Journal of the
Experimental Analysis of Behavior

JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2019, **111,** 225–238 NUMBER 2 (MARCH)

Bayesian data analysis as a tool for behavior analysts

MICHAEL E. YOUNG

KANSAS STATE UNIVERSITY

Bayesian approaches to data analysis are considered within the context of behavior analysis. The paper distinguishes between Bayesian inference, the use of Bayes Factors, and Bayesian data analysis using specialized tools. Given the importance of prior beliefs to these approaches, the review addresses those situations in which priors have a big effect on the outcome (Bayes Factors) versus a smaller effect (parameter estimation). Although there are many advantages to Bayesian data analysis from a philosophical perspective, in many cases a behavior analyst can be reasonably well-served by the adoption of traditional statistical tools as long as the focus is on parameter estimation and model comparison, not null hypothesis significance testing. A strong case for Bayesian analysis exists under specific conditions: When prior beliefs can help narrow parameter estimates (an especially important issue given the small sample sizes common in behavior analysis) and when an analysis cannot easily be conducted using traditional approaches (e.g., repeated measures censored regression).

Key words: Bayesian, data analysis, brms, behavior analysis

The term "Bayesian" is bandied about in various circles of psychology in multiple ways and with too little understanding. Thus, one of my first goals in this article is to clarify the term. A second goal is to consider how and when a Bayesian approach can be of use to behavior analysts. It is not possible to provide a complete tutorial on Bayesian data analysis, but I hope that readers will come away with a greater conceptual understanding that will help them in the future and, perhaps, motivate further learning in order to conduct these types of analyses.

There are two reasons that Bayesian approaches should be in the toolbox for behavior analysts. First, Bayesian approaches are rapidly becoming feasible and tractable; although they have been around for many decades, the required computational power made Bayesian analysis unfeasible except for the simplest analyses. Second, they address important concerns raised in the behavior analytic community regarding the historical use of statistical analysis. Classical null hypothesis significant testing (NHST) arguably enables decision making in the absence of human judgment because the latter is prone to error (Deprospero & Cohen, 1979; Matyas & Greenwood, 1990); results from NHST are deemed significant or not significant based on a field standard for significance (e.g., $p < .05$). This rigid approach has been rightly criticized by many (Cohen, 1994; Cumming, 2014; Fidler, Cumming, Burgman, & Thomason, 2004; Glover & Dixon, 2004; Meehl, 1978), including those within the behavior analytic community (Branch, 1999, 2014; Perone, 1999). In contrast, Bayesian approaches to data analysis emphasize evidence rather than decisions. Traditional approaches can also be repurposed to focus on degrees of evidence, but Bayesian analyses are naturally suited to this approach due to the logic behind their construction.

It is important to emphasize that the criticisms of classical statistics among behavior analysts involve many issues, each of which prompts different solutions. Some have raised significant concerns with analyses that collapse across individuals (Branch, 1999; Perone, 1999; Skinner, 1956). These concerns are valid but addressable through techniques that fit data at the individual level or at individual and group levels simultaneously (using multilevel modeling, Everitt, 1998; Gelman & Hill, 2006). Addressing this issue does not require Bayesian analysis. Relatedly, others have raised concerns that a focus on statistical results means that the scientist will be less in touch with the data (Skinner, 1956; Branch, 1999). This criticism is targeted at the application and quality of the analysis. If data are aggregated across

Address correspondence to: Dr. Michael Young, Department of Psychological Sciences, Kansas State University, 492 Bluemont Hall, Manhattan, KS 66506. E-mail: michaelyoung@ksu.edu

doi: 10.1002/jeab.512

trials or sessions without accurately depicting the range and uncertainty of the observed data, then a statistical analysis does indeed misrepresent the observed data. However, this misrepresentation can be a shortcoming of the analyst rather than the analytical tools being employed. Often, inappropriate statistical aggregation occurs because the analyst does not know the tools that can model the observed relationships (e.g., using nonlinear fitting) or how to appropriately plot the uncertainty in these relationships. These shortcomings occur even with visual representation when data are aggregated before plotting. A final concern is the unnatural logic in classical statistical analysis that prompts misunderstanding of what an analysis really provides. Specifically, the ubiquitous NHST *p*-value assesses the probability of obtaining the observed data under an assumption (that the effect is exactly zero) that is indefensible in the limit because *p* will always approach zero as the sample size increases unless the population effect is exactly zero. Why does this occur? Because larger sample sizes allow greater precision in the estimate of the population effect (e.g., a difference, correlation, or regression weight). If the effect is small but not zero, then increasingly larger samples are able to reject the null for increasingly smaller effects. Bayesian analysis is ideally suited to addressing the problem of only assessing rejection of the null hypothesis because it simultaneously evaluates the relative likelihood of an entire distribution of hypothesized values for the effect, not just a value of zero.

Bayesian Inference

Bayes' theorem was developed in the 1700s by the Reverend Thomas Bayes and is simply a mathematically correct way to integrate prior beliefs with new evidence to create updated beliefs. Psychologists who study Bayesian inference (Gigerenzer & Hoffrage, 1995; Tenenbaum, Griffiths, & Kemp, 2006) typically present subjects with a prior belief in the form of a probability: for example., the base rate of diabetes in the population, $P(diabetic)$. New evidence in the form of an observed result is then provided along with two conditional probabilities involving that evidence. For example, a subject could be told that a patient tested positive for diabetes (the new evidence), the probability of testing positive if a patient is diabetic—the true positive rate, $P(positive \mid diabetic)$—read as "the probability of testing positive given that the person is diabetic"), and the probability of testing positive if a patient is not diabetic—the false positive rate, $P(positive \mid not\ diabetic)$. More concretely, a decision maker may be provided a base rate of diabetes of 9%, a true positive rate of 98% and a false positive rate of 32% and asked to infer the probability that a patient testing positive actually has diabetes (in this case, the correct answer is 23%[1]). Bayes theorem dictates the rules of probability that should be followed to derive this posterior (after the evidence) probability.

Psychologists who study Bayesian inference may study people's inferences involving conditional probability problems, but they may also be positing that a species possesses an evolved Bayesian inferential mechanism shaped by individual or species history. For example, sensory information might update beliefs about stimulus identity based on new evidence using Bayesian inference as implemented in the nervous system (Kersten, Mamassian, & Yuille, 2004). A behavior analyst might posit that a subject's prior experimental history establishes a behavioral distribution that must be updated based on new evidence (e.g., new reinforcement contingencies) to create a new distribution of behaviors that optimally integrates the prior distribution with the new contingency evidence (Courville, Daw, & Touretzky, 2006). However, researchers positing that organisms are naïve Bayesians are making claims about the behavior of the organism under study, not the analysis of any data that they might collect.

Bayesian data analysis involves a researcher who has prior beliefs about hypotheses or parameter values (e.g., the slope of a line), collects evidence by running an experiment, and then updates those beliefs using Bayes theorem (Kruschke, 2014). The particulars of how this updating is historically done fall into

[1]To avoid using the Bayes theorem equation, imagine 100 people tested of which 9 have diabetes (9%). Of those 9, we would expect all 9 to test positive (98% true positive rate), whereas of the 91 without diabetes, about 29 are expected to test positive (32% false positive rate). Thus, of the 38 testing positive (9 true positives and 29 false positives), only 9 of 38 = 23.68% (23.25% without the rounding used here) are expected to have diabetes.

two categories—using Bayes Factors to compare the relative likelihood of two or more hypotheses after data has been collected, and using full-fledged Bayesian data analysis to posit a prior parameter distribution which is then updated based on new experimental data to create a posterior parameter distribution. I will separately consider each of these uses of Bayesian principles in data analysis.

Model Comparison and Bayes Factors

Commonly, statistical analysis involves assessing the parameter estimates for a single model. In a classical *t*-test or analysis of variance, the key parameters estimate the differences between conditions and include the uncertainty in those estimates (the standard error of the estimate). In a regression, the key parameters estimate the slopes of the relationships between a predictor and the outcome. The *t*-test, ANOVA, and regression are all special cases of the general linear model that allows the inclusion of continuous and categorical predictors and their associated parameter estimates. When people adopt NHST, the model comparisons involve individual parameter estimates—is each estimate different from a specific posited point estimate of zero (what Cohen, 1994, calls the nil hypothesis, and Branch, 1999, calls the dumb-null-hypothesis).

A more insidious issue is that the logic of these classical "frequentist" analyses provides evidence in the form of *P*(*data when the null hypothesis is true*) whereas the researcher believes that the analysis is estimating *P*(*the null hypothesis given the data observed*) (Cohen, 1994). This confusion of inverse probabilities is common in the general public (Eddy, 1982), and researchers are not immune (Bakan, 1966). For example, if someone undergoing a cancer screening is told that the probability of testing positive for cancer is 99% if one has cancer (*P*(positive test | cancer) = .99), then the natural conclusion is that a positive test indicates a 99% probability of having cancer (*P*(cancer | positive test) = .99). Bayes theorem, however, dictates that the latter probability cannot be determined without more information regarding the base rates of a positive test and having cancer. If the type of cancer being screened is exceedingly rare, then a positive test is more likely to be a false positive rather than a true positive if *P*(positive test | no cancer) is much greater than zero as it is for many cancer screenings. Thus, a positive test may not substantially increase the likelihood of actually having cancer when the base rate of this type of cancer is very low. In hypothesis testing, the importance of base rates to estimating the evidence in favor of a theory is exemplified by the aphorism "extraordinary claims require extraordinary evidence." If the a priori likelihood of a hypothesis is extremely small (say, 0.001% change of a pigeon having extrasensory perception), then even positive evidence in favor of the hypothesis may not substantially move the needle (e.g., its likelihood may have increased ten-fold but still only be 0.01%).

Disturbingly for most scientists, the actual *P*(*one's favorite theory*) given any set of observed data is routinely very small because an enormously large number of theories are possible—slight variations in functional relationships, predictors, interactions, and so forth, produce a huge array of alternatives most of which are not distinguishable given a finite data set. Although the prospects may thus appear grim for one's favorite theory, any particular research question should involve comparison among a discrete set of alternative theories, $P(theory_1)$ versus $P(theory_2)$ versus $P(theory_3)$ (the method of strong inference, Platt, 1964). For example, a behavior analyst can collect data on delay discounting in rats and compare the relative likelihoods that the relationship is best modeled as a hyperbolic, exponential, or hyperboloid discounting function. If one's favorite theory has much more evidence than the currently popular alternatives, then there is some reason for confidence, at least until better alternatives emerge.

Using classical methods, a researcher can compare two models by calculating whether one model accounts for more variance (R^2) than another model after appropriate adjustment for differences in model complexity. This approach has the advantage of being intuitive. To reduce the comparison to a decision, an associated *p*-value calculation requires that one model is nested within another (i.e., that one model contains a strict subset of the variables used in the comparison model), so when the two models are not nested (e.g., for the hyperbolic vs. exponential example) it is necessary to either use human judgment or an alternative metric. The problem

with using human judgment is that software does not routinely produce an uncertainty estimate for an R^2 value. Thus, an R^2 of .95 for one model might seem much higher than an R^2 of .85 for the alternative, but if the 95% confidence interval for the two R^2 values were [.80, 1.00] and [0.71, 0.96], respectively, a researcher should feel less confident that the first model is superior to the second. To further highlight the issues with R^2, if there are 2, 10, or 200 data points to predict and are perfectly predicted by a model, the R^2 value for all three fits would be 1.0. However, any reasonable person would have much greater confidence regarding the quality of the model with 200 data points than with only 2 (as an aside, note that any best-fit regression model used to predict any 2 nonidentical outcomes would produce an R^2 of 1.0 because the line would go through both points with zero error and have a nonzero slope).

An alternative method of model comparison involves likelihood-based estimates. Simply put, these approaches estimate the relative likelihood that the observed data were produced by one theory versus another—the likelihood ratio. For example, if the estimated likelihood for a hyperbolic model is 5×10^{-6} and that for the exponential is 2×10^{-7}, then the likelihood ratio in favor of the hyperbolic is 25 to 1. A *p*-value is not derivable for this particular comparison because the models are not nested, but a researcher would feel quite confident that their experimental evidence strongly favored the hyperbolic model. I want to emphasize, however, that the field should avoid turning this evidence-based approach into a categorical one in which a particular likelihood ratio must be exceeded in order to claim that one model should be accepted over another. The goal should be to accumulate evidence, not to make discrete decisions for every experiment (Baron & Perone, 1998; Sidman, 1960). For example, the likelihood ratio for a particular study may be relatively weak, say 2:1, due to the considerable variability in the estimates or small sample size. But, if the ratio for 15 other published studies were all in the same direction consistently favoring the first theory over the second but with similarly weak ratios, then the field should have much greater confidence that the body of results provides strong evidence in favor of the first theory. A focus on discrete decisions based on a threshold has historically resulted in weak results not being published (Franco, Malhotra, & Simonovits, 2014). Of course, any conclusion based on likelihood ratios only involves comparisons among the theories being tested, and it is quite likely that a future theory will be favored over any of the ones currently being tested—that is the nature of science.

Two common metrics for model comparison that are based on likelihood but adjust for model complexity are the AIC (Akaike Information Criterion, Akaike, 1974), and the BIC (Bayesian Information Criterion, Burnham & Anderson, 2004; Wagenmakers & Farrell, 2004). The difference in model AIC/BIC values is a measure of the relative evidence for the models being compared (the model with the lower value is more likely to have generated the data). Note that these metrics are on an uninformative scale of evidence and thus only meaningful when compared by examining the difference (not ratio) of the two metrics. The comparison also requires that the models are fitted to the same data. It is convenient for researchers to use these metrics because they are produced by most statistical software (although SPSS provides them for only a subset of common models). Both metrics are based on model likelihood and represent different assumptions about the models being compared; BIC uses a more severe penalty for model complexity than AIC and thus more strongly favors simpler models.

The Bayes Factor or BF operates similarly to a standard likelihood ratio except that the BF incorporates information about the priors of the model parameters. The notion of priors will be explored more in the section on Bayesian parameter estimation, but conceptually priors represent what is believed about parameter values before data collection. Thus, the calculation of the BF depends on what is assumed before the data are collected. The notion of priors makes researchers uncomfortable because the results of the analysis depend on prior beliefs and not only on the data collected in the current experiment. To help ease concerns with the specification of the prior, I will note three issues (also see Kruschke & Liddell, 2018). First, any experimenter has expectations that guide their actions when, for example, a data point is identified as an outlier because it is too extreme, a session is

omitted due to having too few completed trials, or a subject is dropped because of a failure to follow directions. Second, a prior can be based on earlier studies from one's laboratory or across multiple laboratories—the researcher knows the plausible values for the generalized matching law (Baum, 1974; Herrnstein, 1961), for example, based on earlier studies and should question an experiment or analysis that produced a value well outside this range. As a postdoctoral scientist, I was initially surprised when my advisor was routinely able to identify a likely hardware problem (burned out lights, clogged automatic feeders, failing touch screens) based on the data being observed in a session, a byproduct of his years of experience shaping his expectations. Third, it is possible to specify uninformative priors such that the results of an analysis are exclusively driven by the current set of data and functionally identical to those derived from frequentist analyses (the results are not guaranteed to be identical because Bayesian methods are derived from simulation, not computation).

A BF can be approximated using BIC values; this approximation assumes a particular prior called the *unit information prior* (Wagenmakers, 2007). This prior peeks at the data to generate a weakly informed prior centered around the observed value, but the resulting liberal bias caused by peeking at the data is relatively small for reasonable sample sizes. If your software generates a BIC for your models, the BF in favor of H_0 over H_1 is $exp((BIC(H_1) - BIC(H_0))/2)$. If the BIC for a hyperbolic model is -3110 and for an exponential is -3100, the BF favoring the hyperbolic is: $exp((-3100 - -3110)/2)$ or 148:1. Note that BICs can be compared only when both models are based on the same data (missing data for a predictor only present in one model will produce different data for the models) and the same outcome variable (i.e., one model cannot be based on a transformation of the outcome variable whereas the other is not); differences in sample size and the scale of the outcome variable alter the evidence scale in a way unrelated to the model fit. Because using the BIC to approximate a BF has been criticized as being too conservative by overly favoring simple models (Weakliem, 1999), some researchers favor a ratio based on AIC values, but the debate is far from resolved (Wagenmakers & Farrell, 2004).

There are other default priors that can be used to compute Bayes Factors for basic designs (e.g., Rouder, Morey, Speckman, & Province, 2012) or the researcher can specify an informative prior based on previous research or based on reason and then use available on-line calculators to estimate a Bayes Factor for a particular result (e.g., http://pcl.missouri.edu/bayesfactor, https://medstats.github.io/bayesfactor.html). More helpful, this approach allows the researcher to specify a model comparison in which a "null range" can be specified instead of a specific value (e.g., the nil hypothesis). The model comparison then represents an examination of the evidence that the value is greater than "unimportantly small" where the researcher can define the range of values that are unimportant (also see Kruschke & Liddell, 2018).

The use of BFs for model comparison focuses the behavior analyst's attention on the relative strength of evidence in favor of the candidate models. These models could differ in complexity (e.g., by including an interaction or additional parameter), predictors, or functional form. However, remember that scientists should avoid hard decisions based on a threshold, whether a specific *p*-value, AIC difference, or BF. A shift toward model comparison, rather than statistical significance, serves to focus attention on relative evidence in favor of competing theories rather than a single model in isolation (Platt, 1964).

Bayesian Data Analysis

In a fully Bayesian approach, all of the parameters of a model have priors that are updated using an experiment's data, and conducting such an analysis requires specialized software. While classical statistics focus on estimating each parameter's most likely value and the uncertainty of that estimate, a Bayesian analysis assesses a range of plausible parameter values and their relative likelihoods. Thus, the analysis requires that a distribution of possibilities is specified (the prior) which is integrated with the results of the experiment (represented by a likelihood function) to produce a posterior distribution. Figure 1 shows four examples of prior and posterior distributions for the intercept in a nonlinear regression of a delay discounting function (these will be explained more fully in a later section).

Although a plot of the full posterior distribution of each parameter can be provided in the results section of a manuscript, it is also common to only report the 95% credible interval for each parameter.

Before unpacking the general concepts behind Bayesian data analysis, it is important to consider why and when a behavior analyst should care. First, the BF approach noted above is restricted in the types of priors that can be specified. For example, should the prior be a unit information prior (the assumption in the BIC approach), an alternative default prior (e.g., as used in Rouder et al.'s,

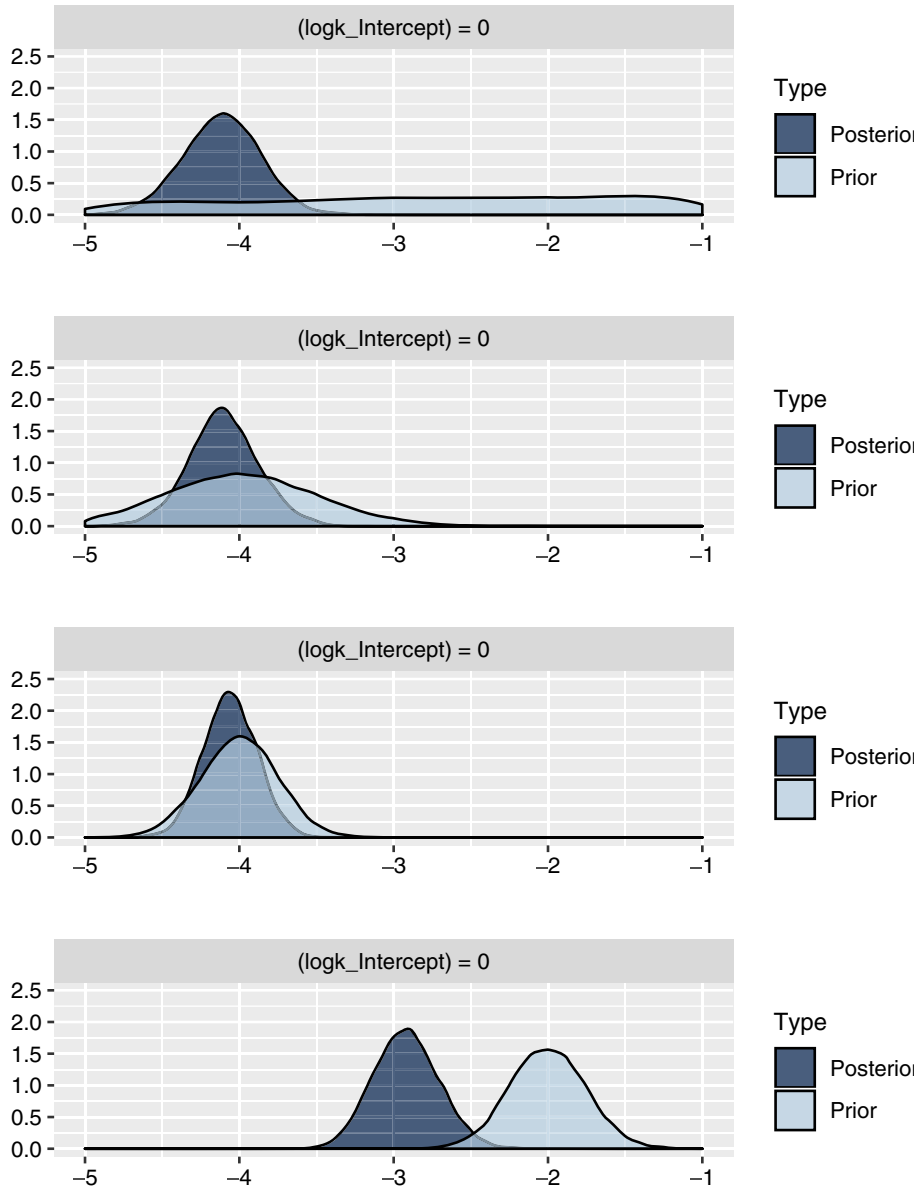

Fig. 1. Prior and posterior distributions of a single parameter using the same data but four different priors. The priors were all normally distributed. The mean and standard deviations for the priors were *N(0.0, 5.0)*, *N(-4.0, 0.5)*, *N(-4.0, 0.25)*, and *N(-2.0, 0.25)*, top to bottom, respectively. The posterior for the top plot is driven nearly entirely by the data.

2012, online calculator), or a more specific prior based on earlier experiments or theory (which requires a full Bayesian analysis)? Second and relatedly, a Bayesian analysis creates a more incremental scientific approach by allowing subsequent experiments to be informed by prior results using a logically sound process rather than by idiosyncratic human judgment. Bayesian analysis can be especially helpful when the new experiment involves a small number of subjects, thus representing a sample size that may be too small for statistical significance to emerge from a classical analysis. Third, Bayesian modeling provides a greater range of models that can be considered—separate distributions can be specified for each parameter with a wide selection of available options. For example, a distribution of latencies could conform to a Gaussian, lognormal, gamma, exGaussian, or Weibull, *inter alia*, the distribution of a slope estimate could be normal, whereas that for a variance may be a half-normal, half-t, etc. This flexibility is especially important for nonlinear models where the parameter variability may be best modeled using a nonnormal distribution (e.g., one that does not allow negative values). Fourth, Bayesian analysis evaluates the relative likelihood of a wide range of plausible values for each parameter conditioned on the data; in other words, the analyst is no longer confined to computing only the most likely value (i.e., the mode) along with a measure of uncertainty. Despite those many advantages, Bayesian analysis for very complex models is not simple and can be quite computationally expensive requiring many hours to run a single analysis and thus compelling the scientist to consider the tradeoffs at hand. This brute force approach is a major part of why Bayesian approaches can be used to fit a greater variety of statistical models, but this flexibility comes at a cost in computational time.

Of the advantages noted, I consider the incremental science issue to be the most important for behavior analysts. This conclusion is based on the much smaller number of subjects used in the field. An experienced behavior analyst will have well-informed expectations concerning the range of parameter estimates that should be observed in a study. There are numerous published studies of delay discounting, the generalized matching law, and variable ratio schedules that will eliminate a wide range of possible values of discount rates, sensitivity to reinforcer rate, and interresponse times (IRTs), *inter alia*. A Bayesian does not have to assume the uninformative priors that are implicit in classical statistics in which parameter estimates are derived solely from the current experiment. By integrating this prior knowledge into the analysis, the posterior parameter estimates can be better informed. Many researchers, however, are uncomfortable with the specification of priors (Rouder & Morey, 2012; Rouder et al., 2012) especially when priors must be specified at the level of individual parameters. Bayesians must, however, be explicit regarding their choice of priors, and their choices can be critiqued by reviewers, editors, and readers.

To illustrate how a Bayesian analysis is approached, I will use a concrete example involving a basic repeated measures design using a graded predictor, trial, and a three-level within-subject categorical predictor, condition. The experiment involved pigeons being trained on a transposition task involving one, two, or four pairs of stimuli (Lazareva, Wasserman, & Young, 2005). My goal was to predict the probability that the response was correct. I used a multilevel model that allowed each subject to have a different learning rate (trial slope), different estimates for the condition effects (condition slopes), and different average accuracy (intercept). Because the analysis is based on individual trial accuracy, the outcome variable was binomial, 0 (incorrect) and 1 (correct).

Figure 2 shows two commands used in R's *brms* (Bayesian regression models using Stan) package to help illustrate what is common to both Bayesian and classical approaches and what is distinct to a Bayesian analysis. The model is a standard multilevel model in which there are three predictors: two main effects (trial and condition) and their interaction (trial × condition). The terms within the inner parentheses of the command represent the random effects that are included to allow individual subject estimates of the trial slope, condition differences, and average accuracy (note that technically the three-level condition variable produces two dummy predictors, thus there are more parameters being estimated than described here). The second line is also standard and simply notes the name of the data set and the distribution family for the

```
fit1<- brm(accuracy ~ trial*condition + (trial + condition | subjectID),
       data=myd, family=bernoulli)

Defaulted to uninformed priors on all parameters, four chains, 1000 warmup, 1000
sampling (i.e., iter = 2000).


fit2<- brm(accuracy ~ trial*condition + (trial + condition | subjectID),
       data=myd, family=bernoulli,
       prior = c(set_prior("normal(0,1)", class="b"),
                 set_prior("lkj(2)", class="cor")),
       warmup = 500, iter = 1000, chains = 3, cores=3,
       sample_prior = TRUE)
```

Fig. 2. Commands using the *brms* package of *R* to perform Bayesian multilevel logistic regression. The trial*condition notation is shorthand that will expand into a full factorial combination of the predictors trial, condition, and trial × condition.

outcome variable. The family specification is omitted if the residuals are assumed to be normally distributed. For fit1, the analysis used the defaults for the Bayesian aspects of the fit including the use of uninformed priors.

The additional lines for the second command, fit2, include an explicit specification of the priors. Here, all beta weights (other than the intercept) are given the same modestly informed prior with a mean of 0 and a standard deviation of 1. All standard deviations relied on a default mildly informed prior. Because the design was within-subject, measures are assumed to be correlated when taken from the same subject, and these correlations were given a weak prior using the LKJ-Correlation prior (Lewandowski, Kurowicka, & Joe, 2009) to ensure that the correlations are within their legal range of -1 to +1. The previous model, fit1, also included correlations between measures but relied on an uninformed default prior but still restricted to the -1 to +1 range. Although priors can be specified at a finer level (e.g., the intercept could have a mean of 0.0 and s.d. of 2.0 whereas the trial slope could have a mean of 0.2 and a s.d. of 0.3), this approach is not represented here.

The next line of the second command requires more explanation regarding the mechanics of a Bayesian analysis. To produce the entire prior and posterior distributions of each parameter as shown in Figure 1, the analysis actually computes the likelihood of obtaining the observed data for a large number of combinations of parameter values and then stores them. This approach is different from a conventional analysis that attempts to find the single most likely value and estimates the uncertainty in that estimate based on the assumptions inherent in the analysis. Needless to say, it is not possible for a Bayesian analysis to estimate all possible combinations of the parameters for continuous variables like beta weights, the standard deviation of the residuals, or the correlations between measurements. Powerful computers make this process more tractable by using specialized algorithms to restrict the search of parameter values and their combinations to the range of values most capable of producing the data. The search process is called MCMC (Markov chain Monte Carlo) sampling. The sampling technique begins with a relatively blind search that is variously called the "burn in" or "warm up" period that informs later sampling. The likelihood data from this period are not included when tracing out the curves shown in Figure 1; these data merely ensure that the later sampling is focused around the values most likely to describe the data. In *brms*, the number of iterations spent during both the exploratory phase (*warmup*) and the total iterations (*iter*, which includes the warmup iterations) are specified. The larger these numbers, the more sampling that is done and the greater the likelihood of obtaining excellent specification of the relative likelihoods of the various parameter values. Of course, this desire for precision must be balanced against the amount of time required to generate the distribution. There is no reason to iterate more than is necessary to satisfy the scientific goals of the experiment. Identifying a sufficient number of warmup and sampling iterations is a function of the specific model being tested and largely determined by experience. For example, if the resulting distributions are not sufficiently smooth, have multiple peaks, or vary across "chains" (i.e., sampling runs) or otherwise show poor

convergence, then a larger number of warm-ups and iterations may address these problems. For a more complete description of MCMC sampling, the reader can consult a number of excellent treatments (e.g., Kruschke, 2014; van Ravenzwaaij, Cassey, & Brown, 2018).

The highest likelihood region for each model parameter represents a credible range for that parameter. In Bayesian analysis, this is called the "credible interval" and is roughly analogous to a confidence interval. I will not outline the distinct differences between credible and confidence intervals except to note that a 95% credible interval is providing precisely the information that most researchers believe that a 95% confidence interval provides, but does not (Hoekstra, Morey, Rouder, & Wagenmakers, 2014); specifically, the unobserved population parameter value is 95% likely to fall within the 95% credible interval. As usual, the accuracy of the estimated interval is rooted in the assumption that the sample is a random sample from a particular population to which the researcher wants to generalize.

Importantly, there are no *p*-values produced by a Bayesian analysis (see the sample output from such an analysis in Fig. 3; it includes estimates, approximate estimate uncertainty, and credible intervals but no *p*-values). Although one could treat the output in a categorical fashion by resorting to judgments regarding whether zero is within the 95% credible interval, this approach is actively discouraged because it is reminiscent of the nil hypothesis test (Cohen, 1994). Instead, the focus should be on the estimated magnitudes and the uncertainty in those estimates. A behavior analyst trained to use Bayesian statistical analysis is free to misuse the results of the analysis in a manner analogous to the misuse of NHST, an easy trap to fall prey to when trying to make the transition from classical statistical analysis to Bayesian analysis. A superior approach is for a behavior analyst to use the results of the analysis to fully describe the observed relationships by using a model that faithfully captures the data. In contrast, a visual representation that aggregates data before graphing is not depicting the variability, the correlation between data points, and the finer changes in behavior across time (Young, 2018b). Although the human eye can easily estimate changes in average performance from a visual plot alone, I would challenge any consumer of cumulative records or response rate graphs to accurately estimate changes in the variability in performance (standard deviation) or the rate at which performance is changing (slope).

The Effect of Priors

Because those new to Bayesian analysis are most nervous about the effect of priors on an analysis, Figure 1 illustrates four situations in which the same data are analyzed with four different sets of priors. The plots show both the prior and posterior distributions for one parameter in an analysis (in this case, the log of the *k* value in a hyperbolic discounting function which is assumed to be normally distributed). The top plot shows the results when using an uninformative prior that was extremely broad

```
Population-Level Effects:

                Estimate Est.Error l-95% CI u-95% CI Eff.Sample  Rhat
Intercept          1.068     0.192    0.687    1.454       1940 1.001
trial              0.000     0.001   -0.001    0.002       3490 1.000
condition1        -0.759     0.229   -1.196   -0.295       1927 1.000
condition2         0.086     0.267   -0.464    0.647       1906 1.001
trial:condition1  -0.000     0.001   -0.002    0.002       4000 0.999
trial:condition2  -0.000     0.001   -0.002    0.002       4000 1.001
```

Fig. 3. Example output from a Bayesian analysis as produced by the model shown in Figure 2; note that the three-level condition predictor created two dummy variables, condition1 and condition2. Only the population level effects are shown here; a multilevel model also estimates between-subject variability for random effects and the correlations between subject parameter estimates. The Estimate column is the most likely value, the Est.Error column is the standard deviation of the posterior distribution, and the l-95% CI and u-95% CI columns represent the upper and lower limits of the most compact range that contains 95% of the posterior distribution—the 95% credible interval. The last two columns provide information on MCMC sampling performance not addressed in this article.

and extended well beyond the range of intercept estimates plotted. Thus, for all practical purposes the posterior distribution is entirely determined by the data from the current experiment (and thus is conceptually similar to what is called the likelihood function in Bayesian analysis).

The second plot in Figure 1 used a modestly informed prior with a mean that was nearly identical to that reported in the current study but with greater uncertainty. This prior could have come from a small pilot study, from the results of published literature using different populations that created the greater variability, or from an earlier study that actually had less parameter uncertainty but for which the researcher felt uncomfortable providing a narrower prior because of differences in the preparation. By comparing the posterior from this analysis to that derived in the top plot using an uninformative prior, the visible effect is that the current posterior is slightly narrower.

The third plot in Figure 1 used a much more informative prior that was actually quite similar to the posterior observed in the top plot. This outcome might occur if the scientist ran 20 subjects in an experiment using an uninformative prior, and then later ran 20 additional subjects with the same preparation while using the results from the first 20 subjects as the prior for analyzing the follow-up. It is important to note that the Bayesian approach would produce a posterior that would be the same as that produced if all 40 subjects were analyzed in a single analysis. The result is that the posterior is notably narrower that that observed in the analysis based on the uninformative prior.

The fourth plot illustrates the situation that most worries scientists unfamiliar with Bayesian data analysis. Here, the prior was decidedly different from that observed in the current experiment and considered relatively precise. The analyst chose an average log k of -2.00 with a standard deviation of 0.25 whereas the average log k from the current study was -4.07 with a standard deviation of 0.26 (as represented in Fig. 1's top plot). Thus, there is no overlap between the prior distribution and that observed in the current experiment. Given that the analysis was confronted with two very different estimates that needed to be combined, and that the uncertainty in both estimates was very similar, the resulting posterior distribution had a mean that was in the middle between the two means. Furthermore, the new posterior distribution does not encompass the means from either the current data nor from the earlier study. While this outcome may raise concerns, the scientist would have been very explicit about the basis for their prior. Readers could then closely examine whether the rationale for the chosen prior was well founded. Perhaps the two samples involved two different rat strains, but the experimenter intended to estimate log k only for the current strain; in this case, the experimenter should have used a much broader, and perhaps uninformative, prior if it is unknown whether these strains should be expected to have similar discounting rates. Or, perhaps the experimenter intended to make claims about the log k for a sample that encompassed both strains; in that case, the new posterior estimate of log k would represent a good estimate for a new sample that encompassed an equal number of both strains.

Figure 4 illustrates the iterative nature of Bayesian analysis. For this example, I used the model shown in Figure 2 to analyze data from a sample of 12 pigeons learning to perform a transposition task involving one, two or four pairs of stimuli (Lazareva et al., 2005). The first row shows the posterior distributions for three key parameters based on analyzing 9 of these 12 birds (flat priors were used in this analysis). The second row shows the results when only the three withheld birds were analyzed but using the posteriors from the first nine birds as priors for this analysis. The breadth of the posterior distribution is noticeably narrower for two of these three parameters because the posterior reflects both the evidence provided by this small sample of three new birds as well as the body of evidence provided by the other set of nine birds. For comparison, the final row shows the posteriors when all 12 birds were analyzed together using flat priors; the similarity between the second and third rows is to be expected.

This example depicts a formal approach to the integration of prior beliefs derived from an earlier study with new evidence to produce an updated set of beliefs. The effect of the prior is determined by the uncertainty in the prior; this uncertainty would be large for a novel study with no prior expectations about

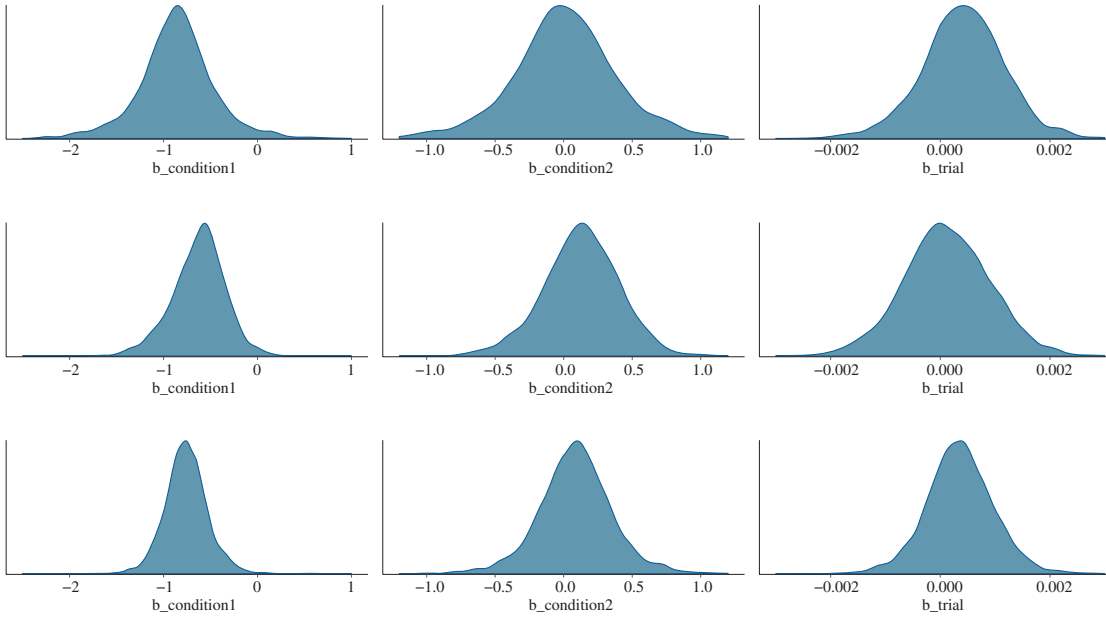

Fig. 4. Posterior parameter distributions of three of the parameters for a multilevel model. (Top) The result of analyzing the first nine subjects using flat priors. (Middle) The result of analyzing the next three subjects using the posteriors from the first nine, shown in in the top row, as priors for analyzing these three subjects. (Bottom) The result of analyzing all 12 subjects in a single analysis.

parameter values, but the uncertainty would be very small if the prior is based on a very large sample of earlier data collected in the same laboratory and using subjects very similar to those used in the new study. Everyday scientific practice lacks this formal approach to the integration of sources of information. Instead, current beliefs about response rates, discounting rates, and key preference are the result of years of experience and reading that is integrated in a much more informal and ad hoc manner. This informality includes the use of arbitrary cutoffs for data inclusion or reporting multiple estimates across studies with little consideration of the precision of those estimates. This informal integration is prone to memory biases, unsystematic averaging that is not appropriately sensitive to differences in sample size, idiosyncratic sampling of the literature, motivated reasoning (whereby one's reasoning is affected by one's motivations, Kunda, 1990), and numerous other human shortcomings. Of course, a Bayesian could still be selective in their choice of earlier results on which to base their priors, but these choices would now be explicit and need to be justified.

A Cautious Case for Bayesian Analysis

Despite the conceptual clarity offered by a Bayesian approach, the ability to specify priors in order to build an incremental science, and the flexibility to test a wider range of models (including nonlinear and ordinal) with proper specification of the behavior of all model parameters, I am a pragmatist and thus usually use classical statistical approaches to analyze my data. My desire to focus on parameter estimates and their uncertainty (rather than *p*-values) predates my learning about Bayesian analysis. Because I have observed that confidence intervals and credible intervals are very similar in most studies conducted in my laboratory and have routinely used the likelihood-based AIC/BIC for model comparison rather than R^2, I experienced little pressing utility to adopt Bayesian methods. Indeed, I strongly believe that all behavior analysts should make the conceptual shift toward parameter estimation and model comparison in their consideration of data in all of their analyses, whether Bayesian or not.

There are situations, however, where the added complexity, run time, and effort are

well compensated by the benefits of a Bayesian approach: when I want to use a model that is not available in my traditional analysis packages, and when I want to use prior information to inform a condition in a new study that is a replication of a previous experiment. An example of the first situation was a recent need to analyze repeated measures latency data that had been censored by the experimental design. I was assessing how long a subject would wait before cashing in to receive a reward that was continuously increasing in value, but each trial had a varying deadline to respond thus resulting in a large proportion of trials producing no response. A proper analysis would not treat these trials as having a missing latency nor substitute the cutoff time as the latency (both approaches create estimation bias, Young & Crumer, 2019), but rather the analysis should faithfully represent that these trials have latencies that are at least as long as the cutoff time and then estimate the likely distribution of the unobserved tail of latencies. A Bayesian approach provided the solution by allowing the use of multilevel censored regression (Young & McCoy, in press).

An example of the second situation involving priors occurs when it is necessary to include a baseline condition or control group that has been previously tested many times but still must be included for comparison purposes. The prior work can have established the likely range of various parameter values including the mean, standard deviation, and correlation between measurements, thus avoiding the implicit use of an uninformative prior that undermines the power of the analysis of new data. Because granting agencies find it increasingly necessary to spend their limited funds wisely, a Bayesian approach can reduce the cost of proposed studies by leveraging prior results to increase the precision of a new study's parameter estimates without resorting to large samples.

Some Practical Guidance

For those situations in which Bayesian analysis offers sufficient utility to prompt the data analyst to learn a new technique, there is a new vocabulary and set of operating procedures, as clearly demonstrated by this article. New software needs to be learned, as well as how to run, interpret, and present the results of an analysis. For traditional statistical approaches involving *t*-tests, analysis of variance, and regression, a researcher can consider Jeffreys's Amazing Statistics Program (*JASP;* https://jasp-stats.org); JASP is public domain and has a graphical user interface. The primary drawback to the use of this software is that it does not currently handle any form of multilevel modeling which is the primary tool for repeated measures analysis involving continuous predictors. In contrast, the public domain *R* software (https://www.r-project.org) has excellent broad functionality, especially for multilevel modeling using the *lme4* package (for examples involving discrete choice data, see Young, 2018a). The *brms* package (for an overview, see https://cran.r-project.org/web/packages/brms/vignettes/brms_overview.pdf) provides a Bayesian implementation of the types of statistical models that are important to behavior analysts including multilevel generalized linear models, censored regression, and ordinal regression. The emergence of *brms* has made Bayesian analysis accessible to a much wider range of scientists; *brms* was used for all of the examples and graphs produced here.

For the most general software to conduct Bayesian analyses, the Just Another Gibbs Sampler (*JAGS*; http://mcmc-jags.sourceforge.net) and *Stan* (http://mc-stan.org) packages are public domain and comprehensive. Proper use of either requires a much stronger background in Bayesian data analysis because the increased functionality of these packages presents more pitfalls for the naïve user. Those people interested in using *JAGS* or *Stan* should consider John Kruschke's text on Bayesian analysis (Kruschke, 2014). Regardless of the software chosen, proper use will require background reading to understand the impact of various choices on an analysis and the interpretation of the results.

Behavior analysts are beginning to embrace statistics as a principled tool for understanding small-N designs as demonstrated by the other contributions to this special issue. The historical shortcomings of statistics were often a combination of limitations in the tools (e.g., the absence of accessible repeated measures logistic regression to analyze the type of choice data often produced in behavioral labs), misuse of current tools that was frequently the result of inertia in scientific practice (Young,

2016), and insufficient training. All of these shortcomings can be addressed without resorting to Bayesian data analysis. Bayesian analysis is targeted at the inability of classical approaches to estimate the support for one theory over another conditioned on the data, and credible intervals rather than the misinterpreted confidence interval. Bayesian approaches also provide a more general solution to testing more complex models. It is prime time for behavior analysts to fully embrace a new approach to data analysis that can put them at the forefront of psychological science, rather than being risk averse and late adopters of an innovation that so nicely addresses the historical concerns of the field regarding classical statistics.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723. https://doi.org/10.1109/TAC.1974.1100705

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437. https://doi.org/10.1037/h0020412

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 45-91). New York: Springer.

Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior, 22*(1), 231-242. https://doi.org/10.1901/jeab.1974.22-231

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *Behavior Analyst, 22*(2), 87-92. https://doi.org/10.1007/BF03391984

Branch, M. N. (2014). Malignant side effects of null-hypothesis significance testing. *Theory and Psychology, 24*, 256-277. https://doi.org/10.1177/0959354314525282

Burnham, K. P., & Anderson, D. R. (2004). Multimodal inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261-304. https://doi.org/10.1177/0049124104268644

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10*, 294-300. https://doi.org/10.1016/j.tics.2006.05.004

Cumming, G. (2014). The new statistics: why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Deprospero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*(4), 573-579. https://doi.org/10.1901/jaba.1979.12-573

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, UK: Cambridge University Press.

Everitt, B. S. (1998). Analysis of longitudinal data: Beyond MANOVA. *British Journal of Psychiatry, 172*, 7-10. https://doi.org/10.1192/bjp.172.1.7

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33*, 615-630. https://doi.org/10.1016/j.socec.2004.09.035

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: unlocking the file drawer. *Science, 345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704. https://doi.org/10.1037/0033-295X.102.4.684

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*, 791-806. https://doi.org/10.3758/BF03196706

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*, 267-272. https://doi.org/10.1901/jeab.1961.4-267

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review, 21*(5), 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kruschke, J. K. (2014). *Doing Bayesian data analysis*. Waltham, MA: Elsevier.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review, 25*(1), 155-177. https://doi.org/10.3758/s13423-017-1272-1

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-498. https://doi.org/10.1037/0033-2909.108.3.480

Lazareva, O. F., Wasserman, E. A., & Young, M. E. (2005). Transposition in pigeons: Reassessing Spence (1937) with multiple discrimination training. *Learning & Behavior, 33*(1), 22-46. https://doi.org/10.3758/BF03196048

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*, 1989-2001. https://doi.org/10.1016/j.jmva.2009.04.008

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*(3), 341-351. https://doi.org/10.1901/jaba.1990.23-341

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834. https://doi.org/10.1037/0022-006X.46.4.806

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *Behavior Analyst, 22*(2), 109-116. https://doi.org/10.1007/BF03391988

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353. https://doi.org/10.1126/science.146.3642.347

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for model selection in regression. *Multivariate Behavioral Research, 47*, 877-903. https://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356-374. https://doi.org/10.1016/j.jmp.2012.08.001

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*, 221-233. https://doi.org/10.1037/h0047662

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*, 309-318. https://doi.org/10.1016/j.tics.2006.05.009

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin and Review, 25*, 143-154. https://doi.org/10.3758/s13423-016-1015-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review, 14*(5), 779-804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192-196. https://doi.org/10.3758/BF03206482

Weakliem, D. L. (1999). A critique of of the Bayesian Information Criterion for model selection. *Sociological Methods and Research, 27*, 359-397. 1

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural Processes, 123*, 43-53. https://doi.org/10.1016/j.beproc.2015.09.009

Young, M. E. (2018a). Discounting: A practical guide to multilevel analysis of choice data. *Journal of the Experimental Analysis of Behavior, 109*(2), 293-312. https://doi.org/10.1002/jeab.316

Young, M. E. (2018b). A place for statistics in behavior analysis. *Behavior Analysis: Research and Practice, 18*, 193-202. https://doi.org/10.1037/bar0000099

Young, M. E., & Crumer, A. (2019). Reaction times. In J. Vonk & T. K. Shackelford (Eds.), *Encyclopedia of animal cognition and behavior.* Springer.

Young, M. E., & McCoy, A. W. (in press). Variations on the balloon analogue risk task: A censored regression analysis. *Behavior Research Methods.* doi.org/10.3758/s13428-018-1094-8

Received: August 7, 2018
Final Acceptance: February 3, 2019
Editor in Chief: Amy Odum
Associate Editor: Amy Odum