The application of Bayesian analysis to issues in developmental research

Lawrence J. Walker, Paul Gustafson and Jeremy A. Frimer

International Journal of Behavioral Development 2007 31: 366
DOI: 10.1177/0165025407077763

International Journal of Behavioral Development
2007, 31 (4), 366–373
http://www.sagepublications.com

DOI: 10.1177/0165025407077763

The application of Bayesian analysis to issues in developmental research

Lawrence J. Walker, Paul Gustafson, and Jeremy A. Frimer
University of British Columbia, Canada

This article reviews the concepts and methods of Bayesian statistical analysis, which can offer innovative and powerful solutions to some challenging analytical problems that characterize developmental research. In this article, we demonstrate the utility of Bayesian analysis, explain its unique adeptness in some circumstances, address some concerns and misconceptions about the approach, and illustrate some applications of Bayesian analysis to issues that frequently arise in developmental research. The illustrations of the approach provided here reflect several important issues within the domain of moral reasoning development (such as assessing patterns of stage change over time); however, the methods are readily applicable across content areas in developmental research.

Keywords: Bayesian statistics; frequentist statistics; moral reasoning

The purpose of this article is (1) to introduce Bayesian statistical analysis to the field of developmental psychology, (2) to demonstrate its utility, (3) to explain its unique adeptness in some circumstances, (4) to address some concerns and misconceptions about the approach, and (5) to illustrate some applications of Bayesian analysis to issues that frequently arise in developmental research. The Bayesian approach offers innovative solutions to some challenging analytical problems that plague research in developmental psychology. The illustrations of the approach provided here are in the domain of moral reasoning development, but the methods are applicable across content areas in developmental research.

1. Introduction to Bayesian analysis

There are numerous books describing the Bayesian approach to statistics (some accessible examples include Congdon, 2003; Gelman, Carlin, Stern, & Rubin, 2004; Gill, 2002). The essence of Bayesian analysis, however, can be expressed succinctly. An investigator wishes to learn about some unknown *parameters*, which we denote as P. Typically, these parameters are attributes of the population under study and, particularly in the developmental context, they would include descriptors of change over time. The investigator starts by formulating a probability distribution over the possible values of P. This *prior distribution* is taken to reflect the initial beliefs about P. These could be the investigator's personal beliefs or, in some instances, the aim might be to formulate a prior distribution representing a diversity of initial opinions among involved parties. Here, there is a philosophical shift in using probability theory to describe belief about fixed but unknown quantities (the parameters) as opposed to physically random processes such as the outcomes of coin tosses. In fact, there is a rich literature on axiomatic development of so-called subjective probability (see Bernardo & Smith, 1994, for an overview). Setting aside the philosophical underpinnings, however, the basic notion is that the possible values of the parameters are weighted in advance. For instance, say x and y represent two possible and competing sets of values of the parameters P. If $P = x$ is considered twice as likely as $P = y$, in advance of observing the forthcoming data, then the chosen prior distribution should assign twice as much probability to x as it does to y.

To be more explicit, the specification of a prior distribution typically boils down to choosing a center (say the mode of the prior distribution) and a spread (say the standard deviation). The center can be regarded as an initial best guess at P and, by varying the spread, the investigator can infuse less or more background knowledge about P into the forthcoming analysis. Indeed, less is often perceived as desirable in striving for scientific objectivity, so that widely spread prior distributions are relatively common in practice.

Next, the investigator collects data (D), thought to be linked to P via a known conditional probability distribution or *statistical model*, for D given P. As a simple example, say D consists of test scores for n participants regarded as a random sample from the population under study, with P being population mean and standard deviation scores. Then, a supposition that the scores are normally distributed across the population would comprise a statistical model for D given P. This part of the inferential process is inherent in all statistical methods, and the science and art surrounding how to specify this model and then check its appropriateness once data are observed cuts across statistical paradigms.

Once the investigator has chosen the prior distribution and statistical model he or she thinks are appropriate and once the data are in hand, there is no ambiguity about how to carry

Correspondence should be sent to Dr Lawrence J. Walker, Department of Psychology, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada; e-mail: lawrence.walker@ubc.ca

The illustrations of some of the applications of Bayesian statistics provided here are based on data originally reported by Walker, Gustafson, and Hennig (2001).

through the Bayesian analysis. Bayes theorem, which is a simple application of the mathematical laws of conditional probability, determines the conditional distribution of the parameters *P* given the data *D*, referred to as the *posterior distribution*. A mathematical statement of Bayes theorem is given in the Appendix. The posterior distribution encapsulates all knowledge about *P* after having seen *D*, and various inferential statements can be "read-off" from this distribution. For instance, some measure of the center of this distribution (i.e., a mean, median, or mode) can be taken as the single best guess for the value of *P* (i.e., a point estimate). Similarly, a summary of the posterior distribution's variability (such as the standard deviation of this distribution) reflects the accuracy of the estimate, while percentiles of this distribution can be used to form interval estimates. For instance, the 2.5th and 97.5th percentiles can be regarded as a *95% credible interval* for *P*, which serves as an alternative to 95% confidence intervals arising from standard non-Bayesian (*frequentist*) statistical techniques.

Whereas Bayesian analysis is relatively simple conceptually, there are many issues, both practical and philosophical, surrounding whether and when this approach to statistical inference is preferable to the more established frequentist approach. We take up some of these issues in section 2, but first we give a simple example of Bayesian analysis applied to developmental data in order to illustrate the basic concepts.

Consider the study of moral reasoning development described by Walker, Gustafson, and Hennig (2001). In this longitudinal study, 64 children and adolescents participated in five annual administrations of Kohlberg's Moral Judgment Interview (MJI; Colby & Kohlberg, 1987). These individual interviews were recorded, transcribed, and then scored for stage of moral reasoning development. Later we consider a more detailed analysis of stage-to-stage transitions on the five moral stages but, as a simple initial illustration, we consider the composite *weighted average score* (Colby & Kohlberg, 1987) to indicate level of moral reasoning development. This score is given by the sum of the products of the percent usage at each stage multiplied by the stage number (and thus has a range of 100–500). So, each participant has a series of five numerical scores, one at each of the five time-points, with each score being on a scale from 100 to 500.

So, to illustrate Bayesian analysis with a straightforward example from this study, consider making inferences about the average score on the baseline interview in this population (i.e., What is the typical level of moral reasoning for children in this age range?). Although this may not be of great interest developmentally, because there is no consideration (yet) of change across time, it does serve to provide some intuition concerning Bayesian analysis. Say we start with a wide prior distribution for this quantity, namely a normal distribution centered at 300 (i.e., equivalent to moral Stage 3) with a *SD* = 75. This distribution is represented as the black curve in the left-hand panel of Figure 1. Now say we obtain some data, in the form of baseline scores for five (randomly chosen) participants. These data are combined with the prior distribution to give the posterior distribution of the underlying mean, portrayed as the least-peaked of the gray curves in the figure. Upon adding more data, the posterior distribution becomes more concentrated, reflecting increased knowledge about the population quantity. The successively more peaked curves in the figure represent the posterior distributions based on data from 10, 15, and 20 participants in total.

Thus, after 20 participants have been considered, the posterior distribution is centered at 265 points, with 95% of the posterior probability concentrated within 8 points of the center. These are Bayesian point and interval estimates for the population quantity of interest. Numerically, they are almost identical to the usual frequentist estimates (sample mean as the point estimate; sample mean ± 1.96 standard errors as the 95% interval estimate). Of course, the statistics of the two methods reflect slightly different phenomena. The Bayesian inferences are framed in terms of probabilistic expression of belief about unknown quantities, so that one can ascribe 95% probability to the interval 265 ± 8 containing the quantity of interest. Conversely, the frequentist inferences are framed in terms of hypothetical repetitions of the experiment, with 265 ± 8 being the result from the one real repetition, and 95% of the hypothetical repetitions producing a correct interval. Notwithstanding the differing interpretation, this sort of numerical agreement between frequentist and Bayesian inferences is not uncommon in *simple* statistical settings.

The right-hand panel in Figure 1 describes the accumulation of Bayesian evidence for the same sequence of data (i.e., the same sequence of 20 participants), but starting with a dramatically different prior distribution. In particular, consider an investigator who started with a prior distribution centered at 400 (i.e., equivalent to moral Stage 4) with a *SD* = 25. With the hindsight of observed data, this investigator's prior convictions turned out to be very poor, as evidenced by a sequence of posterior distributions which march to the left as the data accumulate, and end up being rather similar to those in the

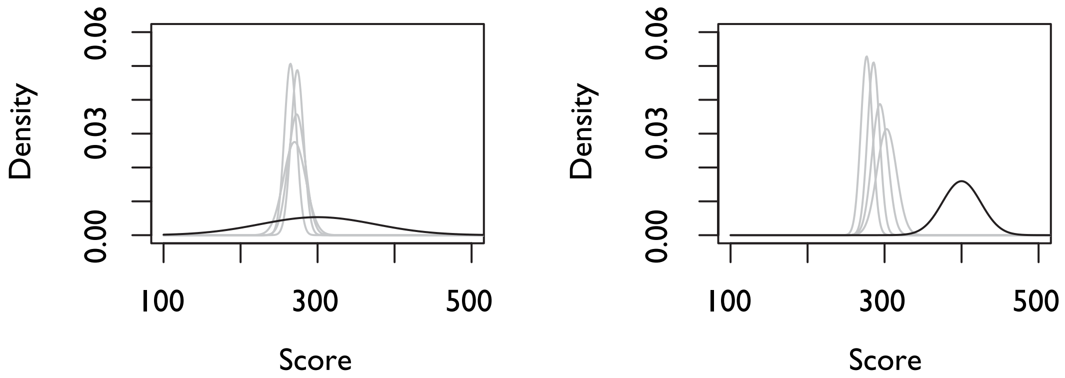


Figure 1. Bayesian inference on the population-average baseline score. In each panel the black curve gives the prior distribution, while the four successively more peaked curves give the posterior distribution after observing data from 5, 10, 15, and 20 participants.

left-hand panel once 20 data-points are in hand. That is, whereas the "bad" prior distribution had some impact on the posterior distribution based on data from 5 participants, this impact is essentially dissipated once we have data from 20 participants. Again this is a common phenomenon, at least in simple settings. Relatively modest amounts of data will "trump" prior beliefs. More precisely, mathematical theorems guarantee that the prior distribution will be "forgotten" as the sample size increases.

We return to the composite scores in section 3, where we carry out a Bayesian analysis of change over time; but first we make some general remarks about the advantages and difficulties of Bayesian analysis.

2. General features of Bayesian analysis

In this section, we discuss several aspects of Bayesian analysis, including that it: (1) is a general, comprehensive, and unified approach; (2) is precise; (3) is more interpretable than frequentist approaches; (4) can now be conducted with computational ease; (5) can account for multiple uncertainties; (6) is natural with hierarchical models and those involving unobserved latent structure; and (7) entails the ability to evaluate prior information.

Perhaps the nicest feature of the Bayesian approach is that it provides a general and complete strategy for the analysis of data. After specifying a model for how the data arise given the parameters and a prior distribution for the parameters, there is no ambiguity about the form of statistical inferences. In contrast, the frequentist statistical toolbox, as represented say by the content of standard statistical software packages, is comparatively scattered (e.g., regression versus analysis-of-variance approaches). There is a tendency for certain methods and certain estimation techniques to be advocated in certain kinds of problems based mostly on historical precedent. For instance, "structural equation modeling" is a widespread technique in some quantitative disciplines (e.g., psychometrics), but uncommon in others (e.g., biostatistics).

Generally speaking, the two "complete" theories for statistical inference, in the sense of providing comprehensive recipes for getting point and interval estimates of target quantities, are Bayesian theory and maximum-likelihood theory. However, various pressures have led to more fracturing in the likelihood camp, so that the literature abounds with mention of methods such as quasi-likelihood, pseudo-likelihood, restricted likelihood, penalized likelihood, and so on. Meanwhile, the Bayesian approach, with its simple and elegant foundation in the laws of conditional probability, remains unified.

In addition to being unified, there are underlying mathematical reasons why Bayesian inferences are especially "good" in terms of being as precise as possible. In particular, Bayesian estimators emerge as optimal inferential procedures under standard decision-theory arguments (for instance, see Berger, 1985). Roughly put, there is a mathematical guarantee that a Bayesian estimate of a parameter has the smallest possible mean-squared-error (in an aggregate sense across all possible values of the parameter, where this aggregation is weighted according to the prior distribution). Admittedly this is a rather technical sense in which Bayesian answers are the best possible but, in fact, it has clear implications. In particular, a Bayesian estimate can never be "clearly beaten" by a non-Bayesian estimate, where "clearly beaten" means having a larger mean-squared error for every possible set of values for the true parameters.

Bayesian inferences are arguably more interpretable than frequentist inferences. As mentioned earlier, in the simple example of section 1, we can say that, with 95% probability, the parameter of interest lies in the interval of 265 ± 8 points. This probability statement describes degree of belief (i.e., the investigator would view a bet involving 19:1 odds that the parameter is in this interval as "fair"). In contrast, the analogous interpretation of a frequentist confidence interval is rather more convoluted: If we repeat the experiment over and over again, then 95% of these repetitions give an interval containing the parameter of interest.

The differences between the approaches are even more extreme when comparing two or more statistical models in light of data, often referred to as *hypothesis testing*. The Bayesian approach extends the concept of a posterior distribution over parameters to that of a posterior distribution over hypotheses and parameters jointly. Thus, one can end up with a claim such as: Given the data, there is only a 3% probability that a simple model (a "null" hypothesis) is true versus a 97% probability that a more complex encompassing model is true (an "alternative" hypothesis). In contrast, if a frequentist hypothesis test of null versus the alternative produces a p-value of .03, then the interpretation is again convoluted: Given that the null hypothesis is true, there is only a 3% chance of observing data as or more extreme than those actually observed. Thus, it is not surprising that significance testing of this sort has engendered considerable angst within psychology in recent years (Harlow, Mulaik, & Steiger, 1997). Whereas one response to critiques of significance testing is to shift focus from significance testing to estimation of effect sizes instead, another response is to adopt Bayesian methods for hypothesis testing.

The unintuitive nature of p-values and classical hypothesis testing, more generally, are notorious for causing students and researchers grief, in part because of the erroneous tendency to interpret the p-value as a probability of the null hypothesis being true. Bayesian analysis gives investigators what they want! Although we do not pursue Bayesian hypothesis testing further in this article, it is worth mentioning in passing that this is *not* one of the simple settings where numerical agreement between Bayesian and frequentist answers is common. In particular, the Bayesian probability that a null hypothesis is true is often substantially *larger* than the corresponding frequentist p-value. This may initially sound negative, in the sense of Bayesian methods being less sensitive to detect real "effects." However, the ease with which frequentist tests reject null hypotheses has been implicated as a driving factor behind misleading claims, which are initially "proved," but then subsequently discredited. In particular, using .05 significance-level testing as the threshold for publishing a "research finding" has been demonstrated to produce a literature in which far more than 5% of published findings are, in fact, false (see Ioannidis, 2005, for a recent and accessible discussion of this point).

If the Bayesian approach is principled and interpretable, then why is it not ubiquitous? Until about 1990, computational limitations were the Achilles' heel of Bayesian inference. The posterior distribution could be written in abstract mathematical form, but could only readily be computed in cases of very simple statistical models and prior distributions. The computation involves the evaluation of integrals which, in general, is a hard numerical problem that suffers from a *curse of dimensionality* – the computational burden grows exponentially with the number of parameters involved. For this reason, the Bayesian approach was viewed as something of a curiosity with the reservation that, regardless of its merits, it could not be implemented across a wide range of problems. The breakthrough came with the development of new algorithms or, more accurately, the adaptation of old algorithms from the statistical physics community, which could break the curse. Roughly put, these algorithms do not quite compute the integrals involved, but they do enough to generate random samples from the posterior distribution which, in turn, is sufficient to give a picture of the posterior distribution. These computer-generated samples can be made arbitrarily large, so that Bayesian estimates can be computed as precisely as desired. These algorithms, usually referred to collectively as Markov chain Monte Carlo (MCMC) algorithms, have continued to be a fulcrum of intense statistical research activity ever since the seminal article of Gelfand and Smith (1990).

Notwithstanding the advent of MCMC algorithms, in some quarters computation is still viewed as a weak link in the Bayesian firmament. It is true that, in relatively simple statistical models, implementing and tweaking MCMC-based answers requires more effort than for most frequentist analyses. However, MCMC algorithms are particularly adept at handling models involving unobserved or latent structure. The *hierarchical models* considered in the next section constitute an example of this. Also, there have been efforts to "package" MCMC algorithms in a user-friendly manner. Most notably, the WinBUGS software (Spiegelhalter, Thomas, Best, & Lunn, 2003) takes a user's prior and model distributions and produces a posterior distribution, while the user is shielded from details of the MCMC computations used behind the scene. This Windows-based software has a graphical interface to facilitate the process of inputting model and prior distributions. One testament to its success is that the traffic on the WinBUGS e-mail discussion list (bugs@jiscmail.ac.uk) is generated by a healthy mix of statisticians and subject-area researchers; that is, the latter group is now interested and willing to go the Bayesian route.

There are various other advantages claimed for the Bayesian approach in particular circumstances. In many situations, Bayesian inferences can be regarded as "more honest" in the manner that they simultaneously account for multiple uncertainties. Other methods often estimate parameters for one part of the model and then feed these in as known true values in another part of the model, resulting in overly confident inferences at the end of the day (i.e., confidence intervals which are unjustifiably narrow). This issue arises commonly in random-effect models. A frequentist confidence interval for an individual (participant-specific) random effect is based on the pretense that the random-effect variance is known when, in fact, it is only estimated.

Also, Bayesian modeling and computation are very natural in models with "unobserved latent structure," as is often postulated in psychometrics contexts, for instance. Here a joint posterior distribution (or density) over latent (unobserved) variables and parameters arises, which is readily computed via MCMC algorithms. In contrast, non-Bayesian approaches require evaluation of the likelihood function for observed variables given parameters. Technically, this involves mathematical integration of the joint density function over the latent variables, which is not always readily implemented in software. This advantage of the Bayesian approach extends to dealing with missing data, as conceptually the missing values are just further unobserved or latent variables whose uncertainty is described by the joint posterior distribution.

Another very natural use of the Bayesian approach arises when data and parameters are organized in a "hierarchical" manner, with an archetype being data from student examinations, whereby students are nested within classes, which are nested within schools, which are nested within school districts, and so on. In fact, we give examples of the Bayesian approach to hierarchical models in the next section.

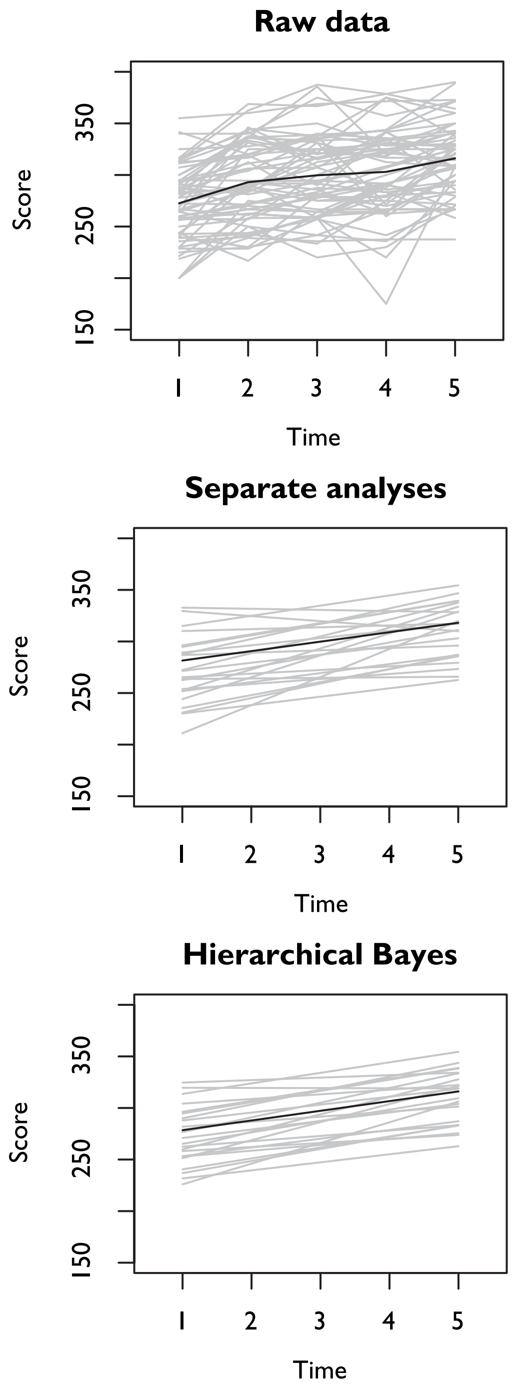
The most contentious part of the Bayesian statistical paradigm is undoubtedly the formulation of prior distributions. Some argue this to be a great strength – they hold that virtually any scientific investigation is carried out against a backdrop of considerable previous work and related knowledge, and that this should be acknowledged explicitly when drawing inferences. Others see it as neutral – they are not generally keen on trying to encapsulate prior knowledge in a very precise way, but are pleased to reap the other benefits of Bayesian analysis. Such investigators tend to use widely spread prior distributions, with the aim of "letting the data speak for themselves." Others are critical – they argue that subjective assessments have no place in science. The standard Bayesian retort is that frequentist statistical methods have their own subjectivities if you merely scratch the surface. In addition, Bayesian inference is often used in a way that acknowledges and accommodates this concern by formulating a few different prior distributions, and then reporting inferences arising from each. If these inferences are, in the end, very similar, then there is confidence that the conclusions hold across a wide range of prior beliefs. If not, at least the important role of prior information in the given problem has been clearly identified. The ability to identify, acknowledge, and evaluate potential biases is a particular strength of the Bayesian approach.

3. Hierarchical Bayesian models

As alluded to in section 1, in simple settings, Bayesian and frequentist inferences often agree closely in numerical terms, despite having different conceptual interpretations. With more complex data structures, however, such agreement is not a foregone conclusion. A common example of this is when the data have *hierarchical* structure. For instance, in the Walker et al. (2001) moral reasoning study, individual item responses are nested within time-points, which are nested within participants.

Again consider Walker et al.'s composite weighted average scores for each participant, but now including all five time-points, rather than just the initial one as in the earlier example. Later we consider these data at the finer level of proportion of moral reasoning at each stage. The raw scores (for the 64 participants with data at all five time-points) are plotted in the top of Figure 2 (gray lines), with the average score at each time-point superimposed (black line). The "jaggedness" in the scores across time is probably attributable to measurement error in the interview scheme, rather than real short-term oscillations in moral reasoning development.

In order to account for measurement error, we postulate that the observed data for each participant arises from an underlying linear (developmental) change, superimposed with random noise. That is, every participant has an underlying or true score which varies linearly over time. Moreover, the magnitude of the random noise is assumed to be inversely proportional to



Raw data

Separate analyses

Hierarchical Bayes

Figure 2. Composite scores over time. The top panel gives the raw scores over time for all 64 participants with complete data. The middle panel gives the result of fitting separate weighted regressions for each participant. Only the fitted lines for every third participant are displayed, to make the plot less crowded. The solid line is the result of a pooled analysis. The bottom panel gives the results of fitting a hierarchical Bayesian model. Here the solid line is the estimated relationship for an average participant.

the number of scoreable items at that time-point (which in this data-set ranges from 3 to 13). In statistical terms, this corresponds to a weighted linear regression model for each participant, that is, regressing score on time. We fit this model to each participant's data separately. That is, we conduct a Bayesian analysis for each participant separately, which roughly can be regarded as achieving the lower part of the hierarchical analysis. The estimated linear relationships for every third participant appear in the middle panel of Figure 2 (including results for all 64 participants makes the plot too crowded). We see there considerable variation in the intercepts of these lines (which represent baseline scores), and some variation in the slopes of these lines (which represent rates of development over time). We now have 64 separate posterior distributions in hand and, thus, can ask if they are somehow related.

At the other extreme from fitting a separate model for each participant (which suggests differing and independent rates of development for each participant), one can postulate that a single linear relationship governs all participants and that all the variation in the data is the result of random noise deviations from this single line. The estimated relationship from this "pooled" analysis is superimposed on the middle panel of Figure 2. Given the variation in the estimated time–score relationships from the 64 separate regression fits, the appropriateness of the pooled analysis seems dubious. Or, more to the point, it seems implausible that all participants have the same underlying score at baseline and the same underlying rate of development over time. By contrast (and key here), one expects some commonality in participants' rates of development, and this is not reflected in fitting a separate model for each participant. One wonders if something intermediary between separate analyses and pooled analysis might be appropriate. An intermediary approach can be achieved via a "mixed model" involving random effects. Although there are approaches to mixed models without an explicitly Bayesian slant, arguably these are simply approximations to hierarchical Bayesian models.

The hierarchical Bayes approach to problems like this does indeed achieve this compromise between separate and pooled analysis. The form of the crucial prior information is that participant-specific parameters (intercepts and slopes in our case) have a certain distribution across the population from which participants are drawn. As this distribution narrows, we get closer to the pooled analysis case; that is, the variation in the participant-specific parameters diminishes. Conversely, a wide distribution leads toward separate analyses without commonality across participants. The nice feature here is that the investigator need not try to judge the width of this distribution in advance. Rather, this width, which initially is viewed as characterizing a prior distribution, is itself assigned a prior distribution (one interpretation is that of hierarchically placing prior distributions upon prior distributions). Thus, the data themselves will answer the question of whether or not there is a relationship by suggesting an appropriate width. Consequently, we can take advantage of the amorphous prior assertion that there is some level of commonality in the participant-specific parameters. The approach is described as "hierarchical" in that it involves a model for observable data given participant-specific parameters, followed by a distribution for the participant-specific parameters given common parameters, followed by a (prior) distribution for the common parameters. The model fitting, however, is done simultaneously, rather than in sequential stages.

It should be mentioned that the use of participant-specific parameters is useful for representing a specific kind of heterogeneity in a population. Another kind of heterogeneity arises from mixture models, where participants belong to one of several clusters, with common parameter values within a cluster and no overt information about which participants belong in which clusters. Many nice features of Bayesian analysis do carry over to mixture model settings; see Marin, Mengersen, and Robert (2005) for a review.

The Appendix gives some further technical details concerning the specification and fitting of this hierarchical model to the present data. We focus on the results of this analysis, as given by the estimated linear relationships (for every third participant, as before), in the bottom panel of Figure 2. The impression from the plot is that the variation in underlying baseline scores is about the same as with separate analyses. The variation in the slopes, however, is less than we see from separate analyses. That is, in aggregate, the lines are closer to parallel now. Overall, then, Bayesian analysis indicates that there should be almost no pooling of the baseline score estimates, but that there should be some degree of pooling for the slope estimates. For instance, the 64 slope estimates from the separate analyses have a *SD* = 14.2 (with a range of –9.7 to 54.4), whereas the corresponding estimates from the hierarchical Bayesian analysis have a *SD* = 11.4 (with a range of only –0.3 to 40.5). There are strong conceptual and theoretical statistical arguments to suggest that, relative to separate analyses, the hierarchical Bayesian analysis gives a better assessment of how much participant-to-participant variation exists in the rate of development over time.

The hierarchical Bayesian approach can be adapted to a more sophisticated modeling of participants' distribution of moral reasoning across stages on Kohlberg's MJI scale. Of course, the hard-stage model that Kohlberg (1984) posited makes the primary claim not of average change in composite moral reasoning scores across individuals, but rather of invariant order in the acquisition of the five moral stages that make up the sequence; that is, that individual development should be irreversibly forward, one modal stage at a time, with no regressions and no stage-skipping (for a fuller explication of Kohlberg's model, see Lapsley, 2006, and Walker, 1988, 1996). Longitudinal data regarding this claim have been impervious to statistical test because of the difficulty in distinguishing real change from random fluctuations and measurement error. Bayesian analysis handles this problem with relative ease.

The basic premise of the hierarchical Bayesian approach to this problem is that, at each time-point, each participant has an underlying distribution of moral reasoning; that is, a certain percentage at Stage 1, a certain percentage at Stage 2, etc. Then the MJI scoring is regarded as a "noisy" reflection of this, via a simple statistical model. For instance, if the underlying distribution across the five moral stages (in percentage terms) is (5, 10, 50, 25, 10), then each scoreable item is regarded as having these percentage chances of being scored at these stages. Technically, then, the scores can be regarded as multinomial data, given the underlying percentages or probabilities. Thus, for a given participant and time-point, we distinguish between the observed data (stage scores in the form of frequencies) and the unknown parameter (true distribution of reasoning over stages, expressed in percentage terms).

It should be noted that our framework for this problem is quite different from a latent variable framework. For instance, a latent transition analysis (LTA) might be envisioned for these data (for a review of LTA, see Lanza, Flaherty, & Collins, 2003). LTA would assume that each participant at each time-point belongs to a single true (but unobservable) stage, and the percentage chances for scoring an item are a function of this stage (but the same for participants at the same stage). Thus, whereas LTA uses a mixture-model approach to capture heterogeneity, the approach we describe uses random effects. Both non-Bayesian and Bayesian approaches might be used to fit LTA models and, in fact, Lanza, Collins, Schafer, and Flaherty (2005) describe the use of Bayesian computational techniques to facilitate standard error calculations in LTA.

Now we use the hierarchical Bayesian idea to postulate structure in how the moral reasoning distribution changes over time. In this instance, we compromise in the extent to which data are pooled across time. At one extreme there is no pooling, so that temporal change in the distribution can be arbitrarily "jagged." At the other extreme, the change is structured and smooth (in a manner which is elaborated upon in the Appendix). As in the previous application of hierarchical modeling, we do not fix the nature of the compromise in advance. Rather, the relevant parameters are assigned a prior distribution, to let the data speak about appropriate values and commensurately appropriate points on the smooth–jagged spectrum.

One simple application of the Bayesian model is to study the change in modal stage across time. The primary claim of Kohlberg's structural-developmental model is that individual development is irreversibly forward, one modal moral stage at a time. If we examine the raw-data estimates of modal stage, we find that there are no instances of stage-skipping from one time-point to the next for any participant. And further, consistent with Kohlberg's model, the estimated modal-stage never declines across time for 57 of the 64 participants (32 of these have estimated modal stages with no variation over time, while the other 25 exhibit at least one increase to the next stage). However, there are seven participants exhibiting a decline in modal stage at some time-point, and the natural question is whether these are real regressions in the underlying moral reasoning or simply artifacts of the assumed (multinomial) noise in the data.

For each participant, our model yields the posterior probability of a nondecreasing modal-stage sequence given the data. These come directly from the MCMC-based model-fitting procedure (with further details given in the Appendix and with the code available for download at www.stat.ubc.ca/~gustaf). They range from 22 to 99%, with 61 of 64 participants having probabilities >50% (confirming the Kohlbergian hypothesis). Among these, 39 participants have individual probabilities exceeding 80%. Consider, for instance, the 53rd participant, one of the seven with raw estimates suggesting a decrease in moral stage (from modal Stage 3 at t_1 to Stage 2 at t_2, and then remaining at that stage until progressing back to Stage 3 at t_5). This participant's raw data (expressed as the distribution of reasoning across stages at each time-point) are displayed in the upper panel of Figure 3. The lower panel of the figure gives Bayesian estimates of the participant's moral reasoning development over time. The analysis suggests a lot more smoothness in the underlying moral development than is evident in the raw data and, in fact, the Bayesian estimates display a constant modal stage over time (viz., at Stage 2).

The figure conveys only the "best guess" at the underlying moral stage sequence, but the model itself is capturing how much uncertainty to attach to this guess. In fact, for this participant, the posterior distribution ascribes 52% probability to the sequence being nondecreasing, and 48% to the sequence having at least one decline. That is, the data are equivocal on the sequentiality question for this participant; but more particularly, the decline seen in the raw-data modal

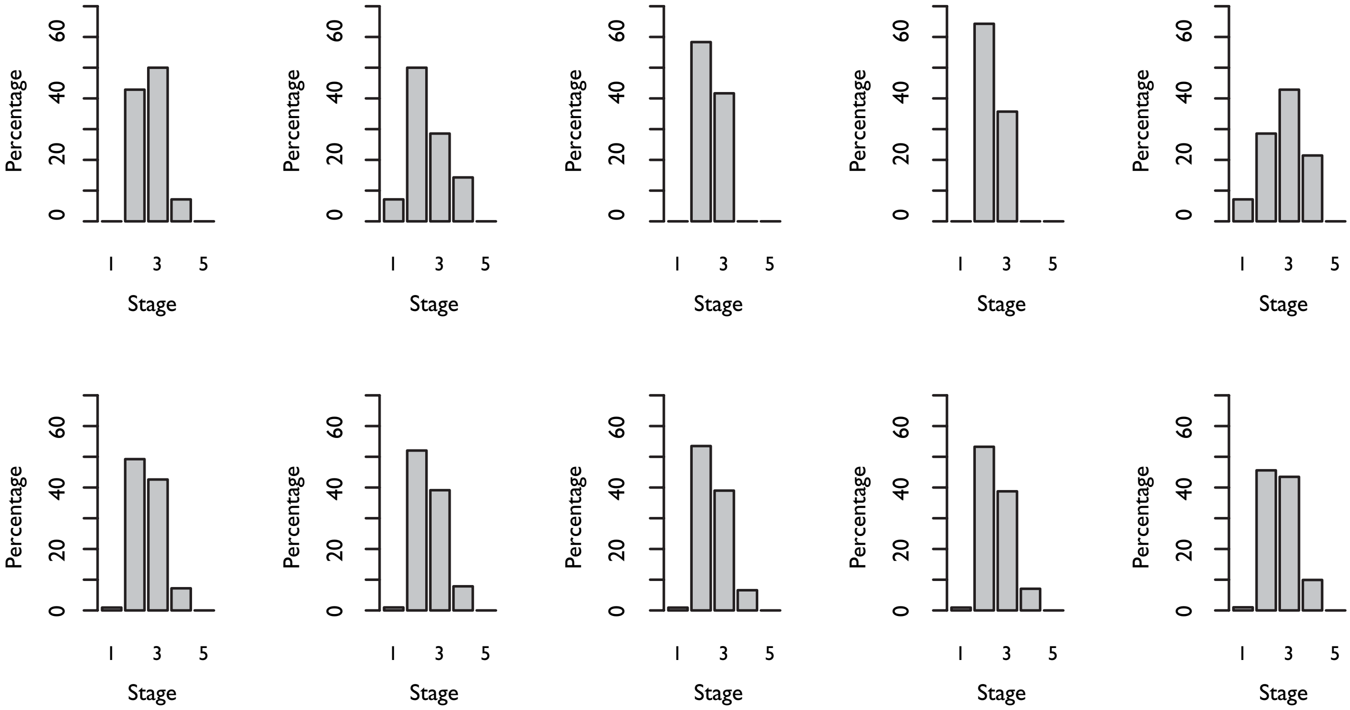


Figure 3. Raw data (top row) and Bayesian estimates (bottom row) for the distribution of reasoning across stages, for the 53rd participant. The five panels per row correspond to the five time-points. The numbers of scoreable items for this participant across time-points are (7,7,6,7,7).

estimates cannot be regarded as evidence of a real regression in moral stage in this case.

Providing posterior probabilities on hypotheses about the moral stage sequence are just one example of how Bayesian inferences can be used to characterize evidence about developmental patterns in a nuanced way, rather than simply checking whether or not raw data sequences match the hypothesized pattern in question. In fact, Walker et al. (2001) use the same idea to evaluate more complex hypotheses about cyclical patterns of consolidation and transition phases in stage progressions.

4. The utility of Bayesian methods for developmental research

There does seem to be undeveloped potential for Bayesian methods to be useful in developmental research. As exemplified in section 3, Bayesian techniques are well-suited to the sharing of information across participants and/or across time. The use of Bayesian methods to infer change over time has been well-explored in health research and other settings (see, for instance, Congdon, 2003), with the infusion of realistic prior information about the magnitude and smoothness of such changes being a central theme. Developmental researchers should consider what such techniques can bring to bear on their research programs.

A particular advantage of Bayesian analysis manifested in the second example of section 3, and also in the Walker et al. (2001) study, involves the assessment of how consistent an empirical sequence of data over time is with a postulated developmental pattern. A raw-data sequence is either consistent or not, but this ignores the measurement error inherent in empirical data. With frequentist statistical methods one can obtain point estimates and confidence intervals for the parameters underlying the raw data. But again, the point estimates will either be consistent or inconsistent with the postulated pattern, and it is not obvious how to account for the uncertainty represented by the confidence intervals. On the Bayesian side, however, one has a posterior distribution over the parameters, which directly implies a posterior probability that the parameters are consistent with the pattern.

Although only mentioned tangentially in this article, another potential advantage of Bayesian methods in developmental contexts involves the comparison of competing hypotheses. For instance, Walker et al. (2001) quantified the evidence *in favor* of a particular hypothesis about temporal patterns in stage progressions against a vague hypothesis of arbitrary change over time. That is, Bayesian methods treat null and alternative hypotheses in a symmetric fashion, whereas frequentist methods reflect only the strength of evidence against a null hypothesis. The limitations of classical hypothesis testing in psychological science contexts has been the subject of considerable commentary (Trafimow, 2003). Because empirical support for specific theories about patterns of development is crucial to the discipline, Bayesian methods deserve consideration.

Another issue which we have side-stepped is that of missing data. For simplicity our example analyses are based only on the 64 participants in the Walker et al. (2001) study who completed all five interviews. A further five participants completed only some of the interviews. As mentioned earlier, though, with some extra effort data from these participants could be added, by treating the missing values as latent variables "inside" the posterior distribution. A nice feature of such an approach is that the posterior distribution then naturally propagates uncertainty about the missing data values to uncertainty about the parameters of interest. Ibrahim, Chen, Lipsitz, and Herring (2005) review Bayesian approaches to missing data. This is clearly an important topic in developmental studies, given the need to obtain longitudinal data with the attendant problems of participant attrition.

On the technical side, it might be argued that Bayesian analysis is more demanding of its users, though advances in algorithms and software have, and will likely continue to improve this situation. In 1975, a prominent statistician famously predicted that while the twentieth century (at least to that point) belonged to frequentist methods, the twenty-first century would be "Bayesian" (Lindley, 1975). While the data are not all in yet, this prediction may turn out to be close to the mark.

References

Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Bernardo, J.M., & Smith, A.F.M. (1994). *Bayesian theory*. New York: Wiley.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment* (Vols 1–2). New York: Cambridge University Press.

Congdon, P. (2003). *Applied Bayesian modelling*. New York: Wiley.

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman and Hall/CRC.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R., & Herring, A.H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. Available [accessed 20 November 2005] http://dx.doi.org/10.1371/journal.pmed.0020124

Kohlberg, L. (1984). *Essays on moral development: Vol. 2. The psychology of moral development*. San Francisco: Harper & Row.

Lanza, S.T., Collins, L.M., Schafer, J.L., & Flaherty, B.P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods*, *10*, 84–100.

Lanza, S.T., Flaherty, B.P., & Collins, L.M. (2003). Latent class and latent transition analysis. In J.A. Schinka & W.F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 663–685). Hoboken, NJ: Wiley.

Lapsley, D.K. (2006). Moral stage theory. In M. Killen & J.G. Smetana (Eds.), *Handbook of moral development* (pp. 37–66). Mahwah, NJ: Erlbaum.

Lindley, D.V. (1975). The future of statistics – A Bayesian 21st century. *Advances in Applied Probability*, *7*(Suppl.), 106–115.

Marin, J.-M., Mengersen, K., & Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. In D.K. Dey & C.R. Rao (Eds.), *Handbook of statistics: Vol. 25. Bayesian thinking, modeling and computation* (pp. 459–507). Amsterdam: Elsevier.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual* (Version 1.4). Cambridge: Cambridge University, MRC Biostatistics Unit. Available [accessed 25 November 2005] http://www.mrc-bsu.cam.ac.uk/bugs

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110*, 526–535.

Walker, L.J. (1988). The development of moral reasoning. *Annals of Child Development*, *5*, 33–78.

Walker, L.J. (1996). Kohlberg's cognitive-developmental contributions to moral psychology. *World Psychology*, *2*, 273–296.

Walker, L.J., Gustafson, P., & Hennig, K.H. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology*, *37*, 187–197.

Appendix

Bayes theorem gives the distribution over a parameter P given data $D = d$ as

$$\Pr\left(P = p \mid D = d\right) = \frac{\Pr(D = d \mid P = p) \times \Pr\left(P = p\right)}{\sum_{p^\star} \Pr(D = d \mid P = p^\star)\Pr\left(P = p^\star\right)},$$

with summation replaced by integration when the set of possible parameter values is continuous rather than discrete.

In the first hierarchical model example of section 3, let Y_{it} denote the composite score of participant i at time-point t, based on n_{it} scoreable items. At the lowest level of the hierarchical model, Y_{it} is modeled as arising from a normal distribution with mean $\alpha_i + \beta_i \times t$ and variance $n_{it}^{-1}\sigma^2$. At the next level, the participant-specific parameters α_i and β_i are modeled as independent and identically distributed realizations from respective normal distributions, having means $\alpha^\star$ and $\beta^\star$, and variances σ_α^2 and σ_β^2. At the third and final level, these two mean parameters and all three variances (σ^2, σ_α^2, and σ_β^2) are assigned very spread-out prior distributions, namely a normal distribution with mean 0 and variance 10 000 for each mean parameter, and an inverse-gamma distribution with parameters (.001, .001) for each variance parameter. A standard MCMC algorithm, the *Gibbs sampler*, is used to fit the model.

In the second hierarchical model, the stage scores for each participant at each time-point are regarded as multinomially distributed data; that is, for each of the n_{it} scoreable items, there is a probability p_{itj} of the item being scored at stage j. The multinomial-logit transformation is used, so that

$$p_{itj} = \frac{\exp\left(\eta_{itj}\right)}{\sum_{k=1}^{5} \exp\left(\eta_{itk}\right)},$$

with say stage 1 taken as the reference category, so that $\eta_{it1} = 0$ is assumed. Then a multivariate normal prior distribution can be assigned to the η_{itj} in a way which smoother temporal trends are more likely a priori. Particularly, for each i and each $j = 2, \ldots, 5$, $(\eta_{i1j}, \ldots, \eta_{i5j})$ is assigned a multivariate normal prior distribution with mean vector comprised of 0s, while the variance matrix is $\tau_i^2 V$, with the matrix V chosen to give more weight to linear (rather than rough) change over time. Each participant-specific variance parameter τ_i^2 governs the overall magnitude of change over time. These parameters are assigned widely spread inverse-gamma prior distributions, as in the previous example. Again, standard MCMC algorithms are used to fit the model.