Neuropsychological Rehabilitation, 2014
Vol. 24, Nos. 3–4, 572–589, http://dx.doi.org/10.1080/09602011.2013.866903

Bayesian analysis of data from single case designs

David Rindskopf
Educational Psychology Program, CUNY Graduate Center, New York, NY, USA

(*Received 24 June 2013; accepted 13 November 2013*)

Bayesian statistical methods have great potential advantages for the analysis of data from single case designs. Bayesian methods combine prior information with data from a study to form a posterior distribution of information about their parameters and functions. The interpretation of results from a Bayesian analysis is more natural than those from classical methods, and there are interpretations of useful quantities that are not possible in classical statistics, such as the probability that an effect size is small, or is greater than zero, or is large enough to be considered important. They are not based on asymptotic theory, so small sample size is not a problem for inference. These methods are implemented on free software, and are similar to non-Bayesian software, so analysts familiar with frequentist methods for multilevel data should find the transition relatively painless.

Keywords: Bayesian; Single case; Single subject; Multilevel models; Hierarchical models.

INTRODUCTION

Single-case designs (SCDs) seem simple in theory: one compares the behaviour of an individual in one experimental condition to his or her behaviour in another condition, and the difference in behaviour (on some scale of

Correspondence should be addressed to David Rindskopf, Educational Psychology Program, CUNY Graduate Center, 365 Fifth Avenue, New York, NY 10016 USA. E-mail: drindskopf@gc.cuny.edu

The research reported here was supported by Grant No. R305D100046 from the Institute of Education Sciences, US 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 US Department of Education.

measurement) represents the difference between the effects of the two conditions. The experimental conditions could be two treatments, or a treatment and a control or placebo condition. Typically more than one individual is involved in a study, and one can see if the effect of treatment varies from one individual to another. If the world (and people's behaviour) were perfect, a simple plot of the data would suffice to demonstrate the effect of treatment. The only statistical test needed would be the famous intra-ocular traumatic test: the effect is so big it hits you between the eyes.

Alas, the world is not so perfect, and for many reasons we need the assistance of statistical tests. Our measures are not continuous, but are often counts, proportions, or rates of behaviour; there is autocorrelation; behaviour is not constant during a phase of treatment, but has outliers, a trend up or down, or an even more complicated shape. These and other problems necessitate the use of statistical methods for analysis of data to supplement the traditional visual analysis. This article describes one such approach, multilevel models using Bayesian inference. First I will discuss the general characteristics of Bayesian inference that are advantageous for these analyses. Then I will discuss hierarchical modelling, the basis for the statistical models for single-case designs. Finally, I will demonstrate Bayesian hierarchical analysis using a real data set; although this data set (and corresponding analysis) does not include all complications of SCDs, it shows a number of reasons why these analyses are useful.

BAYESIAN STATISTICS

Bayesian statistics start with the premise that there are (unknown) parameters, θ, in a model for behaviour, and that we start with some (perhaps very little) information about these parameters. One such parameter may be the average effect size for all respondents; another may be the amount of variation across respondents in effect size. This information is quantified in the prior distribution, which describes our knowledge prior to collecting the data, and is generically represented as $p(\theta)$, the probability density function of θ, where θ can be a single parameter or vector of parameters (unknown constants). We then gather data in a study, and that data provides information about the parameters. What we learn from the data in the experiment is captured in the likelihood (as in some classical statistical methods), represented generically by $L(x|\theta)$, in which x represents the data, and what we know afterwards is called the posterior distribution, which is proportional to $L(x|\theta)\,p(\theta)$, the product of the prior and the likelihood. The posterior distribution thus combines the information we have prior to the study (if any) with the information in the data (captured in the likelihood). We can also make predictions about the future occurrences of events by using the predictive distribution, but we will not discuss this element of Bayesian statistics in this paper. Nor will

we discuss decision theory, which involves the specification of various actions that might be taken on the basis of data and the assignment of utilities to possible outcomes.

Bayesian statistics are sometimes said to be subjective because the capturing of prior information in a distribution often has an aspect of subjectivity, but this can be eliminated (or made essentially irrelevant) by using what is called an uninformative (or noninformative, or flat) prior distribution. These terms are meant to imply that, prior to seeing the data, we have little or no knowledge of the parameter values. In this case the posterior is solely (or mainly) determined by the likelihood, which summarises the information in the data. Then the main difference between classical and Bayesian statistics is the interpretation. Because the posterior distribution is an actual probability distribution, intervals calculated on that distribution are probabilities. For a 95% credible interval (similar to a confidence interval in classical statistics), there is a 95% chance that the parameter is in the interval. The statement applies to this interval, not to a sequence of hypothetical repetitions. We can also calculate other interesting probabilities; for example, the probability that a parameter is larger than zero (or some other important quantity), or the probability that one parameter is larger than another. One can also determine the probability that some parameter (or difference between parameters, or other function) is less than a given quantity, and can therefore be said to be "small".

Thus Bayesian methods are based on a different philosophy than classical/frequentist statistics. Classical statistics make inferences that depend on a hypothetical ability to repeat a study an infinite number of times. Parameters are fixed values, and the only probability statements we can make about them is in terms of the repetition of the study. For example, the interpretation of a 95% confidence interval is that in an infinite number of repetitions of the study, 95% of the confidence intervals will contain the parameter. There is no probability statement associated with the specific confidence interval we calculate. For the Bayesian no such conceptualisation is necessary; the probability statements can apply to a nonrepeatable study. The resulting probability statements are about the belief of the researchers, not about the actual value of the parameters.

Bayesians also do not generally use hypothesis tests (except when trying to show how such testing would be done in the Bayesian framework) because usually parameters are on a continuous scale, so the probability of a specific value (such as 0) would be zero. Bayesians favour probability statements based on the continuous scale, so they talk of probabilities that a parameter is greater than 0, or the probability that a parameter is greater than some value that would be considered large in a practical sense, or the probability that a parameter is in an interval. If they were interested in showing that a parameter was close to some null value (in the frequentist context) they could show that with high probability the parameter is in a short interval that

includes the null value; that is, they would establish that the parameter is probably small (see Rindskopf, 1997, for a fuller discussion of this issue).

The actual result of the analysis is the posterior distribution of the parameters. One can use this posterior to calculate interesting probabilities, as discussed above. If the form of the distribution is known, one can present the parameters and let the reader determine what probabilities are valuable; for example, presenting the mean and variance of the posterior for a normally distributed parameter summarises all the information about the parameter, because a normal distribution is completely characterised by its mean and variance (or standard deviation).

One purpose for which Bayesians do use informal or quasi-frequentist methods is model selection. Decisions must be made about which variables to include in a model (such as regression), or more generally about which form the model should take. I will illustrate some of these methods below. While there are pure Bayesian approaches to these issues (such as Bayesian model averaging), more often Bayesians use a blend of ideas for model selection.

Modern Bayesian computation is done using simulation methods, called Markov Chain Monte Carlo (MCMC) methods, that give samples from the posterior distribution of the parameters. One can imagine the result of these computations as a spreadsheet or data set, with the columns ("variables") being the parameters, and each row ("subject") being the values taken by the parameters on a particular trial (or iteration) of the simulation. Starting with a row of initial values for the parameters, each row is built up based on the data, the model, and the parameter values in the prior row (iteration). Usually the first few hundred (if the procedure settles down quickly) or thousand iterations (called the burn-in) are discarded so that the posterior distribution is accurately represented.

Estimation of a parameter is as simple as finding the average of the column representing that parameter. Estimating percentiles of the posterior distribution of a parameter is likewise simple: find the percentiles in the column of the spreadsheet. To estimate the probability that a parameter is greater than 0 (or any other value) find the proportion of the estimates that are greater than 0 (or any other value).

This spreadsheet format of draws from the posterior distribution makes it possible easily to compute probabilities not only for the parameters themselves, but also for transformations or combinations of parameters, such as ratios and differences. Classical methods for doing this typically require large sample sizes, and use an approximation through the delta method (based on derivatives of the transformation); MCMC methods do not require large samples or the use of approximations. For example, suppose one parameter represents the logit (log odds) of a behaviour occurring. To find the distribution of the probability of the behaviour, the logit is transformed by creating a new variable exp(logit)/(1+exp(logit)), and we have

another column in the spreadsheet which represents the probability. We can find the average of these (to estimate the mean), and similarly find the median and any percentiles or probabilities that we wish.

MULTILEVEL (HIERARCHICAL) MODELS

Single-case designs are, despite their name, not usually studies involving only a single subject. The name comes from the emphasis on establishing causal effects by within-person comparisons instead of across-person comparisons, each of which is then replicated across a small number of respondents. Statistical models deal with this by having one part of the model describing the behaviour of each individual, and another part describing similarities and differences among individuals, explaining those differences (when possible) on the basis of subject characteristics. I will provide an overview of these models here, but for more details see Baek et al. (2013).

Consider the simplest case, in which a continuous outcome variable is measured on each participant each day (or week, etc.) over a period of time, and that at some time point the phase is switched from no treatment or standard treatment (phase A) to a new treatment (phase B). If the behaviour is constant within a phase (except for random variation), the statistical model for each person can be represented in the form:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Phase_{ij} + r_{ij}$$

In this equation Y is the dependent variable, which has a potentially different value for each individual j at each measurement time i; β_{0j} and β_{1j} are the phase A (baseline) value and jump from phase A to phase B (treatment effect), respectively, for individual j; $Phase_{ij}$ is the phase at time i for individual j, coded 0 for phase A and 1 for phase B, and r is a residual term. It is often assumed that r has a normal distribution at each time (and phase) and usually the same variance for each individual and phase, although this is not necessary and can be relaxed if there is evidence that it is false.

Note that β_{0j} and β_{1j} are allowed to be different for each person; these differences may be left unexplained, or may be (at least partially) explained by person-level characteristics, such as age, sex, race, education, diagnostic category, or other variables. The equations for this part of the model, with one explanatory variable W, would be:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}$$

In these equations, the intercept (baseline level) and slope (treatment effect) for each person j is modelled as a constant plus an effect of the person-level variable W, plus a residual u. The variances of the residuals are represented by τ_{00} and τ_{11} and their covariance by τ_{01}. Of course, the same variable W might not explain variation in both the intercept and slope, and the equations can have totally different sets of explanatory variables if needed. Further, there must be enough respondents to justify including explanatory variables.

One advantage of a Bayesian approach to multilevel modelling is that estimation of person-specific parameters (such as β_{0j} and β_{1j}) can be improved by using information from all respondents. Thus we need not take one of the two extreme positions of (i) assuming all respondents behave the same way, or (ii) that only the data from each respondent is relevant for estimating his or her parameters. Rather, we assume that respondents will be similar, but not identical, and that we can "borrow information" from all respondents to estimate each respondent's parameters, by shrinking towards the overall mean. These shrunken estimates will have lower variance (and mean square error) than separate estimates from each person's data.

Many possible complications can be accommodated with multilevel models. First, the dependent variable might not be continuous, or if it is, might not be normally distributed. For example, a t-distribution can be used to accommodate heavy tails or outliers, or an exponential distribution for waiting times. For categorical dependent variables a Poisson or binomial model can be used (depending on whether a count is for some period of time, or is out of a total number of trials); in more complicated cases a negative binomial or other distribution may be appropriate.

Another complication occurs when there are more than two phases, or when the number or order of phases differs from one respondent to another; many variations are possible, and the complexity of the model will necessitate special construction of variables to handle each case. One such case is illustrated in the example below, with some variations.

Autocorrelation is a frequent problem in time series, and can occur in single-case designs. Most series in SCDs are too short to estimate the autocorrelation with any degree of accuracy, although if it can be assumed to be the same across subjects it is sometimes feasible to include it in the model. Unfortunately, there are other factors that can masquerade as autocorrelation, such as omitted variables, and these can happen with SCD data frequently enough that one has to question whether a nonzero *estimate* of an autocorrelation actually is due to autocorrelation, particularly when measures are not close in time. Autocorrelation should not bias parameter estimates, although it can affect standard errors, and therefore affect significance tests (in classical inference) and probability statements (in Bayesian inference).

A more realistic complication is nonlinearity. A sudden shift from one level of behaviour to another when the phase changes is less likely than a

gradual shift from one level to another. (The time period between measurements also plays a role here; short times between measurements are more likely to show a curve from one level to another than would a long time between measurements, which could miss the curve.) This type of nonlinearity is not always accommodated by the standard nonlinear models (generalised linear models, such as logistic regression and ordered logit models), and can occur with continuous outcomes also. For more information on these models, see Rindskopf (in press).

EXAMPLE: AN ABA′B DESIGN

The data I will use for illustration are from Tasky, Rudrud, Schulze, and Rapp (2008), the same data set used by Shadish et al. (2013). The outcome is the number of trials out of six which show on-task behaviour, so the simplest probability model for these data is the binomial distribution. One of the natural models in this situation is the logit model, where the left-hand side of the linear model is the logarithm of the odds of the event occurring.

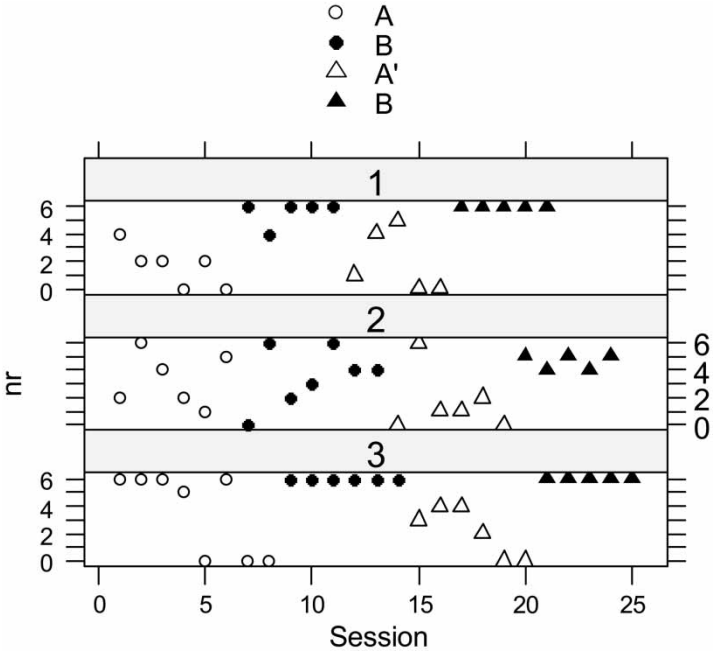

Figure 1. Plot of data from Tasky et al. (2008). The three respondents (Rebecah, Cara, and Amber) were observed for 21, 24, and 25 sessions. The dependent variable is the number of times (out of six) that they were observed making on-task behaviour. The first (A) and second (A′) baseline phases were somewhat different, while the two B phases were under identical conditions.

A plot of the data is in Figure 1, and the means for each subject in each phase are in Table 1.

Two aspects of the plots give us reason to be cautious in modelling. One is that there sometimes appears to be more variation than we might expect in a binomial model with constant (within phase) probability. For example, Amber's first phase has only very high and very low values; this would be very unusual if the probability were constant. The second aspect is the ceiling effect, where during a phase all (or almost all) observations were at the highest level. This can sometimes cause estimation problems, depending on the nature of the model.

The study had four phases, denoted in generalised notation as ABA′B, where the prime indicates that the second A phase is a different version of the first A phase baseline condition. In the simplest plausible outcome model, the behaviour during phase A and phase A′ would be at the same level, and behaviour during both B phases would be at the same level. Other models would allow for differences in the two A phases, the two B phases, or both; I will investigate these models also.

In each session the dependent variable was counted as occurring or not at each of six times. I will model this as a binomial with a particular probability that is constant within phases, but differs across phases. This is undoubtedly an oversimplification, but for the purposes of this overview paper I will not use more complex models.

To represent the phases we will use dummy variables called *a, b, c*, and *d* to represent the four phases each person experiences during the study. If the model has an intercept, then no more than three of these variables are needed. This representation allows us to represent many models easily. For example, the model where the A and A′ periods are at one level, and the two B periods are at a different level, requires one dummy variable distinguishing the B from the A periods. This can be accomplished by creating a new dummy variable equal to the sum of the dummy variables *b* and *d*.

I fit three models to the data. The first is the simplest model, in which the A and A′ periods are estimated to have one probability of showing the

TABLE 1
Means for each phase and subject

	Phase			
Subject	A	B	A′	B
1	1.67	5.60	2.00	6.00
2	3.33	3.57	1.67	4.60
3	3.63	6.00	2.17	6.00

behaviour, and the two B periods are estimated to have a different probability. The second model adds a difference between period A and period A′, and the third is a full factorial, with all four phases having a different probability. One possible coding scheme for the three models is illustrated in Table 2. The first model contains a dummy variable for phase B. The second model adds a dummy variable that allows phase A to differ from phase A′. The third model allows all phases to differ; it does this by representing an initial value for the first (A) phase, and each parameter represents the change from one phase to the next.

The fit of each model is assessed by the deviance, which is −2 times the logarithm of the likelihood. A large deviance indicates a poorly fitting model. Only with grouped discrete data can the absolute deviance value be interpreted; more typically the values are used to compare models. The deviances for the models are 271.0, 258.7, and 257.7. The first deviance is much higher than the second, indicating that the first model does not fit as well as the second; the second does not fit much worse than the third, indicating that the added complexity of the third model is unnecessary.

How are the sizes of differences judged? Here we use a method similar to what would do a frequentist. The difference in deviances has approximately a chi-square distribution. The degrees of freedom for this difference are not always simple to assess, and are sometimes approximated by a numerical calculation whose stability is not certain. For our models we can use a simplified rule that each parameter added or dropped contributes one degree of freedom. Each of the differences would have one degree of freedom, so differences less than 3.84 would not be significant in the classical sense, and a difference of 1 (between the second and third models) is certainly not important. Therefore

TABLE 2
Model coding for the three models for Tasky et al. (2008) data

A	B	A′	B
Model 1			
1	1	1	1
0	1	0	1
Model 2			
1	1	1	1
0	1	0	1
0	0	1	0
Model 3			
1	1	1	1
0	1	1	1
0	0	1	1
0	0	0	1

the second model is chosen as the simplest that fits the data well. (See Spiegelhalter, Best, Carlin, & van der Linde, 2002, for detailed information on model comparison from a Bayesian viewpoint, including degree of freedom calculations for complex models.)

It is interesting that although the second model fits quite a bit better than the first, none of the parameters for the difference between the A and A′ phases is large. All are in the same direction, and the difference for subjects 2 and 3 are nearly significant in the classical sense. Nevertheless, the fit statistics tell us that, taken together, the evidence favours the more complex model 2 over the simpler model 1.

The second model is written so that the parameters represent (i) the first A phase, (ii) the difference between the first A phase and the average B phase, and (iii) the difference between the first A phase and the second A′ phase. As we will see, the average performance of the three subjects does not capture all of the information: there is significant variation among the subjects as well. We will start with the overall effect, and then describe the variation among subjects. The parameters are on the logit (logarithm of the odds) scale; the model is written so that transformations back to the scale of proportions are calculated and tracked as well, as this is the scale that is most easily interpretable. On this scale, the proportion of observations during which on-task behaviour was observed averaged .49 in phase A and .91 in Phase B; the difference averaged .42 with a standard error of .14. Roughly speaking, the average went from on-task behaviour in half the observations during baseline up to 9 out of 10 during treatment, an increase of 4 out of 10 instances of additional on-task behaviour.

Both the classical and Bayesian interpretation of these results are illustrated, not only for comparative purposes, but also because some researchers will want to maintain connections to traditional interpretation. The increase of .42 is almost certainly significant in the classical sense, but if we want to be careful about inference we should see whether the distribution is approximately normal (so that we can use approximately the estimate plus or minus 2 standard errors), or whether it is skewed and we must use a different technique. If the distribution were normal, an approximate 95% interval would be .42 plus or minus 2(.14) = .28, or an interval of (.14, .70).

A plot of the posterior distribution of the difference (called *diff* in the program) is in Figure 2. The distribution is skewed to the left, indicating that high values are less likely than low values. (Another way to demonstrate the skew is to compare the distance from the 2.5th percentile to the 50th percentile, .346, with the distance from the 50th percentile to the 97.5 percentile, which is .226. If the distribution were symmetrical, these two distances would be approximately the same.) The reason for the skewness is probably because when the estimate during phase A is high the difference cannot be too large a positive number, but when the estimate during phase A is lower one can get a

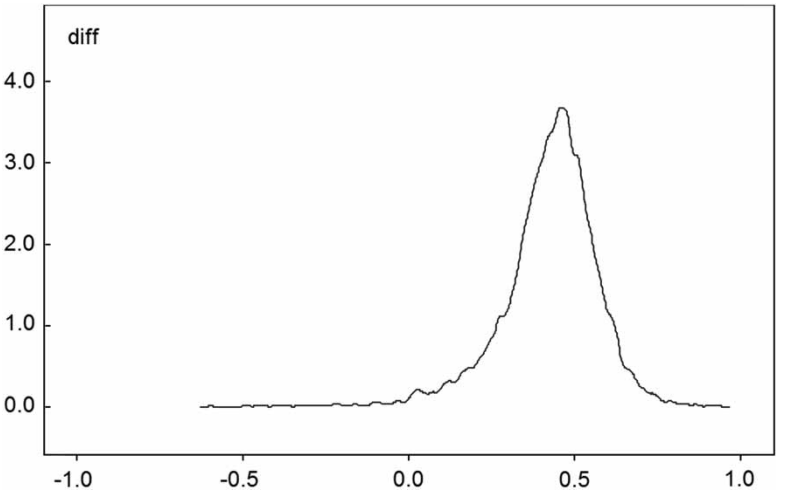

Figure 2. Posterior distribution of the difference between the first A phase and the average B phase. A smoothed plot of a sample of 10,000 observations drawn from the posterior distribution. Note the skew to the left.

wider range of differences. To get a better credible interval, we take the 2.5 and 97.5 percentiles of the posterior distribution, which are .095 and .667. (This is not equivalent to a highest posterior density interval, the optimal interval for a Bayesian, but it is not far off. A highest density region is the region which contains all the values with the highest values of posterior density, so that any point in the region is more likely than any point outside the region. For skewed distribution this may not correspond to the region defined by the 2.5 and 97.5 percentiles. Unequal regions are especially common with bounded parameters such as the variance, where the highest density region may be one-tailed if zero is a very plausible value.) Even when the Bayesian credible interval is similar to the classical confidence interval (due to a noninformative or relatively uninformative prior distribution) the interpretation is different: a Bayesian will say that there is a 95% probability that the parameter is in the interval.

A Bayesian can look at other probabilities that are meaningful in addition to the usual intervals. For example, one can calculate the probability that the difference is positive, which in this case is (for all practical purposes) equal to 1; thus there is almost no chance that the average effect is negative, even though we have only three respondents. A more interesting number is the probability that the average effect is at least moderately large, where some number is designated as representing an effect of practical importance. For example we could calculate the probability that the average difference is larger than .20; that probability is .94. So there is a high probability that

the average effect is at least moderately large. We could also calculate the probability that the average effect is small, say less than .05 in absolute value; this probability is .01, so we can say with some certainty that the effect is not small.

The estimates for individuals (as well as the total that we have just discussed) are in Table 3. Note that person 1 has the largest gain, person 3 the next largest, and person 2 not only has the smallest, but is not likely to be extremely large as are the others. Again, we can calculate useful probabilities that are not part of standard statistics. The probability that the effect for person 1 is greater than the effect for person 3 is .978; and the probability that the effect for person 3 is greater than for person 2 is also .978. Thus, the people are quite different from each other, and we are fairly sure of this even though we have a small sample of observations on each person. Note also that the smaller gain of person 2 is not due to starting high and having a ceiling effect; person 3 started higher, yet gained more.

DETAILS OF THE MODELLING PROCEDURE

This section contains details for those who want to learn how the model is specified and run using the WinBUGS program; it can be skipped without loss of continuity by others. Much of the code will seem natural to those who have used other multilevel software, but there are some differences. One is that the data are usually organised into one file containing both individual-level variables and observation-level variables, unlike some programs that have separate level-1 and level-2 data files. To index subjects, a nested index is used; for example, if the intercept, b0, differs across subjects, one would refer to the intercept of a particular subject by using the notation b0[subject[i]], where *subject* is a variable indicating which subject made the response recorded in observation *i*.

A second major difference between a Bayesian program and frequentist program is the need to specify prior distributions for all of the parameters

TABLE 3
Estimated proportion of times observed in on-task behaviour during A and B phases, and difference between A and B phases with standard error

Person	A	B	B-A	Std. Err. (B-A)
1	.36	.96	.60	.08
2	.50	.68	.18	.09
3	.58	.99	.40	.07
Overall	.49	.91	.42	.14

in the model. Typically there are two types of parameters: Regression coefficients, which have normal priors, and precisions (or variances or standard deviations) which could be given gamma priors (inverse gamma for variances), or uniform distributions.

Another difference is that there are model statements that allow assessment of the probability that an effect is larger than a certain value, or smaller than a certain value, or in an interval; or that one parameter is larger than another. These use the BUGS step() function to create an indicator variable whose value is 1 if the expression is non-negative and 0 if it is negative.

As discussed previously, for the data set analysed here, there were four phases arranged in the order ABA′B, with the prime indicating that the second A phase was slightly different than the first. In the data set I included four indicator variables, one for each phase, called a, b, c, and d. These could be combined in various ways to express different models or expressions of models. For example, the variable b+d creates a dummy variable for the B phases of the design, and could thus test a model that compared the A phases with the B phases. To use effects coding rather than dummy coding more complicated new variables could have been created; for example, the main effect of the A vs B phases could be coded as (b+d)−(a+c), which would create the usual 1/−1 effects coding that equals 1 for the B phases and −1 for the A phases.

One problem in specifying a weakly informative or noninformative prior is that very large values of parameters can be chosen that cause problems in the estimation, particularly of variances. Thus I had to try different types of priors before finding some that would work correctly. Typically the choice of dgamma(.001,.001) for a gamma distribution is used in WinBUGS examples, which has a mean of 1 and a variance of 1000. Sometimes this is too extreme; other possibilities are dgamma(.01,.01) or even dgamma(.1,.1), which also have a mean of 1, but variances of 100 and 10, respectively. If these do not work, a uniform prior on the standard deviation can be used (which may be preferable in any case, being somewhat more noninformative, according to some researchers).

The main part of the model is the following; the complete model is in the appendix:

```
for(i in1:70)
{r[i] ~ dbin(pr[i], 6)
logit(pr[i]) <- b0[s[i]]+b1[s[i]] * (b[i] + d[i])+
   b2[s[i]] * c[i]}
```

The first line specifies that the statements following will apply to all 70 observations (21 sessions for respondent 1, 24 for respondent 2, and 25 for

respondent 3). The second line says that the observed variable $r[i]$, the number of times that a subjects shows on-task behaviour during a session, has a binomial distribution with probability $pr[i]$, out of six trials. The third and fourth lines are the actual model: The logit of the probability of on-task behaviour is an intercept (phase A baseline), plus an effect for phase B (either the first or second B phase, symbolised as $b[i]+d[i]$), plus an effect for the second A phase (represented by $c[i]$). The subject associated with observation i is represented by the variable $s[i]$. Thus, for example, b0[s[i]] is the intercept for the subject whose data is involved in observation i.

Because this is a logit model, the effects are in a scale that is difficult for most people to understand. For this reason additional statements are included in the model to transform important quantities from logits to proportions. The variables *diff* and *diff.i*[*j*] are differences between probabilities of on-task behaviour during the B period and the first A period, *diff* for the average and *diff.i*[*j*] for the three individuals.

Some of the prior distributions are specified as:

$$\text{mu2} \sim \text{dnorm}(0, .04)$$
$$\text{prec0} \sim \text{dgamma}(.1, .1)$$
$$\text{prec1} \sim \text{dunif}(.04, 25)$$

The parameter *mu2,* which can be any real number, is specified using a normal distribution with a mean of 0 and a precision of .04 (which corresponds to a variance of $1/.04 = 25$). The precision of the residual for the intercept, *prec0*, is given a gamma distribution as noted above, and *prec1* is given a uniform distribution. The lower bound of *prec1* is .04, corresponding to an upper bound on the variance of $1/.04 = 25$, or a standard deviation of 5. The upper bound is 25, which corresponds on a lower bound of $1/25 = .04$ for the variance, or .2 for the standard deviation. These values will not be right for all data, and some thought should go into their selection. In all three cases here, the estimated standard deviations were in the range of .8 to 1.9, well within the prior distribution limits.

DISCUSSION

Bayesian models offer several advantages in the analysis of data from single case designs. They allow the use of prior information about model parameters, but without forcing one to make assumptions about them if one knows little or nothing. Through computational techniques they allow one to make inferences about complicated functions of parameters without knowing the theoretical distribution of the function. Perhaps most

importantly, they allow one to make the kinds of statements one wants to make but cannot make in the traditional approach to statistics, such as the probability that a parameter is large, or that the probability is .95 that a parameter (or function of one or more parameters) is between a lower and upper value. An additional advantage that was not demonstrated here is that it is relatively straightforward to handle missing data in the Bayesian computational framework. (Missing data does not mean that subjects are observed at different times or different numbers of times; these are automatically handled by the multilevel model.)

Bayesian methods automatically adjust for uncertainty in the estimates of random effects when estimating the fixed effects. Most important are the level 2 random effects (variation among subjects), which are based on a small number of observations. This method is not used in classical statistics (or in standard multilevel programs such as HLM, SAS Proc MIXED, or others). This means that one can make incorrectly precise claims about fixed effects by underestimating their standard errors. This is especially relevant for SCDs because they usually involve a small number of respondents.

For those who want to learn more about Bayesian methods, good introductory texts are Berry (1996) and Winkler (2003). An introductory book that also covers multilevel models is Kruschke (2011), and at a more advanced level, Gelman and Hill (2007) and Ntzoufras (2009). More advanced still are the books by Gelman, Carlin, Stern, and Rubin (2004) and Congdon (2001). The theory that lies behind the multilevel model is in Lindley and Smith (1972). Sources specifically on WinBUGS by its developers are Lunn, Thomas, Best, and Spiegelhalter (2000), Lunn, Jackson, Best, Thomas, and Spiegelhalter (2012), and Spiegelhalter, Thomas, Best, and Lunn (2003). Andiloro and Rindskopf (2013) offer a simple introduction to WinBUGS for those who currently use HLM. Woodward (2011) provides an Excel interface to simplify setting up and running standard models in WinBUGS.

In this paper I used WinBUGS, which can be complicated to set up and run. There are several alternatives for simple multilevel models, including SAS and R programs; many include generalised linear models. I prefer WinBUGS (or an equivalent alternative such as OpenBUGS or JAGS) because of the flexibility to

(i) allow different residual variances for different individuals or phases (for normally distributed outcomes),
(ii) easily parameterise the model in different ways (e.g., a + bX vs. c(d + X) in logistic models),
(iii) parameterise priors in different ways,
(iv) sample easily from the posterior of transformed values (as in the example where I sampled the difference in probabilities),

(v) sample the complete posterior distribution, thus allowing easy estimation of interesting probabilities such as prob(theta > 0), prob(theta > a), prob(a < theta < b), and so on, where prob() stands for probability, theta for any parameter, and a and b are constants,
(vi) simply specify nonlinear models such as those with floor and ceiling effects for gradual change (Rindskopf, 2010).

One disadvantage of Bayesian methods based on simulation methods is that these methods are sometimes numerically unstable, and will not converge if they are not given somewhat informative priors and "reasonable" start values. But this is a minor inconvenience for a major gain. Bayesian methods started becoming popular largely because of their utility in handling missing data; now they are advancing because they offer many advantages. They should be in the armament of every data analyst.

REFERENCES

Andiloro, N. R., & Rindskopf, D. (2013). How to translate models from HLM to WinBUGS (with hints on preventing and solving common problems). Unpublished manuscript.

Berry, D. A. (1996). *Statistics: A Bayesian perspective*. Belmont, MA: Duxbury Press.

Congdon, P. (2001). *Bayesian statistical modelling*. Chichester: Wiley.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Amsterdam, The Netherlands: Academic Press.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, *34*, 1–41 (with discussion).

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. London: Chapman and Hall.

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.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley.

Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). Mahwah, NJ: Lawrence Erlbaum Associates.

Rindskopf, D. (2010). Logistic regression with floor and ceiling effects. In *JSM Proceedings*, (pp. 806–815). Alexandria, VA: American Statistical Association.

Rindskopf, D. (in press). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2013). A *d*-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation*. doi:10.1080/09602011.2013.819021

Spiegelhalter, D., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, *64*, 583–639.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual,* Version 1.4. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine. Cambridge: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Tasky, K. K., Rudrud, E. H., Schulze, K. A., & Rapp, J. T. (2008). Using choice to increase on-task behavior in individuals with traumatic brain injury. *Journal of Applied Behavior Analysis*, *41*, 261–265.

Winkler, R. L. (2003). *An introduction to Bayesian inference and decision* (2nd ed.). Gainesville, FL: Probabilistic Press.

Woodward, P. (2011). *Bayesian analysis made simple: An Excel GUI for WinBUGS*. Boca Raton, FL: Chapman & Hall/CRC.

APPENDIX

WinBUGS model code for the Tasky data set

```
#-------------------------------------------------------------
# Tasky.model.3a.ocd
# coded for A B A' B design features
#-------------------------------------------------------------
model
{ for (i in 1:70)
{ r[i] ~ dbin(pr[i], 6)
logit(pr[i]) <- b0[s[i]] + b1[s[i]] * (b[i] + d[i]) +
   b2[s[i]] * c[i]
}

for (j in 1:3)
{ b0[j] ~ dnorm(mu0, prec0)
b1[j] ~ dnorm(mu1, prec1)
b2[j] ~ dnorm(mu2, prec2)

eb0[j] <- exp(-b0[j])
eb1[j] <- exp(-1*(b0[j] + b1[j]))
pb0[j] <- 1/(1 + eb0[j])
pb1[j] <- 1/(1 + eb1[j])
diff.i[j] <- pb1[j] - pb0[j] }

mu0 ~ dnorm(0, .04)
mu1 ~ dnorm(0, .04)
mu2 ~ dnorm(0, .04)

prec0 ~ dgamma(.1, .1)
prec1 ~ dunif(.04, 25)
```

```
prec2 ~ dgamma(.1, .1)

var0 <- 1/prec0
var1 <- 1/prec1
var2 <- 1/prec2

sd0 <- sqrt(var0)
sd1 <- sqrt(var1)
sd2 <- sqrt(var2)

e0 <- exp(-mu0)
e1 <- exp(-1*(mu0 + mu1))
p0 <- 1/(1 + e0)
p1 <- 1/(1 + e1)
diff <- p1 - p0

diff2 <- step( diff - .2)
compare13 <- step(diff.i[1] - diff.i[3])
compare32 <- step(diff.i[3] - diff.i[2] )

a[1] ~ dnorm(0,1) # unused variable
}
```