CHILD DEVELOPMENT

Child Development, xxxx 2013, Volume 00, Number 0, Pages 1–19

A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research

Rens van de Schoot
Utrecht University and North-West University

David Kaplan
University of Wisconsin–Madison

Jaap Denissen
University of Tilburg

Jens B. Asendorpf
Humboldt-University Berlin

Franz J. Neyer
Friedrich Schiller University of Jena

Marcel A.G. van Aken
Utrecht University

Bayesian statistical methods are becoming ever more popular in applied and fundamental research. In this study a gentle introduction to Bayesian analysis is provided. It is shown under what circumstances it is attractive to use Bayesian estimation, and how to interpret properly the results. First, the ingredients underlying Bayesian methods are introduced using a simplified example. Thereafter, the advantages and pitfalls of the specification of prior knowledge are discussed. To illustrate Bayesian methods explained in this study, in a second example a series of studies that examine the theoretical framework of dynamic interactionism are considered. In the Discussion the advantages and disadvantages of using Bayesian statistics are reviewed, and guidelines on how to report on Bayesian statistics are provided.

> *… it is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem.*
>
> Jerome Cornfield (in De Finetti, 1974a)

In this study, we provide a gentle introduction to Bayesian analysis and the Bayesian terminology without the use of formulas. We show why it is attractive to adopt a Bayesian perspective and, more practically, how to estimate a model from a Bayesian perspective using background knowledge in the actual data analysis and how to interpret the results.

We would like to thank the following researchers for providing feedback on our manuscript to improve the readability: Geertjan Overbeek, Esmee Verhulp, Mariëlle Zondervan-Zwijnenburg, Christiane Gentzel, Koen Perryck, and Joris Broere. The first author was supported by a grant from The Netherlands Organization for Scientific Research: NWO-VENI-451-11-008. The second author was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to The University of Wisconsin–Madison. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Rens van de Schoot, Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508TC Utrecht, The Netherlands. Electronic mail may be sent to a.g.j.vandeschoot@uu.nl.

Many developmental researchers might never have heard of Bayesian statistics, or if they have, they most likely have never used it for their own data analysis. However, Bayesian statistics is becoming more common in social and behavioral science research. As stated by Kruschke (2011a), in a special issue of *Perspectives on Psychological Science*:

> whereas the 20th century was dominated by NHST [null hypothesis significance testing], the 21st century is becoming Bayesian. (p. 272)

Bayesian methods are also slowly becoming used in developmental research. For example, a number of Bayesian articles have been published in *Child Development* (n = 5), *Developmental Psychology* (n = 7), and *Development and Psychopathology* (n = 5) in the last 5 years (e.g., Meeus, Van de Schoot, Keijsers,

DOI: 10.1111/cdev.12169

Schwartz, & Branje, 2010; Rowe, Raudenbush, & Goldin-Meadow, 2012). The increase in Bayesian applications is not just taking place in developmental psychology but also in psychology in general. This increase is specifically due to the availability of Bayesian computational methods in popular software packages such as Amos (Arbuckle, 2006), Mplus v6 (Muthén & Muthén, 1998–2012; for the Bayesian methods in Mplus see Kaplan & Depaoli, 2012; Muthén & Asparouhov, 2012), WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), and a large number of packages within the R statistical computing environment (Albert, 2009).

Of specific concern to substantive researchers, the Bayesian paradigm offers a very different view of hypothesis testing (e.g., Kaplan & Depaoli, 2012, 2013; Walker, Gustafson, & Frimer, 2007; Zhang, Hamagami, Wang, Grimm, & Nesselroade, 2007). Specifically, Bayesian approaches allow researchers to incorporate background knowledge into their analyses instead of testing essentially the same null hypothesis over and over again, ignoring the lessons of previous studies. In contrast, statistical methods based on the frequentist (classical) paradigm (i.e., the default approach in most software) often involve testing the null hypothesis. In plain terms, the null hypothesis states that "nothing is going on." This hypothesis might be a bad starting point because, based on previous research, it is almost always expected that "something is going on." Replication is an important and indispensible tool in psychology in general (Asendorpf et al., 2013), and Bayesian methods fit within this framework because background knowledge is integrated into the statistical model. As a result, the plausibility of previous research findings can be evaluated in relation to new data, which makes the proposed method an interesting tool for confirmatory strategies.

The organization of this study is as follows: First, we discuss probability in the frequentist and Bayesian framework, followed by a description, in general terms, of the essential ingredients of a Bayesian analysis using a simple example. To illustrate Bayesian inference, we reanalyze a series of studies on the theoretical framework of dynamic interactionism where individuals are believed to develop through a dynamic and reciprocal transaction between personality and the environment. Thereby, we apply the Bayesian approach to a structural equation modeling (SEM) framework within an area of developmental psychology where theory building and replication play a strong role. We conclude with a discussion of the advantages of adopting a Bayesian point of view in the context of developmental research. In the online supporting information appendices we provide an introduction to the computational machinery of Bayesian statistics, and we provide annotated syntax for running Bayesian analysis using Mplus, WinBugs, and Amos in our online supporting information appendices.

Probability

Most researchers recognize the important role that statistical analyses play in addressing research questions. However, not all researchers are aware of the theories of probability that underlie model estimation, as well as the practical differences between these theories. These two theories are referred to as the *frequentist paradigm* and the *subjective probability paradigm.*

Conventional approaches to developmental research derive from the frequentist paradigm of statistics, advocated mainly by R. A. Fisher, Jerzy Neyman, and Egon Pearson. This paradigm associates probability with long-run frequency. The canonical example of long-run frequency is the notion of an infinite coin toss. A sample space of possible outcomes (heads and tails) is enumerated, and probability is the proportion of the outcome (say heads) over the number of coin tosses.

The Bayesian paradigm, in contrast, interprets probability as the subjective experience of uncertainty (De Finetti, 1974b). Bayes' theorem is a model for learning from data, as suggested in the Cornfield quote at the beginning of this study. In this paradigm, the classic example of the subjective experience of uncertainty is the notion of placing a bet. Here, unlike with the frequentist paradigm, there is no notion of infinitely repeating an event of interest. Rather, placing a bet—for example, on a baseball game or horse race—involves using as much prior information as possible as well as personal judgment. Once the outcome is revealed, then prior information is updated. This is the model of learning from experience (data) that is the essence of the Cornfield quote at the beginning of this study. Table 1 provides an overview of similarities and differences between frequentist and Bayesian statistics.

The goal of statistics is to use the data to say something about the population. In estimating, for example, the mean of some variable in a population, the mean of the sample data is a "statistic" (i.e., estimated mean) and the unknown population mean is the actual parameter of interest. Similarly,

Table 1
Overview of the Similarities and Differences Between Frequentist and Bayesian Statistics

	Frequentist statistics	Bayesian statistics
Definition of the *p* value	The probability of observing the same or more extreme data assuming that the null hypothesis is true in the population	The probability of the (null) hypothesis
Large samples needed?	Usually, when normal theory-based methods are used	Not necessarily
Inclusion of prior knowledge possible?	No	Yes
Nature of the parameters in the model	Unknown but fixed	Unknown and therefore random
Population parameter	One true value	A distribution of values reflecting uncertainty
Uncertainty is defined by	The sampling distribution based on the idea of infinite repeated sampling	Probability distribution for the population parameter
Estimated intervals	Confidence interval: Over an infinity of samples taken from the population, 95% of these contain the true population value	Credibility interval: A 95% probability that the population value is within the limits of the interval

the regression coefficients from a regression analysis remain unknown parameters estimated from data. We refer to means, regression coefficients, residual variances, and so on as unknown parameters in a model. Using software like SPSS, Amos, or Mplus, these unknown parameters can be estimated. One can choose the type of estimator for the computation, for example, maximum likelihood (ML) estimation or Bayesian estimation.

The key difference between Bayesian statistical inference and frequentist (e.g., ML estimation) statistical methods concerns the nature of the unknown parameters. In the frequentist framework, a parameter of interest is assumed to be unknown, but fixed. That is, it is assumed that in the population there is only one true population parameter, for example, one true mean or one true regression coefficient. In the Bayesian view of subjective probability, all unknown parameters are treated as uncertain and therefore should be described by a probability distribution.

The Ingredients of Bayesian Statistics

There are three essential ingredients underlying Bayesian statistics first described by T. Bayes in 1774 (Bayes & Price, 1763; Stigler, 1986). Briefly, these ingredients can be described as follows (these will be explained in more detail in the following sections).

The first ingredient is the background knowledge on the parameters of the model being tested. This first ingredient refers to all knowledge available *before* seeing the data and is captured in the so-called *prior distribution*, for example, a normal distribution. The variance of this prior distribution reflects our level of uncertainty about the population value of the parameter of interest: The larger the variance, the more uncertain we are. The prior variance is expressed as *precision*, which is simply the inverse of the variance. The smaller the prior variance, the higher the precision, and the more confident one is that the prior mean reflects the population mean. In this study we will vary the specification of the prior distribution to evaluate its influence on the final results.

The second ingredient is the information in the data themselves. It is the observed evidence expressed in terms of the *likelihood function* of the data given the parameters. In other words, the likelihood function asks:

> Given a set of parameters, such as the mean and/or the variance, what is the likelihood or probability of the data in hand?

The third ingredient is based on combining the first two ingredients, which is called *posterior inference*. Both (1) and (2) are combined via Bayes' theorem (described in more detail in the online Appendix S1) and are summarized by the so-called posterior distribution, which is a compromise of the prior knowledge and the observed evidence. The posterior distribution reflects one's updated knowledge, balancing prior knowledge with observed data.

These three ingredients constitute Bayes' theorem, which states, in words, that our updated understanding of parameters of interest given our current data depends on our prior knowledge about the parameters of interest weighted by the current evidence given those parameters of interest. In online Appendix S1 we elaborate on the theorem. In what follows, we will explain Bayes' theorem and its three ingredients in detail.

Prior Knowledge

Why Define Prior Knowledge?

> The key epistemological reason concerns our view that progress in science generally comes about by learning from previous research findings and incorporating information from these research findings into our present studies. Often information gleaned from previous research is incorporated into our choice of designs, variables to be measured, or conceptual diagrams to be drawn. With the Bayesian methodology our prior beliefs are made explicit, and are moderated by the actual data in hand. (Kaplan & Depaoli, 2013, p. 412)

How to Define Prior Knowledge?

The data we have in our hands moderate our prior beliefs regarding the parameters and thus lead to updated beliefs. But how do we specify priors? The choice of a prior is based on how much information we believe we have prior to the data collection and how accurate we believe that information to be. There are roughly two scenarios. First, in some cases we may not be in possession of enough prior information to aid in drawing posterior inferences. From a Bayesian point of view, this lack of information is still important to consider and incorporate into our statistical specifications.

> In other words, it is equally important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand. (Kaplan & Depaoli, 2013, p. 412)

Second, in some cases we may have considerable prior information regarding the value of a parameter and our sense of the accuracy around that value. For example, after decades of research on the relation between, say, parent socioeconomic status and student achievement, we may, with a bit of effort, be able to provide a fairly accurate prior distribution on the parameter that measures that relation. Prior information can also be obtained from meta-analyses and also previous waves of surveys. These sources of information regarding priors are "objective" in the sense that others can verify the source of the prior information. This should not be confused with the notion of "objective priors," which constitute pure ignorance of background knowledge. Often the so-called *uniform distribution* is used to express an objective prior. For some subjective Bayesians, priors can come from any source: objective or otherwise. The issue just described is referred to as the "elicitation problem" and has been nicely discussed in O'Hagan et al. (2006; see also Rietbergen, Klugkist, Janssen, Moons, & Hoijtink, 2011; Van Wesel, 2011). If one is unsure about the prior distribution, a sensitivity analysis is recommended (e.g., Gelman, Carlin, Stern, & Rubin, 2004). In such an analysis, the results of different prior specifications are compared to inspect the influence of the prior. We will demonstrate sensitivity analyses in our examples.

An Example

Let us use a very simple example to introduce the prior specification. We will only estimate two parameters: the mean and variance of reading skills, for example, measured at entry to kindergarten for children in a state-funded prekindergarten program. To introduce the Bayesian methodology, we will first focus on this extremely simple case, and only thereafter will we consider a more complex (and often more realistic) example. In online Appendices S2–S4 we provide the syntax for analyzing this example using Mplus, WinBugs, and Amos.

The prior reflects our knowledge about the mean reading skills score before observing the current data. Different priors can be constructed reflecting different types of prior knowledge. Throughout the study we will use different priors with different levels of subjectivity to illustrate the effects of using background knowledge. In the section covering our real-life example, we base our prior specification on previous research results, but in the current section we discuss several hypothetical prior specifications. In Figure 1, six different distributions of possible reading skills scores are displayed representing degrees of prior knowledge. These distributions could reflect expert knowledge and/or results from previous similar research studies or meta-analyses.

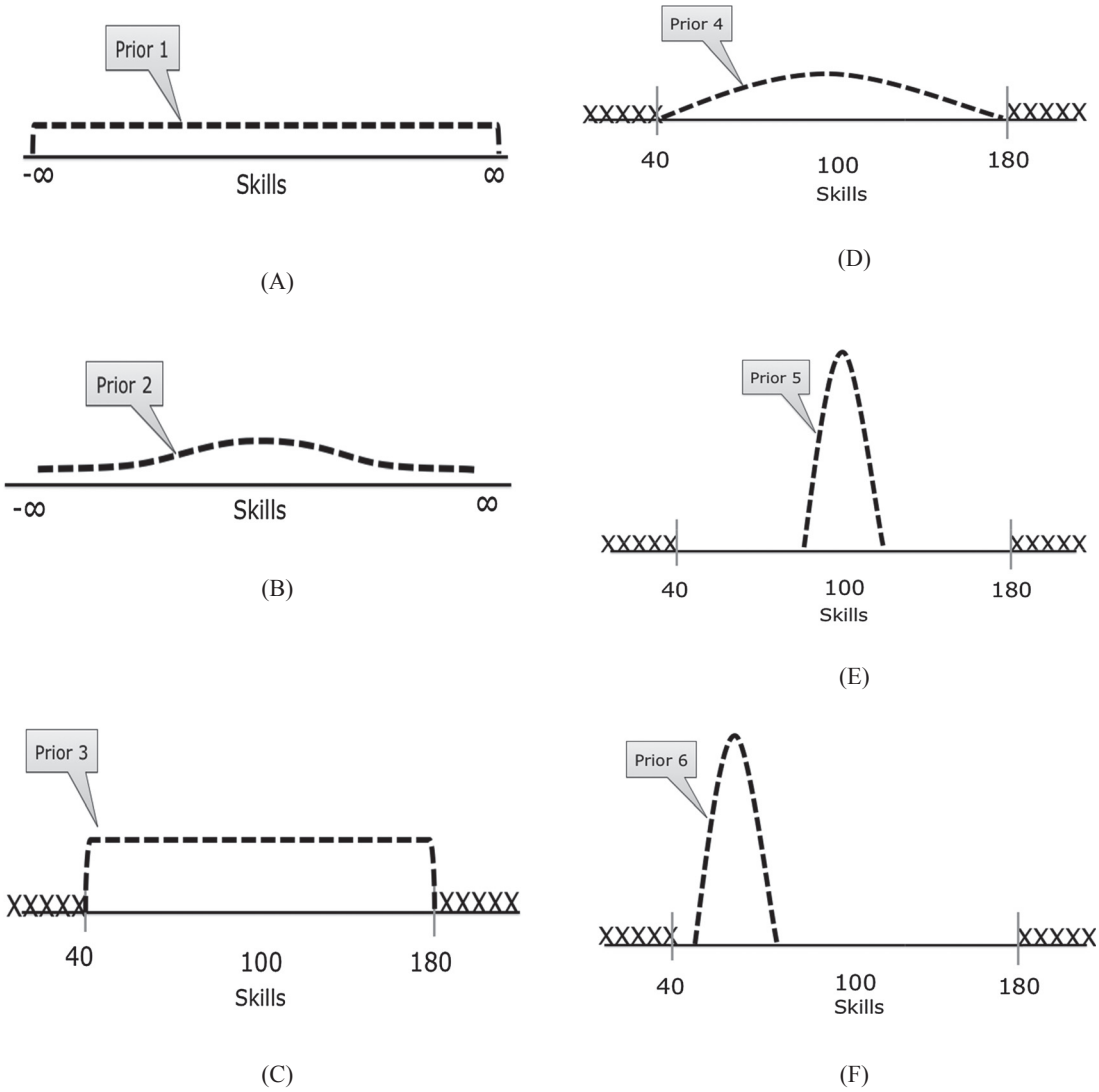

Figure 1. A priori beliefs about the distribution of reading skills scores in the population.

Noninformative Prior Distributions

In Figure 1a, it is assumed that we do not know anything about mean reading skills score and every value of the mean reading skills score in our data between minus infinity and plus infinity is equally likely. Such a distribution is often referred to as a *noninformative prior distribution*.

A frequentist analysis of the problem would ignore our accumulated knowledge and let the data speak for themselves—as if there has never been any prior research on reading skills. However, it could be reasonably argued that empirical psychology has accumulated a considerable amount of empirical information about the distribution of reading skills scores in the population.

Informative Prior Distributions

From a Bayesian perspective, it seems natural to incorporate what has been learned so far into our analysis. This implies that we specify that the mean of reading skills mean has a specific distribution. The parameters of the prior distribution are referred to as *hyperparameters.* If for the mean reading score a normal distribution is specified for the prior distribution, the hyperparameters are the prior mean and the prior precision. Thus, based on previous research one can specify the expected prior mean. If reading skills are assessed by a standardized test with a mean of 100, we hypothesize that reading skills scores close to 100 are more likely to occur in our data than values further away from 100, but every value in the

entire range between minus and plus infinity is still allowed.

Also, the prior precision needs to be specified, which reflects the certainty or uncertainty about the prior mean. The more certain we are, the smaller we can specify the prior variance and, as such, the precision of our prior will increase. Such a prior distribution encodes our existing knowledge and is referred to as a *subjective* or *informative* prior distribution. If a low precision is specified, such as Prior 2 in Figure 1b, it is often referred to as a *low-informative* prior distribution. Note that it is the prior variance of the prior distribution we are considering here and not the variance of the mean of the reading skill score.

One could question how realistic Priors 1 and 2 in Figures 1a and 1b are, if a reading skills score is the variable of interest. What is a negative reading skills score? And can reading skills result in *any* positive score? To assess reading skills, a reading test could be used. What if we use a reading test where the minimum possible score is 40 and the maximum possible score is 180? When using such a test, Priors 1 and 2 are not really sensible. In Figure 1c, a third prior distribution is specified where values outside the range 40–180 are not allowed and within this range obtaining every reading skills score is equally likely.

Perhaps we can include even more information in our prior distribution, with the goal to increase precision and therefore contribute to more accurate estimates. As said before, we assume that our data are obtained from a randomly selected sample from the general population. In that case we might expect a mean of reading skills scores that is close to 100 to be more probable than extremely low or high scores. In Figure 1d, a prior distribution is displayed that represents this expectation. Figure 1e shows that we can increase the precision of our prior distribution by increasing its prior variance.

In Figure 1f, a prior distribution is specified where a very low score of reading skills is expected and we are very certain about obtaining such a mean score in our data. This is reflected by a prior distribution with high precision, that is, a small prior variance. If we sample from the general population, such a prior distribution would be highly unlikely to be supported by the data and, in this case, would be a misspecified prior. If, however, we have specified inclusion criteria for our sample, for example, only children with reading skills scores lower than 80 are included because this is the target group, then Prior 6 is correctly specified and Prior 5 would be misspecified.

To summarize, the prior reflects our knowledge about the parameters of our model *before* observing the current data. If one does not want to specify any prior knowledge, then noninformative priors can be specified and as a result, the final results will not be influenced by the specification of the prior. In the Bayesian literature, this approach to using noninformative priors is referred to as *objective* Bayesian statistics (Press, 2003) because only the data determine the posterior results. Using the objective Bayesian method, one can still benefit from using Bayesian statistics as will be explained throughout the study.

If a low-informative prior is specified, the results are hardly influenced by the specification of the prior, particularly for large samples. The more prior information is added, the more subjective it becomes. *Subjective* priors are beneficial because: (a) findings from previous research can be incorporated into the analyses and (b) Bayesian credible intervals will be smaller. Both benefits will be discussed more thoroughly in the section where we discuss our posterior results. Note that the term *subjective* has been a source of controversy between Bayesians and frequentists. We prefer the term *informative* and argue that the use of any kind of prior be warranted by appealing to empirical evidence. However, for this study, we stay with the term *subjective* because it is more commonly used in the applied and theoretical literature.

Note that for each and every parameter in the model, a prior distribution needs to be specified. As we have specified a prior distribution for the mean of reading skills scores we also have to specify a prior distribution for the variance/standard deviation of reading skills. This is because for Bayesian statistics, we assume a distribution for each and every parameter including (co)variances. As we might have less prior expectations about the variance of reading skills, we might want to specify a low-informative prior distribution. If we specify the prior for the (residual) variance term in such a way that it can only obtain positive values, the obtained posterior distribution can never have negative values, such as a negative (residual) variance.

Observed Evidence

After specifying the prior distribution for all parameters in the model, one can begin analyzing the actual data. Let us say we have information on the reading skills scores for 20 children. We used the software BIEMS (Mulder, Hoijtink, & de Leeuw, 2012) for generating an exact data set where the

mean and standard deviation of reading skills scores were manually specified. The second component of Bayesian analysis is the observed evidence for our parameters in the data (i.e., the sample mean and variance of the reading skills scores). This information is summarized by the likelihood function containing the information about the parameters *given the data set* (i.e., akin to a histogram of possible values). The likelihood is a function reflecting what the most likely values are for the unknown parameters, given the data. Note that the likelihood function is also obtained when non-Bayesian analyses are conducted using ML estimation. In our hypothetical example, the sample mean appears to be 102. So, given the data, a reading skills score of 102 is the most likely value of the population mean; that is, the likelihood function achieves its maximum for this value.

Posterior Distribution

With the prior distribution and current data in hand, these are then combined via Bayes' theorem to form the so-called *posterior distribution.* Specifically, in this case, Bayes' theorem states that our prior knowledge is updated by the current data to yield updated knowledge in the form of the posterior distribution. That is, we can use our prior information in estimating the population mean, variance, and other aspects of the distribution for this sample.

In most cases, obtaining the posterior distribution is done by simulation, using the so-called Markov chain Monte Carlo (MCMC) methods. The general idea of MCMC is that instead of attempting to analytically solve for the point estimates, like with ML estimation, an iterative procedure to estimate the parameters. For a more detailed introduction, see Kruschke (2011b, 2013), and for a more technical introduction, see Lynch (2007) or Gelman et al. (2004). See online Appendix S1 for a brief introduction.

The Posterior Distribution in the Example

The graphs in Figure 2 demonstrate how the prior information and the information in the data are combined in the posterior distribution. The more information is specified in the prior distribution, the smaller the posterior distribution of reading skills becomes. As long as the prior mean is uninformative (see Figure 2a), the result obtained for the mean with ML estimation and the posterior mean will always be approximately similar. If an informative prior is specified, the posterior mean is only similar to the ML mean if the prior mean is (relatively) similar to the ML estimate (see Figures 2b to 2e). If the prior mean is different from the ML mean (Prior 6), the posterior mean will shift toward the prior (see Figure 2f).

The precision of the prior distribution for the reading skills scores influences the posterior distribution. If a noninformative prior is specified, the variance of the posterior distribution is not influenced (see Figure 2a). The more certain one is about the prior, the smaller the variance, and hence more peaked the posterior will be (cf. Figures 2d and 2e).

Posterior Probability Intervals (PPIs)

Let us now take a closer look at the actual parameter estimates. We analyzed our data set with Mplus, Amos, and WinBUGS. Not all prior specifications are available in each software package; this has been indicated in Table 2 by using subscripts. In Mplus, the default prior distributions for means and regression coefficients are normal distributions with a prior mean of zero and an infinitive large prior variance, that is, low precision (see Figure 1b). If the prior precision of a specific parameter is set low enough, then the prior in Figure 1a will be approximated. The other prior specifications in Figure 1 are not available in Mplus. In Amos, however, one can specify a uniform prior, like in Figure 1a, but also normal distributions, like in Figure 1b, and a uniform distribution using the boundaries of the underlying scale, like in Figure 1c. If prior distributions of Figures 1d to 1f are of interest, one needs to switch to WinBUGS. We assumed no prior information for the variance of reading skills scores and we used the default settings in Amos and Mplus, but in WinBUGS we used a low-informative gamma distribution.

In the table, the posterior mean reading skills score and the PPIs are displayed for the six different types of prior specifications for our hypothetical example. Recall that the frequentist confidence interval is based on the assumption of a very large number of repeated samples from the population that are characterized by a fixed and unknown parameter. For any given sample, we can obtain the sample mean and compute, for example, a 95% confidence interval. The correct frequentist interpretation is that 95% of these confidence intervals capture the true parameter under the null hypothesis. Unfortunately, results of the frequentist paradigm are often misunderstood (see

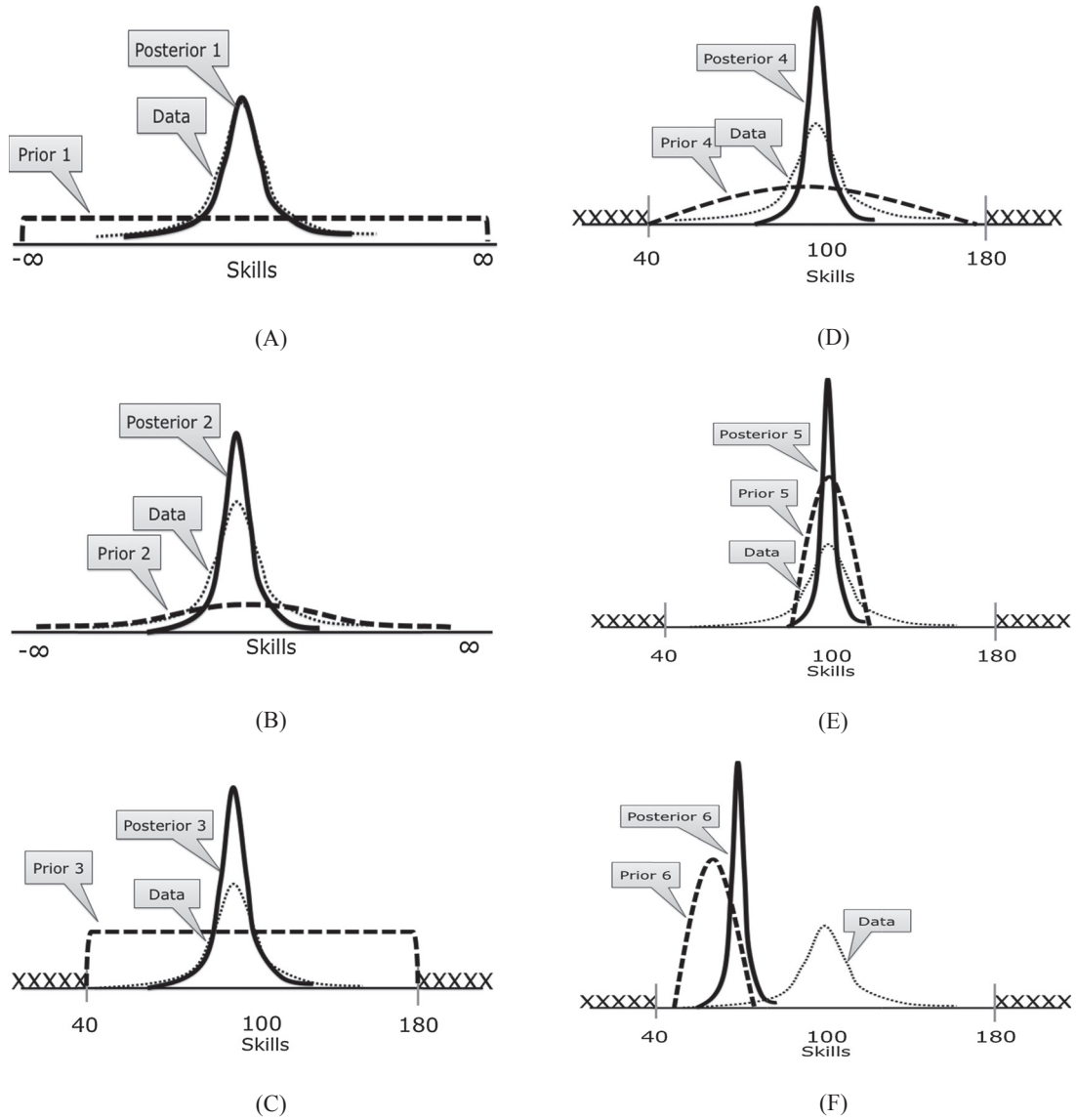

Figure 2. The likelihood function and posterior distributions for six different specifications of the prior distribution.

Gigerenzer, 2004). For example, the frequentist-based 95% confidence interval is often interpreted as meaning that there is a 95% chance that a parameter of interest lies between an upper and lower limit, whereas the correct interpretation is that 95 of 100 replications of exactly the same experiment capture the fixed but unknown parameter, assuming the alternative hypothesis about that parameter is true.

The Bayesian counterpart of the frequentist confidence interval is the PPI, also referred to as the *credibility* interval. The PPI is the 95% probability that in the population the parameter lies between the two values. Note, however, that the PPI and the confidence interval may numerically be similar and might serve related inferential goals, but they are not mathematical equivalent and conceptually quite different. We argue that the PPI is easier to communicate because it is actually the probability that a certain parameter lies between two numbers, which is *not* the definition of a classical confidence interval (see also Table 1).

Posterior Results of the Example

The posterior results are influenced by the prior specification. The higher the prior precession, the smaller the posterior variance and the more certain one can be about the results. Let us examine this relation using our example.

Table 2
Posterior Results Obtained With Mplus, AMOS, or WINBUGS (n = *20)*

Prior type	Prior precision used (prior mean was always 100)	Posterior mean reading skills score	95% CI/PPI
ML		102.00	94.42–109.57
Prior 1_{AW}		101.99	94.35–109.62
Prior $2a_{M\ AW}$	Large variance, i.e., Var. = 100	101.99	94.40–109.42
Prior $2b_{M\ AW}$	Medium variance, i.e., Var. = 10	101.99	94.89–109.07
Prior $2c_{M\ AW}$	Small variance, i.e., Var. = 1	102.00	100.12–103.87
Prior 3_{AW}		102.03	94.22–109.71
Prior 4_{W}	Medium variance, i.e., Var. = 10	102.00	97.76–106.80
Prior 5_{W}	Small variance, i.e., Var. = 1	102.00	100.20–103.90
Prior $6a_{W}$	Large variance, i.e., Var. = 100	99.37	92.47–106.10
Prior $6b_{W}$	Medium variance, i.e., Var. = 10	86.56	80.17–92.47

Note. CI = confidence interval; PPI = posterior probability interval; ML = maximum likelihood results; *SD* = standard deviation; M = posterior mean obtained using Mplus; A = posterior mean obtained using Amos; W = posterior mean obtained using WinBUGS.

When the prior in Figure 2a is used, the posterior distribution is hardly influenced by the prior distribution. The estimates for the mean reading skills score obtained from the likelihood function (i.e., the ML results) and posterior result are close to each other (see the first two rows of Table 2). If a normal distribution for the prior is used, as in Figure 2b, the 95% PPI is only influenced when a high-precision prior is specified; see Table 2 and compare the results of Priors 2a, 2b, and 2c, where only for Prior 2c the resulting PPI is smaller compared to the other priors we discussed so far. This makes sense because for the latter prior we specified a highly informative distribution; that is, the variance of the prior distribution is quite small reflecting strong prior beliefs. If the prior of Figure 2c is used, the results are similar to the ML results. When the prior of Figure 2c is combined with specifying a normal distribution, the PPI decreases again. If we increase the prior precision of the mean even further, for example, for Prior 5, the PPI decreases even more. If the prior mean is misspecified, like in Figure 2f, the posterior mean will be affected; see the results in Table 2 of Priors 6a and 6b. The difference between Priors 6a and 6b reflects the degree of certainty we have about the prior mean. For Prior 6a we are rather sure the mean was 80, which is reflected by a high prior precision. For Prior 6b we are less sure, and we used a low prior precision. The posterior mean of Prior 6b is therefore closer to the ML estimate when compared to the posterior mean of Prior 6a.

To summarize, the more prior information is added to the model, the smaller the 95% PPIs becomes, which is a nice feature of the Bayesian methodology. That is, after confronting the prior knowledge with the data one can be more certain about the obtained results when compared to frequentist method. This way, science can be truly accumulative. However, when the prior is misspecified, the posterior results are affected because the posterior results are always a compromise between the prior distribution and the likelihood function of the data.

An Empirical Example

To illustrate the Bayesian methods explained in this study, we consider a series of articles that study the theoretical framework of dynamic interactionism where individuals are believed to develop through a dynamic and reciprocal transaction between personality and the environment (e.g., quality of social relationships; Caspi, 1998). The main aim of the examined research program was to study the reciprocal associations between personality and relationships over time. In the case of extraversion, for example, an extraverted adolescent might seek out a peer group where extraversion is valued and reinforced, and as such becomes more extraverted.

A theory that explains environmental effects on personality is the social investment theory (Roberts, Wood, & Smith, 2005). This theory predicts that the successful fulfillment of societal roles (in work, relationships, health) leads to strengthening of those personality dimensions that are relevant for this fulfillment. For this study, the social investment theory is important because it can be hypothesized that effects fulfilling societal roles on personality are stronger in emerging adulthood when these roles are more central than in earlier phases of adolescence. At the time of the first article in our series (Asendorpf & Wilpers, 1998), however, the predictions of social

investment theory were not yet published. Instead, the authors started with a theoretical notion by McCrae and Costa (1996) that personality influences would be more important in predicting social relationships than vice versa. At the time, however, the idea did not yet have much support because:

> empirical evidence on the relative strength of personality effects on relationships and vice versa is surprisingly limited. (p. 1532)

Asendorpf and Wilpers (1998) investigated for the first time personality *and* relationships over time in a sample of young students ($N = 132$) after their transition to university. The main conclusion of their analyses was that personality influenced change in social relationships, but not vice versa. Neyer and Asendorpf (2001) studied personality–relationship transactions using now a large representative sample of young adults from all over Germany (age between 18 and 30 years; $N = 489$). Based on the previous results, Neyer and Asendorpf

> hypothesized that personality effects would have a clear superiority over relationships effects. (p. 1193)

Consistent with Asendorpf and Wilpers (1998), Neyer and Asendorpf (2001) concluded that once initial correlations were controlled, personality traits predicted change in various aspects of social relationships, whereas effects of antecedent relationships on personality were rare and restricted to very specific relationships with one's preschool children (p. 1200). Asendorpf and van Aken (2003) continued working on studies into personality–relationship transaction, now on 12-year-olds who were followed up until age 17 ($N = 174$), and tried to replicate key findings of these earlier studies. Asendorpf and van Aken confirmed previous findings and concluded that the stronger effect was an extraversion effect on perceived support from peers. This result replicates, once more, similar findings in adulthood.

Sturaro, Denissen, van Aken, and Asendorpf (2010), once again, investigated the personality–relationship transaction model. The main goal of the 2010 study was to replicate the personality–relationship transaction results in an older sample (17–23 years) compared to the study of Asendorpf and van Aken (2003; 12–17 years). Sturaro et al. found some contradictory results compared to the previously described studies.

> [The five-factor theory] predicts significant paths from personality to change in social relationship quality, whereas it does not predict social relationship quality to have an impact on personality change. Contrary to our expectation, however, personality did not predict changes in relationship quality. (p. 8)

In conclusion, the four articles described above clearly illustrate how theory building works in daily practice. By using the quotes from these articles we have seen that researchers do have prior knowledge in their Introduction and Discussion sections. However, all these articles ignored this prior knowledge because they were based on frequentist statistics that test the null hypothesis that parameters are equal to zero. Using Bayesian statistics, we will include prior knowledge in the analysis by specifying a relevant prior distribution.

Method

Description of the Neyer and Asendorpf (2001) Data

Participants were part of a longitudinal study of young adults. This sample started in 1995 (when participants were 18–30 years old; $M_{age} = 24.4$ years, $SD = 3.7$) with 637 participants who were largely representative of the population of adult Germans. The sample was reassessed 4 years later (return rate = 76%). The longitudinal sample included 489 participants ($N = 226$ females).

To simplify the models we focus here on only two variables: extraversion as an indicator for personality and closeness with/support by friends as an indicator for relationship quality. Quality of relationships was assessed at both occasions using a social network inventory, where respondents were asked to recall those persons who play an important role in their lives. In the present investigation, we reanalyzed the relevant data on the felt closeness with friends. Participants named on average 4.82 friends ($SD = 4.22$) and 5.62 friends ($SD = 4.72$) at the first and second occasions, respectively. Closeness was measured with the item: "How close do you feel to this person?" (1 = *very distant* to 5 = *very close*). The ratings were averaged across all friends. Extraversion was assessed using the German version of the NEO-FFI (Borkenau & Ostendorf, 1993). Internal consistencies at both measurement occasions were .76 and .78, and the rank order stability was $r = .61$.

Description of the Sturaro et al. (2010) and Asendorpf and van Aken (2003) Data

Participants were part of the Munich Longitudinal Study on the Genesis of Individual Competencies (Weinert & Schneider, 1999). This sample started in 1984 (when participants were 3 to 4 years old) with 230 children from the German city of Munich. Participants were selected from a broad range of neighborhoods to ensure representativeness. This study focuses on reassessments of the sample at ages 12, 17, and 23. At age 12, 186 participants were still part of the sample; at age 17 this was true for 174 participants. Because the Asendorpf and van Aken (2003) publication focused on participants with complete data at both waves of data collection, the present analyses focus on the 174 individuals with data at ages 12 and 17. At age 23, a total of 154 participants were still in the sample. For this study, analyses focus on a subset of 148 individuals who provided personality self-ratings.

Measures selected for this study were taken in correspondence with the cited articles. At age 12, support by friends was measured as support from classroom friends. For the category of classroom friends, an average of 3.0 individuals was listed. For each of these individuals, participants rated the supportiveness of the relationship in terms of instrumental help, intimacy, esteem enhancement, and reliability (three items each; items of the first three scales were adapted from the NRI; Furman & Buhrmester, 1985). Ratings were averaged across all friends. At age 17, the same scales were repeated only for the best friend in class. In both years, if participants did not have any classroom friends, they received a score of 1 for support (the lowest possible). At age 23, support was measured using an ego-centered Social Network Questionnaire. Like the Sturaro et al. (2010) article, we focus here on the average quality with same-sex peers because this measure was deemed most comparable with the peer measures at ages 12 and 17.

Extraversion at ages 12 and 17 was assessed with bipolar adjective pairs (Ostendorf, 1990; sample item: unsociable vs. outgoing). At age 23, extraversion was assessed with a scale from the NEO-FFI (Borkenau & Ostendorf, 1993; sample item: "I like to have a lot of people around me"). As reported by Sturaro et al. (2010), in a separate sample of 641 college students, the Ostendorf Scale for Extraversion and the NEO-FFI Scale for Extraversion are correlated almost perfectly after controlling for the unreliability of the scales (r = .92).

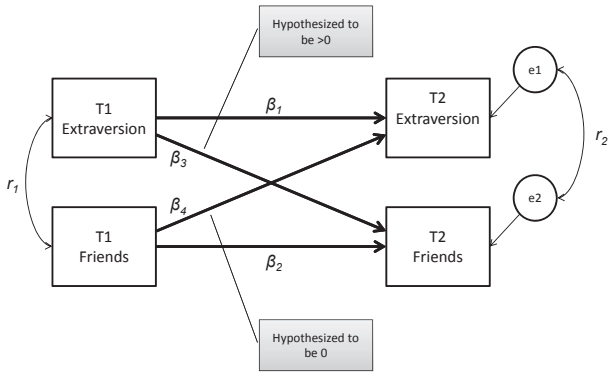

Figure 3. Cross-lagged panel model where r_1 is the correlation between Extraversion measured at Wave 1 and Friends measured at Wave 1, and r_2 is the autocorrelation between the residuals of two variables at Wave 2, β_1 and β_2 are the stability paths, and β_3 and β_4 are the cross-loadings. T1 and T2 refer to ages 12 and 17, respectively, for the Asendorpf and van Aken (2003) data, but to ages 17 and 23, respectively, for the Sturaro, Denissen, van Aken, and Asendorpf (2010) data.

Analytic Strategy

We used Mplus to analyze the model displayed in Figure 3. Two crucial elements when applying Bayesian statistics have to be discussed in the Analytic Strategy section of a Bayesian article: (a) which priors were used and where did these priors came from? And (b) how was convergence assessed (see also online Appendix S1)? Concerning the latter, we used the Gelman–Rubin criterion (for more information, see Gelman et al., 2004) to monitor convergence, which is the default setting of Mplus. However, as recommended by Hox, van de Schoot, and Matthijsse (2012), we set the cutoff value stricter (i.e., bconvergence = .01) than the default value of .05. We also specified a minimum number of iterations by using *biterations = (10,000)*, we requested multiple *chains* of the Gibbs sampler by using *chains = 8*, and we requested starting values based on the ML estimates by using *stvalues = ml*. Moreover, we inspected all the trace plots manually to check whether all chains converged to the same target distribution and whether all iterations used for obtaining the posterior were based on stable chains.

Concerning the specification of the priors, we developed two scenarios. In the first scenario, we only focus on those data sets with similar age groups. Therefore, we first reanalyze the data of Neyer and Asendorpf (2001) without using prior knowledge. Thereafter, we reanalyze the data of Sturaro et al. (2010) using prior information based on the data of Neyer and Asendorpf; both data sets contain young adults between 17 and 30 years of

age. In the second scenario, we assume the relation between personality and social relationships is independent of age and we reanalyze the data of Sturaro et al. using prior information taken from Neyer and Asendorpf *and* from Asendorpf and van Aken (2003). In this second scenario we make a strong assumption, namely, that the cross-lagged effects for young adolescents are equal to the cross-lagged effects of young adults. This assumption implicates similar developmental trajectories across age groups. We come back to these issues in the Discussion section.

Scenario 1

Based on previous research findings, Asendorpf and Wilpers (1998) hypothesized the model shown in Figure 3. As this study described the first attempt to study these variables over time, Asendorpf and Wilpers would probably have specified (had they used Bayesian statistics) an uninformative prior distribution reflecting no prior knowledge (see also Figures 1a and 1b). Neyer and Asendorpf (2001) gathered a general sample from the German population and analyzed their data. As Neyer and Asendorpf used different test–retest intervals as compared to Asendorpf and Wilpers, we cannot use the results from Asendorpf and Wilpers as prior specifications. So, when reanalyzing Neyer and Asendorpf, we will use the default settings of Mplus, that is, noninformative prior distributions; see Figure 1b and see Model 1 (*Neyer & Asendorpf, 2001 | Uninf. Prior*) in the second column of Table 3. Note that "|" means condition on, so the statement is read as the results of the Neyer and Asendorpf (2001) data *condition on* an uninformative prior. Sturaro et al. (2010) continued working on the cross-lagged panel model. In the third column of Table 3, the results of Model 2 (*Sturaro* et al., *2010 | Uninf. prior*) are shown when using noninformative prior distribution (which does not take the previous results obtained by Neyer and Asendorpf into account). What if we used our updating procedure and use the information obtained in Model 1 as the starting point for our current analysis? That is, for Model 3 (*Sturaro* et al., *2010 | Neyer & Asendorpf, 2001*) we used for the regression coefficients the posterior means and standard deviations from Model 1 as prior specifications for Model 3a. Noninformative priors were used for residual variances and for the covariances. This was done because the residuals pick up omitted variables, which almost by definition are unknown. Then, we would have a hard time knowing what their prior relation would be to the outcome or to other variables in model. This way the prior for the subsequent study is a rough approximation to the posterior from the previous study.

As pointed out by one of the reviewers, there is an assumption being made that the multiparameter posterior from a previous study can be accurately represented by independent marginal distributions on each parameter. But the posterior distribution captures correlations between parameters, and in regression models the coefficients can be quite strongly correlated (depending on the data). If one would have strong prior believes on the correlations among parameters, this could be represented in a Bayesian hierarchical model. However, because these correlations are data specific in regression,

Table 3
Posterior Results for Scenario 1

	Model 1: Neyer & Asendorpf (2001) data without prior knowledge		Model 2: Sturaro et al. (2010) data without prior knowledge		Model 3: Sturaro et al. (2010) data with priors based on Model 1	
Parameters	Estimate (*SD*)	95% PPI	Estimate (*SD*)	95% PPI	Estimate (*SD*)	95% PPI
β_1	0.605 (0.037)	0.532–0.676	0.291 (0.063)	0.169–0.424	0.333 (0.060)	0.228–0.449
β_2	0.293 (0.047)	0.199–0.386	0.157 (0.103)	−0.042–0.364	0.168 (0.092)	−0.010–0.352
β_3	**0.131 (0.046)**	**0.043–0.222**	**0.029 (0.079)**	**−0.132–0.180**	**0.044 (0.074)**	**−0.103–0.186**
β_4	**−0.026 (0.039)**	**−0.100–0.051**	**0.303 (0.081)**	**0.144–0.462**	**0.247 (0.075)**	**0.101–0.393**

Model fit	Lower CI	Upper CI	Lower CI	Upper CI	Lower CI	Upper CI
95% CI for difference between observed and replicated chi-square values	−14.398	16.188	−12.595	17.263	−12.735	17.298
ppp value	.534		.453		.473	

Note. See Figure 3 for the model being estimated and the interpretation of the parameters. Posterior *SD* = standard deviation; PPI = posterior probability interval; CI = confidence interval; ppp value = posterior predictive *p* value.

Table 4
Posterior Results for Scenario 2

	Model 4: Asendorpf & van Aken (2003) data without prior knowledge		Model 5: Asendorpf & van Aken (2003) data with priors based on Model 1		Model 6: Sturaro et al. (2010) data with priors based on Model 5	
Parameters	Estimate (*SD*)	95% PPI	Estimate (*SD*)	95% PPI	Estimate (*SD*)	95% PPI
β_1	0.512 (0.069)	0.376–0.649	0.537 (0.059)	0.424–0.654	0.314 (0.061)	0.197–0.441
β_2	0.115 (0.083)	−0.049–0.277	0.139 (0.077)	−0.011–0.288	0.144 (0.096)	−0.039–0.336
β_3	**0.217 (0.106)**	**0.006–0.426**	**0.195 (0.094)**	**0.007–0.380**	**0.044 (0.076)**	**−0.109–0.191**
β_4	**0.072 (0.055)**	**−0.036–0.179**	**0.065 (0.052)**	**−0.040–0.168**	**0.270 (0.075)**	**0.121–0.418**

Model fit	Lower CI	Upper CI	Lower CI	Upper CI	Lower CI	Upper CI
95% CI for difference between observed and replicated chi-square values	−16.253	17.102	−16.041	15.625	−12.712	16.991
ppp value	.515		.517		.473	

Note. See Figure 3 for the model being estimated and the interpretation of the parameters. Posterior *SD* = standard deviation; PPI = posterior probability interval; CI = confidence interval; ppp value = posterior predictive *p* value.

and data and model specific in SEM (see Kaplan & Wenger, 1993), it is unlikely that we would be able to elicit such priors. Therefore, the easiest approach is to specify independent marginal priors and let the posterior capture the empirical correlations.

Scenario 2

Assuming the cross-lagged panel effects to be not age dependent, Asendorpf and van Aken (2003) could have used the results from Neyer and Asendorpf (2001) as the starting point for their own analyses, which in turn could have been the starting point for Sturaro et al. (2010). In the second column of Table 4, the results, without assuming prior knowledge, of Asendorpf and van Aken are displayed, that is, Model 4 (*Asendorpf & van Aken, 2003 | Uninf. prior*). In the third column, that is, Model 5 (*Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001*), the data of Asendorpf and van Aken were updated using prior information taken from Model 1. In the last step, that is, Model 6 (*Sturaro* et al., *2010 | Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001*), the data of Sturaro et al. were updated using the prior information taken from Model 5. In sum, the models tested are as follows:

Uninformative priors

Neyer & Asendorpf, 2001 | Uninf. prior
Asendorpf & van Aken, 2003 | Uninf. prior
Sturaro et al., 2010 | Uninf. prior

Scenario 1: Age specificity when updating knowledge
Sturaro et al., 2010 | Neyer & Asendorpf, 2001

Scenario 2: Age invariance when updating knowledge
Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001
Sturaro et al., 2010 | Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001

Model Fit

When using SEM models to analyze the research questions, one is not interested in a single hypothesis test, but instead in the evaluation of the entire model. Model fit in the Bayesian context relates to assessing the predictive accuracy of a model, and is referred to as *posterior predictive checking* (Gelman et al., 2004). The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model and the actual data itself. Any deviation between the data generated by the model and the actual data suggests possible model misspecification. In essence, posterior predictive checking is a method for assessing the specification quality of the model from the viewpoint of predictive accuracy. A complete discussion of Bayesian model evaluation is beyond the scope of this study; we refer the interested reader to Kaplan and Depaoli (2012, 2013).

One approach to quantifying model fit is to compute Bayesian posterior predictive *p* values (*ppp* value). The model test statistic, the chi-square value, is calculated on the basis of the data is compared to the same test statistic, but then defined for the simulated data. Then, the ppp value is defined as the proportion of chi-square values obtained in

the simulated data that exceed that of the actual data. The ppp values around .50 indicate a well-fitting model.

Posterior Results

Scenario 1

In Table 3 the posterior results are displayed for the first scenario. Consider the posterior regression coefficient for the stability path of Friends (β_2), which is estimated as .293 in Model 1; Models 2 and 3 represent different ways of updating this knowledge. Model 2 ignores the results by Neyer and Asendorpf (2001) and achieves a stability path of .157. Model 3, in contrast, bases the prior distributions on the posterior results of Model 1 (Neyer & Asendorpf, 2001 | Uninf. prior) and arrives at a stability path of .168, which does not differ that much from the original outcome. If we compare the standard deviation of the stability path β_2 of Friends between Model 2 and Model 3 (Sturaro et al., 2010 | Neyer & Asendorpf, 2001), we can observe that the latter is more precise (decrease in variance from .103 to .092). Consequently, the 95% PPI changes from [−.042, .364] in Model 2 to [−.010, .352] in Model 3. Thus, the width of the PPI decreased and, after taking the knowledge gained from Model 1 into account, we are more confident about the results of the stability path of Friends.

The cross-lagged effect between Friends T1 → Extraversion T2 (β_4) is estimated as −.026 in Model 1 (Neyer & Asendorpf, 2001 | Uninf. prior), but as .303 in Model 2 (Sturaro et al., 2010 | Uninf. prior). When Model 1 is used as input for the prior specification for the Sturaro et al. (2010) data, Model 3 (Sturaro et al., 2010 | Neyer & Asendorpf, 2001), the coefficient is influenced by the prior, and the coefficient becomes .247 again with a smaller PPI. Furthermore, in both Models 2 and 3, the cross-lagged effect between Extraversion T1 → Friends T2 (β_3) in Model 3 appears not to be significant.

Scenario 2

Concerning Scenario 2 the results of the updating procedure are shown in Table 4. Compare Models 4 (Asendorpf & van Aken, 2003 | Uninf. prior) and 5 (Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001) where the data of Asendorpf and van Aken (2003) were analyzed with noninformative priors and priors based on Model 1, respectively. Again, in Model 5 the PPIs decreased when compared to Model 4 because of the use of subjective priors. In Model 6 (Sturaro et al., 2010 | Asendorpf & van Aken, 2003 | Neyer & Asendorpf, 2001), the data of Sturaro et al. (2010) were analyzed using priors based on Model 5; consequently, the posterior results of Model 6 are different from the results of Sturaro et al. in Model 2 where no prior knowledge was assumed.

Discussion of Empirical Example

Inspection of the main parameters, the cross-lagged effects, β_3 and β_4, indicate that there are hardly any differences between Scenarios 1 and 2. Apparently, the results of Sturaro et al. (2010) are robust irrespective of the specific updating procedure. However, there are differences between the updated outcomes and the original results. That is, Models 3 and 6 have smaller standard deviations and narrower PPIs compared to Model 2. Thus, using prior knowledge in the analyses led to more certainty about the outcomes of the analyses and we can be more confident in the conclusions, namely, that Sturaro et al. found opposite effects to Neyer and Asendorpf (2001). This should be reassuring for those who might think that Bayesian analysis is too conservative when it comes to revising previous knowledge. Therefore, the bottom line remains that effects occurring between ages 17 and 23 are different from those found when ages 18–30 were used as range. The advantage of using priors is that the confidence intervals became smaller such that the effect of different ages (17–23 vs. 18–30) on the cross-lagged results can be more trusted than before.

Because developmental mechanisms may vary over time, any (reciprocal) effects found between ages 12 and 17 are not necessarily found between ages 17 and 23. Although the Sturaro et al. (2010) study was originally designed as a replication of the Asendorpf and van Aken (2003) study, results turned out to be more consistent with the alternative explanation of the social investment theory of Roberts et al. (2005), namely, that between ages 17 and 23 there might be more change in personality because of significant changes in social roles. In spite of the fact that we have chosen for the exact replication of the Asendorpf and van Aken study (because this was the stated goal of the Sturaro et al., 2010, study), developmental researchers of course should not blindly assume that previous research findings from different age periods can be used to derive priors. After all, development is often multifaceted and complex and looking only for regularity might make the discovery of interest-

ing discontinuities more difficult. In such cases, however, this sense of indetermination needs to be acknowledged explicitly and translated into prior distributions that are flatter than would be typical in research fields in which time periods are more interchangeable.

Discussion

One might wonder when it is useful to use Bayesian methods instead of using the default approach. Indeed, there are circumstances in which both methods produce very similar results, but there are also situation that both methods should produce different outcomes. Advantages of Bayesian statistics over frequentist statistics are well documented in the literature (Jaynes, 2003; Kaplan & Depaoli, 2012, 2013; Kruschke, 2011a, 2011b; Lee & Wagenmakers, 2005; Van de Schoot, Verhoeven, & Hoijtink, 2012; Wagenmakers, 2007) and we will just highlight some of those advantages here.

Theoretical Advantages

When the sample size is large and all parameters are normally distributed, the results between ML estimation and Bayesian estimation are not likely to produce numerically different outcomes. However, as we discussed in our study, there are some theoretical differences.

1. The interpretation of the results is very different; for example, see our discussion on confidence intervals. We believe that Bayesian results are more intuitive because the focus of Bayesian estimation is on predictive accuracy rather than "up or down" significance testing. Also, the Bayesian framework eliminates many of the contradictions associated with conventional hypothesis testing (e.g., Van de Schoot et al., 2011).
2. The Bayesian framework offers a more direct expression of uncertainty, including complete ignorance. A major difference between frequentist and Bayesian methods is that only the latter can incorporate background knowledge (or lack thereof) into the analyses by means of the prior distribution. In our study we have provided several examples on how priors can be specified and we demonstrated how the priors might influence the results.
3. *Updating knowledge:* Another important argument for using Bayesian statistics is that it allows updating knowledge instead of testing a null hypothesis over and over again. One important point is that having to specify priors forces one to better reflect on the similarities and differences between previous studies and one's own study, for example, in terms of age groups and retest interval (not only in terms of length but also in terms of developmental processes). Moreover, the Bayesian paradigm sometimes leads to replicating others' conclusions or even strengthening them (i.e., in our case), but sometimes leads to different or even opposite conclusions. We believe this is what science is all about: updating one's knowledge.

Practical Advantages

In addition to the theoretical advantages, there are also many practical advantages for using Bayesian methods. We will discuss some of them.

1. *Eliminating the worry about small sample sizes*—albeit with possible sensitivity to priors (as it should be). Lee and Song (2004) showed in a simulation study that with ML estimation the sample size should be at least 4 or 5 times the number of parameters, but when Bayes was used this ratio decreased to 2 or 3 times the number of parameters. Also, Hox et al. (2012) showed that in multilevel designs at least 50 clusters are needed on the between level when ML estimation is used, but only 20 for Bayes. In both studies default prior settings were used and the gain in sample size reduction is even larger when subjective priors are specified. It should be noted that the smaller the sample size, the bigger the influence of the prior specification and the more can be gained from specifying subjective priors.
2. When the sample size is small, it is often hard to attain statistical significant or meaningful results (e.g., Button, et al., 2013). In a cumulative series of studies where coefficients fall just below significance, then if all results show a trend in the same direction, Bayesian methods would produce a (slowly) increasing confidence regarding the coefficients—more so than frequentist methods.
3. *Handling of non-normal parameters*: If parameters are not normally distributed, Bayesian methods provide more accurate results as they can deal with asymmetric distributions. An important example is the indirect effect of

a mediation analysis, which is a multiplication of two regression coefficients and therefore always skewed. Therefore, the standard errors and the confidence interval computed with the classical Baron and Kenny method or the Sobel test for mediation analyses are always biased (see Zhao, Lynch, & Chen, 2010, for an in-depth discussion). The same arguments hold for moderation analyses where an interaction variable is computed to represent the moderation effect. Alternatives are bootstrapping, or Bayesian statistics (see Yuan & MacKinnon, 2009). The reason that Bayes outperforms frequentist methods is that the Bayesian method does not assume or require normal distributions underlying the parameters of a model.

4. *Unlikely results*: Using Bayesian statistics it is possible to guard against overinterpreting highly unlikely results. For example, in a study in which one is studying something very unlikely (e.g., extrasensory perception; see the discussion in Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), one can specify the priors accordingly (i.e., coefficient = 0, with high precision). This makes it less likely that a spurious effect is identified. A frequentist study is less specific in this regard. Another example is using small variance priors for cross-loadings in confirmatory factor analyses or in testing for measurement invariance (see Muthén & Asparouhov, 2012). The opposite might also be wanted, consider a study in which one is studying something very likely (e.g., intelligence predicting school achievement). The Bayesian method would now be more conservative when it comes to refuting the association.
5. *Elimination of inadmissible parameters*: With ML estimation it often happens that parameters are estimated with implausible values, for example, negative residual variances or correlations larger than 1. Because of the shape of the prior distribution for variances/covariances, such inadmissible parameters cannot occur. It should be noted, however, that often a negative residual variance is due to overfitting the model and Bayesian estimation does not solve this issue. Bayesian statistics does not provide a "golden solution" to all of one's modeling issues.

In general, we do not want to make the argument for using Bayesian statistics because of its "superiority" but rather one of epistemology. That is, following De Finetti (1974a), we have to come to grips as to what probability is: long-run frequency of a particular result or the uncertainty of our knowledge? This epistemological issue is more fundamental than the divergence of results between the two approaches, which is often less than dramatic.

Limitations and Future Research

Of course, the Bayesian paradigm is not without assumptions and limitations. The most often heard critique is the influence of the prior specification, which might be chosen because of opportune reasons. This could open the door to adjusting results to one's hypotheses by assuming priors consistent with these hypotheses. However, our results might somewhat assuage this critique: The Sturaro et al. (2010) results were upheld even when incorporating priors that assumed an inverse pattern of results. Nevertheless, it is absolutely necessary for a serious article based on Bayesian analysis to be transparent with regard to which priors were used and why. Reviewers and editors should require this information.

Another discussion among Bayesian statisticians is which prior distribution to use. So far, we only discussed the uniform distribution and the normal distribution. Many more distributions are available as an alternative for the normal distribution, for example, a t distribution with heavier tails to deal with outliers (only available in WinBUGS). It might be difficult for nonstatisticians to choose among all these, sometimes exotic, distributions. The default distributions available in Amos/Mplus are suitable for most models. If an analyst requires a nonstandard or unusual distributions, be aware that most distributions are not (yet) available in Amos/Mplus and it might be necessary to switch to other software, such as WinBUGS or programs available in R. Another critique is that in Bayesian analysis we assume that every parameter has a distribution in the population, even (co)variances. Frequentist statisticians simply do not agree on this assumption. They assume that in the population there is only one true fixed parameter value. This discussion is not the scope of our study and we would like to refer interested readers to the philosophical literature—particularly, Howson and Urbach (2006)–for more information.

A practical disadvantage might be that computational time increases because iterative sampling techniques are used. Fortunately, computer processors are becoming more efficient as well as cheaper

to produce. Accordingly, the availability of adequate hardware to run complex models is becoming less of a bottleneck, at least in resource-rich countries. On the other hand, Bayesian analysis is able to handle highly complex models efficiently when frequentist approaches to estimation (i.e., ML) often fail (e.g., McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). This is especially the case for models with categorical data or random effect where Bayes might even be faster than the default numeric integration procedures most often used.

Guidelines

There are a few guidelines one should follow when reporting the analytic strategy and posterior results in a manuscript:

1. Always make clear which priors were used in the analyses so that the results can be replicated. This holds for *all* the parameters in the model.
    a. If the default settings are used, it is necessary to refer to an article/manual where these defaults are specified.
    b. If subjective/informative priors are used, a subsection has to be included in the Analytical Strategy section where the priors are specified and it is should be explicitly stated where they come from. Tables could be used if many different priors are used. If multiple prior specifications are used, as we did in all our examples, include information about the sensitivity analysis.
2. As convergence might be an issue in a Bayesian analysis (see online Appendix S1), and because there are not many convergence indices to rely on, information should be added about convergence, for example, by providing (some of) the trace plots as supplementary materials.

In conclusion, we believe that Bayesian statistical methods are uniquely suited to create cumulative knowledge. Because the availability of proprietary and free software is making it increasingly easy to implement Bayesian statistical methods, we encourage developmental researchers to consider applying them in their research.

References

Albert, J. (2009). *Bayesian computation with R*. London, UK: Springer.

Arbuckle, J. L. (2006). *Amos 7.0 user's guide*. Chicago, IL: SPSS.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi:10.1002/per.1919

Asendorpf, J. B., & van Aken, M. A. G. (2003). Personality-relationship transaction in adolescence: Core versus surface personality characteristics. *Journal of Personality*, *71*, 629–666. doi:10.1111/1467-6494.7104005

Asendorpf, J. B., & Wilpers, S. (1998). Personality effects on social relationships. *Journal of Personality and Social Psychology*, *74*, 1531–1544. doi:0022-3514/98/$3.00

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. doi:10.1098/rstl.1763.0053

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae* [NEO-Five-Factor-Questionnaire as in Costa and McCrae]. Göttingen, Germany: Hogrefe.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafaò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Caspi, A. (1998). Personality development across the life course. In N. Eisenberg (Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 311–388). New York, NY: Wiley.

De Finetti, B. (1974a). Bayesianism: Its unifying role for both the foundations and applications of statistics. *International Statistical Review*, *42*, 117–130.

De Finetti, B. (1974b). *Theory of probability* (Vols. 1 and 2). New York, NY: Wiley.

Furman, W., & Buhrmester, D. (1985). Children's perceptions of the personal relationships in their social networks. *Developmental Psychology*, *21*, 1016–1024. doi:10.1037/0012-1649.21.6.1016

Geiser, C. (2013). *Data analysis with Mplus*. New York, NY: The Guilford Press.

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

Gigerenzer, G. (2004). The irrationality paradox. *Behavioral and Brain Sciences*, *27*, 336–338. doi:10.1017/S0140525X04310083

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago, IL: Open Court.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87–93.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

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

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 407–437). Oxford, UK: Oxford University Press.

Kaplan, D., & Wenger, R. N. (1993). Asymptotic independence and separability in covariance structure models: Implications for specification error, power, and model modification. *Multivariate Behavioral Research*, *28*, 483–498. doi:10.1207/s15327906mbr2804_4

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312. doi:10.1177/1745691611406925

Kruschke, J. K. (2011b). *Doing Bayesian data analysis*. Burlingon, MA: Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603. doi:10.1037/a0029146

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668. doi:10.1037/0033-295X.112.3.662

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*, 653–686.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. doi:10.1007/s11222-008-9100-0

Lynch, S. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*, 126–149. doi:10.1037/a0015857

McCrae, R. R., & Costa, P. T. (1996). Towards a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 51–87). New York, NY: Guilford Press.

Meeus, W., Van de Schoot, R., Keijsers, L., Schwartz, S. J., & Branje, S. (2010). On the progression and stability of adolescent identity formation. A five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Child Development*, *81*, 1565–1581. doi:0009-3920/2010/8105-0018

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). Biems: A Fortran90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*, 2.

Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. doi:10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Neyer, F. J., & Asendorpf, J. B. (2001). Personality-relationship transaction in young adulthood. *Journal of Personality and Social Psychology*, *81*, 1190–1204. doi:10.1037//0022-3514.81.6.1190

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., . . . Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex, UK: Wiley.

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur. Zur Validität des Fünf-Faktoren Modells der Persönlichkeit* [Language and personality structure: Validity of the five-factor model of personality]. Regensburg, Germany: Roderer.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York, NY: Wiley.

Rietbergen, C., Klugkist, I., Janssen, K. J. M., Moons, K. G. M., & Hoijtink, H. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, *32*, 848–855. doi:10.1016/j.cct.2011.06.002

Roberts, B. W., Wood, D., & Smith, J. L. (2005). Evaluating five factor theory and social investment perspectives on personality trait development. *Journal of Research in Personality*, *39*, 166–184. doi:10.1016/j.jrp.2004.08.002

Rowe, M. L., Raudenbush, S. W., & Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, *83*, 508–525. doi:10.1111/j.1467-8624.2011.01710.x

Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, *1*, 359–363.

Sturaro, C., Denissen, J. J. A., van Aken, M. A. G., & Asendorpf, J. B. (2010). Person-environment transactions during emerging adulthood the interplay between personality characteristics and social relationships. *European Psychologist*, *13*, 1–11. doi:10.1027/1016-9040.13.1.1

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W., & Romeijn, J.-W. (2011). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology*, *47*, 203–212. doi:10.1037/a0020957

Van de Schoot, R., Verhoeven, M., & Hoijtink, H. (2012). Bayesian evaluation of informative hypotheses in SEM using Mplus: A black bear story. *European Journal of Developmental Psychology*, *10*, 81–98. doi:10.1080/17405629.2012.732719

Van Wesel, F. (2011, July 1). *Priors & prejudice: using existing knowledge in social science research*. Utrecht Univer-

sity. Prom./coprom.: prof. dr. H.J.A. Hoijtink, dr. I.G. Klugkist & dr. H.R. Boeije.

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

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

Walker, L. J., Gustafson, P., & Frimer, J. A. (2007). The application of Bayesian analysis to issues in developmental research. *International Journal of Behavioral Development*, *31*, 366–373. doi:10.1177/0165025407077763

Weinert, F. E., & Schneider, W. (Eds.). (1999). *Individual development from 3 to 12: Findings from the Munich Longitudinal Study*. New York, NY: Cambridge University Press.

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

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*, 374–383. doi:10.1177/0165025407077764

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*, 197–206. doi:10.1086/651257

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Bayes Theorem in More Details.
Appendix S2. Bayesian Statistics in Mplus.
Appendix S3. Bayesian Satistics in WinBugs.
Appendix S4. Bayesian Statistics in AMOS.