In press at Journal of Mathematical Psychology
This is not the final proof.

Hierarchical models of simple mechanisms underlying confidence in decision making

Edgar C. Merkle
Wichita State University

Michael Smithson
Australian National University

Jay Verkuilen
Graduate Center, City University of New York

Choice confidence is a central measure in psychological decision research, often being reported on a probabilistic scale. Simple mechanisms that describe the psychological processes underlying choice confidence, including those based on error and confirmation biases, have typically received support via fits to data averaged over subjects. While averaged data ease model development, they can also destroy important aspects of the confidence data distribution. In this paper, we develop a hierarchical model of raw confidence judgments using the beta distribution, and we implement two simple confidence mechanisms within it. We use Bayesian methods to fit the hierarchical model to data from a two-alternative confidence experiment, and we use a variety of Bayesian tools to diagnose shortcomings of the simple mechanisms that are overlooked when applied to averaged data. Bugs code for estimating the models is also supplied.

Keywords: Confidence; subjective probability; decision making; Bayesian model checking

In both experimental and applied contexts, people are often required to make choices under uncertainty. Applied examples include students answering test questions, doctors making diagnoses, and jurors entering verdicts. The specific choice that is made often has significant implications, so it is natural to gauge an individual's certainty ("confidence") in his or her choice. As a result, the relationship between choice confidence and choice accuracy has become a popular topic in decision research. The topic is psychologically interesting because it says something about the extent to which judges have "meta-knowledge," and the topic also has implications for interpreting others' reports of confidence. While confidence data are often messier and more subjective than other psychological variables like choice and response time, the utility of research on confidence is well summarized by Koehler and Tversky (1994):

> Unlike the measurement of distance, in which fallible human judgments can be replaced by proper physical measurement, there are no objective procedures for assessing the probability of events such as guilt of a defendant, the success of a business venture, or the outbreak of war. Intuitive judgments of uncertainty, therefore, are bound to play an essential role in people's deliberations and decisions. (p. 565)

In decision research, confidence is often reported on either a probabilistic scale or an ordinal scale, with ordinal confidence receiving more attention in the modeling literature (e.g., Lee & Dry, 2006; Ratcliff & Starns, 2009; Van Zandt, 2000; Vickers, 1979). In probabilistic confidence experiments, the focus of this paper, judges are typically instructed to report *calibrated* probabilities: probabilities that, on average, match the long-run proportion of correct responses. Confidence is then compared to proportion correct in order to determine the "accuracy" of the confidence judgments.

Research on probabilistic confidence tends to show that judges are overconfident, meaning that their average confidence tends to be larger than their proportion correct over sets of items. As a result, decision researchers have proposed a number of psychological mechanisms that may contribute to overconfidence. These mechanisms vary widely in complexity. Focusing on simpler mechanisms, Koriat, Lichtenstein, and Fischhoff (1980) studied the impact of confirmation biases on overconfidence. They proposed that, in assessing confidence, a judge focuses heavily on evidence that supports her choice. Thus, the judge focuses on reasons why her chosen alternative is true and reasons why her unchosen alternative is false, neglecting other evidence. In a similar vein, McKenzie (1997) proposed an alternative underweighting bias, in which all evidence for and against the unchosen alternative is largely neglected. Further, Erev et al. (1994) showed that systematic biases are not necessary to account for overconfidence; the addition of random error to confidence judgments can yield an overconfidence effect.

The authors thank William Batchelder, Michael Lee, Eric-Jan Wagenmakers, Hao Wu, and an anonymous reviewer for comments that helped improve the paper. Correspondence to Edgar C. Merkle, Department of Psychology, Wichita State University, Wichita, KS 67260-0034. Email: edgar.merkle@wichita.edu.

More complex mechanisms underlying probabilistic confidence have been proposed within models such as HyGene (R. P. Thomas, Dougherty, Sprenger, & Harbison, 2008), the Poisson race model (Merkle & Van Zandt, 2006), and the Two-Stage Dynamic Signal Detection model (Pleskac & Busemeyer, 2010). These models generally account for multiple observed measures, including choice, confidence, and response time. HyGene, related to the earlier Minerva-DM (Dougherty, Gettys, & Ogden, 1999), generally describes mechanisms by which individuals generate and evaluate hypotheses in various tasks. Within this context, probability judgments stem from the relative evaluation of candidate hypotheses generated in memory. The Poisson race model, on the other hand, specifies a way by which incoming stimulus information is translated into choice and confidence. Compared to HyGene, it is less specific about the source of the stimulus information but allows for analytic model predictions. Finally, the Two-Stage Dynamic Signal Detection model is an extension of the diffusion model to confidence. It assumes that confidence is based on evidence that accrues after a choice is made. Compared to the other models mentioned, it provides the most complete description of observed response time distributions and their relationships to confidence and choice.

The models described above, both simple and complex, have received support from some combination of experimentation, simulation, and fits to data, leading one to question whether the more complex models are necessary. That is, if the simple models can adequately describe the data, then they would naturally be preferred over the complex models. Importantly, the models have usually been fit to averaged data, neglecting individual differences in subjects, individual differences in items, and other peculiarities in the raw data. By implementing the simple models in a hierarchical framework with raw data, we may study their shortcomings in a more detailed fashion. This can point to areas where the simple models (and corresponding theory) are lacking, and it can also demonstrate the need for more complex models.

In the following pages, we first describe a hierarchical model that allows us to study simple models' abilities to account for individual differences and for trial-by-trial confidence data. This is the same hierarchical approach to modeling individual differences used throughout this special issue, being especially related in spirit to the other decision making models of Nilsson, Rieskamp, and Wagenmakers (this volume) and of van Ravenzwaaij, Dutilh, and Wagenmakers (this volume). The model utilizes the beta distribution to account for the doubly-bound probability scale, and the model is estimated via Bayesian methods. After presenting both the model and approaches for incorporating psychological theories within it, we describe an application to data from a two-alternative decision experiment testing general financial knowledge. We then conduct a thorough study of the fitted model, illustrating the many ways by which the model can be compared to the observed data and detailing the shortcomings of the simple models. Finally, we discuss some general weaknesses of the simple models and describe areas where both theory and models can be improved.

Model

The model described here generally follows the beta regression framework developed by Smithson and Verkuilen (2006). Letting c_{ij} be judge i's reported confidence on item/stimulus j ($i = 1, \ldots, N$; $j = 1, \ldots, M$), we assume that confidence arises from a beta distribution. For example, a simple model assuming that all judges' confidence arises from a single beta distribution is:

$$c_{ij} \sim \text{Beta}(\alpha, \beta). \tag{1}$$

Smithson and Verkuilen's framework influences the model in two major ways. First, instead of the above parameterization, the beta distribution is parameterized with a mean parameter $\mu = \alpha/(\alpha + \beta)$ and a precision parameter $\phi = \alpha + \beta$. As shown below, it is more intuitive to model the mean and precision instead of modeling the traditional beta distribution parameters. Second, because the beta distribution has bounds at (0,1), the c_{ij} must have the same bounds. If this is not the case, then we can take simple linear transformations of the c_{ij} so that they have bounds at (0,1). This allows us to use the beta distribution to generally model doubly-bounded data, so long as the locations of the bounds are known. We next describe the implementation of simple confidence models within the beta framework, and we then describe hierarchical generalizations of the model.

Psychological Theories

We implement two psychological mechanisms of overconfidence using the beta-distributed models: an error mechanism and a confirmation bias mechanism, similar to those described in the introduction.

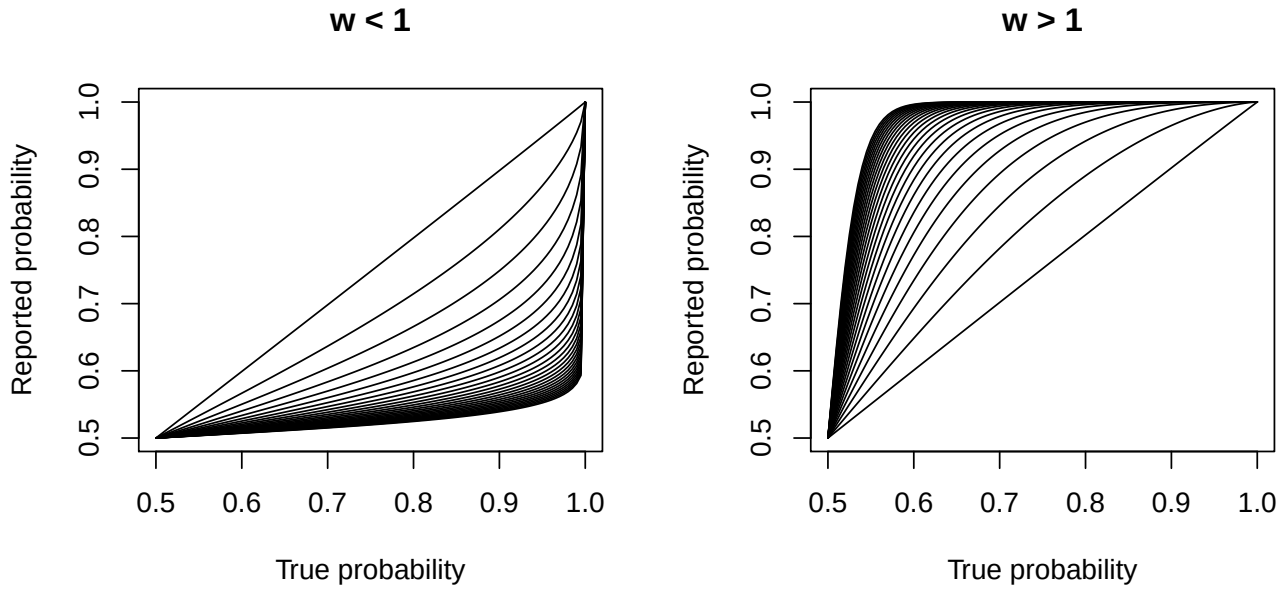
Confirmation Bias We implement a confirmation bias within the above beta distribution by modeling the μ parameter. Starting with the notion that judges would be well calibrated except for the bias, we seek equations that translate calibrated, "internal" confidence judgments into biased, reported confidence judgments. One such equation is:

$$c_{ij} = \frac{p_{ij}^w}{p_{ij}^w + (1 - p_{ij})^w}, \tag{2}$$

where p_{ij} is judge i's calibrated confidence for item j, and $w \in (0, \infty)$ represents a bias parameter. This equation, which was discussed by Karmarkar (1978) and also arises from Tversky and Koehler's (1994) support theory, allows for overconfidence and underconfidence via the w parameter (also see Shlomi & Wallsten, 2010). Figure 1 shows how this equation translates calibrated confidence (x-axis) into reported confidence (y-axis). It can be seen that, for values of w less than 1 (left panel), reported confidence is smaller than calibrated confidence. The opposite is true for values of w greater than 1 (right panel). Finally, if $w = 1$, judges are well calibrated.

The above equation is not the only one we could use (see McKenzie, Wixted, Noelle, & Gyurjyan, 2001 for others),

Figure 1. Curves obtained from Equation (2) for different values of w.



but it has a number of reasonable properties that make it suitable for the purposes of this paper. First, the equation always passes through the points (.5, .5) and (1, 1). Judgments of .5 and 1 usually reflect complete uncertainty and complete certainty, respectively, so the equation implies that judges have knowledge of when they are guessing (i.e., when their true confidence is .5) and when they know the answer (i.e., when their true confidence is 1). Next, values of w greater than one imply a confirmation bias. To be specific, the applications in this paper require $p_{ij} \geq 0.5$ (if not, judges would choose the other alternative). When $w > 1$ in this situation, confidence in the chosen alternative (p_{ij}) is decreased less than is confidence in the unchosen alternative ($1 - p_{ij}$). This implies that confidence in the chosen alternative is weighted more heavily than is confidence in the unchosen alternative, resulting in a confirmation bias. Conversely, when $w < 1$, judges exhibit conservatism (more weight placed on the unchosen alternative).

Finally, while not crucial for our current purposes, it can be shown that the above equation is equivalent to a multiplicative model on the log odds of calibrated confidence. Based on Equation (2), we have that:

$$1 - c_{ij} = \frac{(1 - p_{ij})^w}{p_{ij}^w + (1 - p_{ij})^w}.$$

It is then easily shown that:

$$\text{logit}(c_{ij}) = w\text{logit}(p_{ij}). \tag{3}$$

Random Error Random error can be captured by the ϕ parameter of the beta distribution, with large values implying little error and values close to zero implying considerable error. This is similar to the approach of Erev et al. (1994), who applied normal error to logit(p_{ij}) (i.e., a logit-normal distribution) in order to model c_{ij}. The beta distribution can be advantageous over this approach because it allows us to model a linear transformation of confidence, as opposed to using the nonlinear logit transformation. This can ease interpretation of model parameters. The logit-normal distribution can also exhibit unstable parameter estimates when used to model data on the unit interval.[1]

Summary By incorporating psychological theories in the model, we now have:

$$\begin{aligned} c_{ij} &\sim \text{Beta}(\mu_{ij}, \phi) \\ \mu_{ij} &= p_{ij}^w/(p_{ij}^w + (1 - p_{ij})^w), \end{aligned}$$

[1] The log-normal distribution, a close relative of the logit-normal distribution, is markedly unstable. Schmoyer, Beauchamp, Brandt, and Hoffmann (1996) specifically show that the log-t family, of which the log-normal is a boundary case, has no moments at all with the exception of the log-normal, which is itself not uniquely determined by its moments. Exactly how this carries over to the logit-normal is not entirely clear at this time, but the series expansion of the log-likelihood of the logit-normal contains exponentially diverging terms, whereas the beta has only algebraic terms. Because the sample space is bounded, all of the logit-normal's moments exist. However, we have noticed substantial instability when the logit-normal is used to model values on the unit interval. This is a particular problem when exact boundary observations are replaced with a value strictly inside the unit interval. More research needs to be done on this topic.

with w being a free parameter that represents cognitive bias and ϕ being a free parameter that represents the magnitude of random error in confidence. For simplicity, we assume that judges are calibrated to the group. The assumption entails setting $p_{ij} = \overline{p}_j \; \forall \; i$, where $\overline{p}_j$ is the empirical proportion correct for item j (though see the general discussion for some relaxations of this assumption).[2] This leaves w and ϕ to be estimated.

Hierarchical Models

The above model assumes a single w and ϕ across all judges. A hierarchical extension of the model, allowing each judge to have her own w and ϕ, is implemented in this paper. Such an extension is straightforward via link functions:

$$
\begin{aligned}
c_{ij} &\sim \text{Beta}(\mu_{ij}, \phi_i) && (4)\\
\mu_{ij} &= p_{ij}^{w_i} / (p_{ij}^{w_i} + (1 - p_{ij})^{w_i}) && (5)\\
w_i/20 &\sim \text{Beta}(\mu_w, \phi_w), 0 < w_i \leq 20 && (6)\\
\log(\phi_i) &\sim N(\mu_\phi, \sigma_\phi^2), && (7)
\end{aligned}
$$

where the hyperparameters μ_w, ϕ_w, μ_ϕ, and σ_ϕ^2 all receive prior distributions and are estimated with the model. The above equations show that we assume a normal hierarchical distribution on $\log(\phi_i)$ but not on w_i. Once w_i gets close to 20, the model makes the same substantive prediction, assigning $\mu_{ij} \approx 1$ for almost all values of p_{ij}. Thus, if we take the w_i to be bounded at 0 and 20, we can use a beta hierarchical distribution. This is advantageous because the beta can assume more shapes than the normal distribution, potentially resulting in better models. As described in the application, we can further model μ_w to summarize effects of experimental manipulations.

The above model is most easily estimated in Bugs (e.g., A. Thomas, O'Hara, Ligges, & Sturtz, 2006) via Bayesian methods. For the data described below, sampling speed is reasonable ($\approx$ 5 minutes per 3,000 iterations), and convergence is achieved by about $2,000$ iterations. Judging the fit of the hierarchical model (which reflects the adequacy of the simple mechanisms) is difficult, however, because the error distribution does not have to be bell-shaped (in fact, it is often U-shaped). We illustrate some tools that are useful for gauging the model's fit below; for the most part, these are general tools that can be used in general hierarchical modeling contexts.

Application: Confidence in Financial Knowledge

To illustrate the model, we use data from Experiment 2 of Sieck, Merkle, and Van Zandt (2007). In this experiment, 141 participants completed a 2-alternative, 30-item test of general financial knowledge. Each participant was assigned to one of three conditions differing in the way that participants reported confidence. Based on a verbal theory, the general goal of the experiment was to study confidence elicitation conditions that reduce overconfidence. A specific goal was to determine whether judges report lower confidence if they explicitly consider the unchosen alternative prior to confidence elicitation. We will first describe the experimental conditions, and we will then describe application of the hierarchical beta model.

Methods

All participants completed 30 two-alternative items on general financial knowledge. In the control condition of the experiment, each participant chose an alternative and then reported a probability in (.5,1) that his/her choice was correct. Probabilities were bounded from below at .5 because, if the judge reports a probability less than .5, then she should have chosen the other alternative. In the "choice, independent" (CI) condition, each participant chose an alternative and then reported confidence that each individual alternative was true. Finally, in the "explain, independent" (EI) condition, participants first chose a correct alternative. For each alternative, they then wrote an explanation for why the alternative could be true and reported confidence in that alternative. For both the CI and EI conditions, participants' final choice confidence was obtained by taking P(chosen)/[P(chosen)+P(unchosen)]. The researchers' hypothesis was that, as compared to the standard condition, confidence would be lower in the CI and EI conditions.[3]

Because confidence was bounded at .5 and 1, we needed to transform the confidence judgments so that they lied in (0,1). This was accomplished via $c'_{ij} = (c_{ij} - 0.5)/0.5$, with judgments obtaining values of 0 or 1 perturbed slightly to avoid numerical instability (see Smithson & Verkuilen, 2006, p. 57). This transformation implicitly restricts the p_{ij} to lie in (0.5,1) as well (see Equation (2)); for the current data, this restriction was satisfied as each $\overline{p}_j$ was greater than 0.5 However, this may not generally be the case and reflects one limitation of using Equation (2).

Model

We make one modification to the hierarchical beta model presented in Equation (4) to better conform to this specific experiment. We model μ_w, the mean of the hierarchical beta distribution, as:

$$\text{logit}(\mu_{w,i}) = b_0 + b_1 \text{I}_{\text{CI},i} + b_2 \text{I}_{\text{EI},i}, \qquad (8)$$

where $\text{I}_{\text{CI},i}$ indicates whether or not judge i was in the CI condition and $\text{I}_{\text{EI},i}$ is defined similarly. This equation allows the mean of the hierarchical beta distribution to vary from condition to condition, which implicitly allows the shape of the distribution to vary. We can assess the effects of the experimental conditions through the b_1 and b_2 parameters. Because the ϕ_i represent unsystematic error in our model, we

[2] The simple models implicitly make this assumption in being fit to data averaged over subjects.

[3] Plausibility of the incorrect alternative was also manipulated in the experiment, but we ignore that here for simplicity. Preliminary modeling indicated that inclusion of this manipulation in the model did not result in any improvement.

assume that they are not affected by the experimental conditions. Preliminary modeling indicated that this was a reasonable assumption.

The model was fit in OpenBugs with three chains of parameters being sampled for 7,000 iterations each. The first 2,000 iterations were discarded from each chain as burn-in, with convergence being judged through time series plots, autocorrelation function plots, and Gelman-Rubin statistics.

Results

In examining the model's fit to data, we generally conclude that the model captures general regularities in the empirical data but misses many details. We present five different aspects of the estimated model that help demonstrate this conclusion. They are (1) analyses regarding the effect of experimental conditions; (2) analyses regarding the ability of the model to predict confidence in each condition; (3) analyses regarding the ability of the model to predict confidence for individual subjects; (4) an examination of the utility of the hierarchical distributions within the model; and (5) an examination of the sensitivity of the results to the prior distributions. We employ a variety of "model-checking" tools to demonstrate the model's correspondence to data, including posterior predictive distributions, simulation of data from the fitted model, and Bayes factors.

Experimental Effects The estimated hierarchical distributions on the w_i (see Equations (6) and (8)) are presented in Figure 2 separately for the three experimental conditions. We can see that the distributions are positively skewed and slightly differ in shape across the three conditions (CI and EI being slightly more skewed than Control). The dotted vertical line in each histogram at the point $w = 1$ represents perfect calibration (i.e., perfect mapping from internal confidence to mean reported confidence). It can be seen that, in the hierarchical distributions for the two experimental conditions, there is more density below 1 than there is in the hierarchical distribution for the control condition. This provides evidence that judges in the two experimental conditions reported lower confidence than did judges in the control condition. Posterior intervals for the specific impact of the conditions are available via the b_1 and b_2 parameters; these parameters are both negative, with neither of the 95% intervals containing 0.

Along with posterior intervals, it is possible to calculate a Bayes factor to test hypotheses that the mean value of w equals 1 in each condition. This equality hypothesis is substantively interesting because it reflects perfect calibration. To calculate the Bayes factors, we employ the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Briefly, this entails evaluation of parameters' prior and posterior distributions at the hypothesized value of 1. The ratio of these two distributions then yields the Bayes factor. The hypotheses are slightly more complicated than $\mu_w = 1$ because, within the model, μ_w is the mean of $w/20$ and is itself modeled via a logit transformation. Thus, the following hypotheses are equivalent to $\mu_w = 1$ in the three conditions:

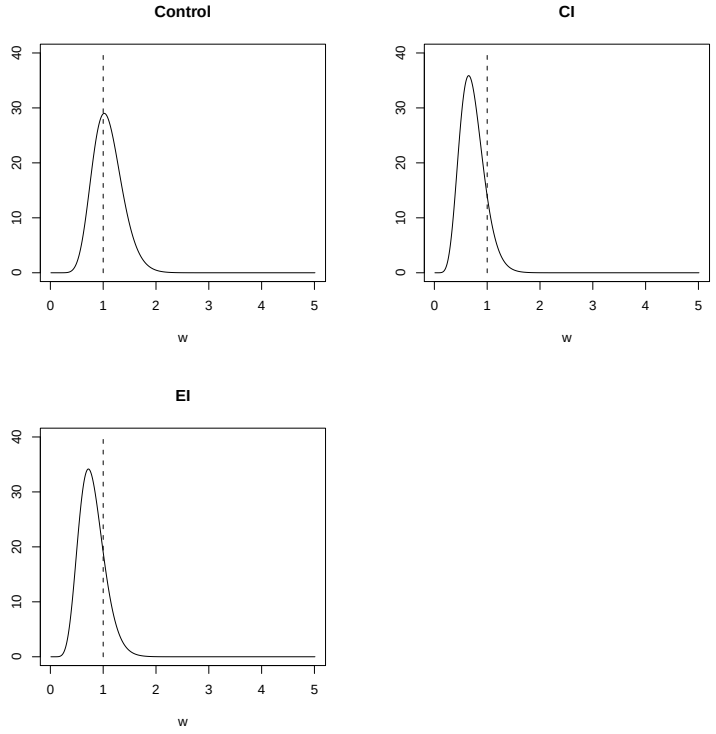
$$\begin{aligned} \mathrm{H}_{\text{control}}: &\quad b_0 = \log(.05/.95) \\ \mathrm{H}_{\text{CI}}: &\quad b_0 + b_1 = \log(.05/.95) \\ \mathrm{H}_{\text{EI}}: &\quad b_0 + b_2 = \log(.05/.95). \end{aligned}$$

The prior distributions for b_0, b_1, and b_2 were all taken to be independent normal, so the implied prior distributions for sums of these parameters follow straightforwardly from the individual priors. Using logspline density estimates for the posterior distributions of b_0, b_0+b_1, and b_0+b_2, we calculate Bayes factors for $\mathrm{H}_{\text{control}}$, H_{CI}, and H_{EI} as 24.4, 0.0001, and 0.006, respectively. As can be seen in Figure 2, the Bayes factors imply that subjects in the control condition were generally well-calibrated, while those in the experimental conditions were not (tending towards conservatism). These results appear to disagree with the observed data, which generally exhibit a greater amount of overconfidence. Specifically, the observed proportions of subjects exhibiting overconfidence[4] in each condition are .77 (control), .48 (CI), and .50 (EI). Within the model, the predicted proportion of overconfident subjects can be obtained by calculating the proportions of the Figure 2 distributions greater than one. These predicted proportions are .60 (control), .12 (CI), and .19 (EI). Thus, on the basis of the w parameter, the model predicts less overconfidence than is observed. The error term (the ϕ_i) can also account for overconfidence, however, so that the model attributes the remaining observed overconfidence (i.e., that not predicted by the w parameter) to the error term. This implies that either (1) random error drives empirical overconfidence (especially in the CI and EI conditions), or (2) the ϕ_i have absorbed some systematic bias that was not accounted for in our model. For further discussion on the difficulty in distinguishing these two explanations, see Juslin, Winman, and Olsson (2000) and Merkle, Sieck, and Van Zandt (2008).

Predictions By Condition We have estimates for b_0, b_1, and b_2, and we can insert those estimates into Equation (8) to obtain a $\widehat{w}$ for each condition. Each $\widehat{w}$ could then be inserted into Equation (4) with other estimated parameters to obtain predicted confidence distributions for each condition. This is misleading because the predictions do not capture the variability inherent in the hierarchical distributions on the w_i (e.g., Gelman, Carlin, Stern, & Rubin, 2004). Further, the hierarchical distributions on the w_i are asymmetric (beta distributed), which implies that means of the distributions may inaccurately reflect model predictions. As a result, we simulate data from the estimated model and compare the simulated predictions to the observed data. We simulated 1,000 experiments of data, with number of subjects in each condition and number of items matching that of the real experiment. For each subject i, we first drew w_i and ϕ_i from the estimated hierarchical distributions. For each item j, we then: (a) obtained μ_{ij} from Equation (2) using w_i and the empirical $\overline{p}_j$; and (b) drew c_{ij} from Beta(μ_{ij}, ϕ_i).

[4] Overconfidence is defined as a subject's mean confidence being larger than his/her proportion correct

Figure 2. Estimated hierarchical distributions on w for each experimental condition. The dotted lines at $w = 1$ reflect perfect calibration.



Summarizing the simulated data by condition, Figure 3 contains quantile-quantile plots comparing observed confidence to the simulated confidence judgments. From left to right, the 5 points in each plot represent 10th, 30th, 50th, 70th, and 90th percentiles, respectively. The extent to which the points fall along the diagonal reflect the extent to which the predicted percentiles match the observed percentiles. The plots show that the model does well close to the bounds of the confidence scale, picking up subjects' (over)use of .5 and 1. The model generally predicts too few confidence judgments in the middle of the scale, however, with the discrepancy being most apparent in the control condition. It is plausible that this discrepancy comes from the frequent occurrence of non-substantive judgments (i.e., those for which the subject is not paying attention to the experiment) at the scale bounds. The theories described in the introduction are largely mute on overuse of the endpoints. The hierarchical beta model is such that, if large proportions of density are assigned to each bound, there cannot be a third "bump" of density in the middle of the scale.

Fit to Subjects In addition to obtaining predictions for experimental conditions, we can examine predictions for individual judges. Figure 4 is a plot of observed vs predicted mean confidence for each judge. The extent to which the points fall along the diagonal reflects the extent to which the predictions match the observations. We see that the model predictions are generally accurate, though they are too large for the means close to .5 and too small for the means close to 1. These "misses" may stem from the fact that the estimated error distributions are generally U-shaped, so that the error can pick up masses of points at either scale bound without shifting the mean all the way to the bound. Further, shrinkage of the w_i to the group may also contribute to the misses.

We can also examine the extent to which the model predicts trial-by-trial data for each subject. It is perhaps most informative to examine extreme subjects that the model cannot fit. Figure 5 contains data from four subjects whose responses differ considerably, with "true confidence" (the p_{ij}'s) on the x-axis and reported confidence on the y-axis. Points reflect individual observations from a judge, and lines reflect the model's predicted mapping from true confidence to mean reported confidence (making use of the estimated w's and Equation 2). We can see that the lines are definitely influenced by the points in the graph, but there is considerable variability around these lines.

As the reader may have already discerned, the previous graphs are inadequate because mean predictions are not the best model summaries. Confidence data from individual subjects are often skewed, U-shaped, and/or non-normal, so means alone do not offer a good summary of the observed data. We must also consider the error distribution around the mean. Thus, as we did for the experimental conditions, it is more useful to examine observed vs predicted quantiles of the confidence distribution. For each subject, these quantiles arise from a mixture of 30 beta distributions (one for each item), so we simulate from the estimated model to obtain predicted quantiles and variability in model predictions. The

Figure 3. Observed vs predicted percentiles of confidence distributions for each experimental condition. The five points represent the 10th, 30th, 50th, 70th, and 90th percentiles, and predicted percentiles are obtained from simulations of the fitted model.

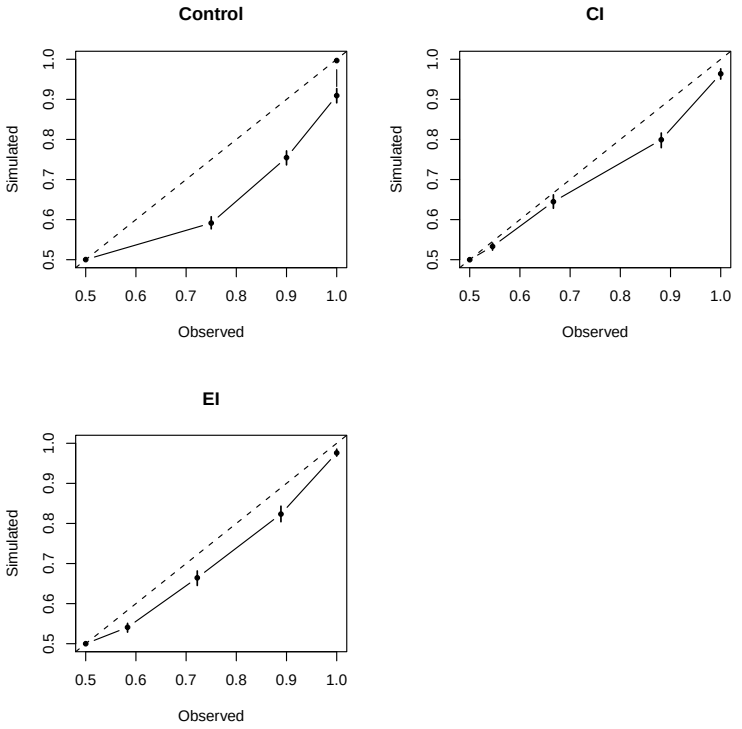


Figure 4. Observed vs predicted mean confidence by subject. Each point reflects a single subject, and the diagonal line reflects accurate predictions.

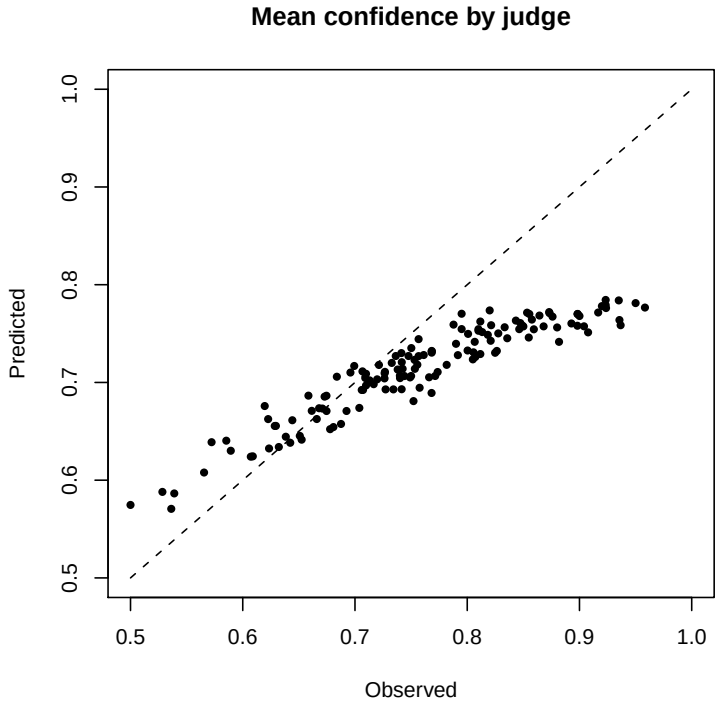
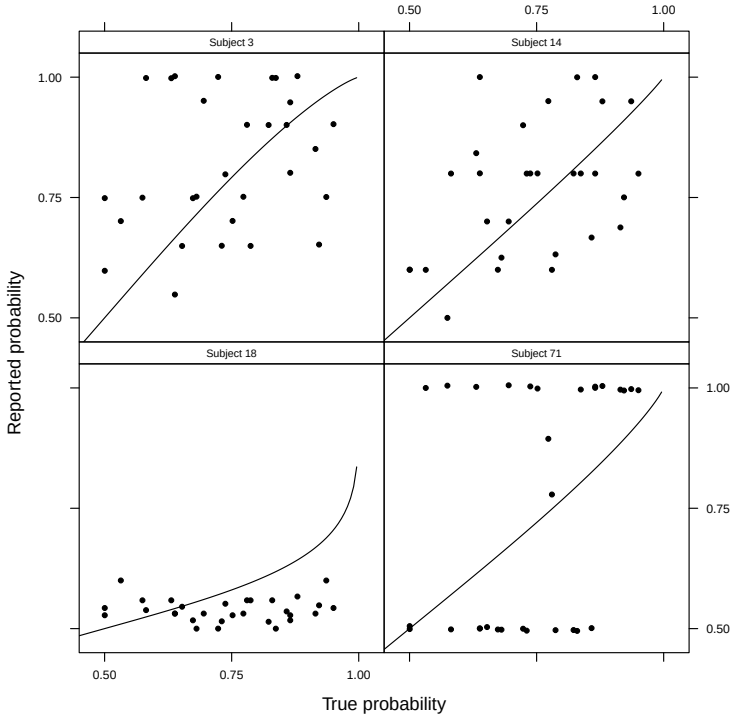


Figure 5. Observed vs predicted mapping from true confidence (p_{ij}) to reported confidence for four selected subjects. Points reflect observed data, and lines reflect model predictions.



simulation proceeded as follows. For subject i responding to item j, we obtained μ_{ij} from Equation (2) using the estimate $\widehat{w_i}$ and the empirical $\overline{p}_j$. We then drew c_{ij} from the Beta($\mu_{ij}, \widehat{\phi_i}$). Subject i's simulated data then consisted of 30 draws from different Beta distributions. The entire procedure was repeated 1,000 times for each subject to observe variability in the model predictions.

Figure 6 contains plots of observed vs predicted quantiles for the same four subjects as before. From left to right, the five points reflect the 10th, 30th, 50th, 70th, and 90th percentiles, respectively. Vertical lines reflect variability in the model predictions (specifically, the middle 90% of the predictions). Across graphs, we can see that the model predictions have considerable variability in the middle percentiles and little variability in the end percentiles. This is because the error distributions are generally U-shaped, meaning that there will always be many judgments at the endpoints and few in the middle.

We chose the four specific subjects here because they reveal some shortcomings of the model. Both Subject 3 and Subject 14 avoided use of 50% confidence judgments, which generally disagrees with most subjects' data. The top two graphs show that the estimated model cannot account for this, with predictions for the 10th and 30th percentiles being lower than the observed data. This is likely due to the general predominance of 50% judgments in the data, which overpowers the individual subjects' data. Next, Subject 18 reports only judgments near 50%. The model now overpredicts the 70th and 90th percentiles because of the general predominance of 100% judgments in the data. Finally, Subject 71 almost exclusively reports 50% or 100% judgments. The model picks up the end percentiles well (the 10th and 30th percentile points overlap), but its prediction for the 50th percentile (median) exhibits considerable variability.

Finally, we can observe the extent to which subjects' estimated w parameters track (mis-)calibration. Figure 7 displays the estimated w_i versus overconfidence for each judge.[5] The vertical dotted line reflects the point of good calibration within the model ($w = 1$), and the horizontal dotted line reflects the point of good calibration empirically (OC = 1). It is observed that larger values of w are related to larger values of OC, which we would expect for sensible model estimates. However, there are many subjects with estimated w's below 1 and OC above 0. This is another demonstration of the contribution of the error terms (ϕ_i) to OC. The correlation between w and OC is .43, with a 95% confidence interval of (.28,.55).

Hierarchical Distributions Our Bayes factors for the experimental conditions (presented in an earlier section) signify that the means of the w distributions do not equal 1 for the CI and EI conditions. This implies that the power equation (Equation (2)) is more useful than a simple error model stating that $\mu_{ij} = p_{ij}\ \forall\ i, j$, with all miscalibration being due to

[5] Overconfidence is defined here as a judge's mean confidence minus proportion correct.

Figure 6. Observed vs predicted percentiles of confidence distributions for four selected subjects. The five points represent the 10th, 30th, 50th, 70th, and 90th percentiles, and predicted percentiles are obtained from simulations of the fitted model. The vertical lines reflect the middle 90% of model predictions for each percentile.

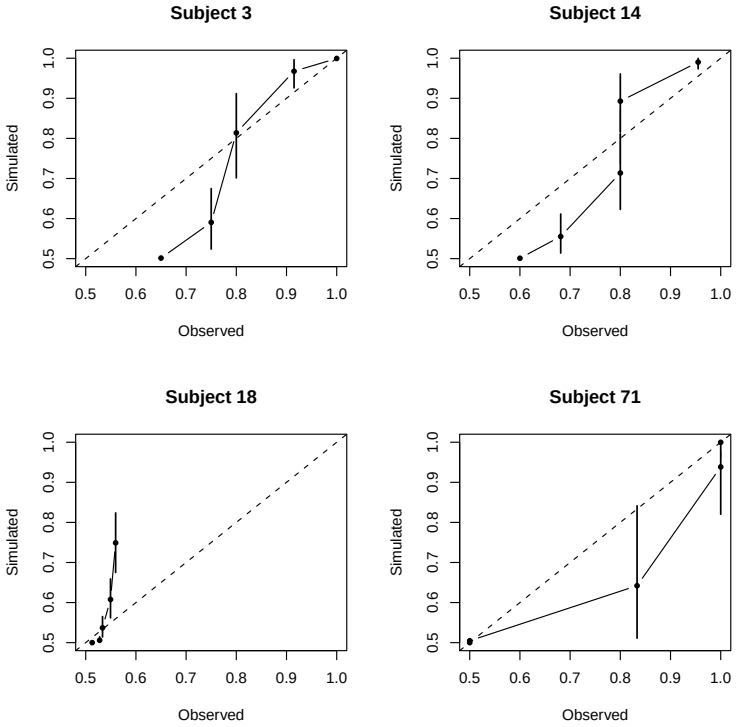
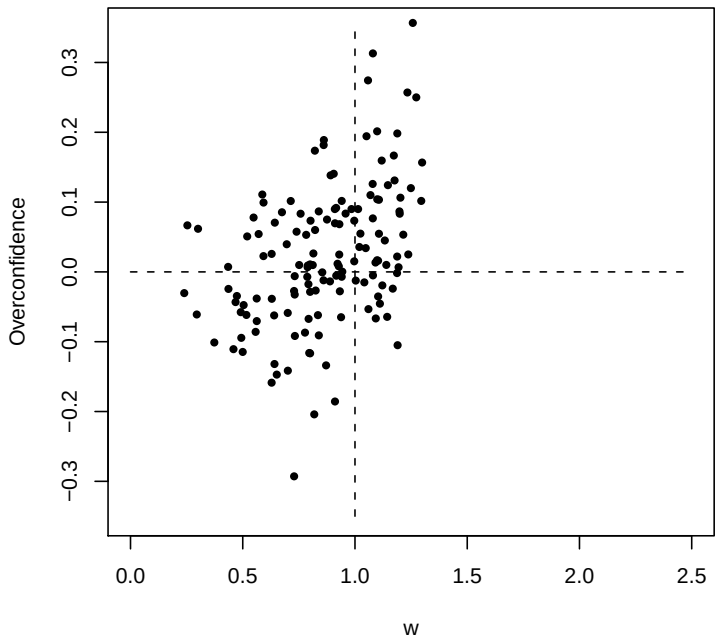


Figure 7. Estimated w_i parameters vs observed overconfidence. Each point represents a subject, the vertical line represents perfect calibration as defined by $w = 1$, and the horizontal line represents perfect observed calibration as defined by OC=0.



the error parameters ϕ_i. We can go on, however, to examine the extent to which the hierarchical aspect of the model is useful. This entails an examination of ϕ_w, the precision parameter of the hierarchical distribution on w, and σ^2_ϕ, the variance of the precision parameter ϕ. Focusing on ϕ_w, large values reflect little variability around μ_w. For example, if $\mu_w = 1$ and $\phi_w = 7500$, then 95% of the w_i will fall between .9 and 1.1. This corresponds to subjects' mean reported confidence falling within .02 units of true confidence, which implies small individual differences among judges. We could use an encompassing prior distribution approach (e.g., Hoijtink, Klugkist, & Boelen, 2008) to test $H_1 : \phi_w > 7500$ vs the unrestricted $H_2 : \phi_w > 0$, but it is unnecessary here. This is because the largest sample from the posterior distribution of ϕ_w is 522.25, which still implies considerable variability in the w_i (for $\mu_w = 1$, 95% of w's lie between .66 and 1.40). The Bayes factor for H_1 vs H_2 is therefore very close to 0.

We next focus on σ^2_ϕ, the variance of the precision parameter for observed confidence (see Equation (4)). The model contains a normal prior distribution on $\log(\phi)$, so we could use the encompassing prior approach to obtain a Bayes factor testing $H_3 : \sigma^2_\phi \approx 0$ vs the unrestricted $H_4 : \sigma^2_\phi \geq 0$. This would involve defining a range of σ^2_ϕ values deemed to be sufficiently close to zero, and then comparing the proportion of the prior and posterior distributions falling in this range. Examining the posterior samples of σ^2_ϕ, however, we find that the smallest sampled value is 0.165. This corresponds, e.g., to an error standard deviation of .35 around a mean confidence judgment of 0.75. Because this reflects considerable error on a scale with bounds at 0.5 and 1, the Bayes factor for H_3 vs H_4 is again close to zero. To summarize more generally, we have found that the hierarchical distributions on w and ϕ are both helpful for accounting for individual differences across subjects.

Sensitivity Analysis Finally, we can examine the sensitivity of the results to the noninformative prior distributions used to estimate the model. The specific prior distributions were:

$$\begin{aligned} b_0 &\sim \text{N}(0, 2.9) \\ b_1 &\sim \text{N}(0, 2.9) \\ b_2 &\sim \text{N}(0, 2.9) \\ \phi_w &\sim \text{U}(0, 2000) \\ \mu_\phi &\sim \text{N}(0, 10^{-6}) \\ \sigma^{-2}_\phi &\sim \text{Gamma}(.001, .001). \end{aligned}$$

The priors on the b parameters look informative here, but they are modeling $\text{logit}(\mu_w)$ and not μ_w directly. When considering μ_w directly, these priors are roughly noninformative.

To examine the sensitivity of the results to the priors, we re-estimated the model using a different set of priors. The new set of priors was designed to assign density only to the space of (what we perceived to be) plausible parameter values. They are taken as:

$$\begin{aligned} b_0 &\sim \text{N}(-2.94, 0.3) \\ b_1 &\sim \text{N}(0, 0.3) \\ b_2 &\sim \text{N}(0, 0.3) \\ \phi_w &\sim \text{U}(0, 700) \\ \mu_\phi &\sim \text{U}(0, 1) \\ \sigma^{-2}_\phi &\sim \text{U}(1, 100). \end{aligned}$$

Parameter estimates and standard errors under the new set of priors nearly all agreed with those under the old priors to two decimal places. The only exception to this was the ϕ_w parameter, which was estimated to be 258.7 (posterior SD=44.4) under the old priors and 259.7 (posterior SD=43.5) under the new priors. The agreement in parameter estimates implies that the data dominate the priors. There was a change in the Bayes factors testing the hypothesis $\mu_w = 1$ in each condition; the more-informative priors generally shifted evidence away from the hypothesis. Under the original priors, logarithms of Bayes factors[6] were 3.19, -9.21, and -5.11 for the control condition, CI condition, and EI condition, respectively. These logarithms are compared to zero instead of to one, where positive numbers favor the hypothesis that $\mu_w = 1$. As a result, we favored $\mu_w = 1$ in the control condition but not in the two other conditions. Under the new priors, the logarithms of Bayes factors are 0.74, -16.75, and -7.01. The log-Bayes factor for the control condition still favors $\mu_w = 1$ in the control condition, but now only moderately so. Conversely, the log-Bayes factors for the other two conditions provide stronger evidence against $\mu_w = 1$. These analyses demonstrate the changes in Bayes factors that can result from different priors (e.g., Gelman et al., 2004; Kass & Raftery, 1995; Liu & Aitkin, 2008).

General Discussion

In this article, we have developed a hierarchical model that allows us to conduct a detailed examination of simple psychological theories of choice confidence. The hierarchical model's use of the beta distribution makes it suitable for probabilistic confidence judgments, and the mean/precision parameterization of the distribution makes it useful for implementing psychological theories within the model. The modeling results show that, when allowing for individual differences and fitting to trial-by-trial data, the confirmation bias and error theories of confidence are incomplete. The results highlight areas where both psychological theory and modeling require further development. We address these areas below. Finally, we provide some general comments on the utility of the hierarchical modeling framework developed here and on the tools used to examine the model's correspondence with the data.

[6] Logarithms of Bayes factors are presented here due to small observed values.

Theory Improvement

In the financial knowledge data presented here and in other probability elicitation experiments we have conducted, there were a large proportion of judgments at the scale bounds. In averaging each subject's data, however, one obtains many numbers near the middle of the scale. Because many psychological theories of confidence elicitation have focused on averaged data, the theories generally neglect the overuse of scale bounds and other "nice" numbers. It is likely the case that these judgments are a mixture of substantive judgments, where the judge assesses confidence, and non-substantive judgments, where the judge does not assess confidence and simply reports a familiar number. Fischhoff and Bruine de Bruin (1999) observed this type of phenomenon in the overuse of 50% probability judgments, where subjects sometimes use 50% not as a probability but as a category conveying that they do not know what number to report. The researchers were able to reduce 50% judgments by giving subjects the option of reporting "absolutely no idea" instead of a probability.[7]

One psychological theory that potentially addresses these issues is fuzzy-trace theory (e.g., Reyna & Brainerd, 1995), which specifies that judges rely on the "least precise level of representation that can be used to accomplish a judgment" (Reyna & Adam, 2003, p. 326). As applied to probabilistic confidence, subjects' least precise level of representation may entail three categories: uncertain, somewhat certain, and certain. The uncertain and certain categories would then map to the scale bounds, and the somewhat certain category would map to numbers between the bounds. While Bouwmeester and Verkoeijen (2010) formally examined some predictions of fuzzy-trace theory in the context of recognition memory, the theory has not been fully formalized mathematically. Thus, there remains some ambiguity concerning its specific predictions in the context of probability judgments.

Modeling Improvement

The theory improvement described above can be augmented with related model improvements. Model improvements may include: (1) implementation of a two-component mixture model; (2) implementation of other explanations of the relationship between calibrated confidence (p_{ij}) and reported confidence (c_{ij}), and (3) allowing the p_{ij} to vary across subjects.

Modification (1) could be used to account for non-substantive judgments at the scale bounds separately from substantive judgments. The estimated beta distributions in this paper were U-shaped, being highly impacted by the masses of judgments at the two bounds. In implementing a two-component mixture model, we might obtain one U-shaped component for overuse of the scale-bounds and one unimodal component for the other judgments. The advantage of this would be the ease by which the unimodal component could be interpreted. For details on mixture approaches to beta models, see Smithson, Merkle, and Verkuilen (in press).

Modification (2) is most applicable to the psychological theory described previously: there are many ways by which c_{ij} can be obtained from p_{ij}. It may be most beneficial to treat p_{ij} as a latent variable and to implement a multivariate model of confidence and accuracy. In this framework, accuracy data would lead to estimates of the latent p_{ij}, which may then be mapped into c_{ij} via many possible transformations (see, e.g., McKenzie et al., 2001). Fuzzy-trace theory may also be implemented in this framework, roughly stating that subjects first represent the latent p_{ij} as ordered categories and the map the ordered categories to probabilities. This appears to offer one method of obtaining probabilistic confidence judgments from existing models of ordinal judgments (e.g., Lee & Dry, 2006; Ratcliff & Starns, 2009; Van Zandt, 2000; Vickers, 1979): assume three ordered categories in the models, map the end categories to the respective bounds of the probability scale, and map the middle category to a distribution in the middle of the probability scale.

Finally, modification (3) addresses the p_{ij} varying across subjects. The key here involves specification of an accuracy model that leads to predictions for the p_{ij}. Some principled models for the p_{ij} stem from item response theory, where the p_{ij} would differ depending on the subjects' ability (a similar idea was described by Budescu & Johnson, 1997). As an initial examination of this, we fit a Rasch model to subjects' accuracy data and used empirical Bayes estimates of the ability parameters to predict the probability (p^*_{ij}) that each judge answers each item correctly. We then refit the hierarchical model described in the paper, using the p^*_{ij} instead of $p_{ij} = \overline{p}_j$. While the Rasch model indicated variability in the subjects' abilities, the results of the hierarchical beta model with the p^*_{ij} did not differ greatly from those of the original model. In particular, the experimental conditions had similar effects on the w_i, and the model still underpredicted the density of some judgments in the middle of the confidence scale. Thus, it appears that subject-specific p_{ij} alone will not improve the model.

Conclusion

Though the specific model used in this paper was unable to account for all aspects of the data, the hierarchical beta framework developed here is generally useful for decision experiments involving probabilistic confidence. One can insert general equations for μ_{ij} in the model, resulting in a wide variety of theories that may be examined in the framework. Further, as described above, the framework can readily accommodate extensions to multivariate models and to subject-specific p_{ij}.

In addition to the model, the diagnostic tools used here to examine the model's correspondence with the data can

[7] Fischhoff and Bruine de Bruin's tasks required subjects to estimate low-probability events, such as being burglarized during a single year. Thus, unlike the financial knowledge task described in this paper, "50%" in their tasks is unlikely to convey a reasonable probability judgment (unless the subject lives in an exceedingly rough neighborhood).

generally be employed in psychological modeling. These include the use of posterior predictive distributions, simulation of data from estimated models, and Bayes factors. The tools emphasize examination of many aspects of the data and model predictions, including distributions for specific judges and items, distributions for experimental conditions, effects of experimental conditions, and correspondence between estimated parameters and empirical data. The tools collectively aid the researcher in developing a thorough understanding of the model and its ability to capture the phenomena of interest, leading to improved models and advances in psychological theory.

References

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2010). Latent variable modeling of cognitive processes in true and false recognition of words: A developmental perspective. *Journal of Experimental Psychology: General*, *139*, 365–381.

Budescu, D. V., & Johnson, T. (1997). On the use of Item Response Theory in the study of probability judgments. Presented at the 30th Annual Meeting of the Society for Mathematical Psychology.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519-527.

Fischhoff, B., & Bruin, W. Bruine de. (1999). Fifty-fifty=50%? *Journal of Behavioral Decision Making*, *12*, 149–163.

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

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384-396.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, *21*, 61–72.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science*, *30*, 1081–1095.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, *71*, 141-160.

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes-no and forced-choice tasks. *Journal of Experimental Psychology: General*, *130*, 140–155.

Merkle, E. C., Sieck, W. R., & Van Zandt, T. (2008). Response error and processing biases in confidence judgment. *Journal of Behavioral Decision Making*, *21*, 428–448.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*, 391–408.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (this volume). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection theory: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.

Ratcliff, R., & Starns, J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*, 59–83.

Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis*, *23*, 325–342.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1–75.

Schmoyer, R. L., Beauchamp, J. J., Brandt, C. C., & Hoffman Jr., F. O. (1996). Difficulties with the lognormal model in mean estimation and testing. *Environmental and Ecological Statistics*, *3*, 81–97.

Shlomi, Y., & Wallsten, T. S. (2010). Subjective recalibration of advisors' probability estimates. *Psychonomic Bulletin and Review*, *17*, 492–498.

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior and Human Decision Processes*, *103*, 68-83.

Smithson, M., Merkle, E. C., & Verkuilen, J. (in press). Beta regression finite mixture models of polarization and priming. *Journal of Educational and Behavioral Statistics*.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*, 54–71.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (this volume). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

Appendix
Bugs Code

The code below fits the hierarchical beta model from the application, which includes effects of experimental conditions. It assumes three pieces of data: `N`, a scalar reflect-

ing the number of subjects; k, a scalar reflecting the number of items; and y, an N×k matrix of transformed confidence judgments (transformed to lie in $(0, 1)$) with rows reflecting subjects and columns reflecting items.

```
model
  {
    for (i in 1:N){
      for (j in 1:k){
        # Specify confidence arising from a beta
        y[i,j] ~ dbeta(alpha[i,j], beta[i,j])

        # Transform alpha and beta to mu and phi
        alpha[i,j] <- mu[i,j]*phi[i]
        beta[i,j] <- phi[i] - mu[i,j]*phi[i]

        # Model for mean confidence
        c[i,j] <- pow(p[j],w[i])/(pow(p[j],w[i]) + pow(1-p[j],w[i]))
        # Transform c[i,j] so its bounds match that of the beta
        mu[i,j] <- (c[i,j] - 0.5)/0.5
      }
      # Hierarchical model for error term
      log(phi[i]) <- b0 + err.phi[i]
      err.phi[i] ~ dnorm(0, inv.sigsq.phi)

      # Hierarchical beta distribution for w
      w[i] <- 20*wtrans[i]
      wtrans[i] ~ dbeta(alpha.w[i], beta.w[i])
      # Transform the beta distribution parameters
      alpha.w[i] <- mu.w[i] * phi.w
      beta.w[i] <- phi.w - mu.w[i]*phi.w
      # Modeling effects of experimental conditions on mu
      logit(mu.w[i]) <- b0.w + b1.w*ci[i] + b2.w*ei[i]
    }

    # Priors:
    b0 ~ dnorm(0, 1.0E-6)
    phi.w ~ dunif(0,2000)
    b0.w ~ dnorm(0, .35)
    b1.w ~ dnorm(0, .35)
    b2.w ~ dnorm(0, .35)
    inv.sigsq.phi ~ dgamma(.001, .001)
  }
```