Routledge
Taylor & Francis Group

Educational Psychologist

ISSN: 0046-1520 (Print) 1532-6985 (Online) Journal homepage: http://www.tandfonline.com/loi/hedp20

Advances in Bayesian Modeling in Educational Research

Roy Levy

To cite this article: Roy Levy (2016) Advances in Bayesian Modeling in Educational Research, Educational Psychologist, 51:3-4, 368-380, DOI: 10.1080/00461520.2016.1207540

To link to this article: http://dx.doi.org/10.1080/00461520.2016.1207540

EDUCATIONAL PSYCHOLOGIST, *51*(3–4), 368–380, 2016

ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2016.1207540

Routledge
Taylor & Francis Group

Advances in Bayesian Modeling in Educational Research

Roy Levy
T. Denny Sanford School of Social and Family Dynamics, Arizona State University

In this article, I provide a conceptually oriented overview of Bayesian approaches to statistical inference and contrast them with frequentist approaches that currently dominate conventional practice in educational research. The features and advantages of Bayesian approaches are illustrated with examples spanning several statistical modeling traditions and contexts in educational research.

Statistical inference in educational research and practice is currently dominated by methods and approaches associated with frequentist perspectives. Bayesian approaches (Gelman et al., 2013; Jackman, 2009) compose less of a share of the market. Broadly, Bayesian approaches are gaining in popularity as researchers recognize the philosophical and pragmatic advantages they afford.

In this article I provide a conceptually oriented overview of such Bayesian approaches and their advantages. I begin by offering an overview of frequentist inference, in particular, as conducted via maximum likelihood (ML). This overview is far from comprehensive and is aimed at being the groundwork for the sections that follow, in which I describe Bayesian inference and offer a high-level contrast between the two approaches. I then describe several examples of Bayesian approaches in educational research, illuminating the key ideas. A brief discussion, highlighting the themes of this work, concludes the article.

OVERVIEW OF FREQUENTIST INFERENCE VIA MAXIMUM LIKELIHOOD

Let y denote the collection of observed data and let θ denote the collection of the unknown model parameters, which are of inferential interest. In frequentist approaches, the parameters are treated as fixed and unknown, and the data are treated as random. A model is constructed by specifying a conditional distribution for the data given the parameters, denoted $p(y \mid \theta)$. The setup here refers to data being stochastically dependent on the parameters, which aligns well with frequentist conceptions of probability, with constant parameters reflecting fixed features of the population and the data varying from sample to sample.

Once we have observed sample data, ML estimation seeks to answer the question, "What are values of the parameters that yield the highest probability of the values of the data that were in fact observed?" To answer this, we input the values that were observed for the data y into $p(y \mid \theta)$ and view the resulting expression as function of the parameters, referred to as the *likelihood*. ML estimation then comes to finding the values for the parameters that maximize this function, which are the ML estimates (MLEs) of the parameters, and associated (possibly asymptotically justified) estimates of standard errors, which can be used to construct confidence intervals.

The results of such analyses are then utilized to do a variety of things as warranted by the application at hand, such as examining model-data fit, comparing models, respecifying models, making predictions for unobserved (e.g., future) data, drawing conclusions, and reporting results.

OVERVIEW OF BAYESIAN INFERENCE

Bayesian approaches invariably conduct inference by invoking Bayes's theorem, which states that

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y \mid \theta)p(\theta)}{p(y)}. \quad (1)$$

The expression on the left-hand side, $p(\theta \mid y)$, is referred to as the *posterior* distribution, which reflects that it

Correspondence should be addressed to Roy Levy, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, PO Box 873701, Tempe, AZ 85287-3701. E-mail: roy.levy@asu.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hedp.

captures what is believed about the parameters after having incorporated the information in the data. The numerator in the first equality in Equation (1) is the joint probability distribution. The second equality follows from the factoring this joint distribution as $p(y, \theta) = p(y \mid \theta)p(\theta)$. As in ML, $p(y \mid \theta)$ is the conditional probability of the data given the parameters, which is the likelihood function for when the data are known. The second term, $p(\theta)$, is the *prior* distribution for the model parameters. The denominator in Equation (1) is the marginal probability of the observed data: $p(y) = \sum_{\theta} p(y \mid \theta)p(\theta)$, with the sum being taken over all possible values of θ for discrete parameters, or $p(y) = \int p(y \mid \theta)p(\theta)d\theta$ in the case of continuous parameters. Note that $p(y)$ does not vary with different values of θ. Dropping $p(y)$, Bayes's theorem is often represented as a proportional relation,

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta). \quad (2)$$

These expressions give a prescription for conducting Bayesian inference. A Bayesian analysis proceeds by first constructing the joint distribution of y and θ, usually by abiding the factoring and specifying the conditional probability of the data, $p(y \mid \theta)$, and the prior, $p(\theta)$, that is, the right-hand side of Equations (1) or (2). Next, the observed data y are conditioned on, yielding the posterior distribution $p(\theta \mid y)$, that is, the left-hand side of Equations (1) or (2). This is the "answer" or "solution" from a Bayesian analysis. As a distribution, it may be expressed, summarized, or communicated in the usual ways. We might use histograms, density plots, and scatterplots as visualizations. We might use the posterior mean (or median, or mode) as a measure of central tendency and the posterior standard deviation as a measure of variability. We might summarize the posterior distribution with an interval indicating the region that delineates a given percentage of the distribution, referred to as a *credibility interval*. For example, the 95% highest posterior density (HPD) credibility interval for θ is the interval in which the probability that θ is inside that interval in the posterior distribution is .95, for which no value outside the interval has a higher posterior density than any value inside the interval. At this point we may continue with the usual story of using the results of fitting a statistical model and proceed to do things like examine model-data fit, compare models, respecify models, make predictions, draw conclusions, report results, and so on.

Historically, applications of Bayesian inference were limited in part due to computational intractability of the posterior in many cases. In situations where there is no closed-form solution for the posterior, we turn to approaches that approximate the posterior. To this end, Markov chain Monte Carlo (MCMC) methods for simulating distributions have opened up new possibilities for Bayesian analyses (Gelman et al., 2013) and has led to an explosion of applications in educational research (Levy, Mislevy, & Behrens, 2011). Briefly, MCMC consists of simulating (generating, drawing) values from a series of distributions. The sequence of draws is linked, forming a chain, in that the current value in the chain depends on the previous value in the chain. In a Bayesian analysis, we construct the chain such that it converges to the posterior distribution, meaning that, in the limit, the draws in the chain may be considered draws from the posterior distribution. In practice, we aim to determine when the chain converges. The subsequent draws are considered draws from the posterior distribution and collectively approximate it. Features of the distribution of interest (e.g., the mean, standard deviation, intervals) may then be empirically approximated by the corresponding features of the collection of draws. In terms of the user experience, most software for conducting MCMC to some extent masks or automates the computational steps (setting up the chain, obtaining draws, assessing convergence), leaving the user the responsibility of specifying the model (in the software's language).

HIGH-LEVEL COMPARISON OF BAYESIAN AND FREQUENTIST INFERENCE

Both Bayesian and ML inference involve the conditional distribution of the data given the model parameters (i.e., the likelihood), $p(y \mid \theta)$. The difference that is most stark on the surface is that Bayesian inference involves the prior distribution, $p(\theta)$, which expresses what is believed, assumed, or stipulated about the modeled situation and the parameters before incorporating the information in the data.

However, a more fundamental distinction that underlies it concerns the way the approaches treat parameters. In frequentist approaches, parameters are treated as *fixed* (constant), and it is inappropriate to discuss them probabilistically. It is the data that are modeled as random variables, and have a distribution. As a function of the (random) data, the ML estimator is therefore also a random variable, with the MLE based on any particular sample of data being a realization of this random variable. The standard error is a measure of the variability of the parameter *estimates* upon repeated sampling of the data from the population. Similarly, the probabilistic interpretation of a confidence interval is ascribed to the process of constructing the interval upon repeated sampling of data. It is important that the notions of variability and likely values refer to the distribution of *parameter estimates* (be they point or interval estimates) upon repeated sampling. A frequentist perspective yields statements of the variability or likely values of the parameter *estimator*, but it does not yield such statements for the parameter *itself*.

In contrast, a fully Bayesian approach treats the model parameters as *random* and uses distributions for them in the

model. In model construction, the parameters are assigned a prior distribution, and the result of a Bayesian analysis is the posterior distribution. The use of distributions for parameters is quite natural from the *subjective* (*degree-of-belief*, or *epistemic*) view of probability (de Finetti, 1974). From this perspective, a probability statement is, at its heart, an expression of belief or uncertainty on the part of a person, rather than a property of the world. Two people may have different beliefs and therefore assert different probability statements for the same event, or the same person may have different beliefs and assert different probability statements at different times.[1]

A Bayesian analysis using a diffuse prior (e.g., a uniform distribution over the real line, a normal distribution with an enormous variance) and/or a large sample size typically yields results that appear to be similar to those from a frequentist analysis. Here, the posterior mean (or mode) tends to be close to the MLE, the posterior standard deviation tends to be close to the standard error, and a posterior credibility interval tends to be close to a confidence interval. It is crucial to recognize that even when posterior means, standard deviations, and credibility intervals are *numerically similar* to their frequentist counterparts, they carry *conceptually different* interpretations. Those arising from a Bayesian analysis afford probabilistic statements and reasoning about the *parameters*, and those arising from a frequentist analysis do not. Bayesian approaches allow us to use the language of probability to *directly* discuss what is of inferential interest—parameters, hypotheses, models, and so on—rather than indirectly as in frequentist approaches. Bayesian approaches allow analysts to "say what they mean and mean what they say" (Jackman, 2009, p. xxviii), and overcome the limitations and avoid the pitfalls of the probabilistic machinery of frequentist inference (Goodman, 2008; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Wagenmakers, 2007).

EXAMPLES OF BAYESIAN INFERENCE

In the following subsections I illustrate Bayesian inference through considerations of several applications, which allows for further discussion of the features of Bayesian inference.

Diagnostic Assessment via Latent Class Analysis

We begin with the simplest situation in which we have two variables, each of which can take on one of two values. Such a situation frequently occurs in assessment situations in which student performances are used in diagnostic or decision-making scenarios to indicate whether the student is a member of group. For this example, suppose we have a variable that captures whether a student has mastered a skill. Whether they are a member of each group (masters or nonmasters of the skill) is not known with certainty. What we do observe and know is another variable that captures whether the student correctly or incorrectly answered an item that measures that skill. Note that the same structure might apply in a variety of situations with different grouping variables (e.g., whether a student had a learning disability, or was depressed), and different observable variables (e.g., a test result indicating pass/fail, or an observation of whether a behavior is present, as in a checklist). Suppose we have a situation where

- if a student has mastered a skill, that student will correctly answer the item with probability .90, and
- if a student has not mastered a skill, that student will incorrectly answer the item with probability .80.

In the diagnostic context, these two features are sometimes referred to as the *sensitivity* and *specificity* of the item, respectively.

In practice, we seek to reason from a student's observed item response to that student's true state of mastery on the skill: Once we observe, say, a correct item response, what are the chances the student actually has mastered the skill? Answering this requires understanding the prevalence of the skill, without appeal to the item. For example, suppose that 15% of the students have mastered the skill. This is known as the *base rate*. Combining the information in the sensitivity, specificity, and the base rate, we can obtain the positive predictive value, which is the probability that a student with a correct response actually has mastered the skill.

Although not always recognized as such, this is an instance of Bayesian inference, where the base rate reflects what we would believe about a student before administering the item and defines the prior probability of mastery, the sensitivity and specificity express the conditional probability of the data given the possible states of mastery, and the positive predictive value is the posterior probability of mastery given a correct item response.

For sake of exposition, we can frame this as a latent class model, where θ is a discrete latent variable capturing student mastery, which can take on either a value of Master or Nonmaster, and y is a discrete observable variable capturing the item response, which can take on

[1]This view of probability is closely associated with Bayesian inference, both historically and currently, and I adopt it throughout this article. It is noteworthy that probability is used in this way in scientific investigations and everyday conversation (Levy & Mislevy, 2016). It is important to note that one need not agree with this perspective in order to employ Bayesian methods.

either a value of Correct or Incorrect. The posterior probability that the student has mastered the skill given a correct item response is, by Bayes's theorem,

$$
\begin{aligned}
&p(\theta = \text{Master} \mid y = \text{Correct}) \\
&\quad = \frac{p(y = \text{Correct} \mid \theta = \text{Master})p(\theta = \text{Master})}{p(y = \text{Correct} \mid \theta = \text{Master})p(\theta = \text{Master}) + p(y = \text{Correct} \mid \theta = \text{Nonmaster})p(\theta = \text{Nonmaster})} \\
&\quad = \frac{(.90)(.15)}{(.90)(.15) + (.20)(.85)} = \frac{.135}{.135 + .17} \approx .44.
\end{aligned}
$$

Proceeding through the same sort of calculations yields the posterior probability that the student has not mastered the skill given a correct item response as $p(\theta = \text{Nonmaster} \mid y = \text{Correct}) \approx .56$, as must be the case because these two probabilities must sum to 1. For another student, who answered the item incorrectly, the same sorts of calculations yield the posterior probability of mastery as $p(\theta = \text{Master} \mid y = \text{Incorrect}) \approx .02$ and the probability of nonmastery as $p(\theta = \text{Nonmaster} \mid y = \text{Incorrect}) \approx .98$. The use of Bayes's theorem to yield probabilities of group membership has become standard in latent class models; see Dayton (1999) and Collins and Lanza (2010) for additional descriptions and examples of its use in this capacity. Additional advantages of adopting a Bayesian perspective may be seen in more complicated scenarios. For example, if the prior probability of mastery (base rate) or the conditional probability structure (sensitivity and specificity) are not known with certainty, they can be assigned prior probability distributions and become subject to posterior inference (Levy & Mislevy, 2016).

Typical latent class models analyze multiple observables (items), synthesizing their evidentiary bearing on the latent variable (mastery). This can be accomplished all at once or one variable at a time, such as in sequential or adaptive assessment environments. Here, a Bayesian approach allows for the incorporation of the results as they arrive. Continuing with the example, suppose we have a second item, with sensitivity .95 and specificity .75. The above calculation reveals that, after observing the correct response on the first item, the probability that the student has mastered the skill is .44. This is a *posterior* probability with respect to the first item but a *prior* probability with respect to the second item. Suppose we now observe a correct response on the second item. Applying Bayes's theorem again, now with the posterior probability from after the first item (.44) serving as the prior probability for this analysis, we obtain the posterior probability of .75. To recap, we began with the probability that the student had mastered the skill of .15. After observing the correct response for the first item, Bayes's theorem yielded an updated probability of .44. After observing a correct response for the second item, Bayes's theorem yielded an updated probability of .75. In this way, Bayesian inference is natural mechanism for synthesizing the data and updating beliefs as the data arrive.

Diagnostic Assessment via Dynamic Bayesian Networks

Extensions of the principles just discussed support assessment in more complex scenarios. To illustrate, I briefly describe a simplified model for a situation in which we seek to diagnose not only whether students have mastered a skill but also *when* they master it, as evidenced by their performances on repeated attempts on tasks.

The example comes from Save Patch, an educational game targeting rational number addition (Chung et al., 2010) in which students complete levels by using math skills to navigate from the beginning of the level to the end. Important to note, the game was designed to promote learning, and feedback occurs in that students know if they successfully completed the level; if they have, they advance to a new level, and if not, they are presented the same level for another attempt. Each student's attempt on each level was evaluated as representing a correct solution or an error of various kinds.

Here I present here a simplified account of a portion of a larger model; see Levy (2014) for details on the full model, including the estimation of the probabilities associated with the network described next.

A Bayesian network structures a system of discrete variables by way of specifying the dependence relations among them. A dynamic Bayesian network (DBN) extends this to model longitudinal data structures and changes over time. A graphical representation of a simplified version of the DBN for Level 19 of Save Patch, which assesses whether the student has mastered the skill *Adding Unit Fractions*, is presented in Figure 1. Here, three time slices are depicted, corresponding to three attempts on the level. Beginning at the top, the Adding Unit Fractions variables are latent variables capturing whether the student has mastered the skill, with the number indicating the time point or attempt on the level. The Level 19 Observable variables capture the student's performance on each attempt.

At each time point, the one-headed arrow from Adding Unit Fractions down to Level 19 Observable indicates that performance is modeled as dependent on mastery of the skill. This dependence is expressed in the conditional probability table in Table 1. The first row indicates that a

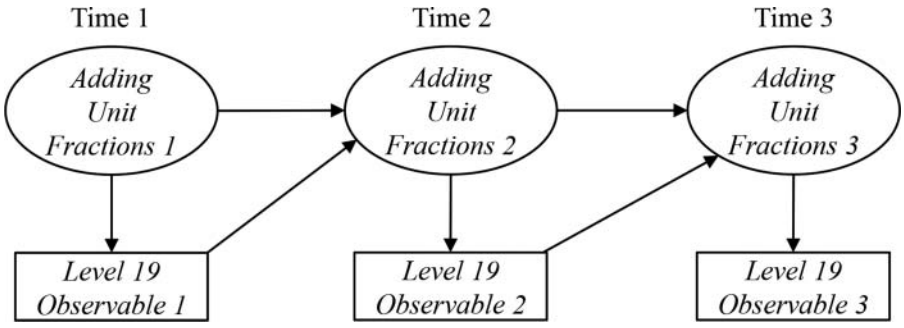

FIGURE 1 Three time slices for the dynamic Bayesian network for performance on Level 19 in Save Patch.

TABLE 1
Conditional Probability Table for Performance on Level 19 in Save Patch

	p(Observable for Level 19 | Adding Unit Fractions)				
Adding Unit Fractions	*Standard Solution*	*Alternate Solution*	*Incomplete Solution*	*Wrong Numerator Error*	*Unknown Error*
Master	0.95	0.00	0.01	0.03	0.01
Nonmaster	0.58	0.02	0.02	0.25	0.13

student who has mastered the Adding Unit Fractions skill will almost certainly complete the level with the Standard Solution, that is, by successfully completing the level in a way that reflects properly adding the fractions. There is a small probability that they will provide an Incomplete Solution, or commit an error of one kind or another, including an error evidencing the student possesses a certain misconception (Wrong Numerator Error; see Levy, 2014, for an expanded DBN that formally models misconceptions) or one not associated with any misconceptions (Unknown Error). The second row in Table 1 indicates that a student who has not mastered the skill is much less likely to successfully complete the level and is more likely to make an error.

The arrows leading from Adding Unit Fractions and Level 19 Observable at one time point to Adding Unit Fractions at the *next* time point convey that whether a student has mastered the skill depends on their previous state of mastery and their previous attempt on the level. The associated conditional probability table is given in Table 2. The first row reflects an assumption that once a student has mastered the skill, he or she will retain it.

TABLE 2
Conditional Probabilities for Mastery of Adding Unit Fractions at Time $t+1$ on Level 19 in Save Patch

	Observable for Level 19 at Time t				
Adding Unit Fractions at Time t	*Standard Solution*	*Alternate Solution*	*Incomplete Solution*	*Wrong Numerator Error*	*Unknown Error*
Master	1.00	1.00	1.00	1.00	1.00
Nonmaster	0.38	0.17	0.19	0.20	0.09

The second row gives the probabilities of transitioning from nonmastery to mastery. These transition probabilities vary depending on the performance on the level on the previous attempt. If a student was not a master at time t but provided a standard solution on that attempt, there is a .38 probability that the student will transition to become a master at time $t + 1$. If a student was not a master at time t but provided an alternate solution on that attempt, there is a .17 probability that the student will transition to become a master at time $t + 1$. This pattern continues, down to the final entry, which expresses that if a student was not a master at time t and provided an unknown error on that attempt, the probability that the student will transition to become a master at the next time point is only .09. These probabilities reflect the chances of a student learning (i.e., in the sense of transitioning from nonmastery to mastery) during the game, as that was indeed what the game was intended to support. To complete the specification of the model in the current illustration, the probability of mastering Adding Unit Fractions at Time 1, before any attempts, is .70.

The networks are named *Bayesian* networks because they support the application of Bayes's theorem across the system of variables, essentially extending the calculations shown in the previous section to multiple variables. To illustrate, Figure 2 contains screenshots from the Netica software package (Norsys, 1999–2012), depicting the situation at four time points.[2] In Netica, there is a bar chart for each variable giving the probability for the variable being at a particular level, which is also represented numerically as a percentage. In Panel (a), no observations have been made. The probability that the student has mastered the skill at Time 1 is .7. The probability distributions for the remaining variables are predictive in that they express what we expect, based on the model.

[2]The Netica file for this example is available online at https://sites.google.com/a/asu.edu/roylevy/papers-software.

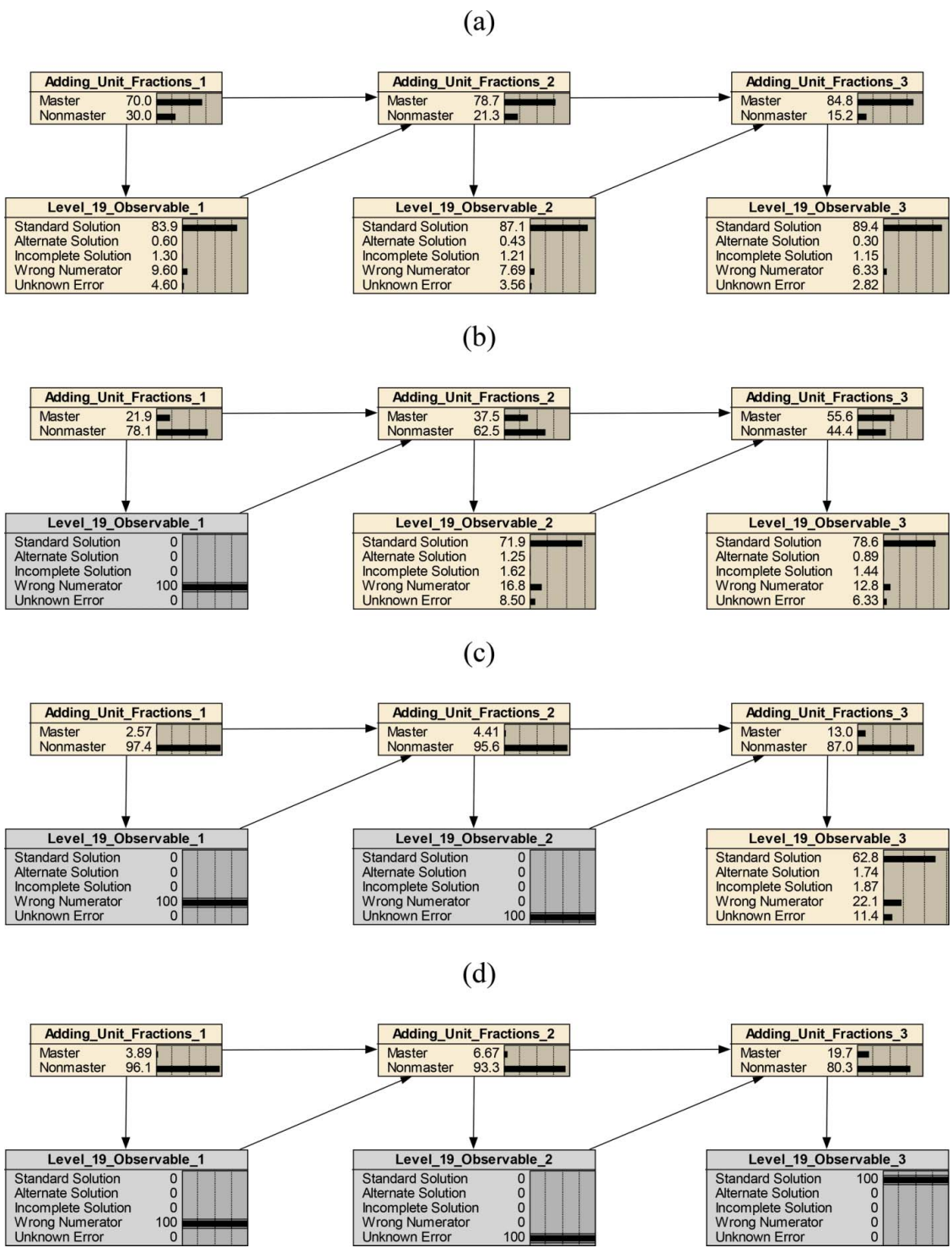

FIGURE 2 States of a portion of the dynamic Bayesian network over multiple time points for performance on Level 19 in Save Patch: (a) prior to observing any attempts, (b) after observing a Wrong Numerator error on the first attempt, (c) after observing an Unknown Error on the second attempt, and (d) after observing a Standard Solution on the third attempt.

Panel (b) represents the state of the network after observing a student commit a Wrong Numerator Error on the first attempt. The posterior probability that the student has mastered the skill at that time is now .22. The probabilities for the remaining nodes are now *posterior* predictive distributions, now posterior to the observation at Time 1.

Panel (c) displays the state of the network based on having observed the student commit an Unknown Error on the second attempt. According to the model it is highly unlikely that that the student had mastered the skill at Time 1 (probability = .03) or at Time 2 (probability = .04), and it is not very likely the student will have mastery at Time 3 (probability = .13).

Panel (d) displays the state of the network based on having observed the student produce the Standard Solution on their third attempt. The posterior probability of mastery at Time 3 is .20. The posterior probabilities that the student had mastered the skill earlier, at Time Points 1 and 2, have increased very slightly.

This example further highlights the use of Bayesian inference to synthesize observations and update beliefs over time as data accrue. Additional examples of Bayesian networks in the context of game and related tutoring contexts can be found in Reye (2004) and Shute, Ventura, Bauer, and Zapata-Rivera (2009). Levy and Mislevy (2016) and González-Brenes, Mislevy, Behrens, Levy, and DiCerbo (in press) provided chapter-length descriptions of Bayesian networks, and Almond, Mislevy, Steinberg, Yan, and Williamson (2015) provided a book-length treatment. See these sources and the references cited therein for additional examples.

Regression

This section introduces an example originally reported by Levy and Crawford (2009) and reanalyzed by Levy and Mislevy (2016), differing slightly from those sources to highlight key ideas for the current purposes. We have scores from three end-of-chapter tests in a course, and we regress scores from the Chapter 3 test on scores from the Chapter 1 and Chapter 2 tests. Table 3 contains summary statistics obtained from $n = 50$ students.

TABLE 3
Summary Statistics for the Three End-of-Chapter Tests for 50 Subjects in the Regression Example

	Chapter 1	*Chapter 2*	*Chapter 3*
No. of Items	16	18	15
M	14.10	14.34	12.22
SD	2.02	3.29	2.96
	Chapter 1	Chapter 2	
Chapter 2	.58		
Chapter 3	.69	.68	

Note. The bottom half of the table gives the correlations between the test scores.

A traditional regression model posits independence of the students and specifies the model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{3}$$

where x_{i1}, $x_{i2,}$ and y_i are the scores on the Chapter 1, 2, and 3 tests (respectively) for student i, and σ_ε^2 is the error variance.

The posterior distribution for the unknown parameters given the data is

$$p(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2 \mid \mathbf{y}, \mathbf{x}) \propto p(\mathbf{y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{x}) p(\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2), \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the full collections of predictors and outcomes, respectively. The first term on the right-hand side is the conditional probability of the data. The model in Equation (3) implies that this is

$$p(\mathbf{y} \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, \mathbf{x}) = \prod_{i=1}^{n} p(y_i \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, x_{i1}, x_{i2}), \tag{5}$$

where

$$y_i \mid \beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2, x_{i1}, x_{i2} \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \sigma_\varepsilon^2). \tag{6}$$

The second term on the right-hand side of Equation (4) is the prior distribution for the unknown parameters. Assuming independence in the prior, I specified normal prior distributions for the intercept and slopes and an inverse-gamma distribution for the error variance:

$$\beta_0 \sim N(0,\ 1000), \beta_j \sim N(0,\ 1000), j = 1,\ 2,$$

$$\sigma_\varepsilon^2 \sim \text{Inv-Gamma}(1, 1).$$

The inverse-gamma is a popular choice as a prior distribution for variances, as it restricts the variance to be positive (as variances ought to be), and can simplify computations when employed for variances of normal distributions (Jackman, 2009). Additional details on this and all the distributions used in these examples may be found at https://sites.google.com/a/asu.edu/roylevy/papers-software. Overall, the priors adopted in this analysis are quite diffuse, reflecting little in the way of prior beliefs regarding the parameters.

The posterior distribution was estimated using MCMC.[3] Figure 3 contains density plots representing the marginal posterior distributions for the intercept (β_0), slope for

[3]Relevant code for this example and the subsequent examples in this and the next two sections, along with accompanying code for the data and initial values used in each example, may be found at https://sites.google.com/a/asu.edu/roylevy/papers-software.

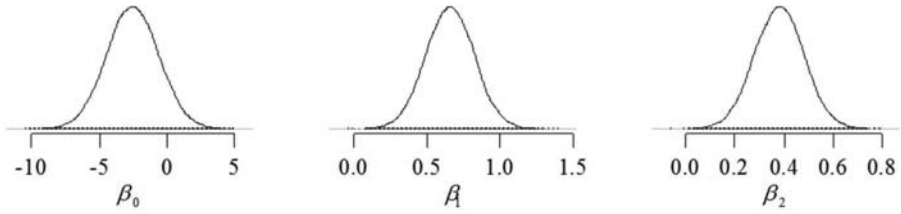

FIGURE 3 Marginal posterior densities for the intercept (β_0) and slopes (β_1, β_2) from the traditional regression model regressing Chapter 3 test scores on Chapter 1 and Chapter 2 test scores.

TABLE 4
Summary of the Results From Frequentist and Bayesian Analyses of the Examples Regressing Chapter 3 Test Scores on Chapter 1 and Chapter 2 Test Scores

	Sample Size of 50									*Sample Size of 1,950*		
	Traditional Model						*Modified Model*			*Traditional Model*		
	Frequentist Analysis			*Bayesian Analysis*			*Bayesian Analysis*			*Frequentist Analysis*		
	Est.	*SE*	*95% CI*	*Posterior M*	*Posterior SD*	*95% HPD Int.*	*Posterior M*	*Posterior SD*	*95% HPD Int.*	*Est.*	*SE*	*95% CI*
β_0	–2.54	1.93	[–6.41, 1.34]	–2.53	1.94	[–6.43, 1.15]	0.92	0.85	[< .01, 2.67]	1.39	0.32	[0.76, 2.02]
β_1	0.66	0.17	[0.33, 0.99]	0.66	0.17	[0.34, 0.99]	0.42	0.12	[.17, .65]	0.39	0.03	[0.33, 0.45]
β_2	0.38	0.10	[0.18, 0.59]	0.38	0.10	[0.17, 0.57]	0.38	0.10	[.18, .59]	0.38	0.02	[0.34, 0.42]

Chapter 1 test score (β_1), and the slope for Chapter 2 test score (β_2). As depicted there, the marginal distributions were unimodal and fairly symmetric. Table 4 contains numerical summaries of the posterior distribution, as well as the results from a frequentist (ML) solution to this traditional model. Note the similarity in the values between the Bayesian and frequentist analysis. This illustrates the point just discussed that, with a diffuse prior, the posterior distribution strongly resembles the likelihood function and the results from the two approaches are *numerically similar*. However, they are *conceptually different* in that the Bayesian analysis yields summaries and probabilistic statements about the *parameters themselves*, which represent our uncertain beliefs about their values. For β_1, the posterior mean of .66 and posterior standard deviation of .17 are descriptions of the distribution for the parameter, not a parameter estimator as in frequentist analyses. Similarly, the credibility interval is interpreted as a probabilistic statement about the parameter: For example, according to this model, there is a .95 probability that β_1 is between .34 and .99.

There was evidence that the Chapter 1 test was the stronger predictor, in the sense that a difference in one point on the Chapter 1 test leads to a higher expected difference on the Chapter 3 test than does a point difference on the Chapter 2 test (holding all else constant; the posterior means were .66 and .38 for β_1 and β_2, respectively). The Bayesian framework allows for a probabilistic statement to this effect. We can estimate the probability that β_1 is greater than β_2 as the proportion of draws from MCMC for which the value for β_1 exceeds that of β_2. In the current analysis, $p(\beta_1 > \beta_2) = .88$.

The results for the intercept are concerning, as the posterior indicates that β_0 is likely negative.[4] This is problematic because the intercept is interpreted as the expected value on the Chapter 3 test for a student who scored 0 on the Chapter 1 and Chapter 2 tests, but a negative score on the Chapter 3 test is impossible. To overcome this limitation, Levy and Crawford (2009) specified a modified regression model, where the prior for the intercept was specified as a uniform distribution bounded by the minimum and maximum possible values of the Chapter 3 test,

$$\beta_0 \sim \text{Uniform}\ (0, 15).$$

This specification expresses that although we may not have any idea where the intercept is between 0 and 15 before seeing the data, we know it shouldn't be below 0 or above 15. Owing to the positive relations among the variables, the lower bound is the critical issue here to keep expected scores from falling below 0. Similarly, expected scores should not exceed 15. To encode this, the modified model altered the specification of the conditional distribution of the data, such that if the expected score for any student based on the model in Equation (3) exceeded 15, it was changed to be 15.

Numerical summaries of the posterior distribution from fitting this modified model are summarized in Table 4. As implied by the prior distribution, the intercept is now positive.

[4]As revealed in Table 4, the same substantive conclusion is reached using the ML solution.

Note also that the modified model tells a different story about predictive strength of the predictors. Under the traditional model, there was evidence that the Chapter 1 test was the stronger predictor, $p(\beta_1 > \beta_2) = .88$. In the modified model, it seems the Chapter 1 and 2 tests have essentially the same predictive power; here, $p(\beta_1 > \beta_2)$ was only .60.

For comparative purposes, a second, larger sample ($n = 1{,}950$) was obtained, and the traditional regression model was analyzed using *frequentist* approaches. The results are also reported in Table 4. We can see that the results here are quite different from those from the traditional model for the original sample, and are quite close to those from the modified regression model on the original sample: The intercept is positive and the influences of an additional point on the first two chapter tests are comparable. Treating this larger sample as more representative of the population, we can conclude that the analysis of the original sample using the modified model provided a better portrait of the situation than the traditional model. The modified model outperformed the traditional model because it augments the information in the data, building substantive knowledge we have into the model. Adopting a Bayesian approach easily allowed for specifications, both in the prior and the likelihood, that allow our substantive understandings of the situation to be brought to bear in the analysis.

Classic accounts of Bayesian regression can be found in Box and Tiao (1973) and Lindley and Smith (1972). More modern treatments and descriptions of examples, informed by advances in computing, can be found in Gelman et al. (2013) and Jackman (2009).

Item Response Theory

Item response theory (IRT) models specify continuous latent variables as underlying discrete observable variables. To illustrate, I analyzed item responses from all 2,000 students to the 18 items that compose the Chapter 2 test described. I specified a three-parameter normal ogive model for the conditional probability of a correct response,

$$P(x_{ij} = 1 \mid \theta_i, d_j, a_j, c_j) = c_j + (1 - c_j)\Phi(a_j\theta_i + d_j), \quad (7)$$

where x_{ij} is the observable response variable (coded as 1 for correct) from student i on item j; θ_i denotes the latent variable for student i; d_j, a_j, and c_j are the location, discrimination, and lower asymptote parameters for item j; and Φ is the normal distribution function. To complete the model, I specified the following prior distributions. For the latent variables, for each examinee,

$$\theta_i \sim N(0, 1), \quad (8)$$

which resolves the indeterminacies in the location and scale. For the item parameters, for each item,

$$d_j \sim N(0, 2), \quad a_j \sim N^+(1, 2), \quad c_j \sim \text{Beta}(6, 16), \quad (9)$$

where N^+ denotes the normal distribution truncated to the positive real line, used here for the a to express the belief that higher values of the latent variable should yield higher probabilities of a correct response for the items. The use of the Beta(6,16) prior for the c parameters restricts the values to between 0 and 1 and places most of the prior density around .25, which accords with the chance of success by guessing the correct answer on the four-option multiple-choice items. Additional details on these distributions may be found in Appendix A available at https://sites.google.com/a/asu.edu/roylevy/papers-software. Overall these priors are somewhat informative and are frequently important for estimating item parameters in IRT, as the parameters may be poorly determined from the data.

The marginal posterior densities obtained by MCMC were unimodal and approximately symmetric; numerical summaries the parameters for Items 11 and 14 are given in Table 5. The results indicate that Item 14 is easier (in terms of d), is slightly less discriminating (in terms of a), and has a higher lower asymptote (in terms of c). The posterior standard deviations and HPD intervals indicate that there is differential uncertainty about many of the parameters; for example, there is more uncertainty regarding c for Item 14 than c for item 11.

The last row in Table 5 summarizes the results for an examinee who correctly answered all the items. An ML approach would yield an estimate for θ (proficiency) for this examinee of infinity, because the likelihood function continually increases as θ increases.[5] Conceptually, ever higher values of θ represent ever better accounts of this student based on ML, which just uses the information in the data. In the Bayesian analysis, the information in the student's data is combined with the prior distribution in Equation (8) to yield a posterior distribution that indicates that the θ for the student is relatively high, near 1.34, but is almost certainly not greater than 3. Despite observing the student responding correctly to all of the items, we are pretty sure that her proficiency is not infinite.

Note that in the fully Bayesian analysis the estimation of students' latent variables occurs concurrently with that of the item parameters. This stands in contrast to conventional approaches that proceed in stages. In the first stage, the item parameters are estimated. In the second stage, those estimates are used as the values for the item parameters when estimating latent variables. This separate-stage approach suffers in that the uncertainty in the estimation in the first stage is ignored when estimating at the second stage. The consequences of doing so may not be dire if estimates from the first stage are sufficiently precise. However, they are likely to be increasingly severe the more uncertainty from the first stage is ignored. The fully Bayesian framework allows us to properly acknowledge and propagate our uncertainty. What's more, it allows us to examine just

[5]Note that this would hold regardless of how the item parameters were estimated.

TABLE 5
Summary of the Marginal Posterior Distributions for Two Items' Parameters and One Examinee's Latent Variable for the Three Parameter Normal Ogive Model of the Chapter 2 Test

Parameter	*M*	*SD*	*95% Highest Posterior Density Interval*
d_{11}	−0.15	0.11	[−0.35, 0.07]
d_{14}	1.09	0.12	[0.83, 1.30]
a_{11}	1.04	0.13	[0.79, 1.30]
a_{14}	0.87	0.13	[0.64, 1.13]
c_{11}	0.19	0.04	[0.11, 0.27]
c_{14}	0.34	0.10	[0.15, 0.53]
θ_{1996}	1.34	0.63	[0.18, 2.63]

what the implications may be if we take the simpler approach that works just with the estimates from an earlier stage.

Additional descriptions and examples of Bayesian approaches to IRT can be found in the didactic article by Kim and Bolt (2007), a chapter-length treatment in Levy and Mislevy (2016), and the book-length treatment by Fox (2010) and the references therein.

Structural Equation Modeling

Structural equation modeling with latent variables may be used to characterize relations among the constructs that latent variables are posited to represent, all the while acknowledging the measurement error present in our observable variables that are indicators of the constructs. To illustrate, I analyzed item responses from all 2,000 examinees to all the items in the three chapter tests described earlier. For each test, I specified a latent variable to represent proficiency on the material in that chapter. The item responses for each test were modeled as dependent on the latent variable for the test, following the three-parameter normal ogive model in Equation (7).

I specified a structural model that formulates the latent variable for Chapter 2 as a mediator between the latent variable for Chapter 1 and the latent variable for Chapter 3:

$$\theta_{i2} = \beta_{21}\theta_{i1} + \delta_{i2}, \quad \delta_{i2} \sim N(0, \psi_2),$$

$$\theta_{i3} = \beta_{31}\theta_{i1} + \beta_{32}\theta_{i2} + \delta_{i3}, \quad \delta_{i3} \sim N(0, \psi_3),$$

where θ_{i1}, θ_{i2}, and θ_{i3} denote the latent variables for Chapter 1, Chapter 2, and Chapter 3 (respectively) for student i. Finally, the Chapter 1 latent variable for each student is modeled as a normal distribution,

$$\theta_{i1} \sim N(0, \phi),$$

where the fixed mean resolves the indeterminacy in location.

Prior distributions for the parameters complete the specification. For the item parameters of the measurement model, the prior distributions in Equation (9) were employed, save that the a for the first item in each test was set equal to 1 to resolve the indeterminacies in the latent variables. For the remaining parameters:

$$\beta_{21} \sim N(0, 1000),\ \beta_{31} \sim N(0, 1000),\ \beta_{32} \sim N(0, 1000),$$

$$\psi_2 \sim \text{Inv-Gamma}(1, 1),\ \psi_3 \sim \text{Inv-Gamma}(1, 1),$$

$$\phi \sim \text{Inv-Gamma}(1, 1).$$

The model was analyzed via MCMC, yielding marginal posterior densities that were unimodal and approximately symmetric. I focus on the results for the latent regression structure, and for ease of interpretation I present the results for a standardized solution.

The standardized β_{21} had a posterior mean of .72 (SD = .02), 95% HPD interval [.68, .76]. The standardized β_{31} had a posterior mean of .39 (SD = .05), 95% HPD interval [.28, .49]. The standardized β_{32} had a posterior mean of .51 (SD = .05), 95% HPD interval [.41, .59]. The mediation effect is given by $\beta_{\text{Ind}} = (\beta_{21})(\beta_{32})$. In the standardized metric, the posterior mean of the indirect effect was .37 (SD = .04), 95% HPD [.29, .43]. The results suggest that there is an indirect effect, but Chapter 2 proficiency does not fully mediate the relation between that of Chapter 1 and Chapter 3. Of importance, the Bayesian approach allows for statistical inference about the indirect effect in a straightforward manner. The posterior for β_{Ind} is easily approximated via MCMC as the distribution of the product of the draws for β_{21} and the draws for β_{32}. This stands in contrast to frequentist approaches, where an indirect effect poses challenges as it is a product of parameters.

Additional treatments of Bayesian approaches to mediation can be found in Yuan and MacKinnon (2009) and Enders, Fairchild, and MacKinnon (2013). More general descriptions, examples, and pointers to additional examples of Bayesian SEM can be found in the chapter-length treatments of Kaplan and Depaoli (2012) and Levy and Choi (2013) and in the book-length treatments of Lee (2007) and Song and Lee (2012).

Additional Examples

In the following subsections, I briefly discuss two additional situations that are common to data in educational research, namely, that data are missing and that data arise from hierarchical or grouping structures. The goal of each discussion is to highlight how conventionally accepted strategies for analyzing such data may be seen as instances of Bayesian reasoning.

Missing data analysis. One popular approach to dealing with missing data involves multiple imputation (Rubin, 1987), which may be seen as proceeding in three phases. First, a complete data set is constructed by imputing a value for each missing data point. This process is repeated multiple times, yielding multiple completed data sets. Second, the desired analysis is conducted on each of these completed data sets. Third, the results from the second stage are pooled to yield a single set of results. In what follows, I give a brief characterization of the crux of the approach, which is how the imputations are obtained in the first phase.

Values for the imputations are obtained by constructing a model relating the variables with missing data to the other variables. As a simple model, we may take a variable with missing values and set it up as an outcome in a regression model, regressing it on other variables. The regression model allows us to then make predictions for the missing values of the outcome variable. There are two key insights that lead to multiple imputation. First, instead of using a *single* prediction for each missing value, we should take *multiple* values, drawn from a distribution. This is motivated by recognizing that, unless our predictors can perfectly predict the outcome, there is some error associated with our prediction; that is, the actual values will vary around the predicted values, forming a distribution. Accordingly, we will need to take multiple values from this distribution to capture its variability. The second insight is that taking such draws would suffice if we knew (with certainty) the regression parameters that relate the variable with missingness to the other variables. But in general this is not the case; the regression parameters are estimated from the data and are therefore not known with certainty. Ignoring our uncertainty regarding the regression parameters, say, by using point estimates of them from a conventional analysis, would understate our uncertainty about them and ultimately the missing data. Multiple imputation procedures formally acknowledge the uncertainty in the regression parameters and incorporate that into the process of taking draws for the missing values, by way of a Bayesian analysis. This is typically accomplished through MCMC. Relative to the general use of MCMC for estimating posterior distributions, generating imputations for missing data amounts to adding the variable with missingness to the set of parameters for which draws are obtained.

Multilevel modeling. Multilevel or hierarchical models extend more basic statistical models, often to account for dependencies been units (e.g., subjects, measures) that are grouped or clustered in a hierarchical or nested structure. Such situations arise frequently in education, such as test scores over time being nested within students, students within classrooms, classrooms within schools, and so on.

To begin, a model is formulated at the lowest level (Level 1) for units within the groups. Applications in education typically employ linear regression-type models at this and other levels. For ease of exposition, we consider a model with random intercepts and no predictors,

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon), \tag{10}$$

where i indexes a unit at the lowest level (e.g., a student), and j indexes the group or cluster (e.g., a classroom). The intercepts in the first level (i.e., the β_{0j}s) are group-specific means that vary over j and are modeled at the second level. For simplicity, we consider the following model,

$$\beta_{0j} = \gamma_{00} + r_{0j}, \quad r_{0j} \sim N(0, \sigma^2_r), \tag{11}$$

where r_{0j} is the Level-2 error for group j. The model described here is akin to an analysis of variance model, where the β_{0j}s are group-specific means that vary around the grand mean, γ_{00}. The model may be extended to any number of levels, as well as having predictors at any of the levels, but this is sufficient for our purposes

Based on the model, there are two immediate options for estimating each β_{0j}. Based on Equation (10), we could estimate β_{0j} as the mean of the y_{ij}, that is, the group mean can be estimated by taking the mean of the scores in the group. Alternatively, based on Equation (11), we could estimate each β_{0j} as being equal to the grand mean, γ_{00}. Conventional approaches to multilevel modeling take a weighted average of these two estimates (see, e.g., Snijders & Bosker, 2012, section 4.8). The weights are the inverse of the variances associated with each, also known as the precision.

This may be recognized as an instance of Bayesian reasoning. Adopting a Bayesian lens, the Level 1 model in Equation (10) specifies the conditional probability of the data (here, y_{ij}) given a parameter (β_{0j}), and the Level 2 model in Equation (11) specifies the prior distribution for the parameter. Under this specification, the estimate resulting from conventional approaches is just the posterior mean for β_{0j}.

On this account, we can view MLM quite naturally from a Bayesian perspective. Their correspondence suggests that the reverse may also be true, and we may view Bayesian models as multilevel models. That is, the general factorization of the joint distribution in Equation (1), $p(y, \theta) = p(y \mid \theta)p(\theta)$, is an explicit multilevel structuring. At a minimum, we have two levels: $p(y \mid \theta)$ composes Level 1, and $p(\theta)$ composes Level 2. This may be extended to any number of levels, of course. If the distribution for θ involves unknown parameters, they too are modeled via a prior distribution, and so on, until all unknowns have been distributionally modeled.

DISCUSSION

In this article I sketched Bayesian approaches to inference and how they contrast with conventional frequentist approaches. These ideas were developed through the descriptions and illustrations of several examples spanning different modeling traditions and contexts. Several of these

revealed that a number of activities that enjoy near-consensus agreement as being the state of the art of conventional practice may be seen as, or were originally derived as, Bayesian procedures.[6]

The preceding underscores the first of several larger themes that capture the advantages of adopting a Bayesian perspective, namely, as it provides a coherent framework. A Bayesian perspective allows us to tackle problems ranging from missing data to diagnoses in assessment to parameter estimation in complex models. Although we may have situation-specific terminology (e.g., multiple imputation, positive predictive probability), they all come down to being an instance of Bayesian reasoning.

A second theme is the use of the probability concept for all unknowns, including model parameters. Adopting a Bayesian perspective in the examples, we could express things such as the probability that a student has mastered a skill, or will at another time, the probability distribution for missing data, and the probability that one slope was larger than another in regression.

A third theme is the alternative ways of conceiving of Bayesian inference. Those steeped in frequentist traditions are apt to see a Bayesian approach as one in which the information in the data is augmented by the information in the prior distribution. This view is natural when considering the roles of the prior in the modified regression model, the prior for the item parameters in IRT, and the prior for inference regarding latent class membership. On the other hand, an alternative conception is one of Bayes as an updating mechanism, where we update our initial beliefs in the prior with the information in the data, yielding the posterior. This conception is particularly useful in situations where data arrive over time, as the latent class example with multiple tests, and even more so in the DBN model.

A fourth theme is one of acknowledging and propagating uncertainty. This was seen in the context of missing data, where we incorporate our uncertainty in the parameters when producing imputations, and in the context of IRT, where we incorporate our uncertainty in the estimation of item parameters into the estimation of latent variables. The propriety of propagating uncertainty from one aspect when understanding another aspect is a key and distinguishing feature of Bayesian inference and is warranted in other situations in which we use the results of model fitting in another activity (e.g., when examining model-data fit; Levy, 2011).

A fifth theme is that of model complexity. Aided by the power of MCMC, Bayesian approaches allow for easier specification of complex models that more closely align with substantive theories. The modified regression model example illustrated how a Bayesian approach easily facilitated the incorporation of knowledge about the situation into the analysis. A related point is that Bayesian analyses typically perform as well as or better than conventional frequentist approaches in situations with small samples, both because of the possibility of affording auxiliary information (as in the regression model here) and because Bayesian approaches do not rely on asymptotic arguments to justify things like point estimates or distributions of test statistics. The example that married an IRT measurement model with a structural model for the latent variables illustrated how we may easily assemble components from varying modeling traditions. Compared to conventional approaches, adopting a Bayesian perspective allows the analyst to specify and analyze a wider class of models. Bayesian approaches have become an attractive option for analysts wishing to push the boundaries of modeling and analysis to better reflect substantive theories (for reviews and discussions of a number of such applications, see Levy & Choi, 2013, and Levy et al., 2011).

Finally, a sixth theme is that Bayesian approaches allow us to more easily express and marshal information from different sources in a coherent way. In the modified regression model, I drew from information in the data, but also my understanding of the situation, to shape the statistical model and ultimately the resulting inferences. As another example, multilevel modeling involves having our beliefs about the group-specific parameters be shaped in part by the information in the group and in part by the information from other groups.

For all of these reasons, Bayesian approaches are an attractive option for statistical modeling and inference in educational research.

[6]These are by no means the only situations where a commonly accepted statistical practice has subsequently been seen to be an instance of, or aligned with, Bayesian reasoning. As another example germane to educational research, Kelley's formula for estimating true scores and the standard error of the estimation of the true score in conventional classical test theory may be seen as just the posterior mean and standard deviation from a Bayesian analysis (Levy & Mislevy, 2016).

ACKNOWLEDGMENTS

I thank the editors and the anonymous reviewers for comments to an earlier version that prompted improvements to the manuscript.

REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010, May). *Testing instructional design variations in a prototype math game*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

de Finetti, B. (1974). *Theory of probability, Volume 1*. New York, NY: Wiley.

Enders, C. K., Fairchild, A. J., & MacKinnon, D. P. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, *48*, 340–369.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

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

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*, 135–140. Retrieved from http://doi.org/10.1053/j.seminhematol.2008.04.003

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. http://doi.org/10.3758/s13423-013-0572-3

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, England: Wiley.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). New York, NY: Guilford Press.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, UK: Wiley.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 663–685. http://doi.org/10.1080/10705511.2011.607723

Levy, R. (2014). *Dynamic Bayesian network modeling of game based diagnostic assessments* (CRESST Report No. 837). Los Angeles: University of California, National Center for Research on Evaluation, Standards, Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R837.pdf

Levy, R., & Choi, J. (2013). Bayesian structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 563–623). Charlotte, NC: Information Age.

Levy, R., & Crawford, A. V. (2009). Incorporating substantive knowledge into regression via a Bayesian approach to modeling. *Multiple Linear Regression Viewpoints*, *35*(2), 4–9.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: Chapman and Hall/CRC.

Levy, R., Mislevy, R. J., & Behrens, J. T. (2011). MCMC in educational research. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo: Methods and applications* (pp. 531–545). London, UK: Chapman and Hall/CRC.

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

Norsys Software Corporation. (1999–2012). *Netica manual*. Vancouver, British Columbia, Canada.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education, 14*, 63–96.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley & Sons.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (Vol. 2, pp. 295–321). Philadelphia, PA: Routledge/LEA.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Song, X.-Y., & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. Chichester, UK: Wiley.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. http://doi.org/10.3758/BF03194105

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322. http://doi.org/10.1037/a0016972