Journal of School Psychology 52 (2014) 179–189

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

# Nonlinear Bayesian analysis for single case designs☆

David Rindskopf

*Educational Psychology, CUNY Graduate Center, 365 Fifth Avenue, New York, NY 10016, United States*

ARTICLE INFO

*Article history:*
Received 28 July 2013
Received in revised form 13 December 2013
Accepted 14 December 2013
Available online 24 January 2014

*Keywords:*
Bayesian
Multilevel
Nonlinear
Single case
Single subject

ABSTRACT

Several authors have suggested the use of multilevel models for the analysis of data from single case designs. Multilevel models are a logical approach to analyzing such data, and deal well with the possible different time points and treatment phases for different subjects. However, they are limited in several ways that are addressed by Bayesian methods. For small samples Bayesian methods fully take into account uncertainty in random effects when estimating fixed effects; the computational methods now in use can fit complex models that represent accurately the behavior being modeled; groups of parameters can be more accurately estimated with shrinkage methods; prior information can be included; and interpretation is more straightforward. The computer programs for Bayesian analysis allow many (nonstandard) nonlinear models to be fit; an example using floor and ceiling effects is discussed here.


## 1. Introduction

Researchers recognize that data from a single case design (SCD) are structured so that multilevel modeling: (see, e.g., Raudenbush & Bryk, 2002) is the natural way to analyze the observations nested within individuals (presuming there are three or more cases, as is typical). Although multilevel modeling provides the basic structure, simple linear models are not always adequate, nor are typical estimation methods as useful as Bayesian methods.

Many SCD studies require nonlinear models either because of the nature of the dependent variable (e.g., counts) or because there is a gradual change in behavior from one phase to another that requires a nonlinear curve. In the first case, the extension to nested data of generalized linear models (GLMs) instead of linear models handles the usual data types, such as counts out of a total (requiring a binomial or more complex distribution) or counts for a fixed time (requiring a Poisson or more complicated distribution). In the case of continuous variables that require a nonlinear curve, or additional complications in generalized linear models such as floor and ceiling effects, nonlinear models are needed that are not standard in form nor implemented in standard computer software.

Because most SCDs consist of only a few cases (typically three to nine), the methods developed for large samples may be inappropriate. Bayesian inference and computations offer advantages in this case. One is that they are exact for small samples; they do not require large numbers of cases. Another is that they implement shrinkage estimators that allow better estimation for each case while making use of information from the other cases. Bayesian statistics also de-emphasize null hypothesis tests and allow more natural probability statements than classical statistical methods (while still allowing more traditional interpretations, as will be illustrated in this article).

This article will explain these issues in more detail and will illustrate the use of nonlinear Bayesian estimation in two case studies, one of each type (GLM and intrinsically nonlinear model). To provide the background for discussing these examples, the next sections discuss multilevel models for SCDs, present an overview of Bayesian inference, and offer basic information about nonlinear models (including generalized linear models.)

☆ The research reported here was supported by Grant No. R305D100046 from the Institute of Education Sciences, U.S. Department of Education. The opinions expressed are those of the author and do not represent the views of the Institute of Education Sciences or the U.S. Department of Education.
*E-mail address:* drindskopf@gc.cuny.edu.
ACTION EDITOR: William Shadish.

0022-4405/$ – see front matter © 2013 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jsp.2013.12.003

## 2. Multilevel models for SCDs

In a typical SCD, each case has a number of observations in a time series, and the cases go through a number of different phases, typically lasting for 3 to 10 sessions (periods of observation). The number of phases may be as small as 2 (for an AB design, or for multiple baseline designs) or as many as 6 to 10 or more for more complex designs. Different cases may have the same structure of phases, or different phase structures; further they may have different numbers of observations in different phases. Some cases may also have missed certain periods of observation due to illness or other problems, causing missing data in the planned pattern of data collection. All of these complications can make the modeling situation difficult.

The multilevel approach tries to model the behavior (the dependent variable) of each case as a function of the phase and possibly the time of measurement. The behavior of each case is then analyzed to see how much variation there is across cases, and if there is much variation whether it can be accounted for in terms of the cases' characteristics. This viewpoint generates two types of equations, one for the individual case's behavior, the other summarizing and accounting for variability across cases.

We start with the simplest possible design, a multiple baseline design. In this design everyone goes through two phases, a baseline phase (A) and a treatment phase (B). The treatment phase starts at a different time for each person, so that history is eliminated as a threat to internal validity. When the first person goes from phase A to phase B, we should see a change in only that person's behavior and not the behavior of other persons. Similarly, when each successive person changes from phase A to B, no change should occur in the behavior of other persons. A complete statistical model for this design would test all phase changes, but for expository purposes, I will use a simpler model that only tests for changes in phase for each person.

Suppose that the dependent variable is continuous and normally distributed and that there is no time trend in either phase. Therefore, except for residual variation, the behavior is flat during baseline and immediately changes to a different (constant) level during the treatment phase. For each case $i$ and time $t$, the model for observations for that case is

$$y_{it} = \beta_{0i} + \beta_{1i} Phase_{it} + r_{it} \tag{1.1}$$

where $y_{it}$ is the dependent variable for case $i$ at time $t$, $\beta_{0i}$ is the level at baseline for case $i$, $\beta_{1i}$ is the change between baseline and treatment phase for case $i$, $Phase_{it}$ is an indicator (dummy) variable that is 0 during phase A and 1 during phase B, and $r_{it}$ is a residual. The variance of $r_{it}$ is $\sigma^2$. Notice that each case can have a different baseline and a different amount of change and that the interpretation of $\beta_{0i}$ and $\beta_{1i}$ depends on the coding of $Phase_{it}$ and would change if the coding was changed, for example, to effect coding (see, e.g., Shadish, Kyse, & Rindskopf, 2013).

The minimal equations at the case level are

$$\beta_{0i} = \gamma_{00} + u_{0i} \tag{1.2}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}. \tag{1.3}$$

The variance of $u_{0i}$ is $\tau_{00}$, the variance of $u_{1i}$ is $\tau_{11}$, and their covariance is $\tau_{01}$. These equations state that the baseline and change for each case are a constant (average value) plus some deviation from the average. In other words, none of the variation around the average is explained. If there are enough cases, and some reasonable predictor variable is available (perhaps the baseline or effect of treatment varies with age or sex), this predictor variable can be added to one or both equations as needed to become, for example,

$$\beta_{0i} = \gamma_{00} + \gamma_{01} \mathrm{Age}_i + u_{0i} \tag{1.4}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} \mathrm{Age}_i + u_{1i}. \tag{1.5}$$

Generally, age would be centered or re-expressed in some other way to keep the interpretation of the intercept in each equation meaningful. (The meaning of the intercept terms will change unless Age is centered around its mean, so care must be used in comparing the models without and with Age.)

Because the data from SCDs are in the form of time series, autocorrelation is a potential problem. This issue is not discussed in detail here, partly because techniques for dealing with autocorrelation are still under development for nonlinear models. It is also problematic to estimate autocorrelations with such short time series as are common in SCDs.

More detailed information on multilevel modeling of SCD data in the non-Bayesian context can be found in Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008), Shadish et al. (in press) and Rindskopf and Ferron (in press), whereas Rindskopf (in press) provides an overview of the Bayesian approach.

## 3. Bayesian inference

### 3.1. Conceptual basis of Bayesian statistics

Bayesian modeling begins with specification of one or more parameters (that is, unknowns whose value about which the researcher wishes to make inferences). In the case discussed here, the parameters will be regression coefficients ($\gamma_{ij}$) and variances ($\tau_{ij}$) or standard deviations or precisions (i.e., the inverse of variances). Before conducting the study, the researcher may have some knowledge of the values of these parameters; this knowledge is expressed in a probability distribution that describes

the researcher's beliefs about what values of the parameter are more and less likely. If the researcher is completely ignorant about a parameter, then the prior distribution is chosen to be noninformative or flat (on some scale); all values are equally likely. The prior distributions are often chosen subjectively, which is viewed by classical statisticians (also called frequentists) as a weakness. But in many cases prior studies or pretest data provide information about the parameters, and if so we should be able to make use of this information. For this article, I will use prior distributions that are nearly uninformative, so that the data, not the prior, affects the inferences.

The prior information is combined with the information in the data (summarized by the likelihood) to create the posterior distribution, which captures our knowledge of the parameters at the end of the study. In the case of non-informative priors, the posterior distribution is totally determined by the likelihood (that is, the information in the data). The posterior distribution is an actual probability distribution, expressing beliefs about the relative probability of different values of the parameter. Like any other probability distribution, areas under the curve (for a continuous-valued parameter) represent probabilities.

When parameters are estimated for a group of cases, as happens here with the baseline and change parameters, one typically has two choices: Constrain the parameters to be equal across cases or estimate the parameter for each case separately with no relationship across cases. If the parameters are constrained to be equal, there is no allowance for individual differences, and the grand mean is used as the estimate for each case, which usually will be erroneous. If they are estimated separately for each case, no use is made of information from other cases, which is a mistake if the cases have something in common. Bayesian statistics is in between these two extremes, using information from all cases but not constraining all cases to be the same. The estimates are shrunk from the individual value for each case toward the mean, with more shrinkage if (a) there is evidence that the cases are very similar to each other or (b) the individual has little data; and less shrinkage if the evidence points to great differences among cases or there is a large amount of data for the individual. (Shrinkage is toward the mean of similar, rather than all, people when the model includes individual-level variables, such as Age in the example above.)

### 3.2. Advantages of Bayesian methods

Bayesian methods have several advantages in the analysis of SCD data (and more generally). Specific to SCDs, Bayesian methods do not require large sample size (either in terms of number of cases or number of observations per case). Because most SCDs consist of few cases, this advantage is relevant for almost all such studies.

More generally, Bayesian statistics considers parameters (unknowns) to have a probability distribution, whereas classical statistics considers parameters as fixed quantities. Classical inference, therefore, depends on the concept of repeated sampling from a population and bases inferences on the sampling distribution of statistics. In Bayesian statistics one can make probability statements such as "The probability that the mean is greater than 0 is .93", whereas in classical statistics such statements cannot be made. (In classical statistics, the mean is a specific value; we just do not know which value it is. Probability statements apply to repetitions of the study, not to the single study being examined. A classical, or frequentist, interpretation is that in a large number of repetitions of the study, 95% of the confidence intervals would contain the parameter.) The Bayesian interpretation seems more natural to most people—so natural that the most common interpretation of a confidence interval ("There is a 95% chance that the parameter is in this interval") is the Bayesian one, which is wrong from the classical point of view.

Bayesians estimate a credible interval that is comparable to a frequentist's confidence interval in many ways, except for the interpretation and the dependence on the prior distribution. But Bayesians calculate other quantities that are not used (and indeed cannot be used) in classical statistics. For example, a Bayesian can calculate the probability that a treatment effect is positive, that it is larger than some given size, or that it is small (between $-c$ and $c$, where the value of $c$ is chosen to be a small effect). One can also estimate the probability that one parameter is larger than another; one use for this method is to estimate the probability that the treatment was more effective for one person than another. One can also estimate the probability that one parameter is larger than another by at least a certain amount or that the (absolute value of the) difference between two parameters is small (i.e., less than a given amount).

Although Bayesians generally do not test hypotheses about parameters, their results can frequently be expressed in those terms so that those who are used to classical results can feel comfortable. Thus, interpreting a parameter divided by the standard error as "significant" when the ratio is larger than 2 is similar to what frequentists do when calculating $z$ or $t$ ratios. Also, many Bayesians mix Bayesian interpretation of parameter estimates with classical methods for model choice, such as when determining which terms to keep in a regression equation.

Other advantages of Bayesian statistics come from the computational methods that are generally used, called Markov Chain Monte Carlo (MCMC). Although the details are irrelevant (and beyond the scope of this article), the method allows one to take a sample of any size from the posterior distribution of the parameters of the model. This method makes it easy to estimate any function of parameters: Just specify the function, and you get a sample from that function as a consequence. This procedure also means that you need not rely on the central limit theorem to get a confidence interval; in most cases, just look at the 2.5 and 97.5 percentiles of the empirical distribution from the sample produced by MCMC and you will have a 95 percent confidence interval. (The exception is extremely skewed variable distributions, such as those for variances that are close to 0; these require more care.)

Bayesian inference has a subtle advantage in the analysis of multilevel data from studies with few cases. The estimates of the fixed effects the parameters $\gamma_{ij}$ in Eqs. (1.2) and (1.3) and their standard errors are affected by the uncertainty in estimates of the random effects (the parameters $\sigma^2$ and $\tau_{ij}$). If the random effects are estimated well, which happens with approximately 20 or more cases, the uncertainty in their estimation will not influence the estimates of the fixed effects or their standard errors. But with few cases, the effects may be larger. In most estimation methods currently used for multilevel data, the fixed effects are estimated based on the point estimates of the random effects, which could be problematic in the situation where there are few

cases. Fully Bayesian estimates of the fixed effects are not subject to this problem and are estimated by averaging over the various values that the random effects could take on. (They are the marginal averages rather than the averages conditional on the point estimates of the random effects.) The Bayesian fixed effects will generally have larger standard errors than empirical Bayes estimates (used in most multilevel model programs) and thus wider confidence intervals. Empirical Bayes estimates will be significant too often because their standard errors are underestimates.

For further information on Bayesian statistics, the basics are contained in Winkler (2002). A more advanced introduction, requiring some statistical knowledge, is Lee (2012). A moderately advanced book, with applications and computing, is Gelman et al. (2013).

## 4. Nonlinear models (and GLMs)

Many SCDs have a dependent variable that is a count, such as the number of times a student initiates a social interaction or the number of times (out of a particular number of observations) that a student displays disruptive behavior. These variables are not usually close enough to normally distributed to use the usual linear additive model. The simplest distributions for these situations are the Poisson (for a count over a period of time) and the binomial (for a count out of a certain fixed number of trials). There are other distributions for counts, such as the negative binomial, which can be used in more complicated cases.

In these cases, the dependent variable is not a linear function of the predictors but some nonlinear transform of the dependent variable. In the case of the Poisson distribution, the natural transform is the logarithm, so that the equation for observations is

$$\ln(\eta_{it}) = \beta_{0i} + \beta_{1i} Phase_{it} \tag{1.6}$$

where $\eta_{it}$ is the expected value of $y_{it}$. We then specify that $y_{it}$ has a Poisson distribution with parameter $\eta_{it}$.

For the binomial distribution, the transform is the logarithm of the odds of the behavior occurring on a trial, and the equation for observations is

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i} + \beta_{1i} Phase_{it}. \tag{1.7}$$

The variable $y_{it}$ has a binomial distribution with mean $n_{it}\pi_{it}$ where $n_{it}$ is the number of trials for that observation, and $\pi_{it}$ is the probability of a response on a given trial. (Often there will be only a single value $n$ for the whole study, but that is not necessary.)

These forms of the equation for observations are nonlinear—but only on the left hand side; the right hand side is the usual linear additive model. These models are called generalized linear models, and except for the transformation and the distribution of the variable, have some of the same characteristics as linear regression models.

If time is not included as a predictor variable (i.e., if the slope is zero within phases) then these models, while technically nonlinear, represent straight flat lines. If time is included as a predictor, these models are not linear but represent curves. For some data (including continuous normal data), this nonlinearity is necessary, because often the change in behavior when the phase changes is not sudden, but gradual, and this gradual change requires a curve rather than a straight line. (Note that a polynomial of small degree will not fit this type of curve.) Bayesian software based on Bayesian inference Using Gibbs Sampling (BUGS; Lunn et al., 2009), such as WinBUGS (Lunn et al., 2000), OpenBUGS, and JAGS (Plummer, 2003; JAGS stands for "Just Another Gibbs Sampler"), allows these models to be easily fit.

**Example 1.** Bayesian generalized linear model for Lambert et al. (2006) data.

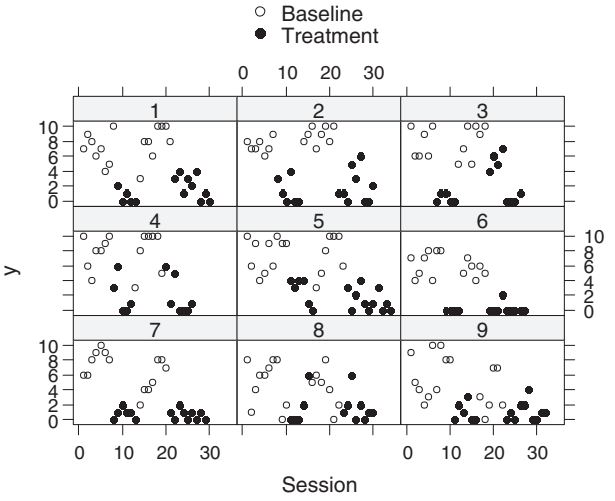


**Fig. 1.** Plots of data from students in article by Lambert et al. (2006). The design was an ABAB design, and each student was observed at 10 time periods per session. The dependent variable (y) was the number of times (out of 10) that the student behaved disruptively.

Lambert et al. (2006) studied the effects of using response cards on the disruptive behavior and academic responses of students during math lessons. Nine students in the fourth grade participated in the study; each was part of one of two larger classrooms. These students had been selected as particularly disruptive students. The design was an ABAB design. The A phase was the usual classroom procedure where the teacher asked a question and called on one student in the class to respond, and the B phase involved the use of cards on which all the students wrote answers to the teacher's questions. During each observation session, each student was observed during 10 15-second periods during which the student either did or did not display disruptive behavior. Thus, each day every student had a score ranging from 0 to 10 on the number of periods of disruptive behavior. The main interest is the proportion of times the student was disruptive.

Because the count was based on a fixed number of trials per day (10), the count is modeled as a binomial distribution with an underlying probability π of showing disruptive behavior on a given trial. Although the model can be made to handle trend lines and differences between the first AB sequence and the second AB sequence, an examination of the plot in Fig. 1 shows that these effects, if they exist at all, would be small compared to the major effect, which is the difference between the A and B phases. (Of course, more detailed analyses may find interesting patterns in these data; see Shadish, Zuur, & Sullivan, 2014–in this issue) Therefore, for illustrative purposes the model I will use is the simple one with no trend and one effect for the difference between A and B phases:

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i} + \beta_{1i} Phase_{it} \tag{1.8}$$

which is the same as Eq. (1.7). The intercept of this equation $\beta_{0i}$ is the log-odds of a disruptive behavior during the baseline (A) phase, and the slope $\beta_{1i}$ is the change in log-odds between the A phase and the B phase. The log-odds of a disruptive behavior during phase B is therefore the sum $\beta_{0i} + \beta_{1i}$. Because most people do not interpret log-odds easily, we will transform back to proportions using the inverse transform exp(x) / (1 + exp(x)), where x is the log-odds.

I will present two versions of the model. The first has no predictors at the individual level, and the second has the classroom as a predictor.

### 4.1. Model 1: no case-level predictors

The equations for individuals for the first model are the same as Eqs. (1.2) and (1.3):

$$\beta_{0i} = \gamma_{00} + u_{0i} \tag{1.9}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}. \tag{1.10}$$

For those who wish to understand the WinBUGS computer code that represents the model, I will describe the main part of it in the next few paragraphs; those who wish can skip these parts without loss of continuity.

The data are arranged in a single file containing values of all the variables, both at the observation level and the person level. Each subject's data is written with one line for each session (time of observation). Each line contains the subject number, the count (dependent variable), a dummy variable for the phase, and the classroom (used in the next model.)

The most important part of the program code for the model is contained in the following lines:

```
model
 { for (i in 1:264){                  # 264 total observations
    logit(theta[i]) <- b0[subj[i]] + b1[subj[i]]*phase[i]# level 1 model
     y[i] ~ dbin(theta[i],10)         # count outcome modeled binomially }
 for (j in 1:9){                       # 9 cases
    b0[j] ~ dnorm(mu0,prec0)          # intercept treated as a random effect
    b1[j] ~ dnorm(mu1,prec1)          # treatment treated as a random effect }
var1 <- 1/prec1                       # var in treatment effect
var0 <- 1/prec0                       # var in baseline
```

The second line shows that the two lines that follow pertain to all 264 observations (an average of about 29 observations for each of the 9 respondents). The next line specifies the equation for observations; the logarithm of the odds, called the logit, is a built-in function in the program. Note that *theta* is used to represent a probability instead of *pi*. The term b0[subj[*i*]] specifies the

b0 value (intercept) for the subject on which observation *i* was made. This structure is known as nested indexing, because it is not the b0 value for observation *i* but the subject having that observation. The same applies to the term b1[subj[*i*]], which represents the difference between the A and B phase for the subject on which observation number *i* was made. The statement following says that the observed dependent variable *y* for observation *i* has a binomial distribution with probability *theta*, out of a total of 10 trials.

The next set of lines is in a "for" loop over the range of individuals rather than observations. The lines of code show that the intercepts and slopes (treatment effects) have a mean and a variation across students. The normal distribution in Bayesian statistics specifies the precision = 1/variance, rather than the variance or standard deviation, for reasons that need not concern us. (Precision is used in most of Bayesian statistics.) We can calculate the variance or standard deviation from the precision, as the statements following the loop illustrate.

Table 1 contains some of the results. The computer programs used for modern Bayesian computation generally implement an algorithm that generates a series of draws from the posterior distribution of the parameters (and functions of them); in this case, we have a sample of 5000 points from the posterior distribution. For each parameter or function of parameters there are several summaries in the default output, including the mean and standard deviation (equivalent to the standard error) of the 5000 draws and three percentiles: the 2.5, 50th (median), and 97.5 percentile. The mean and median are used as point estimates of the parameter (if they differ much, it indicates that the posterior is skewed, so more care is needed, including examination of a plot of the posterior distribution). Confidence intervals, called *credible intervals* in Bayesian statistics, can be either parametric (mean plus or minus 1.96 standard errors for a 95 percent interval, if the sample size is large and the posterior is approximately normal) or nonparametric (2.5 and 97.5 percentile points of the empirically generated sequence of values from the posterior distribution).

In the table *mu0* is the average baseline (phase A) logit (log-odds) of a disruptive behavior during an observation period; as a logit of 0 is equivalent to a probability of .5, we know that mu0 = .83 indicates a much greater than 50% chance of disruptive behavior during baseline. Luckily it is easy to create a variable (p.A; see complete code in Appendix 1) that tracks the probability directly; the mean of p.A is estimated to be .69.

The average difference between phase A and phase B is −2.78, which a very large drop in the log-odds of disruptive behavior. (This difference does not translate directly into a proportion because it is a nonlinear function the proportion, and would be different for different phase A values.) A frequentist would conclude that this value is certainly significantly different from 0; it is more than 10 standard errors from 0, and the 95% credible interval (2.5 and 97.5 percentiles of −3.33 and −2.30) does not include 0. A Bayesian interprets the 95% credible interval as "there is a 95% probability that the parameter is in that interval" and would note that 0 is not a plausible value for the parameter because it does not fall in the interval.

The variable *phaseB* is the sum of *mu0* and *mu1*; it is the logit of disruptive behavior in phase B. The estimate of *phaseB* is −1.95. The variable p.B is the inverse of the logit, which is a proportion; the probability of disruptive behavior during phase B is .13, which is much lower than the .69 observed during phase A. The average reduction in the proportion of observations with disruptive behaviors was p.A–p.B = p.AB, for which the estimate was .57. (Note that .69–.13 = .56, not .57, due to rounding error.)

So the average treatment effect is very large, but how much variation is there across students in the treatment effect? The variable *sig1* measures the standard deviation in the treatment effect (on the log-odds scale); the median value is .57, which we will round to .60 for convenience. Approximately 95% of the students should be within approximately −2.8 plus or minus 1.2 (after rounding both values), or between approximately −1.6 and −4.0. All represent a considerable lessening of the amount of disruptive behavior. Using output not shown, we determined that the reduction in probability ranged from .37 to .63, with the largest changes coming for those who initially showed the most disruptive behavior (those with lower initial levels could not change as much).

One advantage of Bayesian methods is the computation of interesting probabilities that are not defined in classical modeling. We have seen that the average drop in disruptive behaviors is large; what is the probability that the average drop is at least .40? To calculate this probability, we merely count the proportion of times out of 5000 samples from the posterior distribution that the

**Table 1**
Results from model for data from Lambert et al. (2006).

| Node | Mean | Sd | 2.5% | Median | 97.5% |
|---|---|---|---|---|---|
| *mu0* | 0.8334 | 0.2282 | 0.3848 | 0.8323 | 1.306 |
| *mu1* | −2.782 | 0.2529 | −3.327 | −2.771 | −2.301 |
| *phaseB* | −1.949 | 0.3253 | −2.626 | −1.941 | −1.312 |
| *sig0* | 0.614 | 0.1967 | 0.3457 | 0.5759 | 1.096 |
| *sig1* | 0.6208 | 0.2724 | 0.2319 | 0.5724 | 1.285 |
| p.A | 0.6949 | 0.04771 | 0.595 | 0.6968 | 0.7868 |
| p.B | 0.1289 | 0.03631 | 0.06746 | 0.1255 | 0.2122 |
| p.AB | 0.566 | 0.03782 | 0.4868 | 0.5672 | 0.6363 |
| Step.AB | 0.9992 | 0.02827 | 1.0 | 1.0 | 1.0 |

difference between phase A and phase B is at least .40 on the probability scale. This value turns out to be .9992, or very nearly 1, which shows that not only is the point estimate of average change large, but it is nearly certain that it is larger than .40. (Note we could also compute other probabilities of interest, such as the probability the average change is small, for example that it is less than .05.)

### 4.2. Model 2: classroom as a case-level predictor

In the previous model, the intercepts and slopes are allowed to differ across students, but there is no explanation for why they might differ. In the second version of the model, the students' classroom is used to explain differences among student intercepts and slopes. As this article is about statistical methods and not research design, I will not go into the selection of variables. In this case, older students tended to be assigned to the second classroom; perhaps these were students who had repeated a grade, but that information is not available. For descriptive purposes, I will use class as a predictor, although it must be a proxy for some other variable such as retention or age, or perhaps teacher characteristics.

The equations for this model are the same observation-level equation plus the following student-level equations:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} Class_i + u_{0i} \tag{1.11}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} Class_i + u_{1i}. \tag{1.12}$$

This makes the model statement in the program slightly longer:

$$\text{logit(theta[i])} <- b0[\text{subj[i]}] + b1[\text{subj[i]}] * \text{phase[i]} + b2 * \text{class[i]} + \text{b3} * \text{class[i]} * \text{phase[i]}.$$

Here, *b2* is the difference between the two classrooms in baseline disruptions, and *b3* is the difference in the effect of treatment between the two classrooms. Note that unlike *b0* and *b1*, these do not differ across subjects.

A selection of the output is in Table 2. The parameters *mu0* and *mu1* are no longer the overall phase *A* effect and treatment effect, but because classroom was coded as a dummy (0/1) variable, they represent classroom 1.

With two classrooms, there are two sets of all the functions of parameters discussed for the first model, one for each classroom. As it turns out, the classes differ in baseline levels of disruptive behavior, but not in the treatment effect. The difference in baseline levels of disruptive behavior between the two classes is the parameter *b2*, which is large (almost 1 on the logit scale) and from the classical statistician's viewpoint is significant. The difference in treatment effects between classrooms (*b3*) is small – only 1/5 of a logit (.22) – and does not differ significantly from 0. From the Bayesian perspective, the 95 percent credible interval includes many small values, and the probability that the difference between classrooms in treatment effect is medium or larger is not high. Even though *b3* was not clearly important, I left it in the model to illustrate the full version of the model. Another way to examine the effect of including *b2* and *b3* is to notice that *sig0* has been decreased to half the size of the first model, showing that unaccounted for variability between students in baseline behavior has been decreased, although *sig1*, the standard deviation for treatment effect, has not changed.

**Table 2**
Output for the second model for data from Lambert et al. (2006).

| Node | Mean | Sd | 2.5% | Median | 97.5% |
|---|---|---|---|---|---|
| *mu0* | 1.329 | 0.2138 | 0.9197 | 1.325 | 1.763 |
| *mu1* | −2.901 | 0.3519 | −3.596 | −2.901 | −2.205 |
| *b2* | −0.8953 | 0.2651 | −1.446 | −0.8926 | −0.3683 |
| *b3* | 0.217 | 0.4575 | −0.6527 | 0.2331 | 1.088 |
| *sig0* | 0.3831 | 0.1418 | 0.1838 | 0.3567 | 0.7328 |
| *sig1* | 0.6433 | 0.2877 | 0.2295 | 0.6019 | 1.322 |
| phaseB.1 | −1.571 | 0.379 | −2.314 | −1.567 | −0.855 |
| p.A.1 | 0.7886 | 0.03532 | 0.715 | 0.7901 | 0.8536 |
| p.B.1 | 0.1785 | 0.05526 | 0.08997 | 0.1726 | 0.2984 |
| p.AB.1 | 0.6101 | 0.05284 | 0.496 | 0.6141 | 0.6996 |
| Step.AB.1 | 0.9968 | 0.05648 | 1.0 | 1.0 | 1.0 |
| phaseB.2 | −2.25 | 0.3647 | −2.981 | −2. | −1.55 |
| p.A.2 | 0.606 | 0.0436 | 0.5164 | 0.6067 | 0.6934 |
| p.B.2 | 0.09993 | 0.03268 | 0.04828 | 0.09622 | 0.1752 |
| p.AB.2 | 0.5061 | 0.0428 | 0.4185 | 0.5071 | 0.5888 |
| Step.AB.2 | 0.9898 | 0.1005 | 1.0 | 1.0 | 1.0 |

Another interesting point is that p.AB.1, the average change in probability of disruptive behavior in classroom 1, is .61, although p.AB.2, the same parameter for classroom 2, is .51. The apparent difference between the two classrooms in treatment effect is due to the nonlinear nature of the logit transformation: On the logit scale, these differences are the same, but translated back to probabilities, there is a difference.

**Example 2.** Intrinsically nonlinear model for Horner et al. (2005) data.

Horner et al. (2005) presented hypothetical data from three participants in a study to improve performance; the data are plotted in Fig. 2. Although Horner et al. only presented percent correct, I have assumed that there is a fixed number of trials per session. The dependent variable is the number of trials (out of 20) that the subject got a problem correct. Several aspects of the plots are worth noting: (a) the data from the first phase for each subject are at a low level but not necessarily zero, (b) the change between phases is gradual rather than abrupt, and (c) during the second phase the data settle down at a higher level than baseline but not necessarily at 20 (reflecting that all items were correct).

These results suggest a model that resembles a logistic regression but with some additions: There must be parameters to allow the curve to start above a probability of 0 and rise to a probability correct that is below 1. In a way, this curve resembles an item response curve in item response theory; there is a floor ("guessing") parameter, but there is an additional parameter for the ceiling.

Another way in which this curve resembles an item response curve is that we are not interested in the intercept and slope as parameters, but we are interested in the slope and the point at which the curve has risen some substantial fraction of the total rise, perhaps the halfway point. In item response theory (IRT), this form of the model is the usual parameterization in terms of item difficulty (halfway point of rise) and item discrimination (slope). It goes further in having not only the guessing parameter (floor) of a 3-parameter model but also the additional (ceiling) parameter of a 4-parameter model (Waller & Reise, 2010). For the purposes of SCDs, unlike IRT, the most interest is in the difference between the floor and ceiling, as this represent the size of the experimental effect.

We construct the model in several stages to illuminate each of the complications beyond traditional logistic regression. First, consider a traditional logistic regression model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X. \tag{1.13}$$

This equation is the usual intercept and slope model. The point at which the curve rises halfway, to a probability of .5, is when the odds are .5/.5 = 1, and therefore the log-odds is zero. This result happens when $0 = \beta_0 + \beta_1 X$, or after rearranging terms, when $X = -\beta_0/\beta_1$. Using this fact, we can rewrite the logistic regression equation as

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X = \beta_1\left(\frac{\beta_0}{\beta_1} + X\right) = \beta_1(X - \mathrm{H}) \tag{1.14}$$

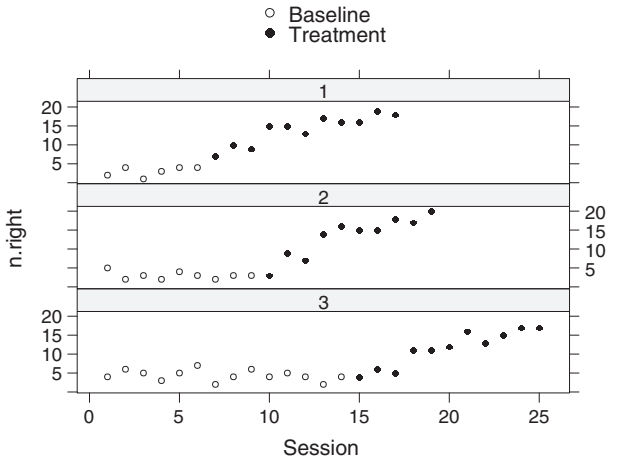


**Fig. 2.** Plots of hypothetical data from Horner et al. (2005). The design was a multiple baseline design, with a baseline phase followed by a treatment phase. The shift to treatment phase was at a different time for each participant. The dependent variable is not fully specified, but for each session, it is assumed to be the number of trials (out of 20) in which a correct behavior or answer occurred.

where H is the "halfway" point $-\beta_0/\beta_1$. Now the parameters of the model are the slope ($\beta_1$) and the halfway point (H). To find the predicted probability for a value of $X$, we exponentiate to get the odds, and divide the odds by the odds plus 1:

$$\pi = \exp(\beta_1(X-\mathrm{H}))/(1+\exp(\beta_1(X-\mathrm{H}))). \tag{1.15}$$

The model is still the same curve as the logistic regression; it starts near 0 for low values of $X$ and rises to nearly 1 for large values of $X$. To make it start above zero and rise to a lower limit than 1, we put additional parameters in for the floor (F) and ceiling (C):

$$\pi = F + (\mathrm{C}-\mathrm{F})\exp(\beta_1(X-\mathrm{H}))/(1+\exp(\beta_1(X-\mathrm{H}))). \tag{1.16}$$

Of course, in addition to the gradual change over sessions, we can also include a jump when the phase changes from baseline to treatment. Each subject can have a different value for all parameters, and those parameters can have a distribution over subjects allowing each subject's estimates to benefit from knowledge of the other subjects' parameters.

The primary interest is in the value of C − F, the change from floor to ceiling. This difference can vary over people, so we can look at the average change across people, the variability across people (variance or standard deviation), and the individual values for each person. For the Horner et al. (2005) data, the effects for the three respondents are .57, .66, and .72, which are quite large. (They are also well below 1, which is the value of C – F in the usual logistic regression.)

Another issue is whether the curves rise in the order in which the phase changes. The first person was the first to have a phase change, followed by the second, and then the third. To assess the effect statistically, we look at the estimates of H for each person: The estimates are 8.5, 12.3, and 18.6, meaning the curves rose to halfway between the floor and ceiling at these sessions for the three people. These estimates are certainly the right order, but we can also ask whether there is a high probability that the halfway point for person 2 is greater than for person 1 and similarly for person 3 compared to person 2. These probabilities are estimated to be very close to 1. Thus, we are fairly certain we have estimated the right order of the subjects.

## 5. Discussion

Bayesian methods are useful in the analysis of data from SCDs because they satisfy the requirements of small-sample theory, are more interpretable than results from classical statistics, and are computationally adapted to track interesting functions of parameters. A Bayesian can obtain a credible interval (comparable to a classical confidence interval), and can calculate probabilities of great interest that are not possible in classical statistics: The probability that a parameter is greater (or less) than zero, large (greater than some specified value), or small. A Bayesian can calculate the probability that one parameter is larger than another or is approximately equal to (within a small distance of) another. Furthermore, Bayesians can calculate probabilities for transformations, such as logits (or their inverses), without using the delta method of classical statistics, which relies on large sample theory. (The delta method uses derivatives of the transformation function to approximate the changes in a transformed value. For small samples or skewed distributions the approximation may be poor.)

Fully Bayesian methods are more appropriate for SCD data than classical or empirical Bayes methods. When estimating a parameter, fully Bayesian methods take into account uncertainty about all other parameters. Thus, standard errors may be larger (and confidence intervals wider) when using fully Bayesian methods, but larger standard errors accurately reflect the totality of the uncertainty about model parameters.

Fully Bayesian methods implemented with MCMC computational methods, such as in WinBUGS, OpenBUGS, and JAGS, can be used to fit complicated models that are outside the scope of standard nonlinear models (e.g., GLMs).

Some of the more advanced aspects of Bayesian model fitting were not illustrated in this article. These aspects include tests for convergence of the algorithm, model fitting, model diagnostics, alternative model parameterizations, selection of prior distributions, and sensitivity analysis. Progress is being made in developing methods for estimating an effect size measure comparable to the standardized effect size $d$ in between-subject designs; Rindskopf et al. (2012) presented a Bayesian approach using WinBUGS that allows not only estimation of $d$ but also probability statements (including credible intervals) for $d$, and Swaminathan, Rogers, & Horner (2014–in this issue) present another approach to this problem.

Bayesian methods are not without weaknesses. The models are more complicated to set up because of the need to specify prior distributions. The prior distributions can be made almost noninformative, but complete ignorance comes at a cost in computational methods: Sometimes the calculations get lost in an area of the parameter space that causes an arithmetic error (some parameter gets too large), and the procedure stops. This arithmetic error happens most frequently with variances or precisions. It can be difficult to sort out and solve these problems.

In spite of the weaknesses, Bayesian methods are likely to become the preferred method of analysis for SCDs. They adequately deal with the small sample sizes, and give estimates of quantities that are of direct interest to researchers. Interpretation of results is much more natural than classical statistics, with probability statements that conform to common sense ways of thinking about events.

## Appendix 1. Lambert et al. (2006) Model

```
model {

for (i in 1:264){                                          # 264 total observations
    logit(theta[i]) <- b0[subj[i]] +
                       b1[subj[i]]*phase[i]        # level 1 model

     class[i] ~  dbin(.5,1)
     y[i] ~ dbin(theta[i],10)       # count outcome modeled binomially
}

for (j in 1:9){                     # 9 cases
    b0[j] ~ dnorm(mu0, prec0)       # intercept treated as a random effect

    b1[j] ~ dnorm(mu1, prec1)       # treatment treated as a random effect

    disrupt.0[j] <- b0[j]          # disruptions expected during baseline

    disrupt.1[j] <- disrupt.0[j] + b1[j]  # disruptions expected during treatment
    disr.0[j] <- 1/(1+exp(-1*disrupt.0[j]))
    disr.1[j] <- 1/(1+exp(-1*disrupt.1[j]))

}

mu0 ~ dnorm(0, .01)                # priors on parameters
mu1 ~ dnorm(0, .01)
prec0 ~ dgamma(.01, .01)
prec1 ~ dgamma(.01, .01)
sig0 <- 1/sqrt(prec0)               # standard deviation of intercept
sig1 <- 1/sqrt(prec1)               # standard deviation of phase effect
var1 <- 1/prec1                     # var in treatment effect
var0 <- 1/prec0                     # var in baseline

phaseB <- mu0 + mu1
p.A <- exp(mu0)/(1+exp(mu0))
p.B <- exp(phaseB)/(1+exp(phaseB))
p.AB <- p.A - p.B
step.AB <- step(p.AB - .4)

}
```

## Appendix 2. Lambert et al. (2006) output

Node statistics

| Node | Mean | Sd | MC error | 2.5% | Median | 97.5% | Start | Sample |
|---|---|---|---|---|---|---|---|---|
| *mu0* | 0.8334 | 0.2282 | 0.003606 | 0.3848 | 0.8323 | 1.306 | 5001 | 5000 |
| *mu1* | −2.782 | 0.2529 | 0.004733 | −3.327 | −2.771 | −2.301 | 5001 | 5000 |
| *sig0* | 0.614 | 0.1967 | 0.003855 | 0.3457 | 0.5759 | 1.096 | 5001 | 5000 |
| *sig1* | 0.6208 | 0.2724 | 0.008672 | 0.2319 | 0.5724 | 1.285 | 5001 | 5000 |
| var1 | 0.4596 | 0.4711 | 0.01204 | 0.05378 | 0.3276 | 1.651 | 5001 | 5000 |
| var0 | 0.4156 | 0.3216 | 0.006079 | 0.1195 | 0.3317 | 1.201 | 5001 | 5000 |
| *phaseB* | −1.949 | 0.3253 | 0.004945 | −2.626 | −1.941 | −1.312 | 5001 | 5000 |
| p.A | 0.6949 | 0.04771 | 7.498E−4 | 0.595 | 0.6968 | 0.7868 | 5001 | 5000 |
| p.B | 0.1289 | 0.03631 | 5.527E−4 | 0.06746 | 0.1255 | 0.2122 | 5001 | 5000 |
| p.AB | 0.566 | 0.03782 | 7.195E−4 | 0.4868 | 0.5672 | 0.6363 | 5001 | 5000 |
| Step.AB | 0.9992 | 0.02827 | 3.898E−4 | 1.0 | 1.0 | 1.0 | 5001 | 5000 |
| Disrupt.0[1] | 1.046 | 0.165 | 0.003055 | 0.7259 | 1.044 | 1.378 | 5001 | 5000 |
| Disrupt.0[2] | 1.412 | 0.1913 | 0.00414 | 1.051 | 1.407 | 1.796 | 5001 | 5000 |
| Disrupt.0[3] | 1.298 | 0.1915 | 0.004737 | 0.9433 | 1.294 | 1.685 | 5001 | 5000 |
| Disrupt.0[4] | 1.292 | 0.1869 | 0.003711 | 0.9397 | 1.289 | 1.677 | 5001 | 5000 |
| Disrupt.0[5] | 1.136 | 0.1652 | 0.003469 | 0.8288 | 1.131 | 1.47 | 5001 | 5000 |
| Disrupt.0[6] | 0.368 | 0.1628 | 0.002744 | 0.04122 | 0.3691 | 0.6882 | 5001 | 5000 |

**Appendix 2** (*continued*)

Node statistics

| Node | Mean | Sd | MC error | 2.5% | Median | 97.5% | Start | Sample |
|---|---|---|---|---|---|---|---|---|
| Disrupt.0[7] | 0.6919 | 0.1694 | 0.003103 | 0.3678 | 0.6909 | 1.036 | 5001 | 5000 |
| Disrupt.0[8] | 0.02085 | 0.1556 | 0.003385 | −0.2835 | 0.02156 | 0.3272 | 5001 | 5000 |
| Disrupt.0[9] | 0.184 | 0.1529 | 0.002972 | −0.1074 | 0.1841 | 0.4905 | 5001 | 5000 |
| Disrupt.1[1] | −1.75 | 0.2232 | 0.003019 | −2.208 | −1.743 | −1.327 | 5001 | 5000 |
| Disrupt.1[2] | −1.548 | 0.2022 | 0.002843 | −1.973 | −1.544 | −1.164 | 5001 | 5000 |
| Disrupt.1[3] | −1.458 | 0.2138 | 0.003183 | −1.886 | −1.457 | −1.056 | 5001 | 5000 |
| Disrupt.1[4] | −1.461 | 0.2186 | 0.00301 | −1.912 | −1.459 | −1.037 | 5001 | 5000 |
| Disrupt.1[5] | −1.531 | 0.1913 | 0.00279 | −1.912 | −1.528 | −1.166 | 5001 | 5000 |
| Disrupt.1[6] | −3.344 | 0.5026 | 0.01371 | −4.467 | −3.297 | −2.508 | 5001 | 5000 |
| Disrupt.1[7] | −2.452 | 0.2815 | 0.004621 | −3.038 | −2.437 | −1.933 | 5001 | 5000 |
| Disrupt.1[8] | −1.87 | 0.2648 | 0.007546 | −2.424 | −1.866 | −1.371 | 5001 | 5000 |
| Disrupt.1[9] | −2.239 | 0.245 | 0.004045 | −2.738 | −2.236 | −1.78 | 5001 | 5000 |
| Disr.0[1] | 0.7387 | 0.03164 | 5.843E−4 | 0.6739 | 0.7396 | 0.7987 | 5001 | 5000 |
| Disr.0[2] | 0.8024 | 0.02998 | 6.409E−4 | 0.741 | 0.8034 | 0.8577 | 5001 | 5000 |
| Disr.0[3] | 0.7837 | 0.03212 | 7.937E−4 | 0.7198 | 0.7848 | 0.8436 | 5001 | 5000 |
| Disr.0[4] | 0.7828 | 0.03144 | 6.201E−4 | 0.719 | 0.784 | 0.8425 | 5001 | 5000 |
| Disr.0[5] | 0.7556 | 0.0303 | 6.321E−4 | 0.6961 | 0.756 | 0.813 | 5001 | 5000 |
| Disr.0[6] | 0.5904 | 0.03916 | 6.602E−4 | 0.5103 | 0.5913 | 0.6656 | 5001 | 5000 |
| Disr.0[7] | 0.6653 | 0.03744 | 6.824E−4 | 0.5909 | 0.6662 | 0.7381 | 5001 | 5000 |
| Disr.0[8] | 0.5052 | 0.03866 | 8.409E−4 | 0.4296 | 0.5054 | 0.5811 | 5001 | 5000 |
| Disr.0[9] | 0.5456 | 0.0377 | 7.329E−4 | 0.4732 | 0.5459 | 0.6202 | 5001 | 5000 |
| Disr.1[1] | 0.1503 | 0.02816 | 3.781E−4 | 0.09901 | 0.1489 | 0.2096 | 5001 | 5000 |
| Disr.1[2] | 0.1773 | 0.02916 | 4.042E−4 | 0.1221 | 0.1759 | 0.2379 | 5001 | 5000 |
| Disr.1[3] | 0.1909 | 0.03273 | 4.86E−4 | 0.1318 | 0.189 | 0.2582 | 5001 | 5000 |
| Disr.1[4] | 0.1905 | 0.03332 | 4.667E−4 | 0.1288 | 0.1887 | 0.2617 | 5001 | 5000 |
| Disr.1[5] | 0.1796 | 0.02794 | 4.102E−4 | 0.1287 | 0.1783 | 0.2376 | 5001 | 5000 |
| Disr.1[6] | 0.03778 | 0.01667 | 4.787E−4 | 0.01136 | 0.03567 | 0.07527 | 5001 | 5000 |
| Disr.1[7] | 0.08171 | 0.02061 | 3.318E−4 | 0.04572 | 0.08036 | 0.1264 | 5001 | 5000 |
| Disr.1[8] | 0.1365 | 0.03066 | 8.53E−4 | 0.08136 | 0.134 | 0.2025 | 5001 | 5000 |
| Disr.1[9] | 0.09839 | 0.02156 | 3.598E−4 | 0.06079 | 0.09658 | 0.1443 | 5001 | 5000 |

## References


Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179.

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.

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Chichester, UK: Wiley.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria.* 1609-395X (Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Plummer.pdf)

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

Rindskopf, D. (2013). Bayesian analysis of data from single case designs. *Neuropsychological Rehabilitation* (in press).

Rindskopf, D., & Ferron, J. (2013). Using multilevel models to analyze single-case design data. In T. Kratochwill, & J. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association (in press).

Rindskopf, D., Shadish, W. R., & Hedges, L. V. (2012, January). *A simple effect size estimator for single case designs using WinBUGS*. Paper presented at the meeting of the Society for Research on Educational Effectiveness, Washington, D.C.

Shadish, W. R., Hedges, L. V., Pustejovsky, J., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2013). Analyzing single-case designs: d, g, hierarchical models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. Kratochwill, & J. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association (in press).

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*, 385–405.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed)models to analyze single case designs. *Journal of School Psychology*, *52*(2), 41–70 (in this issue).

Swaminathan, H., Rodgers, J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2), 105–122 (in this issue).

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

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

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today*, *8*(2), 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*, *3*, 142–151.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Mulitphasic Personality Inventory. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 147–173). Washington, DC: American Psychological Association.

Winkler, R. L. (2002). *An introduction to Bayesian inference and decision* (2nd ed.). Gainesville, FL: Probabilistic Publishing.