Running head: POISSON RACE MODEL

An Application of the Poisson Race Model to Confidence Calibration

Edgar C. Merkle and Trisha Van Zandt

Department of Psychology

Ohio State University

Abstract

In tasks as diverse as making predictions about the stock market to the deliberations of a jury, a person's feelings of confidence in the appropriateness of different choices often impacts that person's final choice. How an estimate of confidence is computed and how accurate it is has been a target of research for many years. Recently-developed mathematical models provide descriptions of psychological processes that may contribute to confidence calibration, which in turn yield predictions of confidence calibration. The current study examines confidence calibration in a simple dual-choice task. The experiments are motivated by an accumulator model in which confidence is modeled in a way similar to Vickers' (1977) balance-of-evidence hypothesis. This model proposes that information supporting each alternative accumulates on separate counters. The observer responds in favor of whichever alternative's counter first hits a designated threshold. Confidence can then be scaled from the difference between the counters at the time that the observer makes a response. We examine the overconfidence result in general and present a new finding: people are less well calibrated under response bias.

An Application of the Poisson Race Model to Confidence Calibration

Judgments of the likelihood of certain events, such as the probability of rain or whether a stock value will go up or down, are made every day, with different degrees of confidence. How confidence is estimated, how it influences decision making, and how accurate a confidence estimate is has implications for real-world problems such as medical diagnoses (Arkes, Dawson, Speroff, Harrell Jr., et al., 1995), jury decision making (Wells, 1981), and financial investing (Wallsten & Budescu, 1983). In particular, researchers have been interested in how people estimate the accuracy of their decisions, or their confidence in having made the appropriate response (Lichtenstein & Fischhoff, 1977; Dawes, 1980; Keren, 1991).

Confidence accuracy is usually described as calibration. Consider an investor whose colleague tells him that there is an 85% chance that stock A will increase in value tomorrow. The investor might interpret the colleague's statement as accurate, or that her confidence level is well calibrated: "The probability that Stock A will go up tomorrow is 0.85". The investor might just as well interpret the statement as inaccurate, or that her confidence is poorly calibrated: "The probability that Stock A will go up tomorrow is greater/less than 0.85". In many situations it is important to know how well calibrated such confidence judgments are. For example, consider the statement, "I am certain that Iraq possesses weapons of mass death."[1] Poorly calibrated confidence estimates may lead to less than optimal or inaccurate decisions.

Research on confidence calibration has focused on simple laboratory decision tasks. The most common task usually consists of two parts: participants first make a response to some two-alternative stimulus (e.g., "Is New York City or Miami farther to the east?"), and they then rate their confidence in this response. Confidence is typically defined as "the belief that one's response is correct," and participants are typically instructed in the

following way:

> If you are absolutely in the dark, no better than flipping a coin, that would correspond to a 50% chance of getting it right. If you are absolutely certain you have the answer, so you think the odds are less than one in a hundred you might be wrong, then that's a 100% chance your answer is right. Please feel free to use any number between 50 and 100 to indicate what you think the chance is that your answer is right. Numbers less than 50 are not allowed, because if you think there's LESS than a 50/50 chance your answer is right, you ought to choose the other answer! (Klayman, Soll, González-Vallejo, & Barlas, 1999)

In these tasks, confidence calibration is determined by comparing confidence rating levels with the proportion of correct responses in each level. If the proportion of correct responses at a particular level is approximately equal to that level, the subject is said to be well-calibrated. If the confidence ratings tend to be greater (less) than the proportion of correct responses at each level, the subject is said to exhibit over- (under-) confidence. For example, a subject might respond to ten different stimuli and rate his confidence in each of his ten responses as 80%. If the subject responds correctly to only 7 of those 10 stimuli, then he would be overconfident because his subjective probability of 0.8 is greater than his objective probability of 0.7.

For general knowledge (or almanac-type) questions, people are generally overconfident: they think that they are correct more often than they actually are (Lichtenstein & Fischhoff, 1977; Wallsten & Budescu, 1983). The extent of overconfidence also seems to depend on task difficulty. This phenomenon is known as the *hard-easy* effect (Lichtenstein, Fischhoff, & Phillips, 1982; Gigerenzer, Hoffrage, & Kleinbölting, 1991) or the *difficulty* effect (Griffin & Tversky, 1992). The hard-easy effect is observed when a

person exhibits more overconfidence for more difficult sets of questions. This effect is pictorially represented in Figure 1, which shows a plot of some hypothetical participants' accuracy as a function of their confidence (calibration curves). The identity line shows perfect calibration, and the two other lines represent hypothetical data. These participants are overconfident because the calibration curves predominantly lie below the identity line, which means that confidence judgments tend to be greater than proportion correct. A hard-easy effect can also be observed in this figure: the dashed line reflects calibration for an easy set of questions, which is higher (better calibrated) than the solid line that reflects calibration for a hard set of questions.

The hard-easy effect has been replicated many times (Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982; Wallsten & Budescu, 1983). For example, in Lichtenstein and Fischhoff's (1977) work, participants answered trivia questions, predicted stock market trends, or predicted the ethnicity of people based upon their handwriting. Following each response, they reported their confidence in responding correctly. It was found that participants were overconfident in all conditions except for very easy tasks, where they were underconfident.

Other important findings concern task differences and cultural differences in overconfidence. Dawes (1980) argued that people may be underconfident in sensory tasks of a psychophysical nature. While his findings and more recent findings (Keren, 1988; Olsson & Winman, 1996; Baranski & Petrusic, 1999) remain inconclusive, it appears that people may at least be less overconfident in sensory tasks than in general knowledge tasks. Furthermore, task details have an effect on confidence calibration. The act of choosing an alternative has been shown to have an effect on confidence (Sniezek, Paese, & Switzer III, 1990), as well as the act of writing reasons why each alternative may be correct (McKenzie, 1997). Finally, calibration differences exist between cultures. A higher degree of overconfidence is generally reported for Asian participants than for American

participants (Yates, Lee, & Bush, 1997).

The preceding paragraphs highlight many results of research on confidence calibration. Recent mathematical models of confidence have been formulated and applied to simple choice tasks, but none have been applied to classic data from calibration studies. In this paper, we apply a Poisson race model to the study of confidence calibration. This model has been previously applied to psychophysical discrimination tasks (Pike, 1971, 1973) and recognition memory tasks (Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004), where it has successfully accommodated response time, accuracy, and confidence. The main purpose of the current study is to determine whether the Poisson race model has a place in modeling confidence calibration.

We start by reviewing other mathematical models of confidence, including the decision variable partition model (Ferrell & McGoey, 1980), the ecological model (Gigerenzer et al., 1991), and the class of stochastic error models (Erev, Wallsten, & Budescu, 1994; Wallsten & González-Vallejo, 1994; Soll, 1996). Next, we describe the Poisson race model in detail and present analytic expressions for the model's predictions. Finally, we apply the model to the results of three calibration experiments. We conclude that the Poisson race model should receive serious consideration as a model of confidence calibration.

The Decision Variable Partition Model

The decision variable partition model (Ferrell & McGoey, 1980) is one of the earlier models of choice confidence. It explains confidence ratings in two-alternative knowledge tasks using a random utility model representation of the choice between alternatives A and B. On a given question, the two alternatives are each assumed to generate some feeling of correctness sampled from a Gaussian distribution. The decision maker chooses the alternative yielding the greater feeling of correctness.

Following selection of an alternative, the decision maker renders a probability judgment by assessing the difference between the feeling of correctness for alternative A and the feeling of correctness for alternative B. It is assumed that the decision maker partitions this difference such that higher differences correspond to higher probability judgments. The hard-easy effect is explained by the difficulty that subjects have in accurately fitting their partitions to the differences in feelings of correctness. Overall, this model was found to fit experimental data reasonably well. Suantak, Bolger, and Ferrell (1996) also found support for this model through both experiment and simulation.

Criticisms of the decision variable partition model mainly revolve around the fact that there are few parallels between the model and the cognitive processes underlying confidence (Keren, 1991). The model explains miscalibration and the hard-easy effect in terms of poor partitions on feelings of correctness, but it is mute on the processes that establish the partitions or cause them to change.

The Ecological Model

Employing Brunswik's (1955) theory of ecological behavior, the ecological model (Gigerenzer et al., 1991) focuses on aspects of the environment that influence decisions. Specifically, it assumes that subjects use pertinent environmental cues to make decisions. Probability judgments are based on the validities of the specific cues used to make the decision. For example, a typical question asks a subject which of two cities ("A" or "B") has a larger population. Assuming that she does not factually know the answer to this question, the subject might select city "A" by realizing that city "A" has a professional baseball team and that city "B" does not. If this "sports team" cue has led to the correct answer 70% of the time in the past, the subject reports her confidence as 70%.

The method by which the ecological model explains calibration effects is considerably different from that of the decision variable partition model. The ecological

model explains calibration errors as occurring when the experimental question set is not representative of the decision maker's past experience. If the experimenter *a priori* selects questions for which common cues fail, then overconfidence will be observed. Hard-easy effects are explained in a similar manner: the experimenter selects as difficult questions ones where the usual cues tend to fail, and she selects as easy questions ones where the usual cues tend to work. A higher degree of overconfidence is thus observed for the set of difficult questions.

Evidence against this model has come from studies such as those of Suantak et al. (1996) and Griffin & Tversky (1992) in which miscalibration was observed for randomly-generated questions. Extensions to perceptual tasks are also difficult to establish for this model.

Stochastic Error Models

Error models have recently gained considerable popularity in the area of confidence calibration. Instead of taking the view that miscalibration is a cognitive bias, these models propose that random error accounts for the majority of observed calibration effects.

Erev, Wallsten, & Budescu (1994) proposed a general class of models incorporating error into probability judgments. In this class, a decision maker's overt confidence judgment depends on both a true judgment and a random error component. The decision maker's true judgment is assumed to be accurate, so it is only random error that causes miscalibration. Erev et al. showed via simulation that this class of models can predict typical overconfidence effects. More specific models within this class have been proposed in the forms of a stochastic judgment model (Wallsten & González-Vallejo, 1994) and an ecological error model (Soll, 1996).

Criticisms of these models include the fact that they typically do not specify a mechanism by which error enters into the response process, nor do they propose a

mechanism whereby the true judgment arises. Furthermore, random error by itself is unable to predict general underconfidence (Suantak et al., 1996).

The Poisson Race Model

Like the models above, the Poisson race model explains over- and underconfidence by the introduction of error into the response process. However, the mechanisms by which error occurs are explicit in this model. Furthermore, the Poisson race model is not specifically targeted at general knowledge tasks (like the ecological model), but at the response selection process more generally. Finally, unlike other models described above, the Poisson race model provides a plausible account of the process giving rise to confidence and yields response time predictions.

Model Overview

In accordance with the experimental tasks to which it has previously been applied, the Poisson race model is primarily geared toward two-alternative tasks. The general idea of the model is that information favoring each response alternative (for example, "A" and "B" in a two-alternative forced-choice task) is stored on a "counter," a mechanism upon which evidence accumulates. Figure 2 shows the levels of information on each counter (denoted $X_H(t)$ and $X_L(t)$) accumulating over time. The counters race to prespecified levels of accumulation (thresholds; denoted K_H and K_L), and the counter that reaches its threshold first wins and determines the response. The time at which the winning counter reaches its threshold determines the response time RT.

For mathematical tractability, the Poisson race model assumes that information accumulates on each counter in discrete units, much as if a decision maker samples the stimulus in discrete packets and classifies each packet as supporting either alternative A or alternative B. A point is added to alternative A's counter each time a packet is judged in

support of alternative A, and vice versa. The decision maker responds in favor of the alternative whose counter first accumulates a prespecified number of packets.

The model makes response time predictions by assuming that the time lag between the accumulation of each packet of evidence is exponentially distributed. The rates, λ_i, at which stimulus information arrives in support of each alternative are different (say, λ_A versus λ_B). For example, two easily-discriminable stimuli might have a high rate of incoming information supporting alternative A and a small rate of incoming information supporting alternative B. The observer samples stimulus packets from the incoming stream according to a Poisson process with parameter $\lambda_A + \lambda_B$. For each packet that arrives, the probability that the packet is judged in favor of alternative A (and thus, the probability that a point accumulates on alternative A's counter) is equal to $\lambda_A/(\lambda_A + \lambda_B)$ (Karlin, 1966; Townsend & Ashby, 1983). The conditional response time given the number of packets to arrive on each counter is then the sum of the exponentially-distributed packet arrival times until one counter hits threshold.

Confidence judgments, the focus of this paper, are modeled based on Vickers' (1979) *balance-of-evidence hypothesis*. Vickers originally proposed that confidence was scaled directly from the difference between the levels of evidence at the time the winning counter reaches its threshold. For example, if both counters' thresholds (K_A and K_B) are set at 5, and Counter A wins at time RT with Counter B having accumulated 2 units of evidence, then the subject's confidence should be related to the value 5-2=3. If the subject is asked to render his judgment on a 0-100 scale (as is typical), troublesome scaling issues arise. Under what constraints is the value 3 scaled to, say, 80? To solve this problem, we modified the absolute balance of evidence formulation (5-2=3) to a relative balance of evidence[2]:

$$Confidence = \frac{K_{win}}{K_{win} + X_{lose}(RT)}, \tag{1}$$

where K_{win} is the threshold of the winning counter and $X_{lose}(RT)$ is the level of evidence

on the losing counter at the time RT of the response. If the threshold $K_{win} = 5$ and $X_{lose}(RT) = 2$, then confidence is $5/(5 + 2) = .71$.

Confidence will depend on the rates of incoming information for each alternative (λ_A and λ_B) and on the thresholds of the counters (K_A and K_B). If λ_A and λ_B are approximately equal, then points should accumulate on both counters approximately equally often and confidence should, on average, be low (.5). If λ_A and λ_B are very different, then points should accumulate on one counter faster than the other and confidence should, on average, be higher. It should be noted here that the thresholds K_A and K_B are not required to be equal. This fact leads to an important new prediction: calibration accuracy will depend on how the thresholds are set. In other words, confidence calibration should change based on the decision maker's response biases.

Our goal in this paper is to examine the Poisson race model's ability to account for classic findings within the area of confidence calibration. We specifically wish to see whether the model can account for miscalibration and the hard-easy effect. This will be accomplished by conducting perceptual judgment experiments with difficulty manipulations. We also wish to test the Poisson race model's novel prediction of calibration effects under response bias.

Analytic Expressions of the Model

To model confidence calibration, we must derive the model's conditional accuracy level given the confidence it predicts. That is, for a certain range of confidence levels, the model will have produced the correct response some proportion of the time. We must compute that proportion and determine how it depends on the model parameters. There are two sets of parameters that are important: those that are determined by task difficulty, and those that reflect response bias.

In the current experiments, decision makers must discriminate between two

distributions of dots: a "high" distribution and a "low" distribution. The high distribution tends to output a large number of dots to the screen, and the low distribution tends to output a smaller number of dots to the screen. For each trial, an array of dots is presented and decision makers give confidence judgments as to which distribution output the array. Within these experiments, task difficulty (stimulus discriminability) is manipulated by varying the amount of overlap between the two distributions. More detail about the experiments will be provided later, but for now we will focus on how the parameters of the model are influenced by the experimental paradigm.

On any given trial in the experiment, participants observe a cluster of dots and make a high or low decision. The model must specify how the number of dots on the monitor maps into information accumulation rates λ_H and λ_L for the two alternatives (high and low). To do this, we assume that a particular number of dots gives rise to some perceived degree of numerosity (X) and that the means of the high and low numerosity distributions are symmetric around 0 (i.e., $\mu_X(\text{High}) = -\mu_X(\text{Low})$). The point 0 represents ambiguous numerosities that do not clearly distinguish a high from a low number of dots. For mathematical convenience, we also assume that the two rates sum to a constant r, so that we need only consider either λ_H or λ_L. A useful transformation is provided by the logistic function, given as

$$\lambda_H = r\left(1 + \exp\{-X\}\right)^{-1}.$$

This relationship states that as X becomes large, λ_H approaches a maximum rate of r (and λ_L approaches 0). As X becomes small, λ_H approaches 0 (and λ_L approaches r). If $X = 0$, both λ_H and λ_L equal $r/2$.

Two probability distributions for λ_H are shown in Figure 3 for $\lambda_H + \lambda_L = 0.01$. The distributions reflect the different values of λ_H that will be obtained under two levels of discriminability between high and low stimuli. Discriminability is represented as a change

in μ, half the distance between the high and low distributions of perceived numerosity. Higher values of μ give wider separations between the high and low numerosity distributions. In Figure 3, the solid curve reflects high discriminability (a large value of μ) and the dashed curve reflects low discriminability (a small value of μ). It can be seen that, for a large value of μ, the rates in support of the high process are generally large (near the maximum of 0.01). For a smaller value of μ, the rates in support of the high process are more spread out and closer to the middle. In the latter case, the rates supporting the high process and the rates supporting the low process are much closer together, reflecting a more difficult task. Using the variable V as an indicator that equals 1 when a stimulus arises from the high distribution and 0 otherwise, the distribution of rates for the high alternative will be denoted $g(\lambda \mid V = 1)$ and the distribution of rates for the low alternative will be denoted $g(\lambda \mid V = 0)$ (see Appendix for more details).

Analytic expressions for the Poisson race model's confidence process, which incorporate the relative balance of evidence hypothesis, are derived in the Appendix. Given that a high response is made, the amount of evidence on the losing counter has a negative binomial distribution with parameters K_H and $\lambda_H/(\lambda_H + \lambda_L)$. Based on this negative binomial distribution, denote the probability that confidence is between levels a and b as $f_H(a, b \mid \lambda)$. We can then show that the calibration curve is given by:

$$P(Correct \mid a \leq Conf < b) =$$

$$\frac{\int_0^r [f_H(a, b \mid \lambda) g(\lambda \mid V = 1) + f_L(a, b \mid \lambda) g(\lambda \mid V = 0)] d\lambda}{\int_0^r [f_H(a, b \mid \lambda) + f_L(a, b \mid \lambda)][g(\lambda \mid V = 0) + g(\lambda \mid V = 1)] d\lambda}. \quad (2)$$

In fitting the Poisson race model, we introduce a slight modification to Equation 2. There is now considerable evidence (e.g., Ratcliff & Smith, 2004) that threshold variability is an important component of accumulation processes such as the Poisson race model. For our purposes, we assumed that thresholds had some small probability (.25) of stepping up or down one increment from their set values. Thus, if the high counter has a threshold set

at 3, with probability .125 on a particular trial it might change to 2. Or, with probability .125, it might change to 4. An example of this distribution is shown in Table 1, where the favored modal threshold value is 3 and the unfavored modal threshold value is 5. A side benefit of variable thresholds is that a wider range of confidence judgments may be observed at lower threshold values.

The threshold distributions slightly change the equation for accuracy as a function of confidence (Equation 2). Since there are 3 possible threshold values for the favored response and 3 possible threshold values for the unfavored response, there are now 9 combinations of threshold values that may be observed. Substituting each of these threshold combinations into Equation 12 in turn, we arrive at 9 different values for accuracy as a function of confidence. In order to arrive at an overall accuracy across the different thresholds, we simply multiply each of the 9 accuracy values by the probability that they occur:

$$P(Correct \mid a \leq C < b) =$$
$$\frac{\sum_k P(Correct \mid a \leq C < b, K = k) P(a \leq C < b \mid K = k) P(K = k)}{\sum_k P(a \leq C < b \mid K = k) P(K = k)}, \tag{3}$$

where K is a vector containing the two threshold parameters and k is a numerical realization of K. Equation 3 is the expression for the calibration curve predicted by the Poisson race model and was fit to the data collected in our experiments.

Experiments

The task that we used required participants to discriminate between two possible systems that generated random patterns of dots (Ratcliff, Van Zandt, & McKoon, 1999). One system tended to produce a large number of dots on a computer monitor and the other tended to produce a smaller number of dots on a computer monitor. Furthermore, due to the fact that the distributions of the number dots produced by each system overlapped, there is no response scheme guaranteed to yield 100% accuracy. The task is,

in essence, a signal detection task, which at its heart is representative of many other simple cognitive decisions, such as psychophysical discrimination, recognition, perceptual matching and lexical decision.

The reason that we chose this particular task was because it lends itself well to modeling the response selection process. As shown by Ratcliff et al. (1999) for the diffusion process, and as we will demonstrate below, the number of dots presented maps nicely into the rate of information accumulation, allowing us to explicitly control accumulation rate in our modeling efforts. Lee and Janke (1964) called this procedure an "externally distributed stimulus" paradigm, which has since been used for similar reasons by many others (Ashby & Gott, 1988; Espinoza-Varas & Watson, 1993; Maddox & Bohil, 2003; Smith & Vickers, 1988; Vickers, 1979; Vickers, Caudrey, & Willson, 1971).

This task is one of perceptual categorization. The observer must categorize random outputs from each process (high and low) based on the perceived numerosity of dots in the display. While the task relies on visual input, there is also a significant learning component: over the course of many trials, the observer must learn the distributions of numerosity produced by the two processes. Generally, participants learn to do this very quickly and reach asymptotic performance in relatively few trials (Ratcliff et al., 1999). This task is somewhat different from other tasks that have been used to study confidence calibration. It is not a general knowledge task, nor is it perceptual in the sense of Baranski and Petrusic's previous research (Baranski & Petrusic, 1994, 1999). Participants in those experiments made judgments which relied heavily on visual input (such as judging which of two lines is longer), and underconfidence was often observed in place of overconfidence [3]. To preview our results, our participants were generally overconfident in their classification judgments, consistent with previous research using general knowledge paradigms.

There are two major aims of our study. As we previously stated, one goal was to

examine whether the Poisson race model could be extended to account for confidence calibration in a simple perceptual judgment task. A first step toward this aim was to study the model's ability to fit observed miscalibration and the hard-easy effect (Experiments 1 and 2). Second, we wished to test a new prediction of the model: bias to respond in favor of a specific alternative affects calibration (Experiment 3). Response bias is incorporated into the model by allowing response thresholds to vary, where one threshold may be set much lower than the other. Unequal thresholds require that the decision maker accrue a larger amount of evidence for an unfavored response alternative than for the favored alternative. Response bias is produced experimentally by manipulating the payoffs that decision makers receive for correct responses. It is assumed that unequal payoffs result in unequal thresholds. The model predicts more overconfidence for favored responses than for unfavored responses.

In the experiments, the number of dots presented in each pattern is randomly drawn from one of two normal distributions with different means (one high and one low). The closer the means are, the more difficult it is to distinguish between dot patterns output by the high process from those output by the low process. If a hard-easy effect obtains for the perceptual categorization task, we should be able to observe it by comparing calibration for a difficult task (with means close together) to calibration for an easy task (with means farther apart).

In terms of the model, decreasing stimulus discriminability results in more equal rates of information accumulation for the two response alternatives. When the two rates of information accumulation are nearly equal, points are nearly equally likely to accumulate on either counter. This means that, across the entire confidence range, proportion correct will decrease. A greater number of confidence responses will tend to be close to 0.5, but the confidence responses close to 1 are as likely to be wrong as they are to be right (e.g., a large amount of evidence is as likely to accumulate on the incorrect

counter as on the correct counter). When the two rates of information accumulation are unequal, evidence is more likely to accumulate on the correct counter than on the incorrect counter. As a result, confidence responses close to 1 are likely to be correct. Overconfidence will be greater for the former case than for the latter case.

General Methods

Apparatus

Stimuli were presented on a Viewsonic 6 video monitor. An IBM PC/SX computer controlled event sequencing and the recording of data. Subjects made confidence responses on a standard IBM keyboard using the bottom row of letter keys ('Z' through '/'). Subjects were instructed to place their index fingers on the 'B' and 'N' keys and to place their other fingers on corresponding keys.

Stimuli

Stimuli consisted of dot patterns presented in the center of the computer screen. The number of dots presented in each pattern was randomly drawn from one of two normal distributions, one with a mean less than 50 and one with a mean greater than 50. After the number was chosen at random from the appropriate (low or high) distribution, that number of dots was placed in uniformly random locations within a 10x10 array.

Procedure

Each trial began with the presentation of a dot pattern on the screen, after which subjects were required to respond how sure they were that the dot pattern was produced by the high process. The bottom row of keys on the keyboard were labeled to correspond to probability judgments of %5, 15%, ..., and 95% "high." Participants were told that judgments above 50% would be considered as supporting the high distribution and that judgments below 50% would be considered as supporting the low distribution. No 50%

judgment was available on the keyboard so that participants were forced to choose between the two distributions. One point was awarded for each correct response, and the participant with the most points at the end of the academic quarter received a $20 gift certificate to a local electronics store.

Participants were provided with trial-by-trial feedback on the accuracy of their responses (no message appeared when they were correct; "Wrong" appeared when they were incorrect). Following the feedback, if any, a 200-millisecond pause preceded the presentation of the next dot pattern. Participants were told to respond "quickly while still being accurate."

Experiments 1 and 2

The first two experiments examined the effects of difficulty on calibration by manipulating the means of the experimental distributions controlling number of dots (high and low). Consistent with previous research on confidence calibration, we expected a hard-easy effect showing more overconfidence for more difficult tasks. Two experiments were conducted in order to examine the model's ability to account for overconfidence under varying levels of difficulty. Specifically, we expected to see the μ parameter of the model (estimated from fits to the data) decrease as experimental difficulty increased.

Specific Methods

Participants. In Experiment 1, thirty-nine Ohio State University undergraduates participated for one 45-min. session in return for course credit. In Experiment 2, forty Ohio State University undergraduates participated for one 45-min. session in return for course credit. All participants had normal or corrected-to-normal vision and were naive to the aims of the experiment.

Stimuli. The distributions of dot numerosity for each experiment are presented in Table 2. Participants were presented first with 450 easy trials and next with 450 hard trials corresponding to the rows of Table 2.

Results

Trials on which response times were less than 200 msec or greater than 10000 msec were excluded from analyses (a total of 6.5% of the trials). Two participants were completely excluded from analyses because they did not follow instructions.

Confidence levels were transformed from a probability judgment about the "high" process to a probability judgment about the selected response ("high" or "low"). The rankings were separated according to whether the response was less than or greater than 50%. All "low" (less than 50%) responses were transformed by subtracting from 100%, so that, for instance, a judgment that the probability of the high process having generated an output was 5% was transformed into a judgment that the probability of the low process having generated the output was 95%. Thus, the probability judgments were collapsed across low and high responses onto a single scale ranging from 55% to 95%, for which accuracy ranged from 50% to 100%.

For the 5 confidence ratings (55%, 65%, 75%, 85%, 95%), accuracy levels were computed for each participant. The resulting accuracies were averaged across participants to produce the calibration curves shown in Figure 4. The solid lines with error bars are the observed data, and the dashed lines are predictions of the fitted model. The solid line with circles represents data from the easy condition, and the solid line with squares represents data from the hard condition.

Calibration. The calibration curve in Figure 4 exhibits some very typical results (Lichtenstein et al., 1982). If participants were well-calibrated, then their responses would lie exactly along the identity line of the graph: their estimated confidence would equal

their accuracy. Because participants' responses lie predominantly below the identity line, confidence is almost always greater than accuracy: an overconfidence effect. Examining overconfidence statistically, we compare average confidence with average accuracy at each level of difficulty. Participants are significantly overconfident in both the easy condition ($t(36) = 5.29, p < .001$) and in the hard condition ($t(36) = 8.16, p < .001$).

The above t-tests reflect an overall measure of overconfidence; to compare overconfidence at specific difficulty levels, a 5 (confidence level) by 2 (difficulty) repeated-measures ANOVA was conducted with overconfidence as the dependent variable. The hard-easy effect showed up as a main effect of difficulty ($F(1, 28) = 47.648, MSE = .010$). There was also a confidence level $\times$ difficulty interaction ($F(4, 112) = 3.587, MSE = .013$), indicating that the hard-easy effect changed as a function of confidence level. The effect is least pronounced at the 0.55 confidence level and tends to increase with confidence level (see Figure 4).

Similar analyses were conducted on the data from Experiment 2. As shown in Figure 5, the overconfidence effect was replicated for both the easy and hard conditions ($t(39) = 4.55, p < .001$ and $t(39) = 9.18, p < .001$, respectively). The hard-easy effect was also replicated ($F(1, 30) = 137.303, MSE = .011$), along with the interaction between confidence level and difficulty ($F(4, 120) = 8.844, MSE = .007$).

Model Fits. The Poisson race model was fit to the curves shown in Figures 4 and 5 by nonlinear least squares, in which the squared difference between the observed and predicted calibration curves was minimized. The predicted curve was computed using Equation 3. There were two completely unrestricted parameters in the model: μ and σ. The two response thresholds were assumed to be equal in the absence of any strong response bias, so a single parameter K represented the average threshold for each response. As described above, the rates of information accumulation were assumed to sum to 0.01 for all fits (i.e., $r = 0.01$).

Experiment 1 and Experiment 2 were very similar to one another, with difficulty being the only difference. Thus, the only parameter that was allowed to vary between the two experiments was the mean separation between the high and low processes, μ. All other parameters were held constant across the two difficulty conditions and the two experiments.

The dotted lines in Figures 4 and 5 are the Poisson race model's fits to the observed data. The model provides excellent fits to the calibration results that we have observed; in terms of goodness of fit tests, we obtained $\chi^2(9) = 7.62, p > 0.5$ for Experiment 1 and $\chi^2(9) = 19.82, p < 0.02$ for Experiment 2. These statistics confirm the good fit of the Experiment 1 data and bring into question the fit of the Experiment 2 data. A residual analysis reveals that the largest contributor to the Experiment 2 χ^2 statistic is the "0.55, easy" point. This point is also the most visually distinctive in terms of lack of fit. Aside from this point, the overall fit of the Experiment 2 data is excellent.

Examining these fits in more detail, the Poisson race model is able to accommodate the larger hard-easy effect in Experiment 2. This is reflected in the μ_H parameter, the mean separation for the "hard" conditions. The value of this parameter is 0.0854 for Experiment 1 and 0.0398 for Experiment 2. Therefore, this single parameter captured the increased difficulty of the second experiment. Because the "easy" conditions were the same in both experiments, we expected that the value of the μ_E parameter, the mean separation for the "easy" conditions, would be nearly the same. The value of this parameter is 0.1589 for Experiment 1 and 0.1490 for Experiment 2, confirming our general expectations.

Discussion

Experiments 1 and 2 accomplished two goals. First, we verified that we are able to obtain the "standard" calibration results in a simple laboratory task of perceptual classification. Participants demonstrated overconfidence: their confidence ratings were

generally higher than their accuracy at each confidence level. Furthermore, a hard-easy effect was observed. When the high and low processes were made less discriminable by decreasing the distance between the means of the numerosity distributions, overconfidence increased.

The Poisson race model was able to fit the data from these two experiments well. Not only did the model show overconfidence, but a hard-easy effect was obtained when accumulation rates were made more equal. The fits of the model were obtained across different conditions by way of changes in a single parameter, the mean distance between the low and high distributions of perceived numerosity. It is noteworthy that the model was able to perform so well by changing this single parameter, and that this parameter represents (theoretically) the changes in difficulty across experimental conditions.

Experiment 3

The experiments up until now have focused solely on difficulty; we wanted to establish that overconfidence and the hard-easy effect can be obtained in a simple choice task, and we also wanted to establish that the Poisson race model can accommodate these classic effects. We noted earlier that the model makes a novel prediction: under conditions of response bias, calibration should be more accurate for unfavored than for biased responses. The purpose of Experiment 3 was to test the model by examining calibration under response bias.

Bias was influenced by manipulating payoffs for correct responses. In the previous two experiments, participants always received one point for each correct response. In Experiment 3, participants received different numbers of points for correct responses at different times. These differential payoffs bias participants to respond in favor of one of the alternatives, which is modeled by changes in the response thresholds for each alternative.

Specific Methods

Participants. Forty Ohio State University undergraduates participated for one 45-min. session in return for course credit. All participants had normal or corrected-to-normal vision and were naive to the aims of the experiment.

Stimuli. For all trials in Experiment 3, the numerosities of "high" and "low" dot patterns were drawn at random from the same distributions as the easy trials in Experiments 1 & 2 ($\mu = 45$ low and $\mu = 55$ high). Experiment 3 was a within-subjects design, where participants responded to 300 trials under each of 3 different point schedules. The point schedules were (2,2), (4,0), and (0,4), respectively, where (x, y) means that the participant received x points for a correct low response and y points for a correct high response.

Procedure. The procedure was identical to Experiment 1 except that participants were also informed that the points they receive for correct responses would change from time to time. Participants were informed before each block of trials what the payoff schedule would be.

Results

Three participants were excluded from analyses for not following directions. Using the point system just described, stimuli were categorized as favored or unfavored alternatives. Favored alternatives were those for which participants received 4 points for a correct response, and unfavored alternatives were those for which participants received 0 points for a correct response. Following the procedures used in the analyses for Experiments 1 and 2, confidence rankings were transformed to reflect confidence on a 55% to 95% scale.

Calibration. Because the model makes different predictions for favored vs. unfavored alternatives (e.g., when threshold values are unequal), we calculated a separate calibration curve for each alternative (high, low) by type (favored, unfavored) combination. Accuracy was computed for each level of reported confidence for each participant, and the accuracies were averaged across participants to give the calibration curves shown in Figures 6 and 7. Figure 6 shows calibration curves under the "low favored" condition, while Figure 7 shows calibration curves under the "high favored" condition.

Collapsing over the two figures, we observe overconfidence for both favored ($t(36) = 12.62, p < .001$) and unfavored responses ($t(36) = 6.52, p < .001$). There is also better overall calibration for the unfavored responses than for the favored responses ($t(36) = 4.48, p < .001$), confirming the qualitative prediction made by the race model.

Better calibration for unfavored responses can be explained by examining how the model makes more or less accurate responses. If participants raise the threshold of the unfavored alternative, a larger amount of evidence will be required to make this response. These responses will therefore tend to be more accurate. In contrast, participants could lower the threshold of the favored response, requiring less evidence to make this response. These responses will therefore tend to be less accurate. It is the contrast between more and less accurate responses that influences the degree of calibration shown in Figures 6 and 7.

Model Fits. Similar to Experiments 1 and 2, the model was fit to the data using nonlinear least squares. The fits are shown as the dashed lines in Figures 6 and 7. Because bias was the only thing varying among the two types of responses (favored, unfavored), the μ and σ model parameters were held constant across conditions while the threshold parameters were allowed to vary. A separate threshold parameter was varied for the 4 types of responses (low favored, low response; low favored, high response; high favored, low response; and high favored, high response).

While the fits of the model do not look as nice as those in Experiments 1 and 2, the fitted lines do not match the observed lines nearly as well. Despite the fact that the model does not provide the best fit, the goodness of fit statistic $\chi^2(18) = 14.11, p > 0.5$. As in the previous fits, most of the discrepancies are due to larger than predicted accuracies for confidence ratings of 0.55.

Discussion

The model predicted that unfavored responses should be better calibrated than favored responses. This prediction was confirmed in Experiment 3. Calibration was better for responses receiving a payoff of 0 points than for those receiving a payoff of 4 points.

In terms of goodness-of-fit, the model's ability to fit the calibration curves was not as visually impressive as in the first two experiments. There are a number of reasons why this might have occurred. For example, the changing point systems throughout the experiment may have caused participants to focus more on responding "high" or "low' than on responding with accurate levels of confidence. There are also issues with the model itself: the Poisson race model is, of course, an oversimplification of a very complicated mental process. Even though the quantitative fits of the model are slightly disappointing, the model accurately predicted the relative levels of calibration for the favored and unfavored responses. While we can tweak the parameters of the model in more elaborate ways in an attempt to improve the fits, it is the qualitative predictions made by the model that will ultimately measure its success or failure.

General Discussion

In this paper, we first reviewed both general findings in confidence calibration research and specific models geared toward explaining calibration data. Models under consideration included the decision variable partition model, the ecological model, and the class of error models. Shortcomings of these models, including the inability to account for

some empirical findings, were noted, and the Poisson race model was then presented as an alternative to the other models. We conducted a series of perceptual categorization experiments to examine whether the Poisson race model can account for typical calibration data and to study a novel model prediction of response bias effects on calibration. Overconfidence and a hard-easy effect were observed in our experimental paradigm, and the model successfully accommodated these findings. The model's prediction of response bias effects was also supported by the data. Overall, our data were consistent with the Poisson race model.

The Poisson race model has some advantages over other models of confidence. First, it describes a plausible process by which error enters into responses. While such processes prove difficult to verify (e.g., Luce, 1995), the model may help us understand the psychological mechanisms underlying confidence elicitation. Furthermore, the fact that the Poisson race model depicts a psychological process gives its parameters convenient interpretations. For example, the setting of unequal threshold parameters in the model is interpreted as the introduction of response bias in an individual. Response bias may then be experimentally induced to examine the model's performance. Compare this parameter interpretation with parameter interpretation of, say, the decision variable partition model (Ferrell & McGoey, 1980). Based on a general signal detection model, the decision variable partition model yields confidence judgments via placement of partitions on the signal and noise distributions of the signal detection model. We may elect to change the decision variable partition model's partition parameters, which will in turn change the predicted confidence calibration. It is hypothesized that difficulty manipulations cause partition parameters to change, but this is only because both difficulty manipulations and partition parameters have an effect on calibration. There is no psychological mechanism that explains why difficulty should cause a change in the partitions.

Calibration Mechanisms. Given that parameters of the Poisson race model have a psychological interpretation, what exactly causes the model to produce overconfidence or underconfidence? In an attempt to answer this question, we explored the effects of individual model parameters on confidence calibration by systematically varying individual parameters while holding the others constant. It appears that the μ parameter is the largest determinant of calibration magnitude and direction (e.g., over-/underconfidence): when the μ parameter is large, as for easily-discriminable stimuli, the model tends to yield underconfidence. When the μ parameter is small, as for indiscriminable stimuli, the model tends to yield overconfidence. These findings are consistent with the hard-easy effect, and the manner in which the model arrives at its predictions is interesting: regardless of the magnitude of the μ parameter, the model always produces a range of confidence judgments between 0.5 and 1. Because the μ parameter is interpreted as task discriminability, we might ultimately think of the μ parameter as determining task performance (e.g., accuracy). When the μ parameter is large, we will tend to see high accuracy levels across the $[0.5, 1]$ confidence range. As a result, we observe underconfidence in the calibration curve. The opposite occurs as well: when the μ parameter is small, we will tend to see low accuracy levels across the confidence range. Overconfidence is the result.

We can similarly explain the calibration differences under response bias predicted in Experiment 3. Our key assumption was that experimental response bias caused model thresholds to be unequal; the threshold for the favored alternative would be set less than the threshold for the unfavored alternative. As a result, an unfavored response would be made only if there was a large amount of evidence supporting the favored response and a small amount of evidence supporting the unfavored response. This leads to higher accuracy across the confidence range for unfavored responses as compared to favored responses. Thus, the accuracy difference yields greater overconfidence for favored responses as compared to unfavored responses.

In summary, confidence calibration is determined more by an individual's accuracy level than by an individual's confidence level. If decision makers use the entire confidence range afforded to them, changes in accuracy across conditions exert a larger influence on calibration than changes in confidence. Experimental conditions promoting high accuracy will result in underconfidence, while conditions preventing high accuracy will result overconfidence.

Random Error in Confidence Judgments. There has been some debate about the extent to which overconfidence and the hard-easy effect are due to random error instead of to cognitive biases. Erev, Wallsten, & Budescu (1994) demonstrated that the perturbation of accurate confidence judgments by random error was enough to produce a standard overconfidence effect. Juslin, Winman, & Olsson (2000) went on to claim that the hard-easy effect was a methodological artifact and that there was little evidence of cognitive bias as a contributor to overconfidence. While we discussed the hard-easy effect above, some clarification may be made here about random error in confidence judgments.

Error exists in all models of confidence calibration. Models such as those used by Erev et al. (1994) specify pure random error, whereas models such as the Poisson race model specify a mechanism by which error enters into confidence judgments. It may be therefore argued that, for all models under consideration, overconfidence arises as a result of error. Continuing with this viewpoint, models are simply vehicles by which error enters into confidence judgments. This view neglects the existence of cognitive biases: error causes overconfidence, regardless of whether the error originates from purely random factors or from a cognitive bias. It is therefore necessary to consider the locus of error and to avoid sweeping statements about overconfidence being due to error and not to cognitive biases. The specification and testing of process models of confidence provide a starting point for determining the nature of error.

The preceding discussion highlights the fact that current definitions of error and of

cognitive biases are unclear. For the overconfidence phenomenon, a loose definition of "cognitive bias" might be "a mental mechanism causing people to think that they know more than they actually do know". The fact that people are often overconfident cannot verify the existence of a cognitive overconfidence bias, and other explanations for overconfidence definitely exist. Erev et al.'s (1994) explanation was that of random error, while a second, related explanation is that people are simply not good at attaching a number to an internal feeling of confidence (e.g., Juslin, Olsson, & Björkman, 1997). It is difficult to know where a cognitive bias ends and where random error begins. Each explanation assumes that bias or error occurs without the individual's conscious awareness, and each explanation is capable of yielding overconfidence.

With respect to the Poisson race model in particular, "cognitive bias" may be defined as "the inability to appropriately scale confidence judgments from stimulus information". Because the Poisson race model yields overconfidence, such a definition supports the existence of a cognitive overconfidence bias in the Poisson race model. The fact that accuracy has a significant impact on confidence calibration brings into question the existence of a cognitive bias, however: people are not necessarily overconfident because they think that they are correct all the time, but because the baseline accuracy level for a given task is low. Perhaps the real cognitive bias is that people tend to give similar confidence judgments for any task, regardless of difficulty.

Conclusions. The Poisson race model has accounted for classical confidence calibration effects, predicted a novel calibration effect, and shed light on factors giving rise to over- and underconfidence. The fact that the model also yields response time predictions means that the model has potential to merge research on response times with research on confidence calibration. On a larger scale, the notion of a cognitive overconfidence bias has been explored in some detail. It is important to recognize that accuracy levels can have a large impact on overconfidence. We must control for this fact if

we wish to understand cognitive biases in judgment, and we must be more specific in defining “cognitive bias”.

References

Arkes, H. R., Dawson, N. V., Speroff, T., Harrell Jr., F. E., et al. (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making*, *15*, 120-131.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33-53.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, *55*, 412-428.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception and Psychophysics*, *61*, 1369-1383.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193-217.

Dawes, R. M. (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of clyde coombs* (p. 327-345). Bern: Han Huber.

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.

Espinoza-Varas, B., & Watson, C. (1993). Effects of decision criterion on latencies of binary decisions. *Perception and Psychophysics*, *55*, 190-203.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes*, *26*, 32-53.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411-435.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189-209.

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.

Karlin, S. (1966). *A first course in stochastic processes.* New York: Academic Press.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, *67*, 95-119.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Psychometrika*, *77*, 217-273.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216-247.

Lee, W., & Janke, M. (1964). Categorizing externally distributed by stimulus samples for three continua. *Journal of Experimental Psychology*, *68*, 376-382.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? the calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The

state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 306-334). Cambridge, England: Cambridge University Press.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*, 1-26.

Maddox, W. T., & Bohil, C. J. (2003). A theoretical framework for understanding the effects of simultaneous base-rate and payoff manipulations on decision criterion learning in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

Olsson, H., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Perception and Psychophysics*, *58*, 374-382.

Pike, R. (1971). The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on simple counting. *Perception and Psychophysics*, *9*, 455-460.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*, 53-68.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333-367.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Comparing connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261-300.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*, 135-168.

Sniezek, J. A., Paese, P. W., & Switzer III, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, *46*, 264-282.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*, 117-137.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201-221.

Taylor, H. M., & Karlin, S. (1998). *An introduction to stochastic modeling* (3rd ed.). Chestnut Hill, MA: Academic Press.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.

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

Van Zandt, T., & Maldonado-Molina, M. (2004). A mechanism for two-choice discrimination: Time-dependent response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

Vickers, D., Caudrey, D., & Willson, R. J. (1971). Discriminating between the frequency of occurrence of two alternative events. *Acta Psychologica*, *35*, 151-172.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*, 152-173.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, *101*, 490-504.

Wells, G. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*, 688-696.

Yates, J. F., Lee, J.-W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and "reality". *Organizational Behavior and Human Decision Processes*, *70*, 87-94.

Appendix

Analytic expressions of the Poisson race model

In this appendix, we derive equations for the Poisson race model's predictions of accuracy given a range of confidence. We first discuss the role that stimulus evidence plays in the model, and we then examine confidence and accuracy.

Let X be a normally-distributed stimulus evidence parameter favoring the high alternative (perceived numerosity) with mean μ and standard deviation σ. The mean μ will differ depending on whether the stimulus is actually high or low; when the stimulus is high, μ is large (positive), and when the stimulus is low, μ is small (negative). We must derive two information accumulation rates based on X: one rate for information favoring the high alternative (λ_H), and one rate for information favoring the low alternative (λ_L). For mathematical simplicity and in accordance with Van Zandt (2000), we assume that the rates of information accumulation for the two alternatives sum to a constant, r. This means that we need only consider the density of $\lambda_H = \lambda$, as $\lambda_L = r - \lambda$.

We can map X into the accumulation rate λ by the logistic transformation: $\lambda = r/(1 + e^{-X})$. If X is large and positive, λ will be close to its maximum of r. If X is negative, λ will be closer to its minimum of 0. The pdf of λ is given by (Van Zandt, 2000):

$$g(\lambda \mid \mu, \sigma) = \frac{r}{\lambda(r-\lambda)} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}([-\ln[\frac{r-\lambda}{\lambda}] - \mu]/\sigma)^2), \tag{4}$$

where both μ and σ may depend on whether the stimulus is high or low. For our analyses, we assumed that σ was the same for both high and low stimuli and that $\mu_L = -\mu_H$.

It is useful to define a Bernoulli random variable, V, which represents the true stimulus identity. V will equal 1 if the stimulus is high and 0 if the stimulus is low. The Bernoulli parameter $P(V = 1)$ is equal to 0.5 because high and low stimuli are equally

likely to occur on any given trial. We can then rewrite Equation 4 as

$$g(\lambda \mid V = v) = \frac{r}{\lambda(r-\lambda)} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}([-\ln[\frac{r-\lambda}{\lambda}] - (\mu_H^v + \mu_L^{1-v} - 1)]/\sigma)^2). \quad (5)$$

Confidence Predictions

To derive the Poisson race model's confidence predictions, we ultimately wish to find an expression for $P(a \leq C < b, \text{High wins} \mid \lambda) = f_H(a, b \mid \lambda)$, where C is the random variable representing confidence. That is, given some accumulation rate λ, we wish to find the probability that confidence C for a high response falls in the range $[a, b)$. Because confidence depends on the information accumulated on the counters at the time a response is made (see Equation 1 in the main article), we can arrive at the probability of interest through the probability of losing counter states. If we restrict ourselves to responses in favor of "High", the state Y of the losing counter (that of the "Low" alternative) has a joint probability given by a negative binomial distribution (Townsend & Ashby, 1983; Van Zandt, 2000):

$$P(Y = y, \text{High wins}) = \left(\begin{array}{c} K_H + y - 1 \\ y \end{array} \right) p^{K_H} (1-p)^y, \quad (6)$$

where K_H is the threshold value of the "High" counter and $p = \lambda_H/(\lambda_H + \lambda_L) = \lambda/r$ is the probability of an observer labelling a count as "High".

Continuing to restrict ourselves to "High" responses, we now examine the number of points on the losing counter that will yield confidence judgments between a and b. In other words, we transform the inequality involving C to an inequality involving Y:

$$P(a \leq C < b, \text{High wins}) = P(a \leq \frac{K_H}{K_H + Y} < b, \text{High wins}) =$$

$$P(\frac{1}{b} < 1 + \frac{Y}{K_H} \leq \frac{1}{a}, \text{High wins}) =$$

$$P(\frac{K_H}{b} - K_H < Y \leq \frac{K_H}{a} - K_H), \text{ for } 0 \leq Y < K_L.$$

The restriction that $0 \leq Y < K_L$ incorporates the (joint) event that the "High" counter wins, meaning that the "low" counter must have fewer than K_L counts. As a result, we can modify the joint probability as:

$$P(\frac{K_H}{b} - K_H < Y \leq \frac{K_H}{a} - K_H, \text{High wins}) =$$

$$P(LB \leq Y \leq UB).$$

In the above formula, the lower and upper bounds LB and UB take the following values.

$$LB = ceil(\frac{K_H}{b} - K_H)$$

$$UB = min[floor(\frac{K_H}{a} - K_H), K_L - 1],$$

where *ceil* stands for the ceiling operator and $floor$ stands for the floor operator.

To obtain the probability of interest, we now use Equation 6 to sum over integers in the above inequality, giving

$$f_H(a, b \mid \lambda) = P(a \leq C < b, \text{High wins} \mid \lambda) =$$

$$\sum_{j=LB}^{UB} \left(\begin{array}{c} K_H + j - 1 \\ j \end{array} \right) \left(\frac{\lambda}{r}\right)^{K_H} \left(\frac{r-\lambda}{r}\right)^{j}, \tag{7}$$

This probability is conditioned on the threshold values K_H and K_L. While not explicitly noted here for simplicity, this fact becomes important when variable thresholds are specified below.

To obtain the joint probability for C when "Low" wins, we: 1) switch the H and L subscripts, and 2) replace λ/r with $(r-\lambda)/r$ both in Equation 7 and in the definitions for LB and UB.

Accuracy Predictions

To generate calibration curves, we must compute the proportion of correct responses at a range of confidence between a and b: $P(Corr \mid a \leq C < b)$. By definition, this

probability may be rewritten as:

$$\frac{P(Correct, a \leq C < b)}{P(a \leq C < b)}.$$

Next, we use the law of total probability (Taylor & Karlin, 1998) to examine "High" responses (denoted "H" below) separately from "Low" responses (denoted "L" below):

$$\frac{P(Correct, a \leq C < b, \text{“H”}) + P(Correct, a \leq C < b, \text{“L”})}{P(a \leq C < b)}. \tag{8}$$

The three probabilities in Equation 8 are derived separately below.

The denominator probability $P(a \leq C < b)$ is a function of $f_H(a, b \mid \lambda)$ and $f_L(a, b \mid \lambda)$. We first use the law of total probability to express $P(a \leq C < b)$ after integrating over all values of λ, so

$$\begin{aligned} P(a \leq C < b) &= \int_0^r P(a \leq C < b, \lambda) d\lambda \\ &= \int_0^r P(a \leq C < b \mid \lambda) g(\lambda) d\lambda. \end{aligned}$$

Applying the law of total probability again, the conditional probability $P(a \leq C < b \mid \lambda)$ can be expanded to yield

$$\begin{aligned} &\int_0^r [P(a \leq C < b, \text{“H”} \mid \lambda) + P(a \leq C < b, \text{“L”} \mid \lambda)] g(\lambda) d\lambda \\ &= \int_0^r [f_H(a, b \mid \lambda) + f_L(a, b \mid \lambda)] g(\lambda) d\lambda. \end{aligned}$$

The marginal density $g(\lambda)$ is obtained by unconditioning Equation 5 over the value of V:

$$\begin{aligned} g(\lambda) &= \sum_{v=0}^{1} g(\lambda \mid V = v) P(V = v) \\ &= \frac{1}{2}[g(\lambda \mid V = 0) + g(\lambda \mid V = 1)]. \end{aligned}$$

We can then write

$$P(a \leq C < b) = \frac{1}{2} \int_0^r [f_H(a, b \mid \lambda) + f_L(a, b \mid \lambda)][g(\lambda \mid V = 0) + g(\lambda \mid V = 1)] d\lambda. \tag{9}$$

Next, we consider the first term in Equation 8, the probability of a correct "High" response with confidence between a and b. Using the Bernoulli variable V,

$$P(Correct, a \leq C < b, \text{“H”}) = P(V = 1, a \leq C < b, \text{“H”}).$$

We now follow the same steps as in the derivation of Equation 9:

$$\begin{aligned} P(V = 1, a \leq C < b, \text{“H”}) &= \int_0^r P(V = 1, a \leq C < b, \text{“H”}, \lambda) d\lambda \\ &= \int_0^r P(a \leq C < b, \text{“H”}, \lambda \mid V = 1) P(V = 1) d\lambda \\ &= \frac{1}{2} \int_0^r P(a \leq C < b, \text{“H”}, \lambda \mid V = 1) d\lambda \\ &= \frac{1}{2} \int_0^r P(a \leq C < b, \text{“H”} \mid \lambda, V = 1) g(\lambda \mid V = 1) d\lambda \\ &= \frac{1}{2} \int_0^r P(a \leq C < b, \text{“H”} \mid \lambda) g(\lambda \mid V = 1) d\lambda \\ &= \frac{1}{2} \int_0^r f_H(a, b \mid \lambda) g(\lambda \mid V = 1) d\lambda. \end{aligned} \tag{10}$$

Similarly,

$$P(Correct, a \leq C < b, \text{“L”}) = \frac{1}{2} \int_0^r f_L(a, b \mid \lambda) g(\lambda \mid V = 0) d\lambda. \tag{11}$$

We arrive at a final formula for the calibration curve by substituting Equations 9, 10, and 11 into Equation 8 (note that the factor 1/2 cancels):

$$P(Correct \mid a \leq C < b) =$$

$$\frac{\int_0^r [f_H(a, b \mid \lambda) g(\lambda \mid V = 1) + f_L(a, b \mid \lambda) g(\lambda \mid V = 0)] d\lambda}{\int_0^r [f_H(a, b \mid \lambda) + f_L(a, b \mid \lambda)][g(\lambda \mid V = 0) + g(\lambda \mid V = 1)] d\lambda}. \tag{12}$$

Variable Thresholds

As alluded to above, Equation 12 is implicitly conditioned on the threshold parameters K_H and K_L being equal to constants. If the threshold parameters have probability distributions associated with them, then we must account for these

distributions in the expression for $P(Correct \mid a \leq C < b)$. We rewrite $P(Correct \mid a \leq C < b)$ in the following manner:

$$P(Correct \mid a \leq C < b) = \frac{P(Correct, a \leq C < b)}{P(a \leq C < b)}$$

$$= \frac{\sum_k P(Correct, a \leq C < b, K = k)}{\sum_k P(a \leq C < b \mid K = k)P(K = k)}$$

$$= \frac{\sum_k P(Correct \mid a \leq C < b, K = k)P(a \leq C < b \mid K = k)P(K = k)}{\sum_k P(a \leq C < b \mid K = k)P(K = k)}, \quad (13)$$

where K is a vector containing the two threshold parameters. The above steps follow by definition of conditional probabilities. The probability $P(Correct \mid a \leq C < b, K = k)$ is given by Equation 12, the probability $P(a \leq C < b \mid K = k)$ is given by Equation 9, and the probability $P(K = k)$ is set by the experimenter.

Intuitively, Equation 13 is simply a weighted average of proportions correct $P(Correct \mid a \leq C < b, K_H = i, K_L = j)$ across all possible threshold combinations. The weights are determined by both the probability that each threshold combination occurs $P(K = k) = P(K_H = i, K_L = j)$ and the probability that confidence falls in the range $(a, b]$ under a certain threshold combination $P(a \leq C < b \mid K_H = i, K_L = j)$.

Author Note

This research was made possible with NSF grants #SBR-0196200 and #SES-0214574. Portions of the work were presented at the 35th annual meeting of the Society for Mathematical Psychology.

Footnotes

[1]New Jersey Representative Robert E. Andrews, October 8, 2002, as published on his website http://www.house.gov/andrews/welcome.htm as the "Message of the Day."

[2]This equation is simply a linear transformation of the strict balance-of-evidence hypothesis.

[3]The notion that people are underconfident in perceptual tasks is somewhat controversial. While Baranski & Petrusic (1994, 1999) found underconfidence in such experiments, Olsson & Winman (1996) have not. The latter researchers claim that underconfidence is an artifact of flawed experimental design.

Table 1

Example probability distributions for favored and unfavored model thresholds.

x	$P(K_{Fav} = x)$[a]	y	$P(K_{Unf} = y)$[b]
2	0.125	4	0.125
3	0.75	5	0.75
4	0.125	6	0.125

[a] K_{Fav}=Favored Threshold

[b] K_{Unf}=Unfavored Threshold

Table 2

Dots Distributions, Experiments 1 & 2.

	LOW	HIGH	d'
Experiment 1			
Easy	$N(\mu = 45, \sigma = 5)$	$N(\mu = 55, \sigma = 5)$	2
Hard	$N(\mu = 47, \sigma = 5)$	$N(\mu = 53, \sigma = 5)$	1.2
Experiment 2			
Easy	$N(\mu = 45, \sigma = 5)$	$N(\mu = 55, \sigma = 5)$	2
Hard	$N(\mu = 48.5, \sigma = 5)$	$N(\mu = 51.5, \sigma = 5)$	.6

Figure Captions

Figure 1. A typical calibration curve.

Figure 2. Information Accumulation on Counters of the Poisson Race Model. See text for more details.

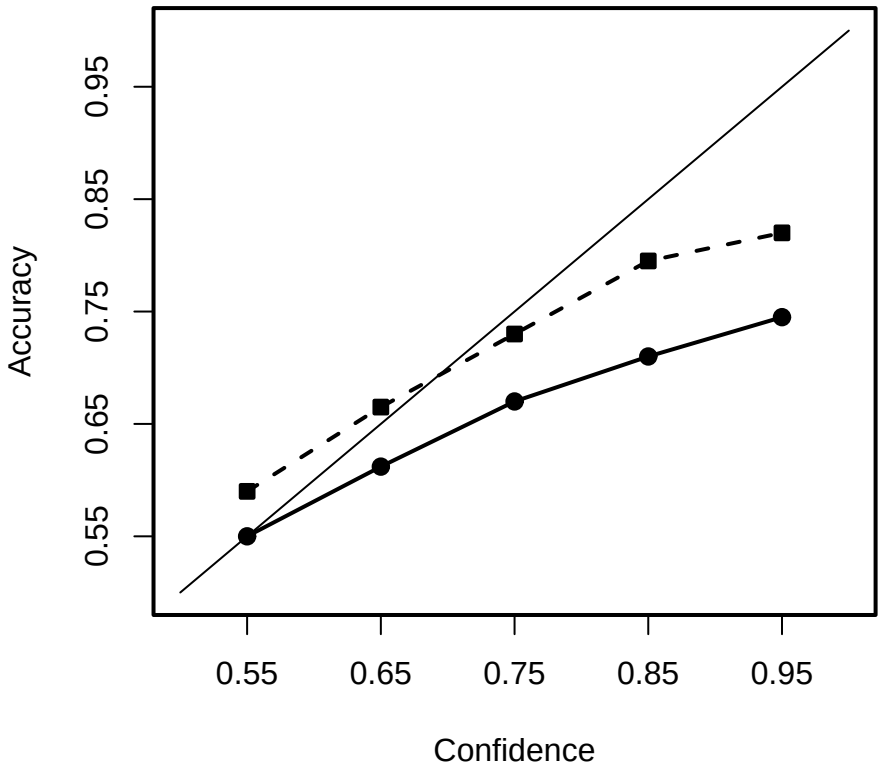
Figure 3. PDFs of λ_H for Two Different Values of μ. See text for more details.

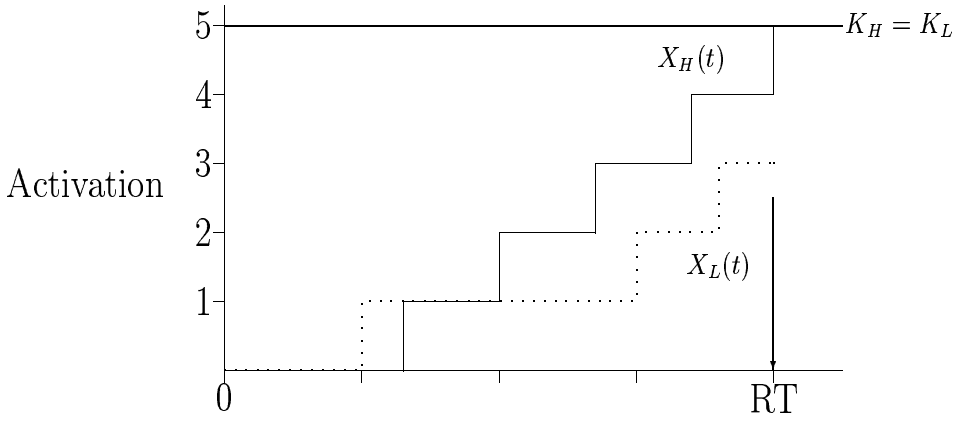
Figure 4. Fitted (dashed lines) & Observed (solid lines) Calibration, Experiment 1. The line with circles is data from the easy condition, and the line with squares is data from the hard condition.

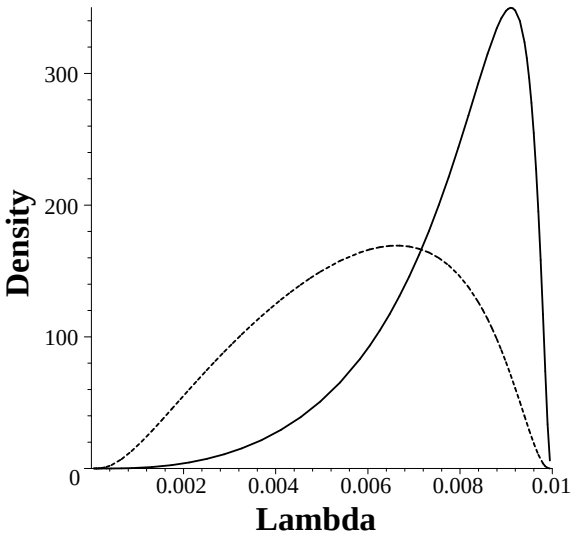
Figure 5. Fitted (dashed lines) & Observed (solid lines) Calibration, Experiment 2. The line with circles is data from the easy condition, and the line with squares is data from the hard condition.

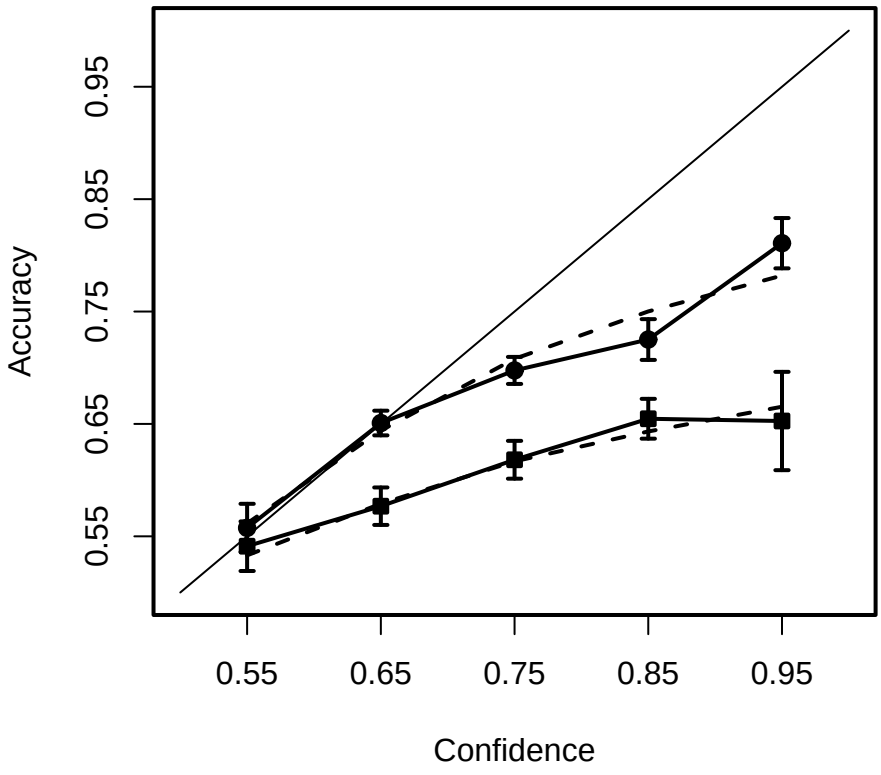
Figure 6. Fitted (broken lines) and Observed (solid lines) Calibration for Low-favored Responses, Experiment 3. Circles denote responses in favor of the high alternative, and squares denote responses in favor of the low alternative. The dashed line is the model fit of responses in favor of the high alternative, and the dotted line is the model fit of responses in favor of the low alternative.

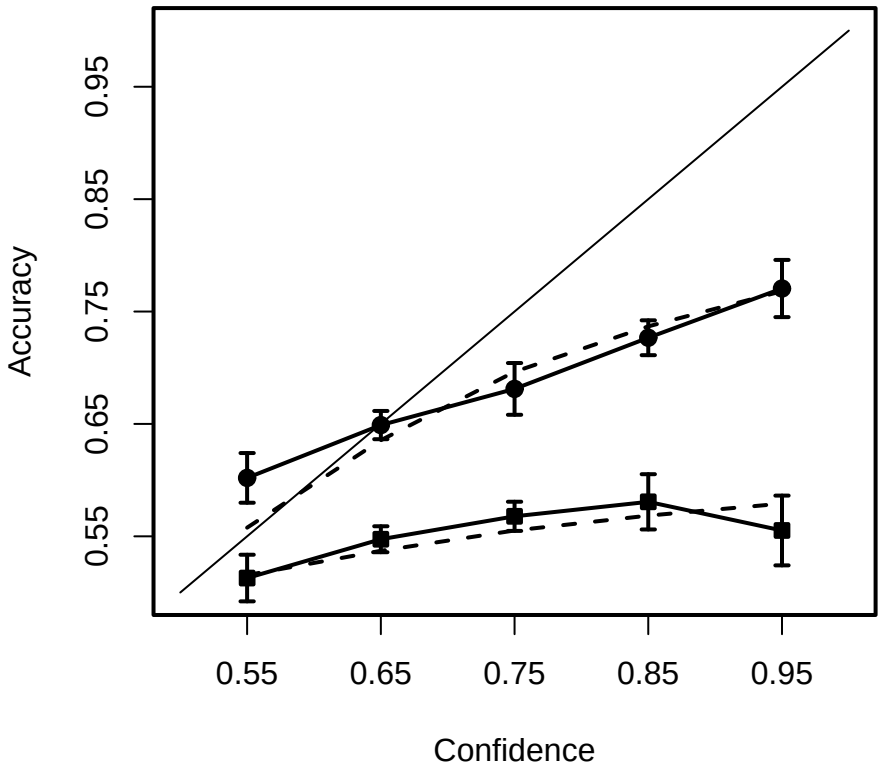
Figure 7. Fitted (broken lines) and Observed (solid lines) Calibration for High-favored Responses, Experiment 3. Circles denote responses in favor of the high alternative, and squares denote responses in favor of the low alternative. The dashed line is the model fit of responses in favor of the high alternative, and the dotted line is the model fit of responses in favor of the low alternative.

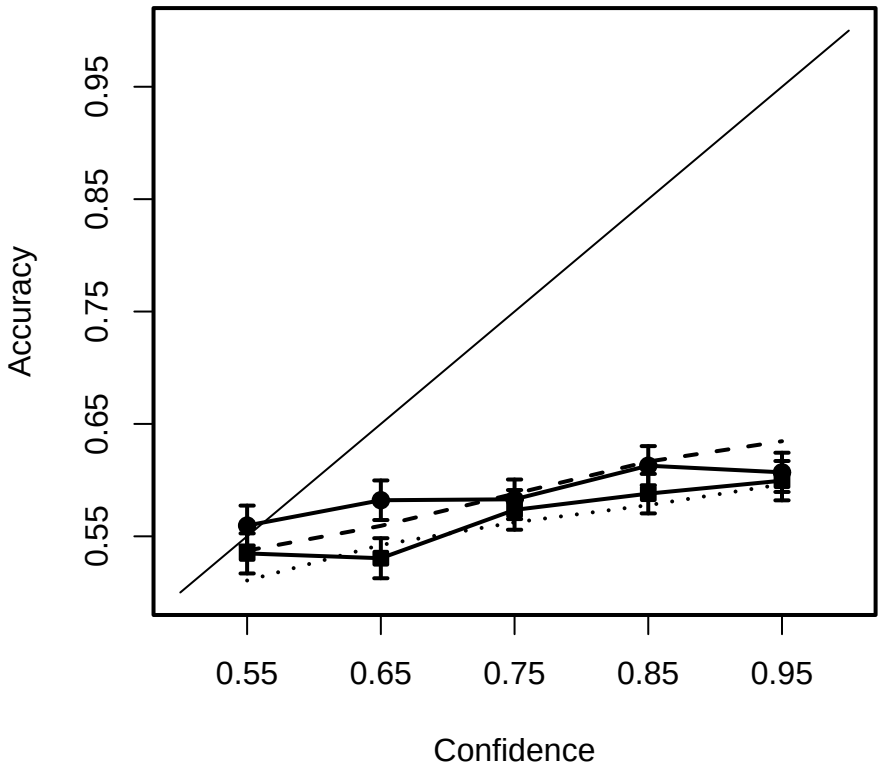

Accuracy
0.55
0.65
0.75
0.85
0.95
0.55
0.65
0.75
0.85
0.95
Confidence

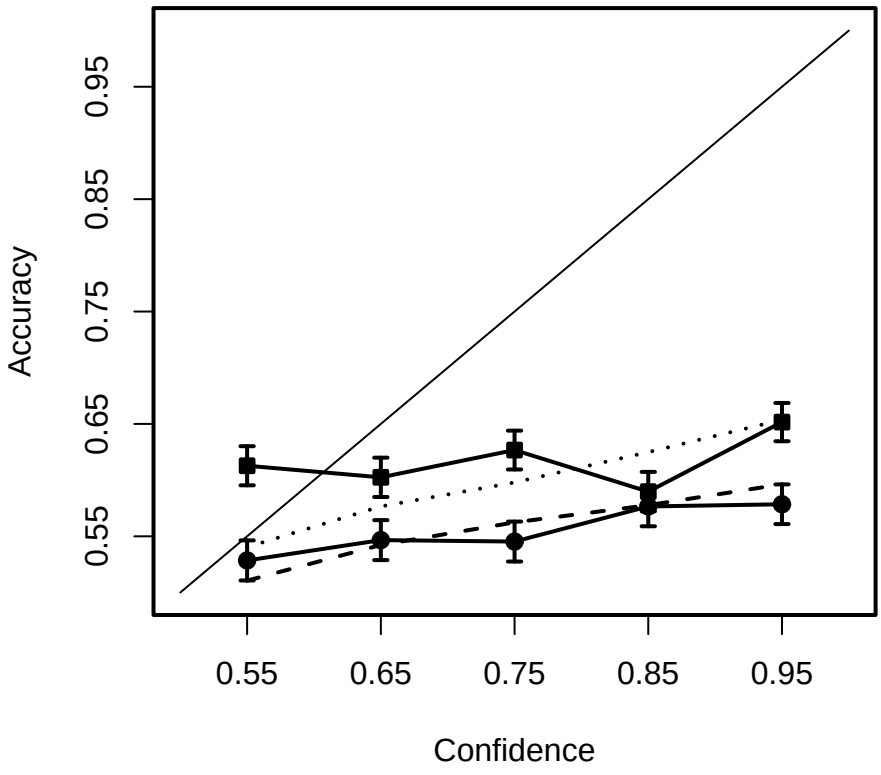

Activation
5
4
3
2
1
0
RT
$K_H = K_L$
$X_H(t)$
$X_L(t)$


Density
300
200
100
0
0.002
0.004
0.006
0.008
0.01
Lambda


Accuracy
0.55
0.65
0.75
0.85
0.95
0.55
0.65
0.75
0.85
0.95
Confidence


Accuracy
0.55
0.65
0.75
0.85
0.95
0.55
0.65
0.75
0.85
0.95
Confidence


Accuracy
0.55
0.65
0.75
0.85
0.95
0.55
0.65
0.75
0.85
0.95
Confidence


Accuracy
0.55
0.65
0.75
0.85
0.95
0.55
0.65
0.75
0.85
0.95
Confidence