Psychological Methods
2013, Vol. 18, No. 3, 385–405

1082-989X/13/$12.00 DOI: 10.1037/a0032964

Analyzing Data From Single-Case Designs Using Multilevel Models: New Applications and Some Agenda Items for Future Research

William R. Shadish
University of California, Merced

Eden Nagler Kyse
Montclair State University

David M. Rindskopf
City University of New York

Several authors have proposed the use of multilevel models to analyze data from single-case designs. This article extends that work in 2 ways. First, examples are given of how to estimate these models when the single-case designs have features that have not been considered by past authors. These include the use of polynomial coefficients to model nonlinear change, the modeling of counts (Poisson distributed) or proportions (binomially distributed) as outcomes, the use of 2 different ways of modeling treatment effects in ABAB designs, and applications of these models to alternating treatment and changing criterion designs. Second, issues that arise when multilevel models are used for the analysis of single-case designs are discussed; such issues can form part of an agenda for future research on this topic. These include statistical power and assumptions, applications to more complex single-case designs, the role of exploratory data analyses, extensions to other kinds of outcome variables and sampling distributions, and other statistical programs that can be used to do such analyses.

Keywords: single-case designs, multilevel model, hierarchical linear model, error covariance structure, error distributions

Single-case designs (SCDs) are widely used to assess the effects of interventions in such areas of research as special education, developmental disabilities, certain kinds of behavioral disorders, instructional strategies aimed at improving the performance of individual students (Shadish & Sullivan, 2011), and medicine (Gabler, Duan, Vohra, & Kravitz, 2011). Many researchers see SCDs as yielding credible and strong causal inferences that ought to contribute to discussions of effective practices and policies (Shadish, Cook, & Campbell, 2002). Some agencies, such as the What Works Clearinghouse funded by the U.S. Department of Education, have promulgated standards that SCDs must meet in order to contribute to discussions of evidence-based practice (Kratochwill et al., 2010). However, one of the unresolved issues in those standards and in the SCD literature more generally is the most appropriate kind of data analysis for SCDs. Some SCD researchers prefer visual analysis to statistical analysis (Kromrey & Foster-Johnson, 1996; Michael, 1974; Olive & Smith, 2005; Parsonson & Baer, 1978, 1992). Although visual analysis has its merits, it can be unreliable and does not allow for quantification of effects (DeProspero & Cohen, 1979). Among the statistical approaches that have been proposed are various effect size estimators, randomization tests, and regression analysis (Houle, 2008; Kratochwill & Levin, 2010; Maggin et al., 2011; Parker, Vannest, & Davis, 2011; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008). None has yet gained wide consensus as the best way to analyze SCD data.

The use of multilevel models for analyzing data from single-case designs is a relatively recent development and one that has been advanced by only a few authors. Van den Noortgate and Onghena (2003a) used SAS PROC MIXED to show how multilevel models are applied to a four-phase (ABAB) design. They discussed one way of coding the ABAB design so that the analysis can estimate heterogeneity of effects over participants, yield empirical Bayes shrunken estimates, and model various kinds of error structures in a two-level model nesting time (i.e., observations) within cases (i.e., participants). Van den Noortgate and Onghena (2003b) applied multilevel modeling to the meta-analysis of regression coefficients from a series of two-phase SCDs; they showed how multilevel models analyze both linear trend and trend by treatment interactions. Van den Noortgate and Onghena (2007) showed how multilevel models can be applied to three-phase (ABC) designs where a baseline phase is followed by a first treatment and then a second treatment. Van den Noortgate and Onghena (2008) extended this to a three-level model in which time

This article was published Online First July 8, 2013.

William R. Shadish, School of Social Sciences, Humanities, and Arts, University of California, Merced; Eden Nagler Kyse, Center for Research and Evaluation on Education and Human Services, Montclair State University; David M. Rindskopf, Department of Educational Psychology and Psychology, The Graduate Center, City University of New York.

This research was supported in part by Grants R305D100046 and R305D100033 from the Institute for Educational Sciences, U.S. Department of Education, and by a grant from the University of California Office of the President to the University of California Educational Evaluation Consortium.

Correspondence concerning this article should be addressed to William R. Shadish, School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 North Lake Road, Merced, CA 95343. E-mail: wshadish@ucmerced.edu

is nested within cases and cases are nested within studies. They also showed how to include covariates and how one might combine results from SCDs with those from between-groups experiments using multilevel models. Jenson, Clark, Kircher, and Kristjansson (2007) and Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009) presented simulations about the use of multilevel models in ABAB and multiple baseline studies, respectively. Gage, Lewis, and Stichter (2012) applied multilevel modeling to the meta-analysis of SCDs testing the effects of functional behavioral assessment-based interventions.

Despite this excellent work, research has only begun to address all the issues that arise when using multilevel models to analyze SCDs. Our purposes in this article are to present new applications of this analytic approach and then to discuss an array of issues that should be studied further before this approach can live up to its potential. The article proceeds in three parts. First, we review the general multilevel modeling framework as it applies to SCDs. This includes the modeling of trend in Level 1 equations, modeling random coefficients from Level 1 in Level 2 equations, adding covariates in Level 2, and hypothesis testing procedures. Second, to extend past applications, we show how multilevel models can be used to address issues not previously covered in that work, including modeling nonlinearities in trend, modeling diverse kinds of outcomes such as counts and proportions, exploring the notion of overdispersion in Poisson and binomial models, demonstrating new ways of coding phase effects in ABAB designs, and showing how to analyze alternating treatment and changing criterion designs. We use the HLM statistical program (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) to analyze these examples. This provides an alternative to SAS PROC MIXED, which was used by the previous work on multilevel models. We sequence the examples from simple to more complex in order to provide a gradual introduction to multilevel modeling for the SCD researcher. The simple models would rarely be considered best practice in SCD research (Kratochwill et al., 2010), but lessons learned from them can be incorporated into the analysis of best practice designs quite easily. Third, we discuss a host of issues that researchers should consider when using multilevel modeling to analyze SCDs, including statistical power, hypothesis testing versus exploratory analyses, statistical assumptions, other outcome metrics, applications to more complex SCDs, and other statistical software packages that can do multilevel modeling with SCDs.

The General Multilevel Framework

Throughout this article we use the notation of Raudenbush and Bryk (2002). SCD data are characterized by multiple time points being nested within cases, typically with multiple cases present within each study. At the most basic level, SCDs can be represented by two equations. The Level 1 equation models time points within cases; a simple example is

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + e_{ti} \quad (1)$$

where Y_{ti} is the observation at measurement occasion t for person i ($i = 1, \ldots, n$), a_{ti} is the time or age at measurement occasion t when the observation was taken for person i, π_{0i} is the person's expected observation at $a_{ti} = 0$, π_{1i} is the rate of change per unit time for person i, and the errors e_{ti}. by default are usually are independent and normally distributed with common variance σ^2. However, other error covariance matrices can be specified in principle, such as a lag-1 error autoregressive model (AR1); in practice, the ability to estimate such models will depend on having sufficient data and will vary somewhat over programs. We discuss this more in the Discussion section.

In every analysis some thought must be given to the scaling of the time variable, so that $a_{ti} = 0$ occurs at a time for which the researcher wants to assess an individual's status. For a one-phase study, as will be illustrated in our first example, it is often useful to have $a_{ti} = 0$ at the end of the study. In other studies, one wants $a_{ti} = 0$ at a change of phases from baseline to treatment. The scaling of time is a special case of centering in multilevel models. Such centering can be done in several ways, and this has implications for the interpretation of both the means and the random effects. For example, one could scale time so that time 0 is either the start of baseline or the end of baseline. In the former case the intercept π_{0i} is the predicted value on the outcome measure at the start of baseline, and in the latter it is the predicted value at the end of baseline. This will, in turn, affect the interpretation of the variances and covariances of the intercept and slope. For instance, the covariance between status at the start of baseline and growth rate may be different than that covariance at the end of baseline. Raudenbush and Bryk (2002, especially pp. 181–198) have provided details and examples. Centering also reduces possible collinearity among predictors (Cheng et al., 2010).

One can extend the Level 1 model in several ways. An example is to incorporate polynomial terms to reflect nonlinear trends in the data:

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + \pi_{2i}a_{ti}^2 + \ \ldots \ + \pi_{Pi}a_{ti}^P + e_{ti} \quad (2)$$

One might also add Level 1 predictors, often called time-varying covariates in the context of time series models (McCoach & Kaniskan, 2010; Raudenbush & Bryk, 2002). These are predictors with values that can change over time, such as whether a person was sick or not on the day a particular observation was taken. In SCDs, treatment is a time-varying covariate because the case receives treatment at some times and not at others. So, additional dummy variables for treatment phase (e.g., 0 = baseline, 1 = treatment) and the treatment by time interaction will be in the Level 1 model. However, even though the covariate changes over time, the coefficient for its effect is a constant over time.

Cases may differ in the value of the dependent variable at time $a_{ti} = 0$, in their rate of change over time, or in both. The Level 2 equations model that variability across cases in the parameters (the πs) of the Level 1 equation. For example, in the simplest case the Level 2 equations for the Level 1 parameters in Equation 1 are

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \end{aligned} \quad (3)$$

Here, β_{00} and β_{10} are fixed effects intercepts, representing the average observation at measurement occasion $t = 0$ over all persons (β_{00}) and the average rate of change in observations over time over all persons (β_{10}), r_{0i} and r_{1i} are random effects, assumed to be normally distributed with a mean of zero, that allow each case to vary from the grand mean randomly. The latter have variances of τ_{00} and τ_{11}, respectively, and covariances of $\tau_{10} = \tau_{01}$, where the subscripts indicate the row and column of the

variance–covariance matrix of the Level 2 errors. Errors in the Level 1 model are assumed to be independent of errors in the Level 2 model.

More generally the Level 2 equations can have predictors at the case level, such as fixed (i.e., not time-varying) person characteristics like gender or height:

$$\pi_{0i} = \beta_{00} + \sum_{q=1}^{Q_0} \beta_{0q} X_{qi} + r_{0i}$$
$$\pi_{1i} = \beta_{10} + \sum_{q=1}^{Q_1} \beta_{1q} X_{qi} + r_{1i} \quad (4)$$

Here, q indexes the covariates from $q = 1$ to Q, and X is a vector of Q covariates that may or may not be the same for predicting π_{0i} (Q_0) and π_{1i} (Q_1). The covariates are frequently identical over all the Level 2 equations, but this need not be the case. We provide examples of both. Note that, in this more general form, β_{00} and β_{10} are not, in general, averages over cases; their interpretation is relative to cases with values of 0 on all X_{qi}.

Null hypothesis tests in HLM for all fixed effects are done using a t ratio that takes the form of $t = \pi_{Pi} / \sqrt{\hat{V}_{\pi_{Pi}}}$, where the denominator is the square root of the sampling variance of the numerator (Raudenbush & Bryk, 2002). Degrees of freedom use the between/within method, where they are a function of the number of cases and predictors. For the Level 1 model, $df = N - Q - 1$ where N is the number of cases and Q is the number of Level 2 predictors (not counting the intercept, which is accounted for by the 1 in the df equation). Tests for the random effects are done using a χ^2 test with the same $df = N - Q - 1$. However, the true values of the variance component parameters under the null hypothesis are sometimes at the boundary of the parameter space; for example, when one tests whether a variance component differs from zero but the observed variance is very close to zero. In that case, the χ^2 test under the null hypothesis does not follow a χ^2 distribution (Cheng, Edwards, Maldonado-Molino, Komro, & Muller, 2010). Rather, it follows a mixture of the χ^2 distributions for the models with and without the parameter, each weighted .50. By default, the HLM computer package generates these mixture chi-squares for this test (Raudenbush & Bryk, 2002; West, Welch, & Galecki, 2007).

Readers should approach the use of multilevel models to analyze SCDs with caution, particularly about two issues: error covariance structures and power. With regard to the former, for example, Gurka, Edwards, and Muller (2011) showed that the results of multilevel models in general can be sensitive to the choice of error covariance structure and that an underspecified structure can seriously distort standard errors. We approach this matter by estimating random effects for intercept, time, treatment, and the interaction of time and treatment, along with the covariances between those random effects. We do not estimate an autoregressive model in addition to or instead of our approach. We outline the issues in this choice in the Discussion. With regard to power, Muller, Edwards, Simpson, and Taylor (2007) showed that standard mixed model tests often have inflated Type I error rates in small samples, which would seem to characterize SCDs. So although the analyses we present would seem to suggest high power to detect effects, it may be that this really reflects high Type I error rates. Again, we comment more in the Discussion.

Finally, in testing these models, the researcher must clearly distinguish between the main statistical hypothesis and any subsequent exploratory analyses. We do so in the examples that follow this section by focusing nearly entirely on testing a clearly identified main hypothesis. Although we have done subsequent exploratory analyses, reported elsewhere, to illustrate the possibilities (Nagler, Rindskopf, & Shadish, 2008), we comment on them only occasionally. The main reason is that SCD data sets often have small numbers of time points and cases that can reduce power. If so, nonsignificant findings for the main statistical hypothesis may tempt the researcher further explore the data for significant results by model respecification. We do not wish to encourage that, especially if such exploratory analyses were not clearly identified. We elaborate on this issue in the Discussion. Next, however, we present examples of the application of multilevel models to SCDs.

Modeling Quadratic Trend and Normally Distributed Data

The first example illustrates how the analysis of SCDs can be done with data that are plausibly normally distributed. Such data are rare in SCD work in psychology and education (Shadish & Sullivan, 2011) but are more common in N-of-1 trials in medicine (Gabler et al., 2011). This example also shows how nonlinear trends can be handled in multilevel modeling; past authors have looked only at linear trends. Stuart (1967) trained eight obese females in self-control techniques to overcome overeating behaviors. Patients were weighed monthly throughout the 12-month program, and these data were graphed individually, as shown in Figure 1. To conduct analyses on these (and subsequent) data, we digitized the graphs using procedures described in detail elsewhere (Nagler et al., 2008; Shadish et al., 2009). Each line in Figure 1 represents the weight loss trend of one patient in the study over time. The graphs suggest that weight loss trends may not be uniform across patients (i.e., the lines are not quite parallel) and that the line of best fit may not be straight but rather might require a quadratic term to account for slight curvature. A good multilevel analysis should, therefore, assess the possibility of a quadratic trend by including such a term in the model. This is less an exploratory analysis than a diagnostic analysis, for the incorrect

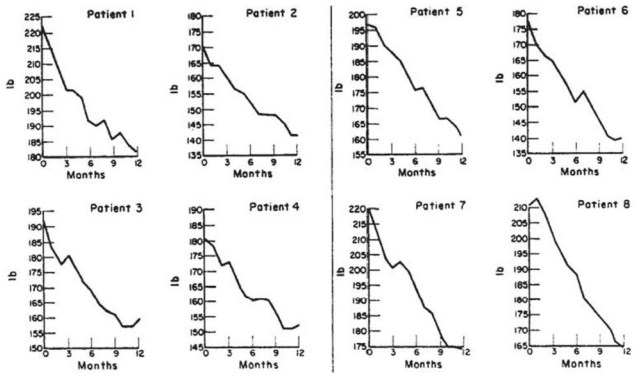

Figure 1. Patient weight loss during a yearlong behavioral treatment for overeating. Adapted from "Behavioral Control of Overeating," by R. B. Stuart, 1967, *Behavior Research and Therapy, 5,* p. 364. Copyright 1967 by Elsevier.

modeling of the functional form of the trend can lead both to bias and to inefficiency in the coefficients in the model.

If, as Figure 1 suggests, the rate of weight loss slows over time, we can model that by including a quadratic transformation of time in the Level 1 equation:

$$\hat{Y}_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + \pi_{2i}a_{ti}^2 \quad (5)$$

In this example, Y_{ti} is the observed weight at time t, a_{ti} is the time when the observation was taken at time t for person i (coded -12, -11, . . . 0), π_{1i} is the rate of weight change per month for person i at the end of the study (i.e., at time $a_{ti} = 0$), π_{2i} is related to the rate of change of slope, a_{ti}^2 is the square of time, and the errors e_{ti} are independent and normally distributed with common variance σ^2. The assumption of a normal distribution makes sense for weight but may make less sense for other kinds of outcomes (e.g., counts, proportions) that we illustrate in subsequent examples. Measurement occasion was scaled as ($t = -12, -11, \ldots 0$) so that the final weight π_{0i} is at the end of treatment for person i ($i = 1, \ldots, n$). In this and subsequent equations where the left side of the equation is a predicted outcome, the error term is omitted because it is represented in the assumption about the distribution of the observed outcome (normal in this case but binomial or Poisson in some later cases).

The Level 2 model is the following:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \end{aligned} \quad (6)$$

That is, each effect from the Level 1 equation is treated as randomly varying over cases, and the variances are τ_{00}, τ_{11}, and τ_{22}, respectively. Further, each random effect can covary, resulting in a variance covariance matrix for the random effects of

$$\mathrm{T} = \begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix}$$

The diagonal variances tell how much cases vary in their intercepts, linear change, and quadratic change. The off-diagonal covariances tell how much these variances are related to each other. Finally, this model does not fit an autocorrelation to the data, so the error variance matrix is assumed to be $\sigma^2 I$.

Results support the need for a quadratic trend. All three Level 2 intercepts were significantly different from zero, including ending weight, $\hat{\beta}_{00} = 158.85$, $t = 29.85$, $df = 7$, $p < .001$; linear rate of change at the end of the study, $\hat{\beta}_{10} = -1.77$, $t = -4.94$, $df = 7$, $p = .001$; and quadratic effect, $\hat{\beta}_{20} = 0.11$, $t = 5.07$, $df = 7$, $p = .001$. Of course, the fact that $\hat{\beta}_{00} = 158.85$ is significantly different from zero is in some sense trivial, for we would not expect treatment to make the person disappear; however, it does provide a way of understanding the total weight loss involved. The estimate $\hat{\beta}_{10} = -1.77$ is the average rate of weight loss per month at the end of the study. The estimate $\hat{\beta}_{20} = 0.11$ suggests that the slope gets about $2 \times 0.11 = .22$ less steep per month (Raudenbush & Bryk, 2002, p. 171); that is, patients lose the most weight per month at the beginning of treatment and less toward the end (the multiplier of 2 for the quadratic term comes from the derivative of the equation, which gives the rate of change). The between-person variance for the quadratic change was not significantly different from zero ($\tau_{22} = .002$, $\chi^2 = 12.85$, $df = 7$, $p = .075$), but the between-person variances for the intercept ($\tau_{00} = 224.63$, $\chi^2 = 814.54$, $df = 7$, $p < .001$) and ending slope ($\tau_{11} = 0.74$, $\chi^2 = 23.99$, $df = 7$, $p = .001$) were both significantly different from zero. That is, people differed significantly in their final weight and in their average weight loss at the last month, but the rate of weight loss slows consistently for all patients over time. Finally, consider the off-diagonal elements of the variance–covariance matrix of the random effects (T). It is customary to present T as a correlation matrix (R) because it is easier to interpret:

$$R = \begin{bmatrix} 1.00 & .073 & .793 \\ .073 & 1.00 & .663 \\ .793 & .663 & 1.00 \end{bmatrix}$$

Results suggest that a case's ending weight is unrelated to linear rate of weight loss at the end of the study, but that slowing of rate of weight loss was greater for those who ended up heavier at the end of the study and for those who ended up with a lower rate of weight loss per month.

These results are consistent with how Stuart (1967) described them in his narrative. For instance, he noted that the patients differed in their overall weights and in the amount of weight they lost each month. Stuart did not address the gradual slowing of weight loss over time, nor the relationships of quadratic change to the other two parameters. So, the multilevel analysis is more nuanced than the interpretation of the original author.

Two-Phase Multiple Baseline Design With Count as Outcome

The second example shows how to analyze data from a two-phase (AB) multiple baseline study. More important, it illustrates ways of dealing with a count as a dependent variable and related issues that may arise during analysis and interpretation. This is a crucial overlooked factor in virtually all analyses of SCDs, because count data is the most prevalent kind of outcome in SCD research (Shadish & Sullivan, 2011) and because results from analyses that assume normality for count data are likely to give incorrect point estimates and standard errors. DiCarlo and Reid (2004) observed the play behavior of five toddlers with disabilities (see Figure 2). Observations took place in an inclusive classroom over approximately forty 10-min sessions. A count of independent pretend-play actions was the target outcome behavior. There were two phases in this multiple baseline study. For the first 16 to 28 sessions (depending on the case), children were observed without intervention (baseline phase). For the remaining sessions, children were prompted and praised for independent pretend-play actions (treatment phase). DiCarlo and Reid concluded that the intervention increased pretend-play actions in all five children.

The dependent variable in this study is a count, which often follows a Poisson distribution, though we discuss some alternatives later. In HLM, one must also specify if exposure is constant or variable. In this example, exposure is the amount of time for each observed session, a constant 10 minutes for each session for all cases. Had times varied over sessions, we would include a

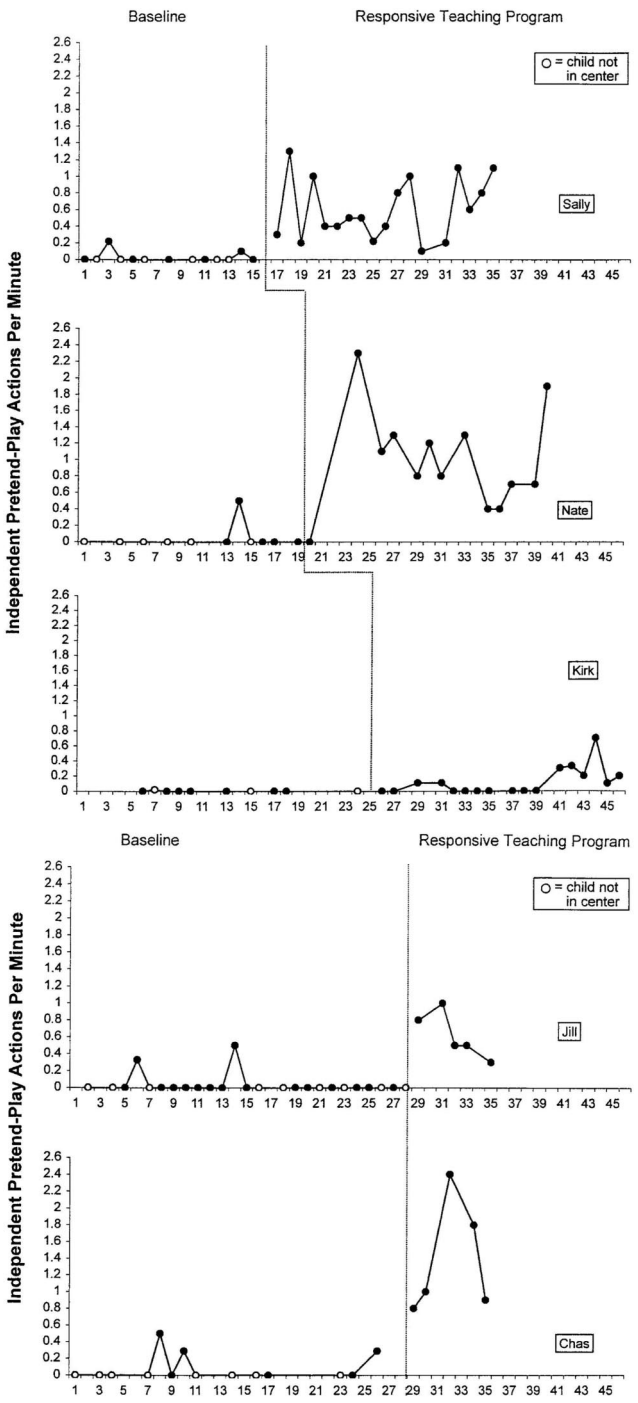

Figure 2. Count of play actions by session and phase for Subjects 1–5. Adapted from "Increasing Pretend Toy Play of Toddlers With Disabilities in an Inclusive Setting," by C. F. DiCarlo and D. H. Reid, 2004, *Journal of Applied Behavior Analysis, 37,* pp. 203–204. Copyright 2004 by the Society for the Experimental Analysis of Behavior, Inc.

variable containing the length of each session. The Poisson model uses a link function (i.e., a transformation of the expected outcome that allows the model to be estimated as a linear model), and it relates the predicted outcome to the observed dependent variable (Cohen, Cohen, West, & Aiken, 2003). The link function for the Poisson model is a log link. The natural model for a count with a Poisson distribution is multiplicative, so taking the logarithm makes it additive (linear).

The outcome in a two-phase design could be either a change in level or a change in slope, with the latter represented by the interaction term in the full Level 1 model:

$$Ln(\hat{Y}_{ti}) = \pi_{0i} + \pi_{1i}a_{1ti} + \pi_{2i}a_{2ti} + \pi_{3i}(a_{1ti}a_{2ti}) \quad (7)$$

where session (a_{1ti} was centered so that 0 represented the session right before the phase change, a_{2ti} is the dummy code for phase (0 = baseline, 1 = treatment), and $a_{1ti}a_{2ti}$ is the product term representing the interaction between phase and session. Consequently, a value of 0 on all Level 1 variables (session, phase, session-by-phase interaction) denotes the count of play acts during the final baseline session. Intercepts for the computed models are then the predicted counts at the phase change. The researcher can, however, center a_{1ti} at any session number. For example, to assess treatment effects at the end of the treatment phase, one should center at the last treatment session number. Equation 7 says that the logarithm of predicted values is the sum of four parts: the predicted value at the intercept (in this case, the final baseline session), plus a term accounting for the rate of change over time, plus a term accounting for the phase change from baseline to treatment, plus an interaction term allowing the time effect to differ across phases. The Level 2 model is

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \end{aligned} \quad (8)$$

This simple model does not include any Level 2 predictors (e.g., child characteristics).

Results were that at the final baseline session, the overall average log count of independent play actions for all students is $\hat{\beta}_{00} = -1.3838$. When the log count is transformed back to counts by exponentiating, the average number of observed independent pretend-play actions during the final baseline session is exp(–1.3838) = 0.2506, $t = -1.99$, $df = 4$, $p = .114$; that is, virtually no independent pretend-play actions. The average rate of change in the log count per session is $\hat{\beta}_{10} = -0.0286$, $t = -0.78$, $df = 4$, $p = .479$; that is, the baseline observations are flat and not changing over time. The average change in log count as a student switches from baseline to treatment phase is $\hat{\beta}_{20} = 2.6680$, $t = 6.07$, $df = 4$, $p < .001$. Thus, the average number of observed independent pretend-play actions per session during phase 2 (treatment) is exp(−1.3838 + 2.6680) = exp(1.2842) = 3.61, significantly higher than during baseline. Last, the average interaction effect, or change in slope between phases, is $\hat{\beta}_{30} = 0.0607$, $t = 0.03$, $df = 4$, $p = .109$. That is, the amount of play during the treatment phase is flat (not changing over time), just as it was during the baseline phase. For these data, then, the best fitting model indicates no effect of time in either phase (i.e., flat slopes) but a significant effect of treatment, predicting more play acts per session in the treatment phase (3.61, on average) than in the baseline phase (0.25, on average). None of the variance compo-

nents were significantly different from zero. The between-person variances were as follows: of intercepts, $\tau_{00} = 1.62$ ($\chi^2 = 3.97$, $df = 3$, $p = .26$); of slopes, $\tau_{11} = 0.003$ ($\chi^2 = 5.72$, $df = 3$, $p = .13$); of the treatment effect, $\tau_{22} = 0.13$ ($\chi^2 = 6.25$, $df = 3$, $p = .10$); and of the interaction, $\tau_{33} = .00003$ ($\chi^2 = 5.79$, $df = 3$, $p = .12$). In fact, the variance–covariance matrix of the random effects suggests extremely high collinearity among the random effects:

$$R = \begin{bmatrix} 1.000 & -1.000 & 1.000 & -.999 \\ -1.000 & 1.000 & -1.000 & .998 \\ 1.000 & -1.000 & 1.000 & -.998 \\ -.999 & .998 & -.998 & 1.000 \end{bmatrix}$$

Clearly, if we were to proceed with further analyses, a prime candidate for change would be fixing some or all of the random effects to zero.

The multilevel model suggests that children all started at the same low level of play acts during baseline, that the treatment significantly increased the overall level of play acts, that increased exposure to treatment over time did not result in increased improvement over time, and that children did not differ significantly from each other in all these things. This is more nuanced but entirely consistent with DiCarlo and Reid's (2004) conclusion that the treatment helped all these children. DiCarlo and Reid also suggested the treatment might have helped one child, Kirk, less than the others, but the multilevel model does not support this conclusion. Kirk's data may appear to be more different from those of the other children than they are when chance is taken into account. In a Poisson distribution, as the mean increases, the variance increases. The nonsignificant variance component for the treatment effect in the paragraph above suggests the differences in means over children during treatment are not distinguishable from chance. So when, say, Nate shows a few rather high data points, that may not reflect better treatment response than Kirk but rather the fact that we would expect more variation in Nate's data by chance around his mean, even though his mean does not differ significantly from Kirk's mean. If these data were graphed on a logarithmic scale to take this into account, the overall consistency in response to treatment across children would be more apparent.

Two-Phase Multiple Baseline Design With Proportion as Outcome and With Overdispersion

A third study (Hunt, Soto, Maier, & Doering, 2003; see Figure 3) extends the illustration of the analysis of data that comes from a two-phase (A-B) study to a dependent variable that is a proportion. In this study, the researchers observed the academic and social participation behavior of six elementary school students in general education classes at two schools. Three of these students had diagnosed severe disabilities; the other three were identified as academically at risk. The study had two phases. For the first three to eight sessions (depending on the student), students were observed without intervention (baseline). Then, teachers, aides, and parents collaborated to plan and implement individualized support plans including academic adaptations, communications, and social supports for each child in the study. The remaining observations were made during this treatment, and they took place in each classroom over several months.

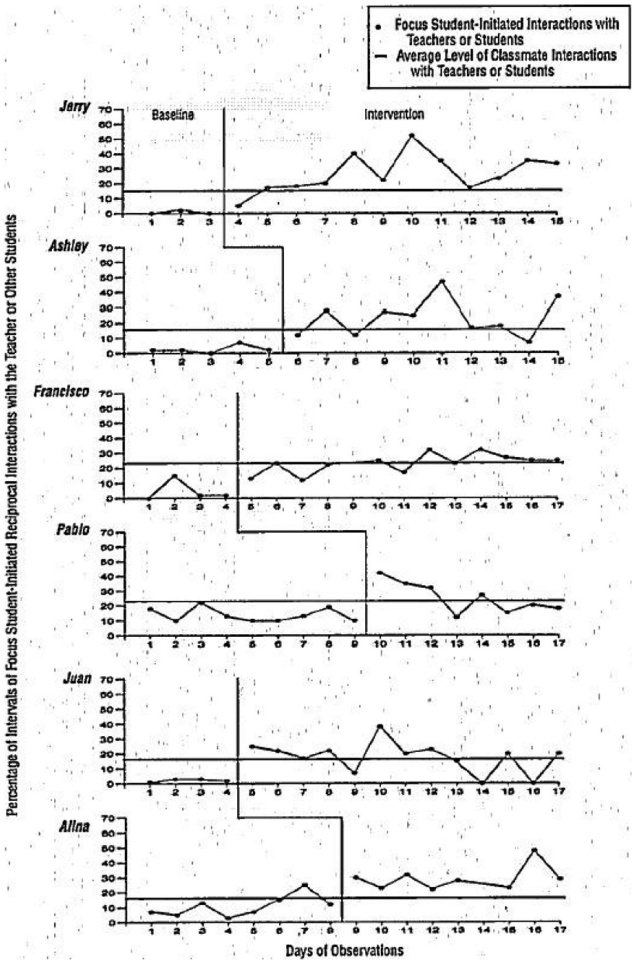

Figure 3. Percentage of intervals of focus student-initiated interactions to the teachers or other students by day and phase for Subjects 1–6. Adapted from "Collaborative Teaming to Support Students at Risk and Students With Severe Disabilities in General Education Classrooms," by P. Hunt, G. Soto, J. Maier, and K. Doering, 2003, *Exceptional Children, 69,* p. 326. Copyright 2003 by the Council for Exceptional Children.

The target behavior was student initiation of interactions with the teacher or other students. Each session was divided into 60 intervals or trials. For each trial, the researcher noted whether or not the student initiated a social interaction with the teacher or other students at least once. The percentage of trials where the student did initiate interactions was computed and recorded as the observation for each session. The dependent variable in this data set is then a proportion (successful trials out of total trials), which must be accommodated in the analyses and in subsequent interpretation. Because this dependent variable is a proportion from a fixed number of binary (0, 1) observations, we used a binomial distribution when analyzing the data. The binomial is used to model the number of events that took place where the total possible number of events is known. In this example, we know that for each session 60 trials were observed (had the number of trials differed across sessions or across children, a simple adaptation to the analysis would be made). A 100% on the dependent measure would indicate that in 60 out of 60 trials the focus student was

observed initiating an interaction; 50% would indicate the student initiated interactions during 30 of the 60 trials on that day. That is, knowledge of the number of events per session allows recovery of the number of successes for each session.

Whereas the Poisson distribution is used to model the frequency of an event in a given period of time, as in DiCarlo and Reid (2004), the binomial distribution is used to model the frequency of a binary (yes, no) event out of a total known number of possible trials (i.e., a proportion or percentage). For both types of distributions, observations are assumed to be independent and identically distributed, so that the outcome of one observation is not expected to affect the outcome of another observation, and the probability of success is the same for all trials. Unlike in normal distributions, in which the variance is completely independent from the mean, in binomial and Poisson distributions the variance is a function of the mean. As noted above, for the Poisson distribution, the mean and the variance are equal; as the mean increases, so does the variance. Counts tend to vary more when their average value is higher (Agresti, 1996). For the binomial, the mean and the variance are related but not equal values; the variance is largest when the mean proportion is .5.

These relations between the mean and variance of the distributions are sometimes violated. When count data (including both rates and proportions) exhibit greater variability than would be predicted by the Poisson or binomial models, the result is called overdispersion. Overdispersion can be caused by statistical dependence or heterogeneity among cases (Agresti, 1996, 2002), or it can occur if the Level 1 model is underspecified (Raudenbush et al., 2004). Overdispersion is measured by the overdispersion parameter σ^2. If the assumptions of the distribution are met, the overdispersion parameter will be approximately $\sigma^2 = 1$. If $\sigma^2 > 1$, overdispersion is present, and if $\sigma^2 < 1$, underdispersion is present. Overdispersion can inflate the χ^2 goodness of fit test and cause the standard errors of the regression coefficients to be too small, resulting in too many Type I errors (Cohen et al., 2003).

If overdispersion is present, two ways exist to deal with it (Cohen et al., 2003). One way assumes that the variance is a constant multiple of the mean, called a quasi-Poisson distribution (sometimes an overdispersed Poisson model, itself a special case of a quasi-likelihood regression model). In this approach, the standard errors are adjusted by the overdispersion parameter, so the excess of Type I errors is reduced. Another way is to fit a negative binomial model, which mixes a Poisson distribution and a gamma distribution to model the extra variation. This option allows the variance to be a nonconstant multiple of the mean. HLM takes the former approach, allowing "estimation of a scalar variance so that the Level 1 variance will be $\sigma^2 w_{ij}$" (Raudenbush et al., 2004, p. 111), an option that must be called specifically. If there is no problem, the Level 1 variance will be $\sigma^2 \approx 1$.

The Level 1 model for this study is

$$\ln\left(\frac{\hat{P}_{ij}}{1-\hat{P}_{ij}}\right) = \pi_{0i} + \pi_{1i}a_{1ti} + \pi_{2i}a_{2ti} + \pi_{3i}(a_{1ti}a_{2ti}) \qquad (9)$$

where $\hat{P}_{ij}$ is the expected proportion of trials within a session in which the behavior was exhibited. When one uses a binomial distribution, HLM estimates are produced on a log odds or logit scale; to interpret them, one typically converts them back to an odds (for the intercept) or odds ratio where 0 is the lower bound, 1 suggests no difference, and infinity is the upper bound. For the remaining coefficients in the Level 1 model, day of observation (a_{1ti}) was centered before analysis, so that 0 represented the session right before the phase change; a_{2ti} is the dummy code for treatment phase (0 = baseline, 1 = treatment); and $a_{1ti}a_{2ti}$ is the product term representing the interaction between phase and day. Consequently, a zero on all Level 1 variables (day, phase, day-by-phase interaction) denotes the proportion of trials in which the target behavior was observed during the final baseline day of observation. Intercepts for the computed models are then the predicted counts at the phase change. The Level 2 model is

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \end{aligned} \qquad (10)$$

Results are in Table 1. Two of four fixed effects are significant. At the final baseline session, the overall average log odds for all

Table 1
Results of Multilevel Model on the Hunt et al. (2003) Data

Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Odds ratio
β_{00}	−2.85	0.43	−6.60	5	0.000	0.06
β_{10}	−0.02	0.08	−0.25	5	0.812	0.98
β_{20}	1.79	0.41	4.37	5	0.008	5.99
β_{30}	0.01	0.08	0.10	5	0.924	1.01
Random effect	Standard deviation	Variance component	*df*	χ^2	*p*	
r_{0i}	0.85	0.72	5	18.13	0.003	
r_{1i}	0.10	0.01	5	4.40	>0.500	
r_{2i}	0.71	−0.51	5	8.02	0.154	
r_{3i}	0.09	0.01	5	3.52	>0.500	
e_{ti}	1.72	2.95				

Note. Values in this table are rounded to two decimals, but computations reported in text, such as conversions of log odds ratios to odds ratios and proportions, were done on results to six decimals. If such conversions are done on the numbers in the table, the results will differ due to rounding error. *df* = degrees of freedom.

students is $\hat{\beta}_{00} = -2.85$, yielding an odds of exp(–2.85) = 0.0576, which converts to a proportion of P = odds/(1 + odds) = .05. That is, the odds at that point of a student initiating interaction are about 1:20, and the probability of an interaction is very low. For the change from the baseline to treatment, the change in odds was exp(1.79) = 5.989. For the average child, then, the odds rose to .0576 * 5.989 = .345, or about 1:3. The terms for the interaction and for sessions are small and not significant, suggesting that there was little consistent trend up or down in general and that treatment did not much change that compared to baseline. Only the variance component for the intercept was significant, suggesting that children varied significantly in how much they initiated interactions but not in their rate of change or their treatment effect. Again, the covariances among the random effects (not presented here) suggested high collinearity; this, combined with some nonsignificant variance components, might suggest dropping some of the random effects. Finally, the Level 1 variance (σ_e^2) provides evidence about whether data are overdispersed. According to Raudenbush and Bryk (2002), a large value of σ^2 serves as evidence of overdispersion: If the binomial model were correct and data were not overdispersed, σ^2 would be close to 1.0. Here, $\sigma^2 = 2.94$, which is far enough from 1.0 for us to assume overdispersion.

The multilevel results are consistent with the conclusions of Hunt et al. (2003). Both the analyses and the original authors conclude that the baseline is stable, that treatment is effective for all students, and that increased exposure to treatment over time does not either increase or decrease the effects of treatment in any consistent fashion. The multilevel model adds the finding that children differed from each other in their overall level of interaction before treatment began, but that has no important impact on the substantive question addressed in this study.

Four-Phase ABAB Designs

A fourth published study extends the illustration of analyses of data from two-phase (AB) studies to a study with four phases (ABAB; Lambert, Cartledge, Howard, & Lo, 2006; see Figure 4). In this study, researchers assessed the effects of a response card program on the disruptive behavior and academic responding of students in two elementary school classes. The data analyzed in this section represent instances of disruptive behavior during baseline single-student responding (phase A), where the teacher called on students one at a time as they raised their hands, and during a response card treatment condition (phase B), where every student wrote a response to each question on a laminated board and presented them simultaneously; both phases were repeated a second time, resulting in an ABAB design. Data collection focused on nine fourth-grade students (four boys, five girls) with a history of disciplinary issues. Each student was observed for ten 10-s intervals during each observation session. The number of intervals during which disruptive behaviors were observed was recorded (with a maximum of 10 for each session). Between five and 10 sessions were recorded for each of the four phases. The dependent variable in this data set is then a proportion (number of trials with occurrences of disruptive behavior out of 10 total trials) for each session. As in the analyses of data from the Hunt et al (2003) study, this type of dependent variable may be accommodated by using a binomial distribution.

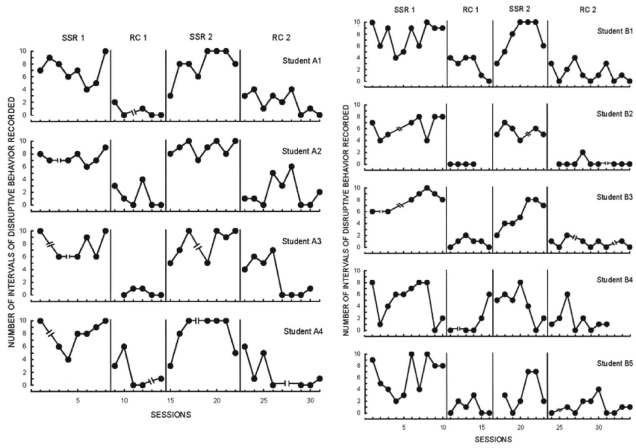

Figure 4. Number of intervals of disruptive behavior recorded during single-student responding (SSR) and response card treatment (RC) conditions. Adapted from "Effects of Response Cards on Disruptive Behavior and Academic Responding During Math Lessons by Fourth-Grade Urban Students," by M. C. Lambert, G. Cartledge, W. L. Heward, and Y. Lo, 2006, *Journal of Positive Behavior Interventions, 8,* pp. 94–95. Copyright 2006 by Sage Publications.

One Way to Code a Four-Phase ABAB Design

The Level 1 model is

$$Ln\left(\frac{\hat{P}_{ti}}{1-\hat{P}_{ti}}\right) = \pi_{0i} + \pi_{1i}a_{1ti} + \pi_{2i}a_{2ti} + \pi_{3i}a_{3ti} + \pi_{4ti}(a_{2ti}a_{3ti}) + \pi_{5ti}(a_{1ti}a_{2ti}) + \pi_{6ti}(a_{1ti}a_{3ti}) + \pi_{7ti}(a_{1ti}a_{2ti}a_{3ti}) \quad (11)$$

where P_{ti} is the proportion of intervals within a session in which a disruptive behavior was exhibited, session (a_{1ti} was centered so that 0 represented the final session of the first baseline phase, a_{2ti} is a dummy code for phase (0 = baseline, 1 = treatment), a_{3ti} is a dummy variable to express whether a phase was part of the first AB pair (0) or the second AB pair (1), and the product terms represent the interactions among these main effects. By virtue of the centering, intercepts for the computed models are predicted proportions at the first phase change.

The unconditional (i.e., without predictors) Level 2 model is then

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \\ \pi_{4i} &= \beta_{40} + r_{4i} \\ \pi_{5i} &= \beta_{50} + r_{5i} \\ \pi_{6i} &= \beta_{60} + r_{6i} \\ \pi_{7i} &= \beta_{70} + r_{7i} \end{aligned} \quad (12)$$

Results are in Table 2, showing a host of significant fixed and random effects. Treatment had a very large effect ($\beta_{20} = -5.97$) in reducing the log odds of disruptive behavior. However, every

Table 2
Results of Multilevel Model on the Lambert et al. (2003) Data

Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Odds ratio	Proportion
			First coding method for ABAB designs				
β_{00} intercept	0.61	0.33	1.83	8	0.10	1.84	.65
β_{10} session	−0.05	0.07	−0.71	8	0.50	0.95	.49
β_{20} treatment	−5.97	1.16	−5.16	8	<0.01	0.00	<.01
β_{30} AB pair	0.75	1.68	0.45	8	0.67	2.12	.68
β_{40} Tmt × AB	5.97	1.57	3.80	8	0.01	393.28	.99
β_{50} Sess × Tmt	0.69	0.30	2.26	8	0.05	1.98	.66
β_{60} Sess × AB	0.01	0.09	0.10	8	0.92	1.01	.50
β_{70} Tmt × AB × Sess	−0.82	0.25	−3.24	8	0.01	0.44	.31

Random effect	Standard deviation	Variance component	*df*	χ^2	*p*
r_{0i} intercept	0.86	0.74	8	26.12	0.001
r_{1i} session	0.18	0.03	8	22.23	0.005
r_{2i} treatment	2.69	7.22	8	22.68	0.004
r_{3i} AB pair	4.49	20.20	8	34.80	<0.001
r_{4i} Tmt × AB	1.36	1.84	8	4.77	>0.500
r_{5i} Sess × Tmt	0.83	0.69	8	35.81	<0.001
r_{6i} Sess × AB	0.14	0.02	8	6.17	>0.500
r_{7i} Tmt × AB × Sess	0.60	0.36	8	13.83	0.086
e_{ti}	1.46	2.15			

Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Odds ratio
			Second coding method for ABAB designs			
β_{00} intercept	0.55	0.24	2.25	8	0.055	1.73
β_{10} A1 to B1	−1.85	0.64	−2.88	8	0.021	0.16
β_{20} B1 to A2	2.57	0.44	5.80	8	<0.001	13.01
β_{30} A2 to B2	−2.28	0.45	−5.12	8	<0.001	0.10
β_{40} session	−0.04	0.05	−0.91	8	0.388	0.97

Random effect	Standard deviation	Variance component	*df*	χ^2	*p*
r_{0i} intercept	0.56	0.31	8	19.19	0.014
r_{1i} A1 to B1	1.65	2.72	8	27.55	0.001
r_{2i} B1 to A2	0.92	0.85	8	13.77	0.087
r_{3i} A2 to B2	0.90	0.80	8	10.91	0.206
r_{4i} session	0.12	0.01	8	17.73	0.023
e_{ti}	1.64	2.69			

Note. Values in this table are rounded to two decimals, but computations reported in text, such as conversions of log odds ratios to odds ratios and proportions, were done on results to six decimals. If such conversions are done on the numbers in the table, the results will differ due to rounding error. *df* = degrees of freedom.

interaction term that includes treatment is also significant, so that the effects of treatment vary depending on AB pair and session in a complex way. Also, nearly every variance component was also significant, suggesting that students varied significantly, not just in their response to treatment but in many other main effects and interactions. This complexity is an accurate rendition, given what we see in the graph. A clear treatment effect does seem apparent, but the size and consistency of that effect also seem to depend on the student, whether one looks at the first or second AB pair, and on the session within those pairs.

This is a case in which some careful exploratory model reduction might be done to see if a simpler conclusion is warranted. This is doubly the case because some odds ratios in Table 2 are suspiciously high or low, raising concern about collinearity. One sensible reduced model, for example, includes the main effect for treatment (A vs. B), the main effect for the first pair of AB phases compared with the second pair of AB phases, and an interaction to see whether the effect of treatment is the same in both AB phase changes of the study. That model found the following. The intercept was significant ($\beta_{00} = 0.599$), which converts to an odds of 1.82. That is, the odds of an interval with disruptive behavior at the end of the first baseline is about 2:1, or two intervals with disruptive behavior for each interval without. Treatment greatly reduces disruptive behavior in general ($\beta_{20} = -2.20$, odds ratio = 0.11), so that the odds of an interval with disruptive behavior during treatment drop to $1.82 * 0.11 = 0.20$, or about 1:5. The order effect is not significant ($\beta_{30} = 0.070$, odds ratio = 1.073). The interaction between treatment and order was also not significant ($\beta_{40} = -0.330$, odds ratio = .719), suggesting the treatment was about as effective in the second AB pair as in the first AB pair.

Only one of the variance components was significant, the one for the order effect ($\tau_{33} = .483$, $df = 8$, $\chi^2 = 18.08$, $p = .021$),

suggesting that the order effect varied over students. The remaining variance components were $\tau_{00} = .063(df = 8, \chi^2 = 10.34, p = .241)$ for the intercept, $\tau_{22} = .507(df = 8, \chi^2 = 13.70, p = .089)$ for treatment, and $\tau_{44} = .751(df = 8, \chi^2 = 8.26, p = .408)$ for the interaction between order and treatment.

Alternative Method of Coding Phases in ABAB Designs

Other methods of coding an ABAB design are possible, depending on what quantities are of interest. Here, we illustrate an alternative and potentially useful coding method, which we call step coding. It resembles dummy coding, in that it uses only the numbers 0 and 1, but the coding is different in other respects. Suppose that we want the intercept to represent behavior during the baseline phase of the study, and we want other effects to measure the changes as we go from one phase to another. That is, one effect should measure the change from A1 (the first A phase) to B1 (the first B phase); another effect should measure the next change, from B1 to A2 (the second A phase); and the final effect should measure the final change, from A2 to B2 (the second and final B phase). This requires three dummy variables that start with the value 0 and then change to 1 with successive changes in phases: the first dummy variable $a_{1ti} = 0$ during phase A1, and then $a_{1ti} = 1$ for phases B1, A2, and B2: variable $a_{2ti} = 0$ for phases A1 and B1, then $a_{2ti} = 1$ for phases A2 and B2; and, finally, $a_{3ti} = 0$ for phases A1, B1, and A2 and $a_{3ti} = 1$ for phase B2. Thus, they form a pattern resembling steps. The meaning of these effects depends on all of them being present in the model; removing any of them changes the meanings of the remaining effects, because they are not orthogonal. In the following models, we also included a term a_{4ti} for session, to allow for time trend, coded so that 0 is the last session in the first phase of the study (phase A1). In addition, we allowed for overdispersion (as we did in fitting previous models).

This coding then allows us to specify the following multilevel model. At Level 1,

$$\ln\left(\frac{\hat{P}_{ti}}{1-\hat{P}_{ti}}\right) = \pi_{0i} + \pi_{1i}a_{1ti} + \pi_{2i}a_{2ti} + \pi_{3i}a_{3ti} + \pi_{4i}a_{4ti} \tag{13}$$

where the terms are as described above. The unconditional Level 2 equations are now

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \\ \pi_{4i} &= \beta_{40} + r_{4i} \end{aligned} \tag{14}$$

This model says that (a) the logarithm of the odds of showing disruptive behavior is a function of a linear trend, as well as changes due to shifts between phases, and (b) each of these effects may vary across individuals.

The results for this model are in Table 2. The average log odds in the first baseline phase was .546. Exponentiating this gives exp(.546) = 1.727, which is the odds of showing disruptive behavior; converting this to a proportion using P = odds/(1 + odds) suggests that one would observe disruptive behavior about 63% of the time, about the same as with the first method of coding. The average change going from one phase to another was significant for each such change. For the change from the first baseline to the first treatment, the change in odds was exp(−1.854) = .157. For the average child, the odds dropped to 1.727 * .157 = .271, or about 1:4. That is, for every observation during which there is a disruptive behavior, there are roughly 4 observations with no disruptive behavior, a huge change from baseline. The change from first treatment to second baseline changes the odds by an average of exp(2.566) = 13.011 times. So, during the second baseline, the odds are 1.727 * .157 * 13.011 = 3.528, suggesting a ratio of disruptive to nondisruptive observations of about 3.5: 1. The final phase change (back to B2) reduces the odds by a factor of exp(−2.281) = .102, so the odds for that phase are 1.727 * .157 * 13.011 * .102 = .360, or about one observation with disruptive behavior for every three without such behavior. The term for sessions is small and not significant, with no consistent linear trend up or down. However, for each of the phases the odds are reduced by about 3% (multiplied by the odds ratio of .97 for session). Finally, the random effects show that intercepts, the effect for session, and the A1-B1 effects all vary significantly across individuals but that B1-A2 and A2-B2 changes do not. The estimate of σ^2 (= 2.69) is well above the expected value of 1 for the model without overdispersion.

Both ways of coding the Lambert et al. (2006) study yield results that are reasonably consistent with the conclusions of the original authors. They concluded that the treatment was effective in general, that the shifts from phase to phase all reflected the changes in outcomes that the step coding above suggested, and that results were somewhat variable over students with larger effects for five students and smaller effects for four others. However, the step coding provides a statistical test of the requirement to demonstrate a treatment effect at least three times what the What Works Clearinghouse Standards (Kratochwill et al., 2010) suggests is best practice for SCDs.

Alternating Treatment Designs With Three Conditions

Resetar and Noell (2008) used an alternating treatment design to compare the effectiveness of (a) no reward (NR), (b) a multiple-stimulus-without-replacement (MSWO) reward condition, and (c) a teacher-selected (TS) reward condition in identifying reinforcers for use to help students improve in mathematics (see Figure 5). The four students were typically developing elementary-school children with performance deficits in mathematics. The outcome was the number of digits correctly answered in 2 minutes on a test of subtraction problems, which should be modeled with the same approach as in the second example, a Poisson distribution having a constant exposure with a log-link function.

If an alternating treatments design has only two conditions, it would be appropriately analyzed with the Level 1 model in Equation 7 and the Level 2 model in Equation 8. But this example has three conditions. This changes the analysis in two ways. First, one could compare all possible pairs of conditions in three separate multilevel analyses: (a) TS versus MSWO, (b) TS versus NR, and (c) MSWO versus NR. In that case, one would still use Equations 7 and 8. Unfortunately, this option has problems. Consider the

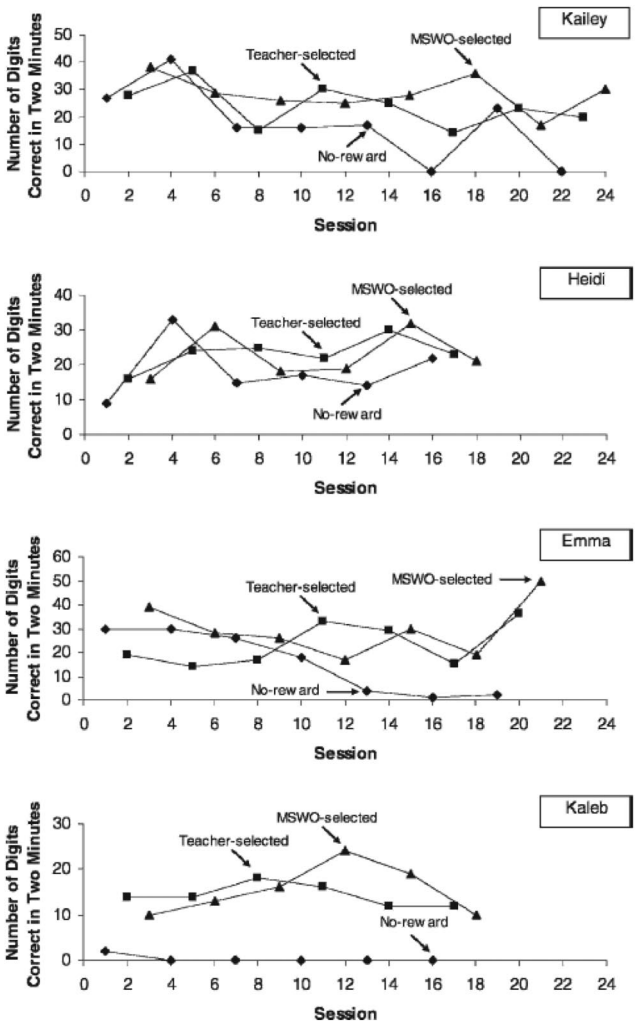

Figure 5. Number of digits correct in two minutes across reward conditions for all participants. MSWO = multiple stimulus without replacement. Adapted from "Evaluating Preference Assessments for Use in the General Education Population," by J. L. Resetar and G. H. Noell, 2008, *Journal of Applied Behavior Analysis, 41,* p. 450. Copyright 2008 by the Society for the Experimental Analysis of Behavior, Inc.

analogy to a one-way analysis of variance in a between-subjects experiment with three conditions. We could analyze such data with three *t* tests, one test for each pair of conditions; but we do not do so because the experiment-wise Type I error rate becomes inflated. Instead, we conduct an omnibus one-way analysis of variance that holds that error rate to the desired level. The same logic applies to the Resetar and Noell SCD data. We reject the all-possible-pairs option, because it would inflate Type I error rates, and instead use an omnibus analysis in which all conditions are analyzed in one model.

Second, the models we propose for analyzing such data each have two terms related to the treatment effect. The analysis should consider the significance of those two terms taken jointly. This can be done in two ways in HLM. One is a general linear hypothesis test that examines whether the two terms are jointly significant, tested with a χ^2 test with $df = 2$. The other is a model comparison test that takes the difference in deviances for two nested models, one with the two treatment terms and one without them, and compares that difference to a χ^2 distribution with $df = 2$. We had difficulty getting useful estimates of the latter test, probably due to small sample sizes, so only the former are reported here.

The overall analysis can be done several ways. One analysis uses the following Level 1 model:

$$Ln(\hat{Y}_{ti}) = \pi_{0i} + \pi_{1i}a_{1ti} + \pi_{2i}a_{2ti} + \pi_{3i}a_{3ti} + e_{ti} \quad (15)$$

where a_{1ti} is session (centered at Session 18 in this case to represent a time near the end of the study for all participants), where $a_{2ti} = 0$ for NR and $a_{2ti} = 1$ for both treatments, and where $a_{3ti} = 0$ for NR and TS and $a_{3ti} = 1$ for the MSWO treatment. Then, the parameter π_{2i} would represent NR versus TS, and π_{3i} would represent TS versus MSWO. The Level 2 model is

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \end{aligned} \quad (16)$$

The Level 1 model says that the outcome is a function of change over time (any trend over sessions) plus an effect from going to the NR (baseline) to TS (first treatment) plus an effect going from TS to MSWO (second treatment), and Model 2 allows these effects to vary over cases.

The results for this model are shown in Table 3. The general linear hypothesis test suggests the two treatment terms are jointly not significant. Just as we would normally not follow a nonsignificant analysis of variance with follow-up tests comparing individual conditions, we should not now interpret the significance to the two individual treatment terms. However, for pedagogical reasons, we can observe that the nonsignificant omnibus test is consistent with the individual coefficients, showing that neither the change from NR to TS nor the one from TS to MSWO affected the outcome significantly. Resetar and Noell (2008) did not draw a conclusion about overall treatment effectiveness like the one in this analysis. Rather, they said that the treatment seemed to work for two children and not for two others. This might be consistent with the significant variance component of $\tau_{22} = 0.63$, which suggests that children vary significantly in their response to TS compared to NR. We will return to this shortly.

A second way to produce an omnibus analysis uses the Level 1 model in Equation 15 but where a_{3ti} is coded using effects coding. This time $a_{3ti} = 0$ for NR, $a_{3ti} = -.5$ for TS, and $a_{3ti} = .5$ for the MSWO treatment. Under this coding, the parameter π_{2i} is the difference between NR and the average of the two treatments, and π_{3i} is the difference between the two treatments. The Level 2 model stays the same. The results for this model are also in Table 3. The interpretation of this model is nearly identical to that from the previous coding. The general linear hypothesis test again suggests the two treatment terms are not jointly significant. Further, outcomes for baseline (NR) do not differ significantly from those for the average of the two treatments, and treatments do not differ from each other. Finally, in both models the estimate of σ^2 (= 3.58) is still above the expected value of 1 for the model without overdispersion.

Resetar and Noell (2008) said that "the mean number of digits correctly answered was greater in the MSWO-selected reward and

Table 3
Results of Multilevel Model on the Resetar and Noell (2008) Data

Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Event rate ratio
		First coding method for alternating treatment designs				
β_{00} intercept	2.16	0.58	3.74	3	.07	8.69
β_{10} session	−0.01	0.01	−1.04	3	.38	0.99
β_{20} NR to TS	0.79	0.42	1.86	3	.15	2.20
β_{30} TS to MSWO	0.12	0.12	1.03	3	.38	1.13
Random effect	Standard deviation	Variance component	*df*	χ^2	*p*	
r_{0i} intercept	1.11	1.24	3	35.16	<.001	
r_{1i} session	0.02	0.00	3	5.09	.164	
r_{2i} NR to TS	0.79	0.63	3	23.85	<.001	
r_{3i} TS to MSWO	0.09	0.01	3	1.20	>.500	
e_{ti}	1.89	3.58				
		Significance of treatment terms taken jointly				
General linear hypothesis test	$\chi^2 = 4.43$	$df = 2$	$p = .107$			
Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Event rate ratio
		Second (effects) coding method for alternating treatment designs				
β_{00} intercept	2.16	0.58	3.70	3	.07	8.65
β_{10} session	−0.01	0.01	−1.04	3	.38	0.99
β_{20} NR vs. Tmt	0.85	0.43	1.97	3	.14	2.35
β_{30} TS vs. MSWO	0.12	0.12	1.02	3	.38	1.13
Random effect	Standard deviation	Variance component	*df*	χ^2	*p*	
r_{0i} intercept	1.12	1.26	3	35.44	<.001	
r_{1i} session	0.02	0.00	3	5.09	.163	
r_{2i} NR vs. Tmt	0.82	0.68	3	27.80	<.001	
r_{3i} TS vs. MSWO	0.09	0.01	3	1.20	>.500	
e_{ti}	1.89	3.58				
		Significance of treatment terms taken jointly				
General linear hypothesis test	$\chi^2 = 4.41$	$df = 2$	$p = .108$			

Note. Values in this table are rounded to two decimals, but computations reported in text such as conversions of log odds ratios to odds ratios and proportions were done on results to six decimals. If such conversions are done on the numbers in the table the results will differ due to rounding error. NR = no reward; TS = teacher selected; MSWO = multiple stimulus without replacement; Tmt = treatment; *df* = degrees of freedom.

the teacher-selected reward conditions relative to the no-reward condition for 2 of the 4 participants" (p. 447), who are Emma and Kaleb in Figure 5. Here, we run an exploratory analysis both to test that claim and to motivate some observations about exploratory analyses in SCDs in the Discussion section. To do this we continue to use effects coding but change the Level 2 model to add a dummy variable x_{2i} to the third equation in (16) that distinguishes Emma and Kaleb from the other two participants:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + r_{0i} \\ \pi_{1i} &= \beta_{10} + r_{1i} \\ \pi_{2i} &= \beta_{20} + \beta_{21}x_{2i} + r_{2i} \\ \pi_{3i} &= \beta_{30} + r_{3i} \end{aligned} \tag{17}$$

The results are essentially unchanged. The treatment is not significantly more effective for Emma and Kaleb than for the other two participants ($\beta_{20} = .10$, $t = 0.84$, $p = .49$).

One could also code this kind of design a third way, with two dummy variables for treatment, where each treatment is then compared to the NR condition. The problem with that option is that the two treatments are not compared to each other. A combination of more than one of the analyses described in this section covers all bases, however. Rindskopf and Ferron (in press) show additional ways to code this design.

Changing Criterion Designs

Ganz and Flores (2008) used a changing criterion design to investigate the use of visual strategies to increase verbal behavior in three children with autistic spectrum disorder when those children were in play groups with typically developing peers. The play group sessions occurred 4–5 times per week over 4 weeks for 30 minutes each. Prior to each play group session, the researcher presented each child with one (or more) statement(s) on a script

card and prompted the child to repeat the statement until he or she was able to do so without prompting. Then, during the play groups, the researcher again prompted the child once with the practiced script card, except during baseline when no prompts occurred. Prompting on that script card stopped when the child recited the scripted phrase, or when the prompting interval ended. In the latter case, the researcher waited one interval before trying again. After baseline, a low criterion for success was set (e.g., that the child respond to one prompt card that was presented during a 20-s interval of play). When the child had successfully done this in three consecutive sessions, the criterion increased to prompt the child to respond to additional prompt cards (e.g., two cards presented in two different intervals).

From videotapes of the sessions, the researchers recorded the occurrence of several outcomes; this analysis uses the outcome called intervals with any speech (see Figure 6). In each of 15 consecutive 20-s intervals during each play session, the researcher recorded whether the child uttered any speech at all. So the count could range from 0 to 15, and from this the researchers computed the percent of intervals during which an outcome occurred. Like the third example in this article, then, this outcome has a binomial distribution, and so we use a logit link function.

These data could be analyzed in two ways (we ignore the generalization phase in Figure 6 in these analyses). One would be to compare outcomes during baseline to those during all of the treatment phases without regard to the changing criteria. That model is the same as in Equations 7 and 8. The term for the interaction between session and the baseline-to-treatments change is useful to test the hypothesis that, compared to baseline, increasing the criterion over time during treatment increases the slope of the data points across the three criterion intervals. Results are in Table 4. None of the four predictors were significant, but two of the four variance components were. The between-person variance of intercepts is $\tau_{00} = 2.17$ ($\chi^2 = 11.03$, $df = 2$, $p = .004$), and the between-person variance of the treatment effect is $\tau_{11} = 5.61$ ($\chi^2 = 20.87$, $df = 2$, $p < .001$). The former suggests that the children displayed significantly different levels of speech at the end of baseline, and they had significantly different responses to treatment even though the effect of treatment was not significant on average. This data set displays virtually no overdispersion. We reran it without the overdispersion option, and the results were nearly identical.

A second way to analyze the data would be to use the step coding from Equations 13 and 14. The models for this example are identical to those two equations, but the interpretation differs. That is, one effect measures the change from baseline to the first treatment criterion phase; another effect measures the change from the first to the second treatment criterion phase; and the final effect measures the change from the second to the third criterion phase. This requires three dummy variables: variable $a_{1ti} = 0$ during baseline and then $a_{1ti} = 1$ for all three treatment phases; variable $a_{2ti} = 0$ for the baseline and first treatment phase and then $a_{2ti} = 1$ for the second and third treatment phases; and, finally, $a_{3ti} = 0$ for baseline and the first two treatment phases and $a_{3ti} = 1$ for the last treatment phase. This model yielded no significant effects for any predictors or variance components except for the between-cases variance in intercepts $\tau_{00} = 1.29$ ($\chi^2 = 7.61$, $df = 2$, $p < .022$). Overdispersion was somewhat higher than it should be.

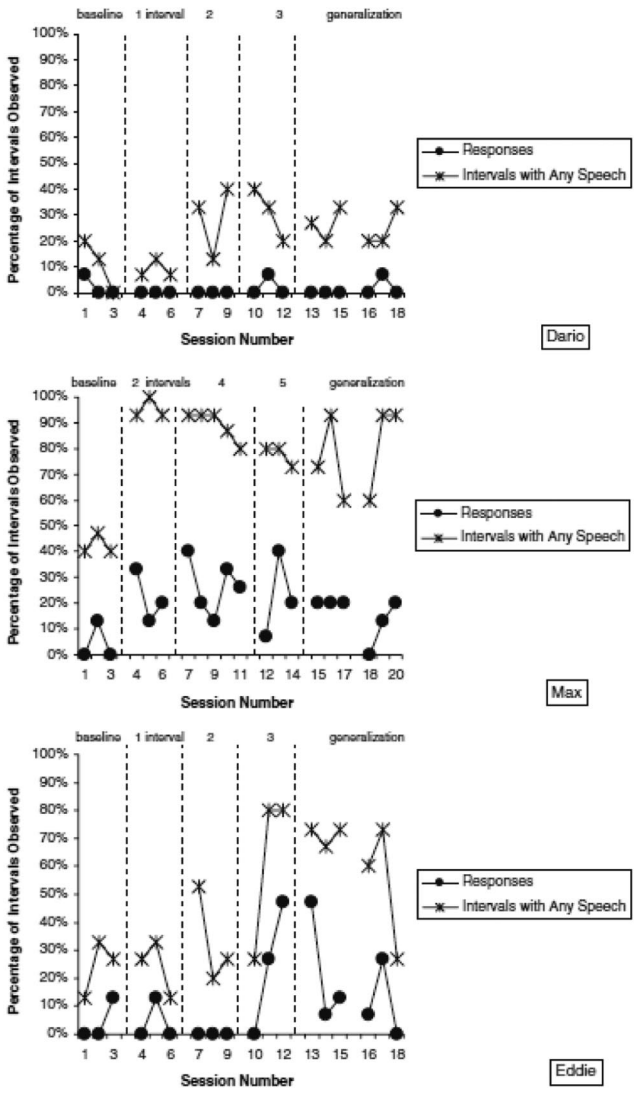

Figure 6. A changing criterion design to test the effect of an intervention to increase verbal behavior in children with autistic spectrum disorder on two outcomes: Responses and Intervals with Any Speech. Adapted from "Effects of the Use of Visual Strategies in Play Groups for Children With Autism Spectrum Disorders and Their Peers," by J. B. Ganz and M. M. Flores, 2008, *Journal of Autism and Developmental Disorders, 38*, p. 936. Copyright 2008 by Springer.

We can compare these findings to three conclusions from Ganz and Flores (2008). First, in the abstract of their article Ganz and Flores concluded that results indicated improvement across all three children in intervals with any speech. Results from the multilevel analysis find no significant effect over the three students. Second, when Ganz and Flores presented the results for this particular outcome in text, they concluded that the treatment was effective for Max but not for the other two children. This is consistent with the results in the first multilevel analysis that the children had significantly different responses to treatment, but the nonsignificant overall treatment effect in the multilevel model cautions us that the treatment effect may not be strong enough in general to be conclusive. Third, Ganz and Flores did not comment

Table 4
Results of Multilevel Model on the Ganz and Flores (2008) Data

Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Odds ratio
		First coding method for alternating treatment designs				
β_{00} intercept	−1.41	0.92	−1.52	2	.27	0.24
β_{10} session	−0.19	0.45	−0.42	2	.72	0.83
β_{20} treatment	1.25	1.45	0.86	2	.48	3.50
β_{30} interaction	0.26	0.49	0.53	2	.65	1.30
Random effect	Standard deviation	Variance component	*df*	χ^2	*p*	
r_{0i} intercept	1.47	2.17	2	10.78	.005	
r_{1i} session	0.63	0.40	2	4.00	.133	
r_{2i} treatment	2.37	5.60	2	20.42	<.001	
r_{3i} interaction	0.72	0.52	2	5.00	.080	
e_{ti}	1.01	1.02				
Fixed effect	Coefficient	Standard error	*t* ratio	*df*	*p*	Odds ratio
		Second (step) coding method for alternating treatment designs				
β_{00} intercept	−1.07	0.73	−1.47	2	.281	0.34
β_{10} session	0.04	0.18	0.20	2	.859	1.04
β_{20} criterion 1	0.40	0.69	0.58	2	.520	1.49
β_{30} criterion 2	0.59	0.92	0.64	2	.586	1.81
β_{30} criterion 3	−0.03	0.99	−0.03	2	.981	0.97
Random effect	Standard deviation	Variance component	*df*	χ^2	*p*	
r_{0i} intercept	1.13	1.29	2	7.61	.022	
r_{1i} session	0.22	0.05	2	1.98	>.500	
r_{2i} criterion 1	0.70	0.49	2	1.47	>.500	
r_{3i} criterion 2	1.21	1.47	2	3.01	.221	
r_{4i} criterion 3	1.41	1.98	2	3.69	.156	
e_{ti}	1.28	1.64				

Note. Values in this table are rounded to two decimals, but computations reported in text, such as conversions of log odds ratios to odds ratios and proportions, were done on results to six decimals. If such conversions are done on the numbers in the table, the results will differ due to rounding error. *df* = degrees of freedom.

on the overall effectiveness of using the changing criterion design. In discussing this outcome variable they did comment on some increases over time consistent with increased criteria; however, they did not discuss any implications of inconsistencies where outcome decreased with increased criteria or increased at a different time than the increased criteria. Results from the multilevel analysis suggest that the changing criteria probably did not systematically change the outcome levels over time.

Discussion

Multilevel models offer attractive features for the statistical analysis of SCDs. They are well developed statistically, provide results that are generally (but not always) consistent with the conclusions of SCD researchers, allow formal tests of how much cases differ from each other through examination of the variance components, and are very flexible in the kinds of data they can address. However, many difficult issues arise in the analysis of even fairly simple-looking data patterns. In this discussion we examine the most salient of those issues in order to suggest the kind of future research that should done in order for multilevel models to reach their full potential in the analysis of SCDs.

Statistical Power

We have located only one direct study of the statistical power of multilevel analyses of SCD data (Jenson et al., 2007). Jensen et al. used computer simulations to test both Type I and Type II error rates in AB and ABAB designs. However, several features of the simulations make them less useful. First, the simulations assumed the data were normally distributed, which is almost never the case in SCDs in the social sciences (Shadish & Sullivan, 2011). Second, the simulations modeled power with 15 cases, 40 cases, and 80 cases. This is more than any of the SCDs published during 2008 in the large Shadish and Sullivan (2011) survey, where the median number of cases was three and the maximum observed was 13. That being said, for 15 cases, power was .47 when the case had five baseline and 10 treatment data points and an autocorrelation of .40 but rose to .76 if the autocorrelation was zero. Increasing the number of baseline (10) and treatment (20) data points increased power to .78 and .99 for autocorrelations of .40 and .00, respectively. Presumably, power would be far lower for more realistic numbers of cases.

The general multilevel modeling literature does contain studies of power. The most common of these concerns power for studies

in which persons are nested within some aggregate like classes (Berkhof & Snijders, 2001), often in the context of a group-randomized trial to test treatment effects (Raudenbush, 1997; Raudenbush & Liu, 2001; Snijders & Bosker, 1993; Zhang & Willson, 2006). In these studies the focus is the power to detect a treatment effect between groups. In general, power in such studies depends on the effect size, the number of clusters in each treatment condition, the number of individuals measured in each cluster, and the intracluster correlation coefficient (Bingenheimer & Raudenbush, 2004). For a fixed budget—that is, where the total number of persons that can be included is fixed and the question is how to divide those persons over aggregates—the general finding is that having more aggregates improves power more than having more persons within each aggregate. Raudenbush and Liu (2001) are the only authors to look at power as a function of frequency of observation over time, assuming a between-groups standardized mean difference of .40. They found that power increased with more time points but appeared to asymptote at about the highest number of time points that they studied; that is, six observations over time for 100 persons in treatment and control groups where power was .43. They concluded that increasing the number of persons improved power more than increasing the number of observations over time. However, six observations over time is considerably shorter than most SCDs (Shadish & Sullivan, 2011).

To the extent that this literature applies directly to SCDs, power in multilevel models in SCDs for a given effect size would depend on the number of cases (n), the number of time points measured within each case (t), the effect size, and the intraclass correlation defined as the ratio of between-case variance to total (within-case plus between-case) variance. Unlike in the between-groups literature, power in SCDs would also depend on the autocorrelation. The recommendation would be that increasing the number of cases will yield more power than increasing the number of time points within cases, although that recommendation is somewhat inconsistent with the intention of SCDs to intensively study relatively few cases over longer times. Further, if that literature applied directly, we might expect a recommendation that the total sample size, $N = t \times n$, would need to fall at least in the range of 500–600, such as 5 cases and 100 time points or 20 cases and 30 time points (Snijders & Bosker, 1993). It might need to be much higher, as much as 500–1,000, using Raudenbush and Liu's (2001) findings. This would be much higher than most SCD studies ever achieve (Shadish & Sullivan, 2011). However, in the group-randomized literature, adding a covariate such as a pretest on the outcome variable can affect power significantly, reducing the total sample size needed to as few as $N = t \times n = 100$ under some conditions (Raudenbush, 1997), such as 5 cases with 20 time points. It may also be the case that adding covariates can improve the power of SCDs, a topic worth further research.

However, past results may not apply well to the SCD context, for several reasons. First, all these studies refer to power to detect a between-groups treatment effect, but SCDs require power to detect a within-person treatment effect. Second, it seems likely that the effect sizes observed in SCDs may be higher than those observed in the typical between-groups cluster-randomized trial. The latter is a highly preliminary suggestion based on recent work to develop a standardized mean difference statistic for SCDs that is in the same metric as those for randomized experiments (Hedges, Pustejovsky, & Shadish, 2012; Shadish, 2012). Larger effects would not be surprising, given that SCD researchers have the luxury of devoting resources intensively to producing a large effect in each case. The likely effect of this is to decrease the total N needed in SCDs, perhaps substantially given that past multilevel power analyses (Raudenbush, 1997; Snijders & Bosker, 1993) have assumed rather small effect sizes. Fourth, we know little about the size of the intraclass correlation in SCDs (as defined above) compared to those that are common in the group-randomized trials literature, and it is difficult to say what the effect of this will be for power.

Most applications of multilevel models to SCDs that we know have found significant treatment effects, suggesting very indirectly that power is not a problem. Unfortunately, that impression may be misleading. Muller et al. (2007) showed that mixed models of the kind used in the present article can have inflated Type I error rates for detecting such effects, especially in small samples and when the error covariance matrix is misspecified. So, the apparent power in all these examples may instead reflect inflated Type I errors. Muller et al. (2007) suggested using multivariate repeated-measures tests instead of mixed models because those tests have very good properties regarding Type I and Type II errors. However, their work was limited to normally distributed data, so we do not know how well it applies to data distributed in other ways like Poisson or binomial, as is usually the case in SCDs.

Finally, all of the preceding discussion refers to the power to detect a treatment effect. Another power issue concerns detecting a significant between-case variance component. In the SCD literature, power to detect between-case variation might be expected to be low, given the small number of cases in most SCDs (Shadish & Sullivan, 2011). That being said, Van den Noortgate and Onghena (2003a) detected significant treatment variance components in a data set with six cases and about 30 time points per case. In the examples in the present article, half the variance components were statistically significant. One might then hypothecate that power is at least not abysmally low to detect variance between cases. Again, the results of Muller et al. (2007) suggest caution in that conclusion, even though Miller et al. did not study power for variances.

Taking all this into account, then, an entire program of research is needed to clarify power when applying multilevel models to SCDs. It should address power to detect treatment effects and variance components, as discussed above, but also to detect covariates at Level 1 or Level 2, more complex phase shifts for the various codings we illustrated in the examples, and how various outcome metrics affect power—all while varying important independent variables such as the number of cases, number of time points, intraclass correlation, and autocorrelation.

Autocorrelation

SCD data consist of consecutive observations over time within cases. As such, the serial correlation of those data points over time must be considered (Weiss, 2005). Huitema (2011) did this in his application of ordinary least squares regression to the analysis of SCDs, testing whether a model with an autoregressive component fits better than one without. We have not done so in the present paper for two reasons. One reason is merely practical, that the HLM computer program cannot do so when the number of cases is smaller than the number of time

points, as will typically be the case in SCD research. Even if it could, it is not clear one could generate good estimates given the small samples involved. The HMLM module within HLM6 can include a first-order autoregressive term in the multilevel model, but we could not get such models to converge in the data sets in this article. Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008) have been more successful in using SAS PROC MIXED to estimate a model with a first-order autoregressive term (note the 2007 article corrects the code from the 2003 articles). However, because most SCD data in psychology and education are either counts or percentages best treated as Poisson or binomially distributed, the proper analysis should use SAS PROC GLIMMIX rather than MIXED. We were unsuccessful in estimating an autoregressive model in GLIMMIX. The WinBUGS computer program will also estimate autoregressive models, and we are currently working on such analyses.

However, a second reason for not modeling an autoregressive model is that the models presented in this paper actually do take into account the correlation between errors, though that may not be immediately apparent (Hedeker & Gibbons, 2006, Chapter 7; Singer & Willett, 2003, Chapter 7). The reason is that a model with random intercept and/or slopes, as was the case for all the models in this paper, induces an error covariance structure for the data. For instance, a model with only a random intercept is essentially like a repeated-measures analysis that assumes sphericity. As one adds more random effects for time varying slopes (time, treatment, interaction), the error covariance structure becomes reasonably complex. At least four issues must be considered in the choice between dealing with serial dependence with autocorrelation, random effects, or both.

First is whether overall model fit, as measured by such criteria as deviance or the various information criteria, much changes over these two choices. Some literature suggests that an autoregressive model and a random effects model tend to fit the data about equally well. Singer and Willett (2003) gave such an example in which the best fitting covariance structure (a Toeplitz structure) fit only 2.2 points better using a Bayesian information criterion (BIC) than the standard random effects error structure we used in this paper, a difference that would be considered weak according to Raftery's (1995) criteria. Of course, determining whether this holds more generally will require further research.

Second is that the researcher is frequently most interested in the fixed effect estimates in the model; for instance, whether the treatment was effective. Those estimates are unbiased under both error covariance structures. Singer and Willett (2003) suggested that for purposes of the fixed effects estimates, refining the error structure is like "rearranging the deck chairs on the *Titanic*" (p. 264).

The third issue is that the precision of the fixed effects estimates can be affected by the choice of error covariance structure. The less accurately the error structure is represented, the more inflated the Type I errors will be (Gurka et al., 2011). On the one hand, ignoring random effects and autocorrelation is clearly problematic. On the other hand, a completely unstructured error covariance matrix will provide the best precision, and a model that allows random intercepts and slopes plus an autoregressive component should also do very well. Estimating such models, however, will often require a large data set from a very well-designed study, and even then estimation may be difficult. In that respect, Hedeker and Gibbons (2006) said that the choice of error covariance structures is "more a matter of avoiding bad models, and selecting a reasonable model from a number of possible alternative models" (p. 124). To judge from Gurka et al. (2011), a model with random intercepts and slopes without an autoregressive component might often be a reasonable model. It tended to produce comparable precision estimates and inferences compared to an unstructured model, an autoregressive model, or the use of sandwich error estimators (which themselves deserve more study in the present context). That being said, so little is known about the different circumstances under which these models will work best that we caution that one cannot simply assume that a random intercept and slopes model will always do well.

The fourth issue is less statistical and more substantive. The random intercept and slopes model implies that the order of the observations is unimportant. In the SCD literature, this might suggest that errors are correlated not because they are adjacent in time but because of some factor within the case that is independent of time. The autoregressive model assumes that the order of observations is important; typically, that observations that are closer to each other are more highly correlated than observations further apart. This might suggest a process that is time dependent rather than person specific. So, the choice between error covariance models should depend on the researcher beliefs about that underlying process. For example, it may be that some children produce more consistent responding than other children, so that it is this between child variability in consistency that is responsible for the correlated errors, not the fact that one observation occurs closer or further away from the other. Of course, we know virtually nothing about these underlying processes for either the random effects or autoregressive cases, but we find both scenarios to be plausible.

Our own belief is that the model should have as many random effects as necessary before trying to model the autocorrelation. Autocorrelation and model specification are very closely related. Omitting a necessary term from the multilevel model, for example, a quadratic trend term at Level 1 or a random effect for slopes, can produce a larger autocorrelation than is really in the data. Because one can never be sure that all the necessary terms have been included, the conundrum is that the researcher can never be sure whether an autocorrelation is due to a true underlying autoregressive covariance structure or to the failure to include the correct fixed effects predictors and random effects in the model. After all fixed and random effects are in the model, adding an autoregressive component tests the right thing, which is the within-subject autocorrelation. Without random intercept and slopes, an autoregressive term will not test a pure within-subject autocorrelation and will be inflated. However, as we said earlier, it may not prove feasible to test models with random effects and autoregressive components, given the sample sizes available to SCD researchers—or indeed to many longitudinal researchers. Hence, it is reassuring that the inferences one obtains are probably plausible as long as one has avoided a badly underspecified error covariance model. Sensitivity analyses are frequently useful in such circumstances, allowing one to see if the effects are consistent over various ways of specifying the model.

More on Outcome Assessment and Data Analysis in SCDs

This article suggests how to model a few of the many kinds of outcomes in SCD research using normal, Poisson, and binomial distributions. For the latter, we have shown cases with constant exposure or constant numbers of trials, but they can also be implemented with variable exposure or trials. An example would be when the researcher changes how many opportunities to respond a case has from session to session, as might happen with an intervention to improve writing where the number of items on a writing assessment varies from session to session. Coding a variable with the number of trials for each session would allow the analysis. Other distributional assumptions for different kinds of SCD outcomes include Bernoulli for binary data and multinomial for outcomes that are either ordered or nominal categories—though the kind of data requiring these models is rare in SCD research. The Bernoulli sampling model applies when each case has only a binary (1/0) response during each session, and the link function is a logit, as in Equation 9. We have not seen such an outcome used in SCD research in surveys of the literature (Shadish & Sullivan, 2011; Smith, 2012). We also have not seen the use of a nominal category as an outcome in SCD research (e.g., whether a case falls into one of several diagnostic categories over time). If such an outcome were to occur, it could be analyzed with a multinomial sampling model with a multinomial logit link function. We have seen the use of ordered categories as outcomes. Jostad, Miltenberger, Kelso, and Knudson (2008), for example, trained children with an intervention designed to prevent them playing with firearms. The outcome was four ordered categories where 0 = touched the firearm, 1 = did not touch the firearm, 2 = did not touch, left the area within 10 seconds, and 3 = did not touch, left the area, and told an adult. The model is multinomial with a cumulative logit link function. Raudenbush and Bryk (2002, Chapter 10) provided details of the model, Raudenbush et al. (2004) showed how the models can be estimated in HLM, and Rindskopf and Ferron (in press) showed an application to SCDs.

Past applications of multilevel modeling to SCDs have not much discussed these matters, certainly not in detail. Nor is the issue specific to multilevel models of SCDs. Proper treatment of outcome metrics is crucial for the other analytic approaches to analysis of SCDs, whether they are ordinary regression, time series analysis, or effect size measures (Houle, 2008; Kratochwill & Levin, 2010; Maggin et al., 2011; Parker et al., 2011; Shadish & Rindskopf, 2007; Shadish et al., 2008). A full treatment of the problems and their solutions is beyond the scope of this article, but we can suggest some of issues that need to be addressed in future research. One is to categorize the many kinds of outcomes used by SCD researchers. For instance, these include frequency counts (e.g., number of words spelled correctly), rate (frequency of behavior divided by time: e.g., average rate of behavior per minute), percent of intervals in which behavior occurred and did not occur (using whole interval, partial interval, or point time sampling methods), duration of behavior (how long a behavior lasts), and latencies (elapsed time from offset of an environment event to the onset of behavior), to name just a few. Examples of such categorizations are in Shadish and Sullivan (2011) and Smith (2012).

This is a start but not enough. Future surveys should code outcome variables guided by the statistical issues pertinent to each. Here is one example. As previously mentioned, using Poisson and binomial models requires some knowledge about the number of trials in each observation session. In the examples we used, the authors reported that explicitly. Sometimes it is not reported. Yet, if we know that percents are based on a constant number of trials, we can usually estimate most important quantities without knowing what that number is. However, with variable numbers of trials (or time periods of measurement), such adjustments are not possible if the author has not reported the number of trials. So, in coding outcomes it would help to distinguish cases that should be treated with a Poisson or binomial sampling model but where the article provides no information about exposure or number of trials.

In addition to considering outcome metric, the categorization should consider the method used for sampling outcome data. Here are examples. Altmann (1974), a primatologist rather than a SCD researcher, provided an informative comparison of the benefits and problems associated with different behavior sampling methods. Mann, Ten Have, Plunkett, and Meisels (1991) discussed different approaches to time sampling to estimate frequency or proportions of time that some behavior occurs, showing that some approaches to time sampling are likely to produce erroneous conclusions. Rapp et al. (2007) compared various methods of continuous duration recording with such approximations as partial interval recording and momentary time sampling, each with different durations, to identify which approximations yielded results closest to the continuous duration recording they viewed as the best method. Rogosa and Ghandour (1991) outlined statistical models for empirical rates of behavior, empirical proportions or relative frequencies of a type of behavior, empirical prevalence (proportion of time the behavior occurs), and empirical event duration. They compared these to continuous sampling, and also discussed sources of unreliability such as finite observation time, recorder errors, and instability of observations over occasions. So, in summary, a good deal of work must to be done to know what SCD researchers are actually doing with outcome measurement in order to provide the best statistical advice about how the resulting data should be analyzed.

Exploratory Analyses

In the examples in the present article, we have not reported many exploratory analyses for two reasons. First, given how little we know about the power of statistical tests in multilevel models of SCDs, it is difficult to have much confidence in the extent to which exploratory or confirmatory results suffer from Type I and Type II errors. Second, especially for studies where power turns out to be low, initially nonsignificant results may encourage the researcher to do exploratory analyses to find significant ones by, for example, testing reduced models in which nonsignificant terms in the main model are dropped to improve overall model fit or searching through a set of predictors to find those that are significant. Just as in all research, the issue is not that exploratory analyses are wrong. Rather, it is that the researcher is obligated to clearly report and distinguish between the primary and exploratory analyses and that interpretation of such exploratory analysis must be quite cautious. Third, just as is often the case in other literatures (e.g., Francis, 2012; Ioannidis, 2005, 2008; Ioannidis & Lau, 2001; Ioannidis & Panagiotou, 2011; Ioannidis & Trikalinos, 2007; Kyzas, Loizou, & Ioannidis, 2005; Renkewitz, Fuchs, & Fiedler,

2011; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), certain biases in the conduct of science, such as confirmation biases and excitement about new treatments that might really be Type I errors, can too easily result from those exploratory analyses. The issue is complex, as the following examples will show.

Consider the exploratory analysis in the Resetar and Noell (2008) example in which we examined whether the treatment was more effective for two of the cases than for the remaining cases. Resetar and Noell used visual analysis to conclude that this was the case. The statistical analysis did not support that conclusion, with the results of treatment not being significantly different for the two sets of cases. Both are exploratory analyses, of course, but one can see the interpretational dilemma. Given how little we know about power in these analyses, is the nonsignificant statistical finding merely a result of low power? Did Resetar and Noell (2008) engage in the visual equivalent of fishing for results, given that the treatment in that study did not seem to have an overall effect?

Consider also the second exploratory analysis in this article, the reduced model analysis of the Lambert et al. (2006) data. The motivation for that analysis was to simplify the interpretation of a very complex set of two- and three-factor interactions (where power was obviously not an issue, as they were significant). It was not to fish for significant results; nor was it to eliminate nonsignificant results from the model in a backwards stepwise fashion. Further, the results of the reduced model were consistent with the full model in general tenor, though the full model implied that the effects of treatment depended on session and on whether the observation occurred in the first or second AB phase. For parsimony, some researchers will prefer the reduced model, but a good argument can be made that the full model more accurately reflects the complexity that visual analysis suggests may be present in that data set. Is this kind of exploratory analysis a good thing?

Finally, common practice in multilevel modeling is to start with an unconditional model and then add predictors at each level to account for the variability implied when some variance components are significant. This may be less well conceptualized as exploratory analysis than it is as using the unconditional model to establish a baseline regarding whether such variability even exists. Those predictors, however, should be planned where possible and acknowledged as exploratory otherwise. Is this to be encouraged or not?

Baer (1977) argued that SCD researchers prefer "very low probabilities of Type 1 errors, and correspondingly high probabilities of Type 2 errors" (p. 167); that is, they are willing to overlook potentially effective treatments with small effects because they are interested in detecting powerful treatment effects. Such SCD researchers may also believe that statistical analyses detect weak treatment effects that are not practically important. Two issues arise. One is that SCD researchers could use this as an argument for more extensive use of exploratory analyses in SCDs. Even if that were the case, however, one might distinguish between analyses that do or do not make an obvious (even if exploratory) contribution to theory or behavior technology development. For example, the example from Resetar and Noell (2008) does not seem to make any such contribution, but it might have done so if it had been tied to some characteristics of the cases of theoretical interest, such as age or a pertinent ability.

The other issue is whether it is factually true that statistical analyses detect weak treatment effects that are not practically important. To judge from the six examples in this paper, the statistical analyses tended to agree with the authors' conclusions most of the time, even though the analyses presented a more nuanced view of the results. In the two cases where the results disagreed, what happened is the opposite of Baer's (1977) claim. The statistical analysis found no significant effect when the authors concluded otherwise. Submitting a much larger sample of SCD studies to statistical analyses might provide better evidence about Baer's concerns. That being said, having this discussion is difficult until formal power analyses for SCD analyses are available so that we can tell which analyses are likely to produce Type I and II errors under which circumstance.

Other Single-Case Research Designs

This paper shows how to apply multilevel models to the four most common designs used in SCD research—phase change designs, multiple baseline designs, alternating treatment designs, and changing criterion designs—which together accounted for 74% of the designs found in Shadish and Sullivan's (2011) survey of the SCD literature. The remainder were designs that combined two or more of these basic four. An example would be a study that started with a simple baseline phase but then moved to an alternating treatment design at the phase change, thus combining phase change and alternating treatment designs. The only example of such more complex designs in the previous examples is the Resetar and Noell (2008) study, which used an alternating treatment design with three rather than the usual two conditions. In general, it is possible to analyze all of these with multilevel models, but the details of each analysis will depend on the idiosyncrasies of the exact design used in each study. Finally, some SCDs use some form of random assignment of conditions to time (Kratochwill & Levin, 2010). Most of these designs will fall into one of the four basic design types, so the analyses will not be novel, even though the benefits of randomization for the logic of causal inference may be substantial.

Other Multilevel Modeling Statistical Programs

Multilevel modeling of SCD research has been done with HLM in the present article and with SAS PROC MIXED by Van den Noortgate and Onghena (2003a, 2003b, 2008). Examples of other programs that can do the same thing include but are not limited to the SPSS Mixed command (Peugh & Enders, 2005), the Stata .xtmixed command (Rabe-Hesketh & Skrondal, 2005), and the R lme4 and nlme packages, as well as specialty programs such as Latent GOLD, Mplus, MLwiN, SuperMix, and WinBUGS. Each of these programs will undoubtedly prove to have its own advantages and disadvantages, but no one has yet cataloged them. For example, one disadvantage of HLM is the previously mentioned difficulty in using autoregressive models. Another is that HLM uses the between/within method of computing degrees of freedom based on counts of the variables and observations, whereas SAS PROC MIXED offers a wider array of options (Satterthwaite, containment, Kenward/Roger, between/within, residual). Simulations by Ferron et al. (2009) suggest that Satterthwaite and Kenward/Roger are more accurate than the other methods, although

their simulation did not include time as a variable in the regression, instead assuming no trend and no treatment by trend interaction. However, none of these methods are perfect with unequal number of observations per subject. It would be useful to investigate a Bayesian approach using Markov chain Monte Carlo, which might give more accurate results in spite of small samples and unequal sample sizes per respondent. SAS PROC MIXED uses a default option of a *t* test to test whether variance components are nonzero, but it includes an option to use a more appropriate chi-square test that is the default in HLM. Compared to Stata, HLM allows more covariance structures to be modeled, and it allows estimation with penalized quasi-likelihood generalized linear models or higher order Laplace for some generalized linear models.

Another issue concerns the appropriateness of the Poisson distribution that we used to model count data. In theory, other distributions might be more appropriate for this kind of data. For example, for cases where overdispersion is present, a negative binomial model allows for correction of overdispersion. A generalized negative binomial model allows the dispersion to vary observation by observation. A generalized event count model allows for modeling both over- and underdispersion. Zero-inflated models can be appropriate when the data contain an excessive number of zeros, which is not an uncommon experience in SCD research. HLM does not allow any of these options at present. Mplus allows fitting a negative binomial model. No program we know of, however, allows estimation of a multilevel zero inflated negative binomial model. So, in view of all these examples, a thorough review of the strengths and weaknesses of the various multilevel modeling programs for use in SCD research would be highly valuable.

Multilevel Modeling and the What Works Clearinghouse Standards

We noted earlier that the What Works Clearinghouse (WWC) has developed preliminary standards for SCDs to meet evidence standards about what treatments are effective (Kratochwill et al., 2010). For example, in a multiple baseline design, the standards call for having three cases (in order to provide three opportunities to demonstrate the effect) and five time points in both baseline and treatment. However, minimally meeting these criteria would result in a set of SCDs that is smaller than any of the examples used in this article. Because we know so little about the power of multilevel models in SCDs, we cannot be sure that studies that meet the WWC criteria could feasibly be analyzed with multilevel models.

Fortunately, most SCDs are on average larger than the minimum WWC criteria require to meet evidence standards. Both Shadish and Sullivan (2011) and Smith (2012), for example, found that the average SCD had more cases and more time points than the minimum required by the SCD. In addition, the WWC standards explicitly state that no consensus exists about what is the best analytic method for SCDs. Presumably, as more consensus on that matter emerges, the WWC standards may be changed to encourage sufficiently large SCD studies to allow proper data analysis—though the SCD community is quite diverse, and not all SCD researchers agree that statistical analysis is a desirable goal (Shadish et al., in press). Last, we know very little about the power and other characteristics of all the other analytic methods that have been proposed for SCDs, so the same dilemma applies to them as well. In all these respect and more, the question of data analysis for SCDs is an extremely fertile one for interested researchers.

References

Agresti, A. (1996). *Categorical data analysis*. New York, NY: Wiley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour, 49,* 227–266. doi:10.1163/156853974X00534

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10,* 167–172. doi:10.1901/jaba.1977.10-167

Berkhof, J., & Snijders, T. A. D. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26,* 133–152. doi:10.3102/10769986026002133

Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Review of Public Health, 25,* 53–77. doi:10.1146/annurev.publhealth.25.050503.153925

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., & Muller, K. E. (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine, 29,* 504–520. doi:10.10002/sim.3775

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579. doi:10.1901/jaba.1979.12-573

DiCarlo, C. E., & Reid, D. H. (2004). Increasing pretend toy play of toddlers with disabilities in an inclusive setting. *Journal of Applied Behavior Analysis, 37,* 197–207. doi:10.1901/jaba.2004.37-197

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41,* 372–384. doi:10.3758/BRM.41.2.372

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19,* 151–156. doi:10.3758/s13423-012-0227-9

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care, 49,* 761–768. doi:10.1097/MLR.0b013e318215d90d

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders, 37,* 55–77.

Ganz, J. B., & Flores, M. M. (2008). Effects of the use of visual strategies in play groups for children with autism spectrum disorders and their peers. *Journal of Autism and Developmental Disorders, 38,* 926–940. doi:10.1007/s10803-007-0463-4

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine, 30,* 2696–2707. doi:10.1002/sim.4293

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York, NY: Wiley.

Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods, 3,* 224–239. doi:10.1002/jrsm.1052

Houle, T. T. (2008). Statistical analysis of single-case experimental designs. In D. Barlow, M. Nock, & M. Hersen (Eds.), *Single-case experimental designs* (3rd ed., pp. 285–324). Chicago, IL: Allyn & Bacon.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies* (2nd ed.). New York, NY: Wiley. doi:10.1002/9781118067475

Hunt, P., Soto, G., Maier, J., & Doering, K. (2003). Collaborative teaming to support students at risk and students with severe disabilities in general education classrooms. *Exceptional Children, 69,* 315–332.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2,* e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Perfect study, poor evidence: Interpretation of biases preceding study design. *Seminars in Hematology, 45,* 160–166. doi:10.1053/j.seminhematol.2008.04.010

Ioannidis, J., & Lau, J. (2001). Evolution of treatment effects over time: Empirical insight from recursive cumulative meta-analyses. *Proceedings of the National Academy of Science USA, 98,* 831–836. doi:10.1073/pnas.98.3.831

Ioannidis, J. P. A., & Panagiotou, O. A. (2011). Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA: Journal of the American Medical Association, 305,* 2200–2210. doi:10.1001/jama.2011.713

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. doi:10.1177/1740774507079441

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analysis, and single subject designs. *Psychology in the Schools, 44,* 483–493. doi:10.1002/pits.20240

Jostad, C. M., Miltenberger, R. G., Kelso, P., & Knudson, P. (2008). Peer tutoring to prevent firearm play: Acquisition, generalization, and long-term maintenance of safety skills. *Journal of Applied Behavior Analysis, 41,* 117–123. doi:10.1901/jaba.2008.41-117

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from the What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124–144. doi:10.1037/a0017736

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65,* 73–93. doi:10.1080/00220973.1996.9943464

Kyzas, P. A., Loizou, K. T., & Ioannidis, J. P. (2005). Selective reporting biases in cancer prognostic factor studies. *Journal of the National Cancer Institute, 97,* 1043–1055. doi:10.1093/jnci/dji184

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8,* 88–99. doi:10.1177/10983007060080020701

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301–321. doi:10.1016/j.jsp.2011.03.004

Mann, J., Ten Have, T., Plunkett, J. W., & Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development, 62,* 227–241. doi:10.2307/1130999

McCoach, D. B., & Kaniskan, B. (2010). Using time-varying covariates in multilevel growth models. *Frontiers in Psychology, 1,* Article 17. doi:10.3389/fpsyg.2010.00017

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647–653. doi:10.1901/jaba.1974.7-647

Muller, K. E., Edwards, L. J., Simpson, S. L., & Taylor, D. J. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine, 26,* 3639–3660. doi:10.1002/sim.2827

Nagler, E., Rindskopf, D., & Shadish, W. (2008). *Analyzing data from small N designs using multilevel models*: A procedural handbook. Unpublished manuscript.

Olive, M. L., & Smith, B. W. (2005). Effect size calculation and single subject designs. *Educational Psychology, 25,* 313–324. doi:10.1080/0144341042000301238

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303–322. doi:10.1177/0145445511399147

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single-subject research* (pp. 101–165). New York, NY: Academic Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement, 65,* 717–741. doi:10.1177/0013164405278558

Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. doi:10.2307/271063

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions, 22,* 319–345. doi:10.1002/bin.239

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6,* 387–401.

Renkewitz, R., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6,* 870–881.

Resetar, J. L., & Noell, G. H. (2008). Evaluating preference assessments for use in the general education population. *Journal of Applied Behavior Analysis, 41,* 447–451. doi:10.1901/jaba.2008.41-447

Rindskopf, D., & Ferron, J. (in press). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances. Washington,* DC: American Psychological Association.

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics, 16,* 157–252. doi:10.2307/1165191

Shadish, W. R. (2012, June). *Meta-analysis of single-case designs*. Address presented at the meeting of the Society for Research Synthesis Methodology, Aix en Provence, France.

Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods, 41,* 177–183. doi:10.3758/BRM.41.1.177

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., Hedges, L. V., Pustejovsky, J., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (in press). Analyzing single-case designs: d, G, multilevel models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances. Washington,* DC: American Psychological Association.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation method: Building the evidence base for method choice in government sponsored evaluation* (pp. 95–109). San Francisco, CA: Jossey-Bass. doi:10.1002/ev.217

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2,* 188–196. doi:10.1080/17489530802581603

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2012). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012–0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43,* 971–980. doi:10.3758/s13428-011-0111-y

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195152968.001.0001

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*. Advance online publication. doi:10.1037/a0029312

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237–259. doi:10.2307/1165134

Stuart, R. B. (1967). Behavioral control of overeating. *Behaviour Research and Therapy, 5,* 357–365. doi:10.1016/0005-7967(67)90027-7

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18,* 325–346. doi:10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35,* 1–10. doi:10.3758/BF03195492

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today, 8,* 196–209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2,* 142–151. doi:10.1080/17489530802505362

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way that they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

Weiss, R. (2005). *Modeling longitudinal data*. New York, NY: Springer-Verlag.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: Chapman & Hall.

Zhang, D., & Willson, V. L. (2006). Comparing empirical power of multilevel structural equation models and hierarchical linear models: Understanding cross-level interactions. *Structural Equation Modeling, 13,* 615–630.

Received April 4, 2011
Revision received March 9, 2013
Accepted April 7, 2013 ■