Judgment and Decision Making, Vol. 13, No. 6, November 2018, pp. 622–635

Bayesian methods for analyzing true-and-error models

Michael D. Lee*

Abstract

Birnbaum and Quispe-Torreblanca (2018) evaluated a set of six models developed under true-and-error theory against data in which people made choices in repeated gambles. They concluded the three models based on expected utility theory were inadequate accounts of the behavioral data, and argued in favor of the simplest of the remaining three more general models. To reach these conclusions, they used non-Bayesian statistical methods: frequentist point estimation of parameters, bootstrapped confidence intervals of parameters, and null hypothesis significance testing of models. We address the same research goals, based on the same models and the same data, using Bayesian methods. We implement the models as graphical models in JAGS to allow for computational Bayesian analysis. Our results are based on posterior distribution of parameters, posterior predictive checks of descriptive adequacy, and Bayes factors for model comparison. We compare the Bayesian results with those of Birnbaum and Quispe-Torreblanca (2018). We conclude that, while the very general conclusions of the two approaches agree, the Bayesian approach offers better detailed answers, especially for the key question of the evidence the data provide for and against the competing models. Finally, we discuss the conceptual and practical advantages of using Bayesian methods in judgment and decision making research highlighted by this case study.

Keywords: true-and-error theory, expected utility theory, Bayesian methods, graphical models, Bayes factors

1 Introduction

Birnbaum & Quispe-Torreblanca (2018) present an application of true-and-error theory to a simple decision-making task. In the task, a person chooses between two gambles for each of two problems, and has to answer each problem twice. For each individual problem, the choices are identified as safe (S) and risky (R) choices.

To demonstrate true-and-error theory, and to illustrate the use of their computer program, TEMAP2.R, Birnbaum & Quispe-Torreblanca (2018) use data from 107 subjects reported by Birnbaum et al. (2017, Experiment 2, Sample 2). These data are shown in Table 1. The most common responding pattern, produced by 43 subjects, is $RS'RS'$. These subjects chose the risky option for the first problem and the safe option for the second problem, and did this for both replicates of the problems. The second most common responding pattern is $SS'SS'$. These subjects always chose the safe option for both replicates of problems. Over all of the subjects and responses, the risky option is chosen more often (142 times) than the safe option (72 times) for the first problem, but the safe option is chosen more often (177 times) than the risky option (37 times) for the second problem. Over both problems, the safe option is chosen 249 times and the risky option is chosen 179 times.

Table 1: Data from Birnbaum et al. (2017, Experiment 2, Sample 2).

First	Second Replicate			
Replicate	RR'	RS'	SR'	SS'
RR'	4	8	2	0
RS'	4	43	2	8
SR'	1	0	2	4
SS'	1	10	0	18

Figure 1 shows the most general true-and-error theory model used by Birnbaum & Quispe-Torreblanca (2018). The basic assumption is that there is a probability a person is in a risky state or a safe state for the first problem and in a risky state or a safe state for the second problem. These probabilities are formalized by response-state parameters $p_{RR'}$, $p_{RS'}$, $p_{SR'}$, and $p_{SS'}$, and are assumed to be the same for both the first and second replicate of each problem.

The choices made by the person are based on their state, but also depend on response error probabilities. In the most general model shown in Figure 1, there are separate response error probabilities for both states of both problems. Thus, for example, if a person is in the risky state for the first problem (R), they choose the risky option R with (high) probability $1-e$, and the safe option S with (low) probability e. Similarly, if a person is in the safe state for the second problem (S'),

I am very grateful to Michael Birnbaum and Edika Quispe-Torreblanca for their constructive comments, and for providing their code, data, and modeling results. I also thank Percy Mistry and Marc Jekel for useful comments. An Open Science Framework project page for this article at https://osf.io/n6z97/ contains supplementary material, including the complete JAGS graphical model scripts.

*Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, 92697-5100. Email: mdlee@uci.edu.

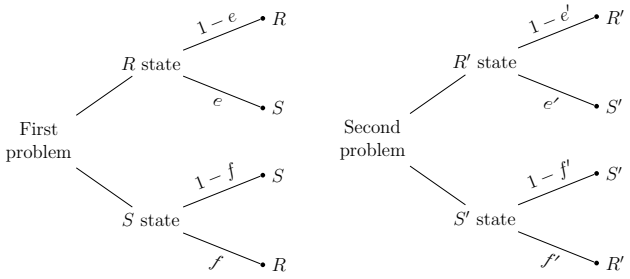

FIGURE 1: General framework for true-and-error models of two binary choice problems. For each problem, the decision maker can be in a risky state or a safe state. The response they generate follows this state according to a response error probability.

they choose the safe option S' with (high) probability $1 - f'$, and the risky option R' with (low) probability f.

Birnbaum & Quispe-Torreblanca (2018) call the model in Figure 1 TE-4, corresponding to the true-and-error theory model with four response error parameters. They consider an additional five models that correspond to special cases of the general TE-4 model. The TE-2 model assumes that the response error probabilities remain dependent on the problem, but are no longer dependent on the response state. This theoretical assumption corresponds to the statistical restrictions $e = f$ and $e' = f'$. The TE-1 model assumes that response error probabilities are the same for both problems and states, so that $e = f = e' = f'$.

The other three models considered by Birnbaum & Quispe-Torreblanca (2018) formalize the key assumption of expected utility theory that the response state is the same for both problems. This theoretical assumption corresponds to the statistical restrictions that $p_{RS'} = 0$ and $p_{SR'} = 0$. Imposing just these restrictions, but allowing four response error probabilities is called the EU-4 model. Following the construction of the TE-2 and TE-1 models, the additional restrictions $e = f$ and $e' = f'$ applied to the EU-4 model produces the EU-2 model, and assuming a single response-error probability $e = f = e' = f'$ produces the EU-1 model.

Given the two major theories, the six specific models, and the behavioral data, there is a sequence of obvious research questions. What evidence do the data provide for and against the competing theories and models? How well do the theories and mdoels fare in describing the observed patterns of decisions? If there are models that do describe the data well, what can we learn about the underlying decision processes involved from inferences about their response state and response error parameters?

Birnbaum & Quispe-Torreblanca (2018) demonstrate their TEMAP2.R software in addressing these sorts of research questions. The statistical methods used are non-Bayesian: frequentist point estimation of parameters, bootstrapped confidence intervals of parameters, and null hypothesis significance testing (NHST) of models. In the last decade or so, however, there has been an increase in the use of Bayesian statistical methods for relating cognitive models to data. Bayesian methods are emphasized in textbooks (e.g., Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2013), edited volumes summarizing the field (e.g., Wixted & Wagenmakers, 2018), and the journal *Psychonomic Bulletin & Review* recently published an extensive special issue entirely devoted to Bayesian methods (Vandekerckhove et al., 2018)

Accordingly, in this article we develop and demonstrate an alternative Bayesian approach to answering the same research questions addressed by Birnbaum & Quispe-Torreblanca (2018). We first present the Bayesian approach, and then compare its results to the ones they presented. We conclude with a discussion of the merits of the two approaches, and draw conclusions about their usefulness in cognitive modeling more generally.

2 Bayesian Analysis

2.1 Model Specification

Specifying the TE and EU models in Bayesian terms requires placing priors on the model parameters. Given the nature of the parameters as probabilities for determining response states and response errors, and the theoretically meaningful bounds of the parameters, it seems reasonable to assume uniform priors based on these bounds. Accordingly, we use

$$\begin{aligned} p_{RR'}, p_{RS'}, p_{SR'}, p_{SS'} &\sim \text{Uniform}(0, 1) \\ e, f, e', f' &\sim \text{Uniform}(0, \frac{1}{2}) \end{aligned} \qquad \text{(M1)}$$

for the TE-4 model, and place the appropriate equality and zero restrictions for the other models. For example, the EU-2 model has the restrictions $e = f$, $e' = f'$, $p_{RS'} = 0$, and $p_{SR'} = 0$. The labeling of Equation M1 reflects that it is a modeling assumption.

The Bayesian approach to defining the likelihood is the same as that used by Birnbaum & Quispe-Torreblanca (2018). For the ith response pattern, a probability θ_i of a subject producing that pattern for a given model and parameterization follows from the decision trees in Figure 1. For example, the response pattern $RR'RR'$, in which the subject chooses the risky option for both gambles both times, could be generated from them being in the risky response state for both problems, and executing their decisions without response error. The probability of this is $p_{RR'}\left[(1 - e)(1 - e')\right]^2$. The same response pattern could also be generated from the other response states, with various response errors. Following this logic, the response

probabilities for all of the possible choice patterns are:

$$\begin{aligned}
\theta_1 &= p_{RR'}\left[(1-e)(1-e')\right]^2 + p_{RS'}\left[(1-e)f'\right]^2 + \\
&\quad p_{SR'}\left[f(1-e')\right]^2 + p_{SS'}\left[ff'\right]^2 \\
\theta_2 &= p_{RR'}(1-e)^2 e'(1-e') + p_{RS'}(1-e)^2 f'(1-f') + \\
&\quad p_{SR'}e'(1-e')f^2 + p_{SS'}f^2 f'(1-f') \\
&\ldots \\
\theta_{16} &= p_{RR'}\left[ee'\right]^2 + p_{RS'}\left[e(1-f')\right]^2 + p_{SR'}\left[(1-f)e'\right]^2 + \\
&\quad p_{SS'}\left[(1-f)(1-f')\right]^2 . \qquad \text{(M2)}
\end{aligned}$$

The observed data, given by the count $\mathbf{y} = (y_1, \ldots, y_{16})$ of how many of n subjects produced each of the 16 response patterns thus has likelihood

$$p(\mathbf{y} \mid p_{RR'}, p_{RS'}, p_{SR'}, p_{SS'}, e, f, e', f', \mathcal{M}) = \prod_{i=1}^{16} \theta_i^{y_i}, \qquad \text{(M3)}$$

where $\mathcal{M}$ denotes the model. This can also be written as $\mathbf{y} \sim \text{Multinomial}(\theta, n)$, where $\theta = (\theta_1, \ldots, \theta_{16})$.

2.2 Bayesian Inference

In the Bayesian framework, it is conceptually helpful to think of the model as a data-generating process. The likelihood $p(\mathbf{y} \mid \theta, \mathcal{M})$ given by the model assumptions in Equations M2 and M3 measures how likely the data are under a specific parameterization of the model. The prior $p(\theta, \mathcal{M})$ given by the modeling assumption in Equation M1 measures how probable each specific parameterization is, according to the theory being formalized by the model, before data have been seen. As emphasized by Wagenmakers et al. (2016), the predictions in Equation M3 form the basis for inference about parameters. Bayes rule provides the logical way, based on probability theory, to transform the predictions about how likely data are given specific parameters into the inference of how likely those specific parameters are, given the data that were observed. This inference about parameters is expressed by the posterior distribution

$$p(\theta \mid \mathbf{y}, \mathcal{M}) = \frac{p(\mathbf{y} \mid \theta, \mathcal{M})\, p(\theta, \mathcal{M})}{p(\mathbf{y})}. \qquad \text{(B1)}$$

The posterior distribution can be interpreted as updating the prior distribution to give more probability to those parameter values that predicted the data, and less probability to those parameter values that did not predict the data. The labeling of Equation B1 reflects that it is required by the adherence of the Bayesian framework to probability theory, and applies for any modeling assumptions about the likelihood and prior.

2.3 Posterior Distributions of Parameters

To infer posterior distributions, we used JAGS (Plummer, 2003), which is standard and free software. JAGS allows probabilistic generative models to be defined in a simple scripting language, and then automatically applies computational methods to sample from the joint posterior distribution. For an introduction to Bayesian graphical models using JAGS aimed at cognitive scientists, see Lee & Wagenmakers (2013).

We implemented each of the six models separately in JAGS. The following script is an excerpt from the full JAGS script for the EU-2 model. The various response-state probability parameters are in the matrix variable `p` and the response-error parameters are in the matrix variable `e`. The definitions of the `theta` variables are produced by separate code that simply enumerates the likelihood for the full TE-4 model. The reduction to the EU2 model is done by the equality and zero constraints in the definition of the priors. Notice that the script also collects samples from the posterior predictive distribution in the variable `yPred`. The scripts for the other five models are constructed similarly, and are provided in the on-line supplementary material.

```
# EU2
model{

# Data
y ~ dmulti(theta, nSubjects)
yPred ~ dmulti(theta, nSubjects)

# Priors
for (i in 1:4) { eTmp[i] ~ dunif(0,0.5) }
e[1,1] = eTmp[1]   # R state,problem 1  (i.e.,e)
e[2,1] = eTmp[1]   # S state,problem 1  (i.e.,f)
e[1,2] = eTmp[2]   # R' state,problem 2 (i.e.,e')
e[2,2] = eTmp[2]   # S' state,problem 2 (i.e.,f')

pTmp ~ dbeta(1,1)
p[1,1] = pTmp      # p_RR'
p[1,2] = 0         # p_RS'
p[2,1] = 0         # p_SR'
p[2,2] = 1-pTmp    # p_SS'

# Likelihood (auto-generated)
# RR'RR'
theta[ 1] =
  p[1,1]*(1-e[1,1])*(1-e[1,2])*(1-e[1,1])*(1-e[1,2])
+ p[1,2]*(1-e[1,1])*e[2,2]*(1-e[1,1])*e[2,2]
+ p[2,1]*e[2,1]*(1-e[1,2])*e[2,1]*(1-e[1,2])
+ p[2,2]*e[2,1]*e[2,2]*e[2,1]*e[2,2]
# RR'RS'
theta[ 2] =
  p[1,1]*(1-e[1,1])*(1-e[1,2])*(1-e[1,1])*e[1,2]
+ p[1,2]*(1-e[1,1])*e[2,2]*(1-e[1,1])*(1-e[2,2])
+ p[2,1]*e[2,1]*(1-e[1,2])*e[2,1]*e[1,2]
+ p[2,2]*e[2,1]*e[2,2]*e[2,1]*(1-e[2,2])
```

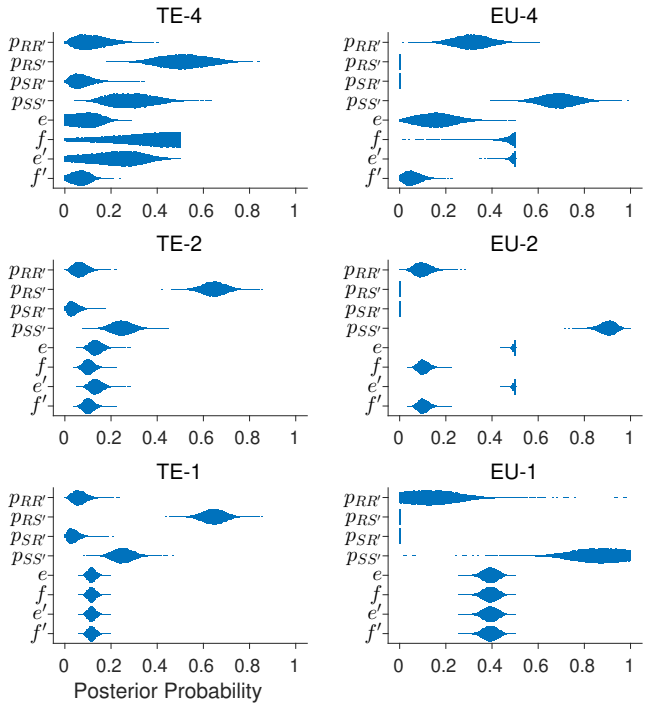

FIGURE 2: Posterior distributions of model parameters. Each panel corresponds to a model, with the parameters on the y-axis. The violin plots show the posterior distribution of each parameter.

```
# theta[ 3] to theta[15] removed for brevity

# SS'SS'
theta[16] =
  p[1,1]*e[1,1]*e[1,2]*e[1,1]*e[1,2]
+ p[1,2]*e[1,1]*(1-e[2,2])*e[1,1]*(1-e[2,2])
+ p[2,1]*(1-e[2,1])*e[1,2]*(1-e[2,1])*e[1,2]
+ p[2,2]*(1-e[2,1])*(1-e[2,2])*(1-e[2,1])*
  (1-e[2,2])
}
```

Figure 2 shows the marginal posterior distribution of every model parameter for every model.[1] These posterior distributions quantify the relative probability that each value is the true value of the parameter, given the assumptions of the model, and the information provided by the data. The constraints on the parameters that define the various models are clear. It is interesting to note that the progression from the TE-4 model to the TE-2 model to the TE-1 model, as parameter constraints are added, does not lead to qualitatively different inferences. The probability of the $p_{RS'}$ response state is always high, followed by the $p_{SS'}$ state, and there is little probability given to the other two states. The certainty of these inferences, however, does improve as the models become more constrained. The same is not true for the progression of EU models. The inferred values for response state and response error parameters are often different between the EU models. For example, the 95% credible intervals, based on the 2.5% and 97.5% percentiles, for the $p_{RR'}$ response state parameter are $[0.17, 0.46]$, $[0.04, 0.17]$, and $[0.01, 0.37]$, under the EU-4, EU-2, and EU-1 models, respectively.

2.4 Posterior Predictive Analysis

The posterior distributions in Figure 2 quantify the relative probability that each value is the true value of the parameter, given the assumptions of the model. Conditioning on the modeling assumptions means that if they are inappropriate the posterior distributions are not useful. A standard Bayesian method for evaluating the adequacy of modeling assumptions is posterior predictive checking (Gelman et al., 2004; Shiffrin et al., 2008). The posterior predictive distribution is the distribution of data generated by a model, based on the posterior distribution of parameters found by conditioning on observed data.

Figure 3 shows a posterior predictive analysis for each of the models. The violin plots show the predictive mass each model gives to how many subjects show each of the 16 possible data patterns. The observed numbers of subjects who produced each pattern are shown by square markers. To the extent that the predictive distribution gives large mass to the observed data, the model is able to "fit" the data. By this standard, Figure 3 shows that all of the TE models pass the basic test of descriptive adequacy. The observed number of subjects producing each response pattern is given significant mass in the posterior predictive distribution. All of the EU models, however, fail to pass the test of descriptive adequacy. The most common response pattern $RS'RS'$ is not one easily produced by the EU models, since the change in response states that would most easily produce the pattern is not permitted by the theory. This incompatibility is visually evident from the posterior predictive distributions for the response patterns. Figure 3 shows a number of other cases of EU models not giving large posterior predictive mass to the data. The basic conclusions is that the TE models are descriptively adequate while the EU models are not.

It is important to understand that posterior predictive checking does *not* involve evaluating the ability of a model to predict data. This is because the posterior predictive distribution is not a genuine prediction.[2] Genuine predictions are made *before* data are observed. The posterior predictive analysis tests the ability of a model to re-describe data

[1] Technical details: the results are are based on applying each model to the data by collecting 5000 samples from the joint posterior from each of 6 independent chains, after a burn-in period of 5000 discarded samples per chain, and with a thinning factor of 10. The convergence of the chains was assessed using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997) and by visual inspection.

[2] The word "predictive" in "posterior predictive" comes from statistics, where it essentially means "over the data space".

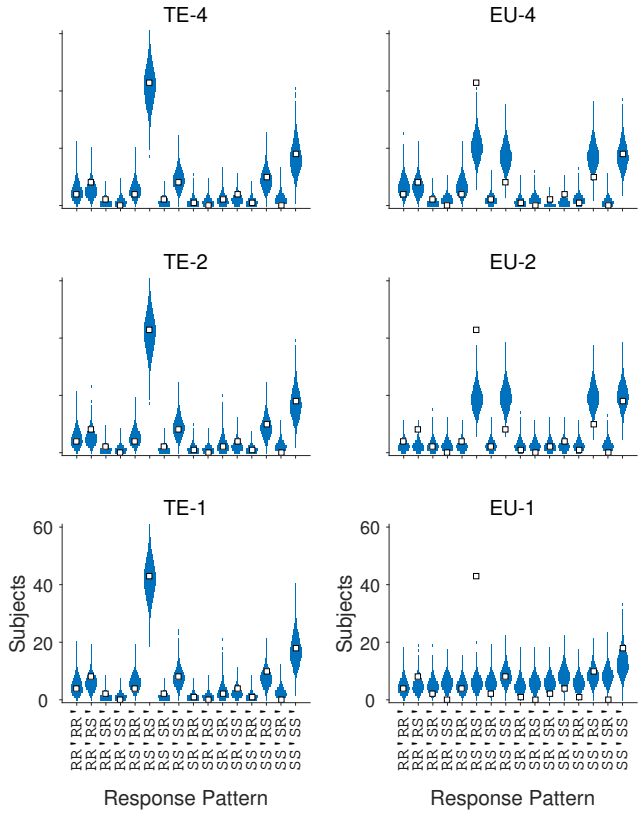

FIGURE 3: Posterior predictive assessment of descriptive adequacy. Each panel corresponds to a model, with the possible response patterns on the x-axis, and the count of the number of subjects showing that pattern on the y-axis. The violin plots show the posterior predictive distributions for the model for each response pattern. The numbers of subjects producing that pattern are shown by square markers.

that have already been seen and incorporated into the posterior distribution. Thus, posterior predictive analysis is best understood as an assessment of descriptive adequacy. It provides a mechanism for assessing whether it is reasonable to interpret posterior distributions for parameters, but not for assessing the predictive adequacy of models.

2.5 Model Comparison

In the Bayesian framework, the comparison and selection of models is based on testing the relative accuracy of their predictions. This involves their prior predictive distributions, which is the analogue to the posterior predictive distribution, but based on the prior assumptions about parameters specified by the models. The prior predictive distribution measures the average likelihood of the data under the model. Formally, this requires integrating the likelihood of the data over all of the parameterizations of the model, weighted by how likely each parameterizations is, as formalized by the prior:

$$p(\mathbf{y} \mid \mathcal{M}) = \int p(\mathbf{y} \mid \theta, \mathcal{M})\, p(\theta, \mathcal{M})\ \mathrm{d}\theta. \tag{B2}$$

The resulting marginal probability $p(\mathbf{y} \mid \mathcal{M})$ can be interpreted as how likely the data $\mathbf{y}$ are to be predicted by model $\mathcal{M}$. It is then natural to compare the predictions of two or more competing models. This ratio

$$\frac{p(\mathbf{y} \mid \mathcal{M}_a)}{p(\mathbf{y} \mid \mathcal{M}_b)} \tag{B3}$$

for two models $\mathcal{M}_a$ and $\mathcal{M}_b$ is called the Bayes factor, and is a standard Bayesian measure for comparing models (Kass & Raftery, 1995; Lee & Wagenmakers, 2013, Ch. 7). Because it is an odds ratio, the Bayes factor has a natural scale for interpretation of significance, calibrated by betting. A Bayes factor of 10, for example, means that the data are 10 times better predicted by one model than the other.

A complementary interpretation of the Bayes factor is as the evidence that updates prior knowledge about models to posterior knowledge. Formally, this updating is given by

$$\overbrace{\frac{p(\mathcal{M}_a \mid \mathbf{y})}{p(\mathcal{M}_b \mid \mathbf{y})}}^{\text{posterior odds}} = \overbrace{\frac{p(\mathbf{y} \mid \mathcal{M}_a)}{p(\mathbf{y} \mid \mathcal{M}_b)}}^{\text{Bayes factor}} \overbrace{\frac{p(\mathcal{M}_a)}{p(\mathcal{M}_b)}}^{\text{prior odds}}. \tag{B4}$$

and is conceptually analogous to the updating of knowledge about parameters in Equation B1 that defined the posterior distribution. In both equations, priors knowledge about parameters or models is updated by data, according to how well the parameter or model predicted the data, to give posterior knowledge about parameters or models. Under this interpretation, a Bayes factor of 10 means that that the data provide 10 times more evidence for one model than the other.

To estimate the Bayes factors between the six models, we used a standard latent-mixture approach — also known as the product-space method — based on the inference of posterior model probabilities (Lee, 2016; Lodewyckx et al., 2011). The approach uses a single discrete parameter z that indexes each of the six models, and controls which model is assumed to generate the data. Formally

$$\mathbf{y} \sim \text{Multinomial}\big(\theta_z, n\big) \tag{M4}$$

where $\theta_z = (\theta_{1z}, \ldots, \theta_{16z})$ are the probabilities generated by the zth model for each of the possible response patterns. We place a simple prior on z

$$z \sim \text{Categorical}\big(\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}\big) \tag{M5}$$

that makes each of the six models equally likely a priori. The posterior distribution $p(z \mid \mathbf{y})$ for the model index parameter thus estimates the posterior probabilities of the models. We

again implemented the latent-mixture model in JAGS. This requires only one script, an excerpt of which focusing on the additional code follows. The key is that the data are now distributed `y~dmulti(theta[:, z], nSubjects)`, with the predictions of each model calculated as a column in the matrix `theta`.

```
Latent-mixture model selection
model{

  # Data
  y ~ dmulti(theta[:, z], nSubjects)

  # Model indicator prior
  z ~ dcat(phi)
  for (modelIdx in 1:nModels){
    phi[modelIdx] = 1
  }

  # TE4 MODEL

  # Priors
   for (i in 1:4) { eTmp[i,1] ~ dunif(0,0.5) }
   e[1,1, ] = eTmp[1,1]  # R state, problem 1  (i.e., e)
   e[2,1,1] = eTmp[2,1]  # S state, problem 1  (i.e., f)
   e[1,2,1] = eTmp[3,1]  # R' state, problem 2 (i.e., e')
   e[2,2,1] = eTmp[4,1]  # S' state, problem 2 (i.e., f')

   pTmp1 ~ ddirch(c(1,1,1, ))
   p[1,1,1] = pTmp1[1]   # p_RR'
   p[1,2,1] = pTmp1[2]   # p_RS'
   p[2,1,1] = pTmp1[3]   # p_SR'
   p[2,2,1] = pTmp1[4]   # p_SS'

 # RR'RR'
 theta[1,1] =
   p[1,1,1]*(1-e[1,1,1])*(1-e[1,2,1])
   *(1-e[1,1,1])*(1-e[1,2,1])
 + p[1,2,1]*(1-e[1,1,1])*e[2,2,1]*(1-e[1,1,1])*e[2,2,1]
 + p[2,1,1]*e[2,1,1]*(1-e[1,2,1])*e[2,1,1]*(1-e[1,2,1])
 + p[2,2,1]*e[2,1,1]*e[2,2,1]*e[2,1,1]*e[2,2,1]

 # theta[2,1] to theta[16,1] removed for brevity

 # EU4, TE2, EU2, TE1, EU1 models removed for brevity

}
```

Figure 4 shows the results of the latent-mixture analysis, giving the posterior probabilities for each of the six models.[3] Only the three TE models have non-negligible posterior probability. Since the prior probabilities of each model that led to the estimated posterior probabilities are given by the modeling assumptions in Equation M5, it is straightforward to estimate Bayes factors between any pair of models. In this case, given the choice of equal prior probabilities,

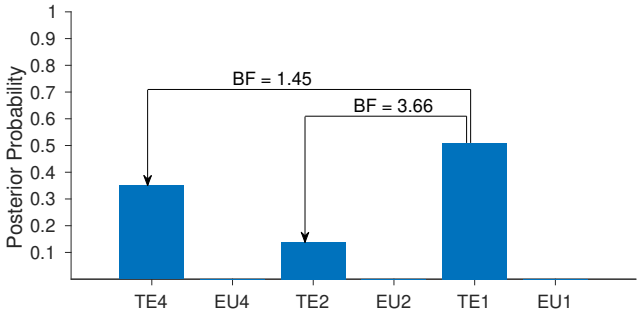

FIGURE 4: Posterior model probabilities for each of the six models, found using the latent-mixture approach. The labelled arrows show the resulting Bayes factors between the TE-1 model, and TE-4 and TE-2 models.

the Bayes factors can be directly measured from the posterior probabilities. For example, the posterior probability of TE-1 is about 0.51 and the posterior probability of TE-4 is about 0.35, so the Bayes factor for TE-1 over TE-4 is about $0.51/0.35 \approx 1.45$. Meanwhile, the Bayes factor in favor of TE-1 over TE-2 is about 3.66. The Bayes factor between any pair of models with non-negligible posterior probability can be found in the same way.[4]

These Bayes factors measure the evidence the data provide for each model in a way that automatically combines goodness-of-fit with a complete and principled measure of the statistical complexity of the models (Myung et al., 2000; Pitt et al., 2002). The basic conclusion is that the TE-1 model receives the most evidence from the data, followed by the TE-4 and TE-2 models. This is an interesting pattern of results, showing a non-obvious trade-off between the goodness-of-fit and complexity of three nested and descriptively adequate models. The Bayes factors, however, are not large, and the evidence in favor of one TE model over the other is extremely weak. In the language used by Jeffreys (1961), the Bayes factors for and against the three TE models shown in Figure 4 are "not worth more than a bare mention." The scientific conclusion we reach is that additional evidence is needed to make decisions about the relative merits of the three TE models. The Bayes factors quantify the lack of discriminating evidence provided by the current data.

2.6 Summary of Bayesian Analysis

The basic research questions outlined in introducing the models and data are naturally and directly addressed by the Bayesian analysis. The evidence the data provide for and

[3]Technical details: the latent-mixture model was applied to the data by collecting 10,000 samples from the joint posterior from each of 6 independent chains, after a burn-in period of 10,000 discarded samples per chain, and with a thinning factor of 50. This more conservative sampling strategy was chosen because latent-mixture models often require additional samples and thinning for convergence. We again used $\hat{R}$ statistic and visual inspection to check convergence.

[4]It would also be possible to estimate Bayes factors involving the EU models by setting priors favoring them in the latent-mixture model, and accounting for the priors in the Bayes factor calculation (Lodewyckx et al., 2011). The result, however, would obviously be overwhelming evidence against these models, and it is not clear what the utility in precise quantification of this overwhelming evidence would be.

against the various models are quantified by Bayes factors, as shown in Figure 4. The descriptive adequacy of models can be assessed from their posterior predictive agreement with the data, as shown in Figure 3. The inferences about latent parameters, for models favored by the data and having descriptive adequacy, are quantified by posterior distributions, as shown in Figure 2.

3 Comparison to Non-Bayesian Analysis

3.1 Parameter Point Estimates

For each of the models, Birnbaum & Quispe-Torreblanca (2018) find point estimates of the free parameters that minimize either a χ^2 or G^2 measure of agreement between the model predictions and the data. Any specific combination of values of response state and response error parameters produces response probabilities $\theta_1, \ldots, \theta_{16}$ following Equation M2. These probabilities, in turn, corresponds to predictions of $\hat{y}_i = n\theta_i$ of the n subjects showing that pattern. Given these prediction, Birnbaum & Quispe-Torreblanca (2018) find the combination of parameters, subject to the equality and zero constraints appropriate for the model, that minimize either

$$\chi^2 = \sum_{i=1}^{16} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i},$$

or

$$G^2 = 2 \sum_{i=1}^{16} y_i \ln \frac{y_i}{\hat{y}_i}.$$

Figure 5 shows the point estimates of parameters found this way, in relation to the Bayesian posterior distributions. It is clear that the point estimates for the same parameter can be meaningfully different depending on the optimization criterion used. This is especially the case for the TE-4 and EU-1 models. It is also clear that, while the point estimates always fall in a region of non-negligible posterior density, they do not always correspond to standard point summaries of the posteriors. Typically point summaries use the mode of the posterior distribution, corresponding to the maximum a posteriori value (optimizing zero-one loss), or the mean of the posterior distribution, corresponding to the expectation (optimizing quadratic loss). Many of the point estimates based on χ^2 or G^2 in Figure 5, again especially for the TE-4 and EU-1 models, lie in the tails of the posteriors and differ significantly from the mode and mean.

Collectively, these observations show that the conclusions about the underlying psychological variables represented by the parameters can depend on the choice of optimization criterion, and can differ from those found by Bayesian methods. These discrepancies are especially relevant for the TE-4 model, which provides a descriptively adequate account of the behavioral data, and so could be argued to be a reasonable model on which to base inferences about parameters.

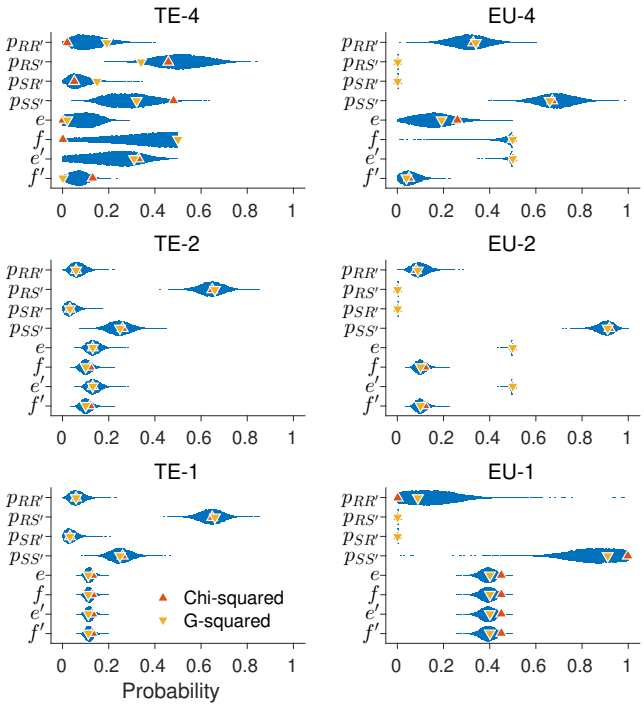

FIGURE 5: Comparison of Bayesian posterior distributions and frequentist point estimates. Each panel corresponds to a model, with the response preference and response error probability parameters on the y-axis and their values on the x-axis. The violin plots show the posterior distribution of each parameter. The point estimates found using χ^2 and G^2 optimization are shown by upward and downward triangle markers, respectively.

3.2 Parameter Distributions

To quantify the uncertainty about parameter estimates, Birnbaum & Quispe-Torreblanca (2018) use a bootstrapping approach that generates distributions of parameters. This standard procedure involves generating alternative data sets based on the observed data, and estimating parameters for each of these newly generated data sets (Efron & Tibshirani, 1986). The distribution of these estimates is then used as a quantification of uncertainty.

Figures 6 and 7 compare bootstrap distributions based, respectively, on the χ^2 and G^2 criterion with the Bayesian posterior distributions. The upper-half of each violin plot corresponds to the bootstrap distribution, and the lower-half to the Bayesian posterior. Thus, the distributions are the same to the extent they have mirror symmetry. There appear to be at least three general ways in which the bootstrap dis-

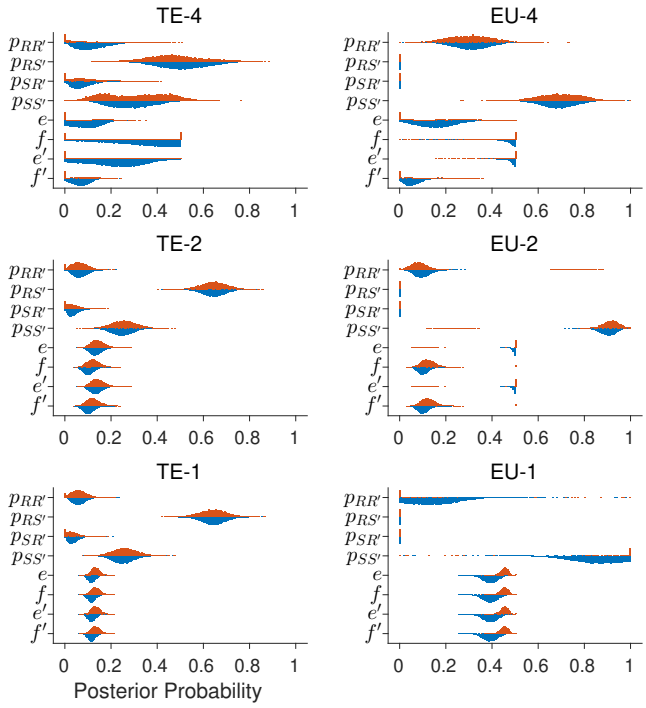

FIGURE 6: Comparison of Bayesian posterior distributions and bootstrap distribution based on the χ^2 criterion. Each panel corresponds to a model, with the response preference and response error probability parameters on the y-axis and their values on the x-axis. The lower-half of each violin plot shows the posterior distribution from the Bayesian analysis while the upper-half shows the bootstrap distribution of the parameter.

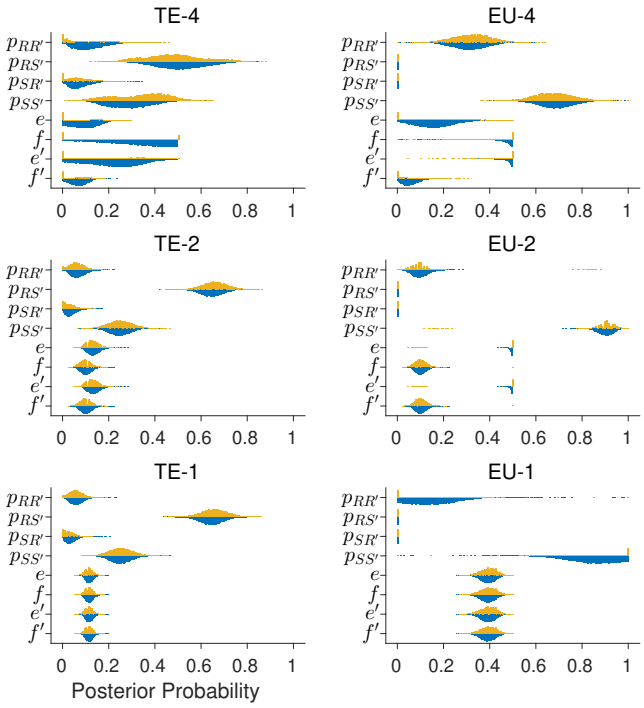

FIGURE 7: Comparison of Bayesian posterior distributions and bootstrap distribution based on the G^2 criterion. Each panel corresponds to a model, with the response preference and response error probability parameters on the y-axis and their values on the x-axis. The lower-half of each violin plot shows the posterior distribution from the Bayesian analysis while the upper-half shows the bootstrap distribution of the parameter.

tributions differ from the posterior distributions. First, the bootstrap and posterior distributions sometimes give probability to different non-overlapping ranges of parameter values. Examples include the response error parameters under the EU-1 model and the χ^2 criterion. Secondly, sometimes the ranges are consistent, but the relative probabilities across the range are not. Examples include the response state parameters under the TE-4 model using both the χ^2 and G^2 criteria. The difference for $p_{SS'}$ response state parameter of the TE-4 model and the χ^2 criterion is especially striking, with a multi-modal bootstrap distribution. Thirdly, sometimes almost all of the bootstrap distribution collapses against a bound for a parameter. This occurs for many parameters for many models with the χ^2 criterion, and occasionally for the G^2 criterion. Clear examples are the response error parameters for the TE-4 model using the χ^2 criterion.

As was the case for the comparison of point estimates of parameters, the basic conclusion is that the bootstrap distribution of parameter values, and the summaries like 95% confidence intervals they generate, depend on the criterion used, and do not always agree with the Bayesian posterior distribution. There is additional evidence that the bootstrap distributions sometimes collapse to a point in the parameter space, rather than representing a distribution of possible values that are consistent with the observed data.

3.3 Model Adequacy

Figure 8 compares the assessment of descriptive adequacy made by the Bayesian and frequentist approaches. The posterior predictive distribution for each model and data pattern from Figure 3 are shown again. The frequentist best-fit predictions, of the type detailed in Table 4 and Table 5 of Birnbaum & Quispe-Torreblanca (2018), are shown as triangular markers.[5] It is clear the descriptions of the data

[5]We note that Birnbaum & Quispe-Torreblanca (2018) encounter similar terminological problems in discussing the assessment of descriptive adequacy as those caused by the Bayesian term "predictive". For example Birnbaum & Quispe-Torreblanca (2018) say "[t]o gain insight into the performance of a model, it is useful to compare predictions against the empirical data" but, of course, the values in Figure 8 are not predictions in any reasonable scientific or everyday sense of the word. They are based on the observed data, and so are not genuine predictions about the data. Pre-

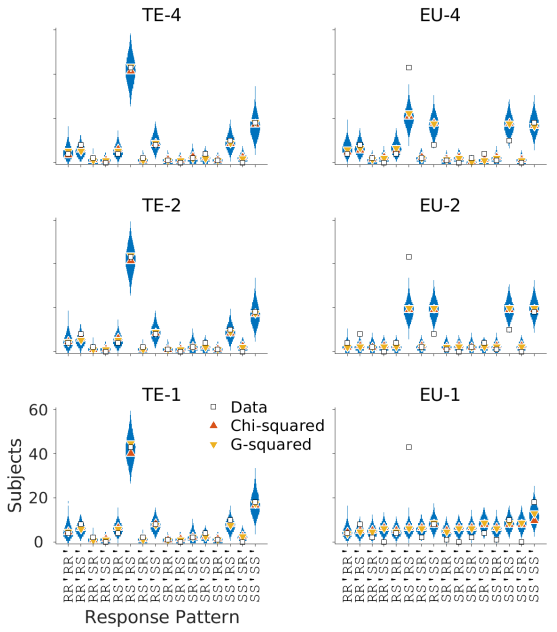

FIGURE 8: Comparison of Bayesian posterior predictive distribution and frequentist point estimates. Each panel corresponds to a model, with the possible response patterns on the x-axis, and the count of the number of subjects showing that pattern on the y-axis. The violin plots show the posterior predictive distributions for the model for each response pattern. The numbers of subjects producing that pattern are shown by square markers. The point estimates found using χ^2 and G^2 optimization are shown, respectively, by upward and downward triangle markers.

provided by both χ^2 and G^2 optimization agree closely with each other, and are completely consistent with the Bayesian posterior predictive distributions.

Birnbaum & Quispe-Torreblanca (2018) also reach the same substantive conclusions about the descriptive adequacy of each model. That is, all three TE model are adequate, but none of the three EU models are. Part of this assessment is based on more formal measures than were used in the Bayesian analysis, using null hypothesis significance testing (NHST) to find that the TE models have p values greater than 0.1, but the EU models have p values less than 0.01. The logic is these p values demonstrate a significant difference between the EU models and the data, but such a significant difference fails to be found for the TE models so they "fit acceptably, by conventional standards."

sumably the phrase "best-fit predictions" captures the important distinction between between genuine prediction and the fitted description of observed data.

3.4 Model Comparison

Birnbaum & Quispe-Torreblanca (2018) use two approaches to compare the models to each other. One approach is based on the difference in the χ^2 distributions measuring the goodness-of-fit to the data for pairs of models. If there are significant differences, following NHST logic, then one model is judged to be better than the other. This approach uses both asymptotic values of the test statistic and Monte Carlo values based on re-fit methods. Using this approach, the basic conclusion is that all of the TE models are significantly better than EU models, and that there are no significant differences between the TE models themselves.

The other approach to model comparison used by Birnbaum & Quispe-Torreblanca (2018) is based on examining the measures of absolute goodness-of-fit provided by χ^2 and G^2 statistics together with the numbers of parameters or degrees of freedom of the models. The logic is that better fitting models should be preferred, especially if they have fewer parameters than another model. The basic conclusion from this approach is that the TE-1 model is to be preferred. Specifically, Birnbaum & Quispe-Torreblanca (2018) conclude: "TE-4 does not fit much better than TE-2 or TE-1 . . . so there are no reasons to reject TE-1 in favor of TE-2 or TE-4."

These findings are consistent with those reached by the Bayesian approach. There does not seem to be any direct equivalent of Bayes factors, providing a quantitative measure of how much evidence the data provide for one model over another. Birnbaum & Quispe-Torreblanca (2018) make more qualitative claims, involving deciding whether or not one model is significantly better than another. These could be quantified in terms of the strength of rejection of null hypotheses, but not in terms of relative evidence for the models being compared.

4 Discussion

At a very general level of comparison, the Bayesian analysis developed here and the non-Bayesian analysis presented by Birnbaum & Quispe-Torreblanca (2018) reach the same conclusions. The TE models are found to be clearly superior to the EU models, because the TE models are able to describe the data whereas the EU models are not. To the extent that a single model is favored, the tentative conclusion in both cases is that this is the TE-1 model, but there is little basis to choose one TE model over another. The values of the response state and response error parameters are generally inferred to have similar values under both methods.

Once the comparison becomes more detailed, however, differences between the two analyses emerge. The frequentist point estimates, which differ depending on whether χ^2 of G^2 is used, do not precisely correspond to standard point

summaries of the Bayesian posterior distributions. The bootstrap distributions also differ according to whether χ^2 of G^2 is used, and often deviate from the the posterior distributions. The model comparison conclusions based on NHST and bootstrap differences in test statistics reach only binary conclusions about which of a pair of models is preferred given the data. They not quantify the level of evidence or confidence in the differences between models in the way that Bayes factors do.

We begin our discussion by examining the reasons for each these differences in parameter estimation and model selection. We then address some standard criticisms of the Bayesian approach, before concluding by examining the prospects of dealing with extensions of the current models.

4.1 Parameter Estimates

The reason the point estimates of parameters do not match the Bayesian posterior distributions involves the optimization criteria used. Both the χ^2 and G^2 measures differ from the likelihood in Equation M3 that is used by the Bayesian approach. It is common to view both the χ^2 and G^2 measures measures as approximations to the multinomial (McDonald, 2009). As approximations, they have limitations and possible undesirable behavior, especially in relation to small or extreme data samples (Jaynes, 2003, Section 9.12, provides a good example of the limitations of χ^2). Given the availability of modern computing capabilities to use the multinomial likelihood, it is not clear what the rationale for using approximations might be. If there is some aspect of the psychological model that differs from the multinomial, that should be built into the model, not introduced implicitly through using an approximate optimization criterion.

Beyond point estimates, it seems clear that there is a need to represent uncertainty associated with the model parameters. For example, the TE-4 model could be viewed a serious theoretical competitor, capable of describing the data, and not obviously out-performed by any other model. But, as the Bayesian analysis resulting in the posteriors in Figure 2 shows, there is large uncertainty in key parameters inferred from the data using this model. For example, the $p_{RS'}$ response state probability could be anywhere between about 0.4 and 0.8. There is no way to summarize this information as a point estimate, without the summarizing process losing valuable information. A fundamental advantage of the Bayesian approach is that it is predicated on the representation and updating of uncertainty about parameters and models at all stages and in all aspects of statistical analysis.

Birnbaum & Quispe-Torreblanca (2018) address the small-sample limitations of the χ^2 and G^2 measures, and the need to represent uncertainty, using bootstrap methods. Bootstrapping is a standard frequentist technique (Efron & Tibshirani, 1986), but involves generating new data sets based on the single data set actually observed, and basing inference on those generated data. The results in Figures 6 and 7 suggest that the data set generating methods used in bootstrapping sometimes produce answers that are difficult to interpret. In particular, the cases in which parameter distributions collapse to a single value on the edge of parameter space seem inappropriate. As a concrete example, consider the $p_{RR'}$ parameter under the TE-4 model. Using both χ^2 and G^2 measures, the bootstrap inference is that the value is near 0. But, as the Bayesian posterior distribution shows, there are other values of the response state, permitted by the TE-4 theory, that are as consistent or more consistent with the data. Of course, it might be possible to address these issues with further refinement of the data set generation procedure. From a Bayesian perspective, none of this methodological inventiveness is needed. The posterior distributions of parameters represent uncertainty about their values coherently and completely, and follow automatically from the application of Bayes rule. Posterior distributions are validly defined and calculated in the same way for any sample size, without the need to any sort of correction or alternative procedure for small numbers of data.

4.2 Model Comparison

Birnbaum & Quispe-Torreblanca (2018) make some use of p values to compare models. Criticisms of p values and NHST as a method for hypothesis testing are widely documented and understood in psychological data analysis and modeling (Wagenmakers, 2007; Wagenmakers et al., 2016). Perhaps most importantly for the analyses reported by Birnbaum & Quispe-Torreblanca (2018), NHST cannot find evidence in favor of the null hypothesis. Thus, concluding that a model is adequate because it has a p value greater than 0.1 is problematic. This p value simply is not evidence of sameness, even if it is often used that way.

The other approach to model comparison used by Birnbaum & Quispe-Torreblanca (2018) involves measures of goodness-of-fit based on the χ^2 and G^2 criteria, together with counts of the number of free parameters or degrees of freedom in the models. There often seems to be an (implicit) assumption assumed that a parameter count provides an adequate measure of model complexity to counter the improved goodness-of-fit achieved by models with more free parameters. For example, Birnbaum & Quispe-Torreblanca (2018) say "[b]ecause TE-2 and EU-4 have the same number of degrees of freedom, one is tempted to compare these to see if EU with four errors might come off better than a model that has fewer errors but rejects EU" and "[b]ut if EU-4 can be rejected in the context of TE-4 it means that one cannot save the simpler decision model even with the extra parameters". Concretely, the test of significance in the difference in χ^2 values for the TE-4 and EU-4 models is based on two degrees of freedom, arising from the difference in counts of their free parameters.

An emphasis of the modern Bayesian literature on model comparison in psychology is that counting the number of parameters is, at best, a crude approximation to model complexity. At worst, it is entirely misleading (Myung & Pitt, 1997; Pitt et al., 2002). It is possible for two models with the same number of parameters to have very different levels of complexity. It is possible for models with many more parameters to be much simpler than models with fewer parameters. In fact, this happens regularly in hierarchical modeling (Lee & Vanpaemel, 2018). It is possible for the introduction of an additional parameter to create a model that is simpler than the original model, even if the new model reduces to the original at a specific value of the newly-introduced parameter. An example is provided by the introduction of a determinism parameter to Luce choice-response models commonly used in decision models (Lee & Vanpaemel, 2018; Vanpaemel, 2016).

Rather that counting parameters or degrees of freedom, the complexity of a model is best understood as the extent of the predictions it is able to make about data (Myung et al., 2000; Wagenmakers et al., 2016). This notion of statistical complexity depends on the number of parameters, the range of values they can take, and the functional form of their interaction. For the current models, the different ranges of values and the complexity inherent in the interaction of the response state and response error probability parameters is ignored by counting parameters. It is extremely unlikely that each of the response state and response error parameters in the TE and EU model contributes equal complexity. This will depend on the range of values of the parameter can take, and whether other parameters in the model have equality or zero-value constraints.

>From a Bayesian perspective, the right way to choose between models is using the Bayes factor. One way to think of a Bayes factor is as a likelihood ratio extended to integrate over the parameters of the models being compared. Another way is as a test of how relatively likely the observed data are, given the prior predictive distributions of the models. Whatever the conception, the Bayes factor has the statistical property automatically accounts for both goodness-of-fit and all forms of model complexity, and is able to compare any set of probabilistic models (Myung et al., 2000; Wagenmakers et al., 2016). In addition, the Bayes factor provides a meaningful quantitative measure of the evidence that the data provide for and against the models being compared. The Bayes factors in Figure 4 quantify the level of evidence the data provide in favor of the TE-1 model over the TE-2 and TE-4 models.

Birnbaum & Quispe-Torreblanca (2018) do not appear to produce any measure of evidence analogous to the Bayes factor. They report binary significance findings in favor of the TE models over all of the EU models, but not between TE models. How much evidence there is for these conclusions, and how much confidence should be placed in them, is never quantified by the p values or goodness-of-fit and parameter-counting measures. In reaching conclusions about model comparison, Birnbaum & Quispe-Torreblanca (2018) point out that "[o]f course, making a scientific decision to prefer one theory over another should depend on more than just comparing indices of fit and numbers of parameters." We agree completely with this point, but it is not a justification for an incomplete statistical analysis of the evidence. Measures like Bayes factors provide only one source of evidence in making scientific decisions about models, but they are an important source of evidence. The failure of the non-Bayesian methods used by Birnbaum & Quispe-Torreblanca (2018) to provide an analogous non-Bayesian version of this sort of evidence is an important limitation.

4.3 The Nature and Role of Priors

The advantages in Bayesian analysis just described, in terms of posterior distributions for inferences about parameters and Bayes factors for inferences about models, both require the specification of prior distributions. Indeed, it is the existence of these distributions that makes an analysis Bayesian (Lindley, 1972). Prior distributions are also the source of the greatest resistance to Bayesian methods. Thus, part of advocating for the use of Bayesian methods for the current problem requires a discussion the nature and role of priors.

Jaynes (2003) argued that without priors the scientific problem of inference is ill-defined. One cannot know the state of knowledge about parameters and models after data are observed if one did not formalize what was known before the data were observed. We think it is natural to require a model to formalize that initial uncertainty. The prior distribution quantifies theoretical assumptions about the prior plausibility of the psychological variables the parameters represent. We suspect, as argued insightfully by Leamer (1983), that the discomfort in specifying priors arises because they present a new modeling challenge. But the nature of the challenge is identical to the ones always faced in specifying other parts of a model, such as the likelihood. The goal is simply to make good creative judgments about the psychological variables and processes that are being assumed to generate behavior. Assumptions about psychological processes typically are formalized by likelihoods, and priors provide a natural modeling vehicle to formalize statistical assumptions about psychological variables (Lee & Vanpaemel, 2018).

Indeed, theoretically-motivated constraints on parameters are a central part of the model specification of Birnbaum & Quispe-Torreblanca (2018). The defining property of the EU models is that some response state parameters are constrained to be zero, while other response state parameter are free to lie in the interval $(0, 1)$. The response error probabilities to be inferred are assumed to lie in the interval $(0, \frac{1}{2})$, consistent with their interpretation as imperfect executions

of deterministic response processes in binary choice. These constraints on the parameter thus capture meaning, and constrain the predictions about data that the models make. But the constraints and equalities are not sufficient for a model to make complete quantitative predictions about the outcomes of the behavioral experiment, in the sense of predicting the probability for each response pattern. To make these detailed quantitative predictions requires a full prior *distribution* on the parameters.

Other choices of priors in our Bayesian analysis would have been reasonable, and perhaps even preferable. For example, it might be better to assume that response error rates are more likely to be near zero than one-half, since the structure of the model in Figure 1 is nullified as response error rates approach one-half. Different prior assumptions about plausible response rates will lead to different inferences than the ones we report. This is not surprising, and it is desirable. The priors formalize theoretical assumptions and different theories should, in general, yield different conclusions when applied to the same data.

4.4 Extensions of the Models

Finally, we discuss an advantage of the Bayesian approach to analysis that goes beyond the specific models and data considered by Birnbaum & Quispe-Torreblanca (2018). Bayesian methods are well suited to models that have richer structures than simply mapping a set of parameters to data. In particular, Bayesian methods work effectively, in both theory and practice, for models with hierarchical, latent-mixture, and common-cause structures (Lee, 2011, 2018). The motivating ambitions of true-and-error theory, in their individual (*i*TET) and group (*g*TET) forms, seem likely to require some of these sorts of extended modeling structures. Describing the underlying theory, Birnbaum & Quispe-Torreblanca (2018) say "[i]n the case of *i*TET, it is assumed that a mixture of true preference patterns can arise over the course of many sessions because a person may change personal parameters over time between sessions" and "[i]n *g*TET, it is assumed that a mixture of true preference patterns can arise from individual differences among people, who may have different parameters or different decision rules for making the choices."

Both of these theories involve heterogeneity that is not captured by the six specific models considered here and by Birnbaum & Quispe-Torreblanca (2018). They all assume there is one single set of response state and response error parameters that generates the aggregate data in Table 1. There is no allowance for inter-individual or intra-individual differences. Developing models that allow for inter-individual differences, as per the motivation for *g*TET, requires hierarchical (also known as multi-level) model structures, which have been widely used in cognitive modeling to incorporate individual differences (e.g., Farrell & Lewandowsky, 2018; Lee & Newell, 2011; Oravecz et al., 2015; Rouder & Lu, 2005; Rouder et al., 2003; Shiffrin et al., 2008). A central theme in this literature is the conceptual and practical ease with which Bayesian methods extend to the new model structures. All of the principles of inference remain the same, and the additional of the hierarchical structure requires just a few more lines of JAGS code. Meanwhile, developing models that allow for intra-individual switches in preferences, such as through changes to the response state, requires the ability to infer when and where these changes occur. Once again, there is a large Bayesian literature on change-point detection (e.g., Barry & Hartigan, 1993; Chib, 1998; Fearnhead, 2006; Green, 1995; Stephens, 1994), including recent developments aimed at cognitive modeling and implemented in JAGS (Lee, in press).

Testing these extended models probably requires richer behavioral data, so that enough is known about each individual to infer differences between them, and an individual is tracked for long enough to detect changes in preferences. But, given those data, it should be straightforward to implement the appropriate models as graphical models, and continue to use the same Bayesian approach to their analysis demonstrated here. Thus, it seems reasonable to argue that the fuller model-based development of the true-and-error theory could benefit from the use of Bayesian methods.

5 Conclusion

The Bayesian method we presented is conceptually simple, statistically principled, and easily implemented and applied. It forces assumptions about the psychological variables represented by parameters to be made explicit, and so allows the models to make predictions about data before they are observed. It formalizes inferences about parameters and models using posterior distributions and posterior probabilities, updates the predictions about future data based on those inferences, and supports models comparison via Bayes factors. All of these Bayesian inferences are founded on the complete, consistent, and coherent foundations of probability theory.

The approach we have used is standard and general. The same mechanisms, involving the Equations B1, B2, B3, and B4 that defined posterior distributions, posterior predictive evaluation, Bayes factors, and posterior model probabilities, and can be used in the same way to apply and evaluate any probabilistic model of cognition. There is nothing in our Bayesian approach that leads to the modeling assumptions in Equations M1, M2, M3, M4, and M5 that formalize the priors and likelihoods for the choice models. Accordingly, there is a clean conceptual separation between modeling assumptions that propose psychological variables and processes that lead to behavior, and Bayesian methods of statistical inference that simply process the consequences of

those assumptions in the light of data. We think this state of affairs contrasts favorably with the set of non-Bayesian methods used by Birnbaum & Quispe-Torreblanca (2018). These involve approximate optimization criteria and measures of model complexity, the need to switch between asymptotic and Monte Carlo exact methods, and various procedures for generating modified data sets to introduce variability and uncertainty in inferences.

Given the similarity, at a very general level of analysis, of the results obtained by both methods, it is reasonable to ask whether it is worth investing in learning and using new Bayesian methods. If none of the detailed differences matter, and theoretical elegance and simplicity are not motivating factors, it is hard to counter that attitude. But this does not, in our view, place the Bayesian and non-Bayesian approaches on an equal footing. If roles were reversed, and the Bayesian analysis presented first was standard, it is near impossible to imagine anyone could muster any enthusiasm for the non-Bayesian analysis that followed. In statistical terms, it is less coherent, less consistent, and less complete.

Perhaps most importantly, we believe that the Bayesian approach is more intuitive. The Bayesian approach starts with a model that makes predictions about data and represents what is known and unknown about the model and its parameters. Given data, the Bayesian approach then updates the knowledge about models and parameters, by a simple application of the laws of probability. The process of theory evaluation by testing model predictions is exactly the account of empirical science advocated so effectively by Feynman (1994). In that sense, Bayesian methods follow the intuitions many researchers in psychology have about the purpose and goals of testing models against data.

References

Barry, D. & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, *88*, 309–319.

Birnbaum, M. H. & Quispe-Torreblanca, E. G. (2018). TEMAP2.R: True and error model analysis program in R. *Judgment and Decision Making*, *13*, 428–440.

Birnbaum, M. H., Schmidt, U., & Schneider, M. D. (2017). Testing independence conditions in the presence of errors and splitting effects. *Journal of Risk and Uncertainty*, *54*, 61–85.

Brooks, S. P. & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*, 221–241.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measuresof statistical accuracy. *Statistical Science*, *1*, 54–75.

Farrell, S. & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, *16*, 203–213.

Feynman, R. (1994). *The character of physical law*. Modern Press.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

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

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*, 31–43.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*, 29–41.

Lee, M. D. (2018). Bayesian methods in cognitive modeling. In J. Wixted & E.-J. Wagenmakers (Eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. Volume 5: Methodology* (pp. 37–84). John Wiley & Sons, fourth edition.

Lee, M. D. (in press). A simple and flexible Bayesian method for inferring step changes in cognition. *Behavior Research Methods*.

Lee, M. D. & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making*, *6*, 832–842.

Lee, M. D. & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lee, M. D. & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia, PA: SIAM.

Lodewyckx, T., Kim, W., Tuerlinckx, F., Kuppens, P., Lee, M. D., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

McDonald, J. H. (2009). *Handbook of biological statistics*, volume *2*. Sparky House Publishing.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.

Myung, I. J. & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika*, *80*, 341–364.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

Rouder, J. N. & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Stephens, D. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, (pp. 159–178).

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Vanpaemel, W. (2016). Prototypes, examplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, *72*, 183–190.

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

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wixted, J. & Wagenmakers, E.-J., Eds. (2018). *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. Volume 5: Methodology*. John Wiley & Sons, fourth edition.