($x_{ij}$) and a centered version of the time variable ($t_{ij} - k_j$):

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \beta_{2j} (t_{ij} - k_j) + \beta_{3j} x_{ij} (t_{ij} - k_j) + r_{ij}. \quad (7.7)$$

A visual representation of the parameters in this model is provided in Figure 7.3. For this illustration we coded the variables as follows: $t_{ij}$ is coded 0 at the beginning of the study, and $k_j$ corresponds to the value of $t_{ij}$ for the first observation in the treatment phase for person $j$, so that $t_{ij} - k_j = 0$ for the first treatment observation. With this coding, $\beta_{0j}$ is the projected baseline value for person $j$ at the time point that coincides with the first treatment observation, $\beta_{1j}$ is the expected immediate shift in behavior for person $j$ (i.e., the vertical gap between the extension of the baseline trajectory and the treatment phase trajectory at the time of the first treatment observation), $\beta_{2j}$ is the slope during baseline for person $j$, and $\beta_{3j}$ is the change in slope that occurs with treatment for person $j$.

A trend will not continue forever and so it is an oversimplification to use this model; we will see that nonlinear models sometimes offer a better solution. If we do use this oversimplified model, we have to consider what measure of treatment effect we want. In our illustration we have two parameters related to the treatment effect: $\beta_{1j}$ is the immediate treatment effect, and $\beta_{3j}$ provides information on how the treatment effect changes with time. In Figure 7.3 we see that the gap between the baseline extension and the treatment phase trajectory increases with time, indicating that the effect of the treatment is greater at the end of the study then it was one observation into treatment. At what point in time should we measure the effect of treatment? Should we have measured the effect at the end of all treatment sessions as opposed to measuring it after a
single treatment session? Perhaps in many cases, five sessions into treatment would be reasonable, but in all cases some judgment is required on the part of the researcher. Whatever the choice, it is implemented by centering time at the point decided on; for example, if the phase change is at Session 7 for person \( j \) and we want to go 5 sessions beyond to measure treatment effect, then we code \( t_{ij} - k_j \) as \( t_{ij} - 12 \).

Changing time effects within a phase involves problems, both statistical and conceptual. In general, it is difficult to think of a trend that would continue for any length of time; generally one expects behavior to level off. Also, many measures are limited by zero at the lower end, and frequently have upper limits as well. For a short time period, behavior may show a linear trend, but we cannot project that linear behavior very far into the future.

### Other Designs

With a more complicated design, handling the coding of time becomes more complicated also. The most common design besides the multiple-baseline is ABAB, which we will write as A1 B1 A2 B2 in order to track the change from the first AB sequence to the second. The most obvious way to code phases in this design is to have one dummy variable for baseline versus treatment, another dummy variable for first sequence versus second sequence, and the product of these for the interaction. If there is no trend, or if trend is constant across phases, then this works well; if there is varying trend, then this coding is problematic. Some options are considered later.

Another common SCD is the alternating treatment design. In this design, treatments are changed frequently, perhaps chosen by a coin flip for each session. There are no long phases, and so one cannot model change in trend within phases. The simplest trend would be an overall trend line, or in more complicated cases perhaps a quadratic (degree two polynomial) to allow a curve over time. A model with linear trend is analogous to an analysis of covariance; the lines for the different treatments are parallel, so the difference between the lines
(effect of treatment) is a constant. If the lines are not parallel, the analyst must decide at what time point to measure the treatment effect. For example, if there are 15 time points, and the desired assessment of treatment effect is the end of the study, compute $\text{sess15} = \text{session} - 15$, and use sess15 instead of session as the time variable. More complicated cases would require treatment-by-trend interactions to allow the lines to have different slopes for different treatments.

**Issues About Coding Phases of Treatment**

With ABAB (and more complicated) designs the most obvious way to code phases is to have a main effect for treatment (A vs. B), a main effect for whether the first AB pair is observed or second AB pair, and an interaction. These effects are traditionally coded using effects coding ($1/0/0$ coding in general, and $1/1/0$ with only two conditions).

The coding for this design would be

$$
\begin{bmatrix}
\mu_{A1} \\
\mu_{B1} \\
\mu_{A2} \\
\mu_{B2}
\end{bmatrix} =
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}
$$

where the four columns of the model matrix represent the intercept, the treatment main effect, the main effect of first versus second set of phases, and the interaction of treatment and phase set.

Such coding can be useful, but other methods should be considered as well. First, suppose that we want to represent the difference between the two baselines, and the difference between the two treatment effects. To represent the coding, we will use matrix notation first, then expand that to individual equations to better see the interpretation. The coding is written in matrix form as

$$
\begin{bmatrix}
\mu_{A1} \\
\mu_{B1} \\
\mu_{A2} \\
\mu_{B2}
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 1 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}
$$

The separate equations are

$$
\begin{align*}
\mu_{A1} &= \beta_0, \\
\mu_{B1} &= \beta_0 + \beta_1, \\
\mu_{A2} &= \beta_0 + \beta_2, \\
\mu_{B2} &= \beta_0 + \beta_1 + \beta_2 + \beta_3.
\end{align*}
$$
We recognize this coding as the usual main effects and interaction, but using dummy coding rather than effects coding. \( \beta_0 \) is the first baseline average; \( \beta_1 \) the jump at the introduction of the first treatment; \( \beta_2 \) is the difference between A1 and A2 (the two baseline phases); and through rearrangement and substitution of terms \( \beta_3 \) can be seen to be the difference between the first treatment effect (B1 – A1) and the second treatment effect (B2 – A2):

\[
(\mu_{B2} - \mu_{A2}) - (\mu_{B1} - \mu_{A1}) = \\
\mu_{A1} - \mu_{B1} - \mu_{A2} + \mu_{B2} = \\
\beta_0 - (\beta_0 + \beta_1) - (\beta_0 + \beta_2) + (\beta_0 + \beta_1 + \beta_2 + \beta_3) = \\
\beta_3.
\]

As this example illustrates, changing from one type of coding to another changes the interpretation of the parameters. For this reason, one should always proceed from the effects wanted to the coding; by first specifying what the parameters should mean, one can write the coding to represent them correctly (see, e.g., Serlin & Levin, 1985).

Next consider coding that would represent differences between successive phases. For successive changes, the coding is

\[
\begin{bmatrix}
\mu_{A1} \\
\mu_{B1} \\
\mu_{A2} \\
\mu_{B2}
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}.
\]

(7.14)

When written as separate equations, this becomes

\[
\mu_{A1} = \beta_0, \quad (7.15)
\]

\[
\mu_{B1} = \beta_0 + \beta_1, \quad (7.16)
\]

\[
\mu_{A2} = \beta_0 + \beta_1 + \beta_2, \text{ and} \quad (7.17)
\]

\[
\mu_{B2} = \beta_0 + \beta_1 + \beta_2 + \beta_3. \quad (7.18)
\]

From this representation, it is clear that \( \beta_0 \) represents the level during Phase A1; \( \beta_1 \) is the change in going from Phase A1 to Phase B1; \( \beta_2 \) the change from B1 to A2; and \( \beta_3 \) the change from A2 to B2.

For alternating treatment designs with three or more conditions, there is a choice of coding methods. For example, if there is a baseline and two treatment conditions (B, T1, and T2), one might code two dummy variables, one for T1 compared to baseline and the other for T2 compared to baseline. Another possibility would be to have one dummy variable for treatment versus baseline (with T1 and T2 coded 1, baseline coded 0), and an effect-coded term comparing
treatment 1 with treatment 2 (with B coded 0, T1 coded 1/2, and T2 coded −1/2). This would result in the model

\[ \begin{align*}
\mu_B &= \beta_0 \\
\mu_{T1} &= \beta_0 + \beta_1 + 0.5\beta_2, \\
\mu_{T2} &= \beta_0 + \beta_1 - 0.5\beta_2
\end{align*} \] (7.19)

where it is easy to see that \( \beta_0 = \mu_B \), and adding the second and third equations and dividing by two, and subtracting equation one, one sees that \( \beta_1 = \frac{\mu_{T1} + \mu_{T2}}{2} - \mu_B \) and subtracting equation three from equation two results in \( \beta_2 = \mu_{T1} - \mu_{T2} \).

**Issues in Coding Subjects as Fixed Effects**

With few participants, we are faced with a choice of using multilevel models (in which case variations among subjects are random effects), or models in which subjects are fixed effects. In the latter case we must code the subjects to reflect differences among them. Suppose there are three subjects; we could use dummy coding (0/1 coding) to represent subjects (e.g., let S1 = 1 if Subject 1, 0 if Subject 2 or 3). If an intercept is included in the model, we would use only two of the three dummy variables, say S2 and S3, letting the intercept represent S1 (and S2 and S3 the difference between Subjects 2 and 3 and Subject 1).

As seen in an example here, it sometimes makes more sense to use effects (1/0/−1) coding, even though it is not as transparent in interpretation. For example, let the three subjects be coded as follows:

\[
\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & -1 & -1
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2
\end{bmatrix}.
\] (7.20)

Written out as separate equations, this is

\[ \begin{align*}
\mu_1 &= \beta_0 + \beta_1 \\
\mu_2 &= \beta_0 + \beta_2, \\
\mu_3 &= \beta_0 - \beta_1 - \beta_2
\end{align*} \] (7.21)

Adding the three equations together and dividing by 3, we see that \( \beta_0 \) is the average for the three subjects. Taking the first equation and subtracting \( \beta_0 \) from both sides, we see that \( \beta_1 \) is the amount by which Subject 1 differs from the average of all subjects; similarly \( \beta_2 \) is the amount by which Subject 2 differs from the overall average. This coding is particularly useful when we want to compare the random effects multilevel model to the comparable fixed effects model.
Dependent Variables That Are Counts

A frequent outcome variable in SCD research is counts, either over a time period or out of some total number of trials. Count outcomes are often disguised as rates or proportions and treated as continuous. Under some circumstances, treating counts (or transforms of counts) as continuous does not create severe problems, but other times it results in major errors in analysis. First, the measures should not result in floor or ceiling effects. For example, counts that cluster around zero, or percentages near zero or 100 percent, result in nonlinearity and heterogeneity of variance. Counts can display heterogeneity of variance if small counts have small variance and large counts have larger variance. Because there are so many ways in which counts can create problems, it is best to treat counts correctly rather than assume normality and analyze them as continuous measures.

First let us consider the case of counts for a period of time, for example, the number of times that a child hits another child during the period of observation. We suppose for this example that the time of observation is the same for each period and the same for each child, but if it is not, then a simple adjustment can be made to this method. Consider again the multiple-baseline design with no effect of time (zero slope). The model is now expressed in terms of the natural logarithm of the counts, $F_{ij}$, for individual $i$ observed in session $j$:

$$\ln(F_{ij}) = \beta_0 + \beta_1 x_{ij}, \quad (7.22)$$

whereas before $x_{ij}$ indicates whether this child is in the baseline or treatment phase. There is no residual in this equation because $F_{ij}$ is the expected frequency, and the observed count $y_{ij}$ has a Poisson distribution with mean $F_{ij}$.

The parameters are now interpreted in the logarithmic scale; to translate back into the original scale we must exponentiate. For example, if $\beta_0 = 3$ for child $j$, and $\beta_1 = -2$, then the average baseline count for the child would be $\exp(3) = 20.09$, and the average during treatment phase would be $\exp(3-2) = \exp(1) = 2.72$. Another way to express this is to note that

$$F_{ij} = \exp(\beta_0 + \beta_1 x_{ij})$$

$$= \exp(\beta_0) \exp(\beta_1 x_{ij})$$

$$= \exp(3) \exp(-2)$$

$$= (20.09)(0.13533) = 2.718$$

In other words, the original behavior was about 20 hits per session, which was reduced by treatment to .135 of its original value, becoming a rate of under three hits per session.

For data in the form of percentages or proportions, the modeling is of the raw counts. One must go back to the form of actual counts (e.g., 6 events out of 10 trials, not 60 %). After all, 60 percent could be 3 out of 5, or 300 out of 500,
and one is much less precisely estimated than the other. The statistical model for each individual is expressed in the form

$$
\pi_{ij} = \frac{F_{ij}}{n_{ij}} = \ln \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_{0j} + \beta_{1j} x_{ij},
$$

where is $n_{ij}$ the number of trials, and $F_{ij}$ is the number of trials on which the behavior of interest was observed. As with continuous outcomes, we can add a time trend, or we can extend the model to handle an ABAB design.

Another type of dependent variable closely related to count data is ordinal data. Perhaps each respondent is rated on a scale from 0 to 3, with 0 being no problems and 3 being major problems. This type of data is easily analyzed by common multilevel programs; the assumption is that a continuous normally distributed unobserved variable underlies the observed categorical variable. It is an extension of the logistic regression model, for which each possible cut of “low” and “high” is used as a dependent variable. For example, the possible divisions for a 0 to 3 scale would be {0} vs. {1, 2, 3}, {0, 1} vs. {2, 3}, and {0, 1, 2} vs. {3}. A logistic regression is conducted for each possible division; the usual restriction made is that all of the logistic regressions have the same slopes, but different intercepts. The model is usually written as

$$
\ln \left( \frac{\pi_{low}}{1 - \pi_{low}} \right) = \beta_{0j} - \eta,
$$

where $\eta$ is a linear combination of the predictor variables, and where “low” indicates being at or below category $j$ of the outcome variable. Note that $\eta$ is subtracted instead of added, because the model is (usually) expressed in terms of responding in a low category rather than a higher category, which is usually an undesirable response (in the example below, a low grade rather than a higher grade). A positive value of $\eta$ means a lower logit, and thus a lower probability of a low grade (a higher probability of a higher grade).

An example comes from Hall, Cristler, Cranston, and Tucker (1970), who studied the effect of treatment (after-school tutoring for students getting a D or F on daily assignments) on quiz score in French assignments. Quizzes were graded with the usual grades of A to F that were coded for data analysis as $A = 4$, $B = 3$, $C = 2$, $D = 1$, and $F = 0$. The study involved a multiple-baseline design with three participants, and each participant was observed for 25 days. The data and code for the analyses using the statistical package R are presented in Appendix 7.1. The data are plotted in Figure 7.4.

A visual analysis of the data in Figure 7.4 leads to the conclusion that there is a functional relationship between the intervention and grades. The treatment effect is demonstrated at three separate points in time, first for Dave, then for Roy, and finally for Deb. The data patterns are similar for the three participants. During the baseline phase each participant primarily has grades of D and F with an occasional higher grade and no trend toward improvement.
The implementation of intervention is followed immediately by a shift to higher grades, with no intervention phase grades below C for any of the participants. Multilevel modeling provides a complement to the visual analysis by allowing us to quantify the size of the intervention effect and to index the uncertainty in our effect inferences.

We first fit two simple models to these data; one is a random effects model, the other a fixed effects model. Both assume that the effect of treatment is the same for all participants. The random effects model assumes that the intercepts may differ across the participants, so that their baselines may differ. The results are presented in Table 7.1.

The threshold coefficients are the intercepts for different dichotomizations of the response variable. The treatment effect is 5.7 on the logit scale, which is $\exp(5.7) = 297$; this is the change in the odds of having a higher versus lower grade (for any dividing point), a very large treatment effect. For example, in the baseline period the “average” participant would have had an odds of $\exp(1.97) = 7.17$ of having a grade of D or F (i.e., 1 | 2 = grade division of (0, 1) vs. (2, 3, 4)); this corresponds to a probability of $\exp(1.97)/(1 + \exp(1.97)) = 7.17/8.17 = .88$. During treatment, that changes to a logit of $1.97 - 5.695 = -3.725$, an odds of $\exp(-3.725) = .024$, and a probability of $0.024/1.024 = .024$ of

getting a D or F. From the complementary perspective, during baseline students had a probability of .12 of getting a C or higher; during treatment this increased to a probability of .976.

The variance of the intercepts among participants is .202, corresponding to a standard deviation of .45. If we take seriously the idea that participants might have been randomly selected from some population, then about 95 percent would be within .90 logits (i.e., 2 standard deviations) of the average logit.

The fixed effects version of this model uses effects coding for the respondents (1/0/-1); this version is to ensure that the intercepts (thresholds) represent the average person. The results for the model with main effects only for subjects are presented in Table 7.2. Note that the treatment effect and intercepts have nearly the same estimates, standard errors, and \( t \) values as in the random effects model, and so the interpretation is the same. The effect for \( e1 \) is Participant 1, who is greater (than the average participant) by about .81, meaning a greater probability of a lower grade during both baseline and treatment. (Remember that these effects are subtracted, changing the sign, so it will enter the equation as \(-.81\)). To determine the effect for Participant 3, note that the effects must sum to zero, so the effect is \(-(.81-.05) = -.76\). To get the standard error, we would have to change the coding to include Participant 3; to get an approximate standard error, we could assume that it would be about the same as for the other participants’ effect-coding standard errors, or about .33.

As a test of whether the effect of treatment is the same for all three participants, we will fit the fixed effects model with treatment by subject interactions. The results are presented in Table 7.3. The quantity labeled “Residual Deviance” can be thought of as measure of goodness of fit (more properly, badness of fit: big is worse). By itself, it cannot be interpreted intuitively, but if we subtract the deviance of a model that is less restricted from the deviance from a model

---

**Table 7.1. Random Effects Analysis of French Quiz Grade Data**

| Random effects: |
|-----------------|-----------|
|                 | Var | SD   |
| id              | 0.202 | 0.4494 |

| Coefficients: |
|-----------------|----------|
|                 | Estimate | SE | \( z \) value | \( Pr(>|z|) \) |
| trt             | 5.695    | 1.129 | 5.042 | .0001 |

| Threshold coefficients: |
|-----------------|----------|
|                 | Estimate | SE | \( z \) value |
| 0|11 | 0.5743    | 0.4205 | 1.366 |
| 1|2 | 1.9717    | 0.5472 | 3.603 |
| 2|3 | 4.9382    | 1.1247 | 4.391 |
| 3|4 | 6.2675    | 1.1747 | 5.335 |
that is more restrictive, the difference has a chi-square distribution with degrees of freedom equal to the number of parameters dropped in the more restricted model. Comparing this model with the model that omits interaction terms gives a result of $146.52 - 140.14 = 6.38$, which just exceeds the critical value of 5.99 with two degrees of freedom. This overall test tells us that the treatment effects differ among respondents. Examining the two effects individually, we see that

Table 7.2. Fixed Effects Results for French Quiz Data

<table>
<thead>
<tr>
<th>Coefficients:</th>
<th>Value</th>
<th>SE</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>5.62426</td>
<td>1.1298</td>
<td>4.9783</td>
</tr>
<tr>
<td>e1</td>
<td>0.80655</td>
<td>1.129</td>
<td>2.3988</td>
</tr>
<tr>
<td>e2</td>
<td>-0.05175</td>
<td>0.3244</td>
<td>-0.1595</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intercepts:</th>
<th>Value</th>
<th>SE</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.4992</td>
<td>0.3334</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1.9126</td>
<td>0.4841</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.9467</td>
<td>1.1023</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6.3646</td>
<td>1.1556</td>
</tr>
</tbody>
</table>

Residual Deviance: 146.5151
AIC: 160.5151

Table 7.3. Fixed Effects Model With Treatment \times Subject Interactions for French Quiz Data

<table>
<thead>
<tr>
<th>Coefficients:</th>
<th>Value</th>
<th>SE</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>5.87684</td>
<td>1.1526</td>
<td>5.0989</td>
</tr>
<tr>
<td>e1</td>
<td>-0.06885</td>
<td>0.5196</td>
<td>-0.1325</td>
</tr>
<tr>
<td>e2</td>
<td>0.35371</td>
<td>0.4490</td>
<td>0.7878</td>
</tr>
<tr>
<td>trt:e1</td>
<td>1.80586</td>
<td>0.7517</td>
<td>2.4024</td>
</tr>
<tr>
<td>trt:e2</td>
<td>-0.49750</td>
<td>0.6819</td>
<td>-0.7296</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intercepts:</th>
<th>Value</th>
<th>SE</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.6256</td>
<td>0.3367</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2.0184</td>
<td>0.4876</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5.3118</td>
<td>1.1602</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7.0531</td>
<td>1.2588</td>
</tr>
</tbody>
</table>

Residual Deviance: 140.1422
AIC: 158.1422
Participant t 1 (Deb) is clearly different from the average in her treatment effect. Of course, going from an average that is already so big does not have much of a practical effect.

**Bayesian Models**

Bayesian ideas provide many benefits, both theoretical and practical. One important aspect in which Bayesian methods improve prediction is in estimating parameters with small amounts of data but when there is auxiliary evidence to aid prediction. For example, sometimes similar quantities must be estimated for all participants, and although each has a separate estimate, the estimates for all participants might be similar. The effect of treatment on each individual is one such quantity. We can “borrow strength” from the other participants when estimating each participant’s value of the parameter. In classical statistics we would have to choose between using each participant’s data by themselves (with the small sample size) or pooling the data to get one estimate for all participants (losing individual estimates). This issue was referred to above in the discussion of fixed versus random effects.

Another way in which Bayesian estimation is an improvement is by its taking into account all sources of uncertainty in estimation. In most estimation procedures for multilevel models, some quantities are estimated first (usually variances), and then other quantities (typically the fixed effects) are estimated conditional on the variances. This process assumes that the variances are estimated well, which is not so with small sample sizes (number of participants or cases). Autocorrelation is another quantity that is typically estimated and then assumed known in the next stage of estimation, even though autocorrelations are usually poorly estimated in SCDs. Bayesian methods do not require these assumptions.

One does not have to know a lot about Bayesian theory to use Bayesian methods. Two aspects are crucial: prior distributions and posterior distributions. The likelihood function that is used in classical statistics also plays a central role in Bayesian methods. The prior distribution (prior) of a parameter summarizes the analyst’s beliefs, before seeing any data, about possible values of the parameter. We can make the prior noninformative, meaning that we have no idea about likely values of the parameter, to make the analysis as objective (and similar to the classical analysis) as possible. (In practice, we must make the priors nearly, but not totally, noninformative; this has no practical effect on the model.)

The posterior distribution (posterior) combines information about the data (expressed through the likelihood) with information from the prior; if the prior is (relatively) noninformative, the posterior resembles the likelihood. The difference is the interpretation. Because the prior and posterior summarize one’s beliefs about a parameter (or several parameters), they are probability distributions and can be interpreted as such. For example, one can calculate the probability that the parameter is greater than zero, or the probability that one treatment is more effective than another, and a 95% interval (called a credible interval in Bayesian terminology) can be interpreted as “there is a 95% probability that the parameter is in the interval.”
One practical benefit of Bayesian modeling using variants of the BUGS program (WinBUGS [Lunn, Thomas, Best, & Spiegelhalter, 2000]; OpenBUGS [Lunn, Spiegelhalter, Thomas, & Best, 2009]; JAGS [Plummer, 2003]) is that one can fit a wide variety of models that are not standard (see, e.g., Chapter 8, this volume). This option includes the nonlinear model with floor and ceiling effects that are described in the next section. It is often simple to write a special model that is appropriate for the situation but is not available in standard statistical software.

For categorical data, Bayesian models have the advantage of being able to deal with extreme results (if somewhat informative, rather than uninformative, priors are used). Extreme results (e.g., all zero counts during baseline) can lead to infinite parameter estimates in several kinds of models; a slightly informative prior can modify this to a finite number. This condition not only affects the estimate of the baseline, but also the estimate of the treatment effect, which is the difference between the two phases.

Another practical benefit of these software programs is their ability to keep track of derived quantities, and not only get an estimate of them, but also a standard error (and for that matter to investigate their entire distribution, which may not be normal). As an example, if one wanted to estimate the final level in a multiple-baseline design, one could track the values of $\beta_0 + \beta_1$ as separate quantities. Other examples are given in the following section on nonlinear models.

**Nonlinear Models**

Seldom do data immediately jump to a new steady level when the phase changes. If only one or two sessions are required for the change, we can adapt by leaving those data points out of the analysis. If this is too wasteful of data, or if the change is very gradual, we must consider different models. For example, when the outcome is binomial, as in a previous section, we might expect the data to start at one proportion, then gradually shift and settle down at a new proportion. The flat lines that the proportions settle down to are not likely to be zero and one, so a typical logistic regression model will not work. Instead the lower level (the floor) and the higher level (the ceiling) will have to be estimated in the model. Analysts who are familiar with item response theory (IRT) will recognize the floor effect; it is analogous to the guessing parameter in a three-parameter IRT model. The ceiling effect is similar, but at the top.

Consider the data plotted in Figure 7.5; they are based on data from Horner et al. (2005). The figure shows data from three participants in a multiple-baseline design. A visual analysis of the data in Figure 7.5 leads to the conclusion that there is a functional relationship between the intervention and the proportion of correct trials. The data patterns for the three participants are similar. Each baseline shows a relatively low proportion of correct trials, little variation, and no trend toward improvement. Once the intervention is implemented there is a change in the trend, with each participant showing gradual increases in the proportion of correct trials. Furthermore, the upward trend occurs first with Participant A, second with Participant B, and third with Participant C, mirroring the order that
the intervention was implemented. Multilevel modeling allows us to quantify the size of the intervention effects and also to quantify the uncertainty in our inferences. For example, we would like to be able to index the uncertainty in the effect estimates and the uncertainty in the inference that the order the participants changed matches the order that the intervention was implemented.

To build the model, we start with the usual way of writing the dependent variable:

\[ L_{ij} = \beta_{0ij} + \beta_{1ij} X_{ij} \]  

(7.25)

(In the usual logistic regression model, \( L \) would be the logit, or log-odds, and it would be that here if it were not for floor and ceiling effects.)

The equation for individual behavior can be expressed as

\[ \pi_{ij} = f_j + (c_j - f_j) \exp(L_{ij}) / \{1 + \exp(L_{ij})\}, \]  

(7.26)

where \( f_j \) is the floor, \( c_j \) is the ceiling, and \( \pi_{ij} \) is the probability of observing the behavior of interest.

This way of writing the model (Equation 7.25) would be fine if we were not interested in the place on the curve where the response occurs. As in IRT, we are interested in the time for which the response has gone halfway between the floor and ceiling. In IRT this is the item difficulty; here we will use \( H \) for “halfway point.” The halfway point occurs at the value of \( X \) that makes \( L = 0 \); solving for \( X \) gives \( X = -\beta_{0ij} / \beta_{1ij} \). We would like to rewrite the equation so that this is a parameter of the model. Rewriting Equation 7.25 gives

\[ L_{ij} = \beta_{1ij} \{ (\beta_{0ij}/\beta_{1ij}) + X_{ij} \} = \beta_{1ij} (X_{ij} - H_j) \]  

(7.27)

Again, this is similar to the form of IRT models. Notice that \( H \) is important for determining whether the change occurs at the right time, but that \( c - f \) is important because it indicates the size of the effect (change from baseline level to treatment condition level.)

Table 7.4 contains the results of fitting the model with floor and ceiling effects to the Horner et al. data. The ceilings are all around .8 or so, and the floors around .2, with little variation across respondents in either. The treatment effect is the rise from the floor to the ceiling; these effects vary from .57 to .72, again showing similar and large values. An important indicator that change was due to treatment is that the halfway points are in the right places and in the right order. For the first participant the halfway point is between Sessions 8 and 9; for the second, about Session 12; and for the third between Sessions 18 and 19; these are all consistent with when the phase changed. Further, the 95% credible intervals (similar to the confidence interval in classical inference) are fairly narrow, so there is little possibility that the ordering is wrong. This outcome can be directly checked in Bayesian models; the quantities \( g_{t21}, g_{t31}, \) and \( g_{t32} \) are based on the number of times in the 25,000 simulations that the
halfway point for Participant 2 is greater than for Participant 1, and so on. The proportion of times is calculated (and rounded to two decimal places) as 1.00 in all cases, showing near certainty that these participants responded in the right order.

**Final Perspectives**

The use of statistical methods (in general) and multilevel models (in particular) can offer SCD researchers many benefits. They can make valid probability statements about their results and generate quantitative estimates of the effects of their treatment. They can capture the important qualities of their results in these models. They can account for influences that would be difficult or impossible without statistics, such as autocorrelation or the ability to predict the performance of respondents over time (and help determine how bad such prediction might be). They can test whether trends are really in the data, or whether they are illusions.

Statistical methods have drawbacks, but generally these drawbacks are as bad or worse for visual-analysis methods. Phases can begin with points that are nonrepresentative, due to a sudden change in conditions. Unmeasured influ-

**Table 7.4.** Results of Fitting a Nonlinear Model With Floor and Ceiling to Horner et al.

<table>
<thead>
<tr>
<th>Node</th>
<th>$M$</th>
<th>$SD$</th>
<th>MC error</th>
<th>2.5%</th>
<th>Mdn</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ceiling[1]</td>
<td>0.8322</td>
<td>0.04401</td>
<td>0.002131</td>
<td>0.7457</td>
<td>0.8315</td>
<td>0.9203</td>
</tr>
<tr>
<td>ceiling[2]</td>
<td>0.874</td>
<td>0.0464</td>
<td>0.002514</td>
<td>0.7828</td>
<td>0.874</td>
<td>0.9637</td>
</tr>
<tr>
<td>ceiling[3]</td>
<td>0.7941</td>
<td>0.05725</td>
<td>0.002796</td>
<td>0.6893</td>
<td>0.7919</td>
<td>0.9103</td>
</tr>
<tr>
<td>floor[1]</td>
<td>0.168</td>
<td>0.03486</td>
<td>6.979E-4</td>
<td>0.1055</td>
<td>0.1657</td>
<td>0.2412</td>
</tr>
<tr>
<td>floor[2]</td>
<td>0.1566</td>
<td>0.02644</td>
<td>4.972E-4</td>
<td>0.1075</td>
<td>0.1553</td>
<td>0.2119</td>
</tr>
<tr>
<td>floor[3]</td>
<td>0.2199</td>
<td>0.024</td>
<td>3.586E-4</td>
<td>0.1748</td>
<td>0.2193</td>
<td>0.2683</td>
</tr>
<tr>
<td>effect[1]</td>
<td>0.6642</td>
<td>0.05588</td>
<td>0.002501</td>
<td>0.552</td>
<td>0.6649</td>
<td>0.7726</td>
</tr>
<tr>
<td>effect[2]</td>
<td>0.7175</td>
<td>0.05381</td>
<td>0.002845</td>
<td>0.6097</td>
<td>0.7187</td>
<td>0.8194</td>
</tr>
<tr>
<td>effect[3]</td>
<td>0.5743</td>
<td>0.0616</td>
<td>0.002982</td>
<td>0.4586</td>
<td>0.5722</td>
<td>0.6991</td>
</tr>
<tr>
<td>half[1]</td>
<td>8.547</td>
<td>0.6156</td>
<td>0.01462</td>
<td>7.325</td>
<td>8.529</td>
<td>9.769</td>
</tr>
<tr>
<td>half[2]</td>
<td>12.29</td>
<td>0.5014</td>
<td>0.01175</td>
<td>11.33</td>
<td>12.3</td>
<td>13.31</td>
</tr>
<tr>
<td>half[3]</td>
<td>18.59</td>
<td>0.7197</td>
<td>0.02454</td>
<td>17.22</td>
<td>18.56</td>
<td>20.05</td>
</tr>
<tr>
<td>gt21</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0E-12</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>gt31</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0E-12</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>gt32</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0E-12</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>mu0</td>
<td>-19.28</td>
<td>5.7</td>
<td>0.3198</td>
<td>-29.05</td>
<td>-19.63</td>
<td>-8.429</td>
</tr>
<tr>
<td>mu1</td>
<td>0.8945</td>
<td>0.4117</td>
<td>0.03188</td>
<td>0.4128</td>
<td>0.7776</td>
<td>1.815</td>
</tr>
<tr>
<td>mu2</td>
<td>8.01</td>
<td>6.122</td>
<td>0.5549</td>
<td>2.542</td>
<td>9.472</td>
<td>18.74</td>
</tr>
<tr>
<td>sd0</td>
<td>5.7</td>
<td>8.523</td>
<td>0.4003</td>
<td>0.0663</td>
<td>3.774</td>
<td>24.79</td>
</tr>
<tr>
<td>sd1</td>
<td>0.2293</td>
<td>0.4204</td>
<td>0.01807</td>
<td>0.02791</td>
<td>0.1259</td>
<td>1.024</td>
</tr>
<tr>
<td>sd2</td>
<td>2.315</td>
<td>4.388</td>
<td>0.2346</td>
<td>0.03385</td>
<td>0.9758</td>
<td>11.95</td>
</tr>
</tbody>
</table>

*Note.* Data from “The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education,” by R. H. Horner et al., 2005, *Exceptional Children*, 2. Copyright 2005 by the Council of Exceptional Children.
ences can occur (such as illnesses not noted, or events that make the participants more or less active generally). If these influences are known or suspected, they can be modeled. But in the end, statistical models are ways to assess what we do and do not know, as well as the precision with which we have estimated important quantities. Such capabilities are beyond the scope of most visual analyses.

**Appendix 7.1: Data and Code for the Analyses Using the Statistical Package R**

**R Program for Ordered Data**

```r
# SCD ordered categorical data

# quiz score grades, 0=F, 4=A

dave <- c(0,0,1,2,0,1,0,0,0, 4,3,3,4,4,4,3,4,4,4,3,4,4)
roy <- c(0,2,0,0,4,1,0,0,0,1,1,2,0,0, 3,3,3,2,2,2,2,3,4)
deb <- c(1,0,0,0,0,0,0,1,1,0,2,0,0,0,1,1,0,0,0,2,2,2,2,3)

trt1 <- c(rep(0,9),rep(1,16))
trt2 <- c(rep(0,15),rep(1,10))
trt3 <- c(rep(0,20),rep(1,5))

id <- rep(1:3,each=25)
sess <- rep(1:25,3)

s2 <- as.numeric(id==2)
s3 <- as.numeric(id==3)

y <- c(dave,roy,deb)
trt <- c(trt1,trt2,trt3)

data <- c(id,y,sess,trt,s2,s3)
data.m <- matrix(data,ncol=6,byrow=F)
data.fr <- data.frame(data.m)
attach(data.fr)

# Fixed effects models

library(polr)
summary(polr(as.ordered(y) ~ trt + s2 + s3))

summary(polr(as.ordered(y) ~ trt*s2 + trt*s3))

# Effects Coding
```

1ST PAGES
\[ e1 = (id == 1) - (id == 3) \]
\[ e2 = (id == 2) - (id == 3) \]

\[
\text{summary(polr(as.ordered(y) - trt + e1 + e2))} \\
\text{summary(polr(as.ordered(y) - trt*e1 + trt*e2))}
\]

# Random effects model

library(ordinal) 
library(ucminf) 

model.2 <- clmm(as.ordered(y) ~ trt + (1 | id)) 
summary(model.2)

**WinBUGS Program for Nonlinear Model**

```
model {
  for (i in 1:61) {
    n.right[i] <- p.right[i]/5 # change percent to number 
    n.right[i] ~ dbin(prob[i], 20)
    logit(yhat[i]) <- b0[person[i]] + b1[person[i]] * sess.10[i] +
    b2[person[i]] * phase[i]
    prob[i] <- floor[person[i]] + (ceiling[person[i]] - floor[person[i]]) * yhat[i]
    n.right[i] ~ dbin(prob[i], 20)
  }

  for (j in 1:3) {
    b0[j] ~ dnorm(mu0, prec0) 
    b1[j] ~ dnorm(mu1, prec1) 
    b2[j] ~ dnorm(mu2, prec2) 
    floor[j] ~ dbeta(2,1) 
    ceiling[j] ~ dbeta(1,2) 
    effect[j] <- ceiling[j] - floor[j] 
    half[j] <- 1*(b0[j] + b2[j])/b1[j] 
    mu0 ~ dnorm(.1,.001) 
    mu1 ~ dnorm(-.1,.001) 
    mu2 ~ dnorm(.02,.001) 
    prec0 ~ dgamma(.001,.001) 
    prec1 ~ dgamma(.001,.001) 
    prec2 ~ dgamma(.001,.001)
    sd0 <- 1/sqrt(prec0) 
    sd1 <- 1/sqrt(prec1) 
    sd2 <- 1/sqrt(prec2)
```
# how often is halfway point higher in person 2 than 1, etc.?
gt21 <- step(half[2] - half[1])
gt32 <- step(half[3] - half[2])
gt31 <- step(half[3] - half[1])

References


