When Learning is Detrimental: SESAM and Outcome Feedback

Henrik Olsson (henrik.olsson@psyk.uu.se)
Department of Psychology, Uppsala University
Box 1225, SE-751 42 Uppsala, Sweden

Peter Juslin (peter.juslin@psyk.uu.se)
Department of Psychology, Uppsala University
Box 1225, SE-751 42 Uppsala, Sweden

Abstract

The *sensory sampling model* (*SESAM*; P. Juslin & H. Olsson, 1997) accounts for the underconfidence observed in sensory discriminations with pair-comparisons. In the present study the model is applied to a single-stimulus task and a comparison is made with pair-comparisons. The model predicts that in the single-stimulus condition training with feedback should lead to *poorer* calibration with more underconfidence. In pair-comparison the feedback should have little or no effect on calibration. The results confirm these predictions.

Introduction

A common presumption is that experience should improve the quality of our judgments and foster insights into the limitations of our knowledge. Indeed, in cognitive tasks there is ample of evidence that experts' confidence judgments are more realistic than those of novices (for a review, see Yates, 1990). The evidence on sensory discrimination is less clear-cut. Some studies report little or no improvement in realism of confidence when participants are provided with outcome feedback (Winman & Juslin, 1993), while other studies report improvement for difficult task sets and worse calibration in easy task sets (Petrusic & Baranski, 1997).

In Juslin and Olsson (1997), it was suggested that this and other discrepancies between inferential and sensory discrimination tasks arise because two different sources of uncertainty are involved. The *sensory sampling model* (*SESAM*) was developed to elucidate confidence in sensory discrimination. This paper extends the work to the case of a single-stimulus task where the participant is to decide whether a presented line is longer than a specified but not seen reference length (e.g., a Swedish twenty-kronor note). We will concentrate on two counter-intuitive predictions by *SESAM*: (a) The underconfidence observed with pair-comparisons will be *unaffected even by prolonged sessions of outcome feedback*. (b) The realism of confidence in a single-stimulus task will deteriorate with outcome feedback, leading to *poorer calibration with more underconfidence*. The experiment reported below provides a test of these two predictions.

Realism of Confidence in Sensory Discrimination

Realism of confidence, or *calibration*, is commonly investigated by presenting participants with a large set of two-alternative decision tasks. For each task-item, the participant decides on one of the two alternatives and assesses his or her confidence in the correctness of this decision as a subjective probability between .5 (random choice) and 1.0 (certainty). Participants are said to be realistic, or well calibrated, to the extent that items assigned subjective probability .xx are correct with relative frequency .xx. An index of *over/underconfidence* is obtained by subtracting the overall proportion of correct decisions from the mean subjective probability, where a positive difference is overconfidence.

In a review of early psychophysical studies, Björkman, Juslin, and Winman (1993) observed that these studies often suggest *underconfidence* (although interpretations in terms of calibration are problematic, because confidence was not assessed as subjective probabilities). For one-hundred years, results such as these has led researchers to speculate about *subconscious mental processes* (Fullerton & Cattell, 1892), or *implicit perception* (Kihlstrom, Barnhardt, & Tataryn, 1992). More recently, underconfidence in sensory discrimination has been replicated in studies with the modern calibration paradigm (for a review, see Juslin & Olsson, 1997). In a meta-analysis (Juslin, Olsson, & Winman, 1998), which aggregated data from 21 sensory discrimination tasks and 44 inferential tasks, a clear main effect of sensory versus inferential tasks revealed more underconfidence for sensory discrimination. This was true even when the effect of proportion correct was removed as a co-variate. These results refute the claim that there is no difference between confidence in inferential and sensory-discrimination tasks (Baranski & Petrusic, 1994; Ferrell, 1995).

The Sensory Sampling Model (SESAM)

Consider yourself as a participant in a difficult pair-comparison task, such as deciding which of two almost equivalent lines is the longest. It may take some time before you reach a decision. The presence of neural noise makes computation of an error-free estimate of the 'true' difference μ between the two lines impossible, where the *true difference* refers to the estimate that would result if there was no noise in the processing of the sensory information. *SESAM* proposes that the nervous system repeatedly computes new estimated values X_i of μ that vary from moment to moment due to the neural noise. The X_i are assumed to be a Normally and Independently Distributed (NID) random variable with mean μ and variance σ_X^2, where the parameter μ is defined in the unit variance of X_i. *Sensory discrimination* refers to a perceptual task where erroneous decisions emanate primarily from neural noise (or approximations to these situations).

In *SESAM*, the decision about the difference μ is based on the participant's overall impression as modeled by a statistical aggregate; the mean sensation $\bar{X}_i$ (where the

index i denotes that the mean is computed after sensation X_i). However, this mean is computed only from the n last estimates X_i that are still contained in a memory window that represents a limitation in the processing capacity of the nervous system. The number of sensations n contained in the memory window at any moment is the *sample size parameter*. When a new sensation enters, the oldest one is pushed out and for every new X_i a new mean $\bar{X}_i$ is computed from the last n sensations. Thus, the process is modeled as a capacity limited sequential sampling process. A decision cannot be made unless the absolute value of the mean $\bar{X}_i$ exceeds a *response threshold* $\phi_{\bar{X}}$, where the threshold corresponds to a definite experience of a difference between the stimuli. When the estimation of μ is that $\mu>0$ then the decision will be that the left line is longer, and vice versa.

Note two consequences: First, the number of sensations X_i sampled before a decision is made provides predictions of response times (after a linear transformation). Second, occasionally there will be an erroneous decision due to the sampling error in the estimates X_i. The probability of a correct decision is determined by the sampling error variance $\sigma_{\bar{X}}$ of the mean used to make the decision. In the case of non-sequential or static sampling of n sensations (i.e., obtained by setting the response threshold $\phi_{\bar{X}}$ at 0), this variance takes the familiar form of $\sigma^2_{\bar{X}} = \sigma^2_X / n$. Finally, the *stopping parameter* N_{max} is the maximum number of iterated computations of X_i before the process is ended. If no noticeable difference is detected within the allotted time, a random decision with subjective probability .5 is made.

When making a conditional probability assessment, the participant first makes a decision and then an assessment of the probability that the decision is correct. Say that the participant decides that $\mu>0$, that is, the left line is longer. According to *SESAM*, confidence reflects the consistency of the information contained in the sample used for the decision. It is proposed that the subjective probability is computed as the proportion of the last n sensations that support the decision. If, for instance, the decision is $\mu>0$ and 70% of the X_i are larger than zero, the subjective probability is .7 (and vice versa when the decision is $\mu<0$). This means that subjective probability is based on a proportion defined by the variance σ^2_X of the sensations.

While confidence reflects the variability of single sensations (σ^2_X), decisions benefit from the greater precision of a statistical aggregate obtained across n sensations (defined by $\sigma^2_{\bar{X}}$). The most immediate test of these assumptions is that when both proportions of decisions $\mu>0$ and mean subjective probability that $\mu>0$ is plotted against (negative and positive) physical stimulus differences, both functions should approximate normal ogive functions, although the ogive for subjective probability should be less steep (larger variance) than the ogive for decisions. This, indeed, is the common finding (e.g., Johnson, 1939, Juslin & Olsson, 1997). A second implication is that in studies with conditional probability assessment, there will be a disposition towards underconfidence, in particular, for stimulus differences with moderate and high proportions correct. *SESAM* provides a good account of subjective probability distributions, hit-rates, and response times in a line discrimination task (Juslin & Olsson, 1997).

The disposition towards underconfidence is alleviated in three circumstances: First, when the physical stimulus difference is zero or close to zero; Second, when the sequential sampling process is constrained by time pressure (Baranski & Petrusic, 1994; Olsson & Juslin, 1998); Third, when the processing is dominated by a perceptual bias.

SESAM with Perceptual Bias

One of the most striking demonstrations of perceptual bias in a calibration task of the kind modeled by *SESAM* is Experiment 2 in Baranski and Petrusic (1994). The participants' task was to indicate which of two vertical lines was located farther from a vertical central referent line. In this task, Baranski and Petrusic observed a "left-looks-farther" effect: When the left line was closer to the central referent line (i.e., the left line was located 296 pixels to the right of the central referent and the right line was located 300 pixels to the right of the central referent) it was perceived as being farther away. The proportion of correct decisions in the 296, 300 order was only .26 and the proportion correct in the 300, 296 order was .87.

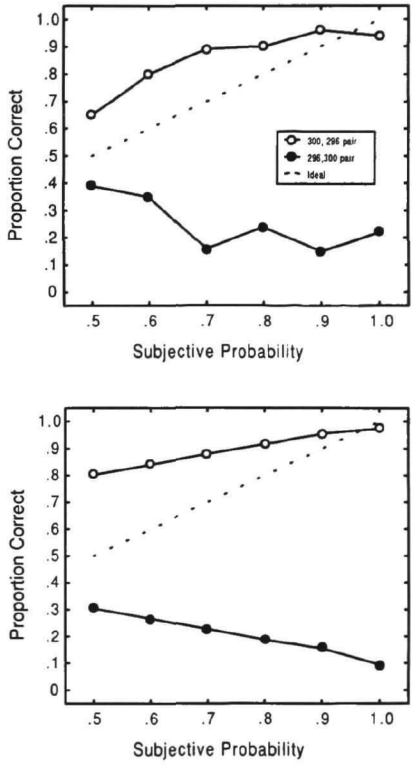

Figure 1: Top panel: Empirical calibration curves from Experiment 2 in Baranski and Petrusic (1994). Lower panel: Calibration curves simulated with *SESAM*.

In realism of confidence studies, data are sometimes presented in calibration curves where the proportion of correct decisions are plotted against the levels of subjective probability (in a conditional probability task with two alterna-

tives these are .5, .6, .7, .8, .9, and 1.0). The effect of perceptual bias is dramatically illustrated in the calibration curve in the top panel of Figure 1. For the 296, 300 order, the participants in Baranski and Petrusic's experiment had a lower proportion of correct decisions when they were absolutely certain (1.0) than when they were guessing (.5).

The effects of perceptual bias can be accounted for by adding a perceptual bias parameter b (with a permissible range of $-\infty$ to ∞) to μ in *SESAM* (Olsson, 1999). For example, consider a line discrimination task with two lines A and B. If we assume that $\mu = .1$ and $b = -.2$, the participant has a tendency to falsely perceive line B to be longer than line A. This means that compared to an unbiased version of the same task, the proportion of correct answers will be lower. If the bias is large the proportion of correct answers can approach 0. The lower panel in Figure 1 shows simulated calibration curves obtained with perceptual bias. The only parameter difference between the two calibration curves is the sign of the bias parameter (.17 and -.17). The other parameters were: $\mu = .06$, $\phi_{\bar{x}} = .35$, and n = 10. The stopping parameter N_{max} was not used in this simulation. It can be seen that the model is successful in accounting for the empirical calibration curves from Baranski and Petrusic (1994) (see Olsson 1999, for a detailed account).

Pair-Comparison and Single-Stimulus

The idea of a distorted perceptual representation can not only be applied to a pair-comparison task but also to a single-stimulus task where the reference is a memory representation. Consider a variation of the pair-comparison task used by Juslin and Olsson (1997). Instead of two lines you are looking at a single line and you are to decide if this line is longer or shorter than a stated reference length, say the diameter of a compact disc. Even if you do not know the exact length of this reference, you will have some apprehension of it, presumably by retrieval of a memory representation. When *SESAM* is applied to this single-stimulus task the nervous system computes the subjective difference between the memory representation and the stimulus line. Whereas in pair-comparison, all error is assumed to stem from neural noise, in the single stimulus case there will be one additional source of error because your memory representation of the reference length may be biased.

Effects of Feedback In a pair-comparison task dominated by neural noise, *SESAM* suggests that feedback should have little or no effect. Both decisions and probabilities fundamentally arise from the finite precision of the sensory system, and this precision is not likely to be affected by the amount of feedback provided in a single experimental session. Aside from the stimulus units with very low proportion correct, we thus expect underconfidence both before and after a long feedback session. This is illustrated in the top panel of Figure 2 (see Winman & Juslin, 1993, for empirical evidence).

For single-stimulus, on the other hand, we expect the provision of feedback to allow the participants to detect the bias in their memory representation of the reference. Initially, there will be small underconfidence or overconfidence depending on the size of the bias. The reduction of bias will make the proportion correct rise and subjective probability will not be much affected. These predictions are illustrated in the lower panel of Figure 2, where it is assumed that the bias is decreased by a constant fraction. These two effects taken together imply that there will be a change towards underconfidence. If the bias is completely eliminated, the task becomes identical to the pair-comparison task and the prediction is thus the same: Underconfidence.

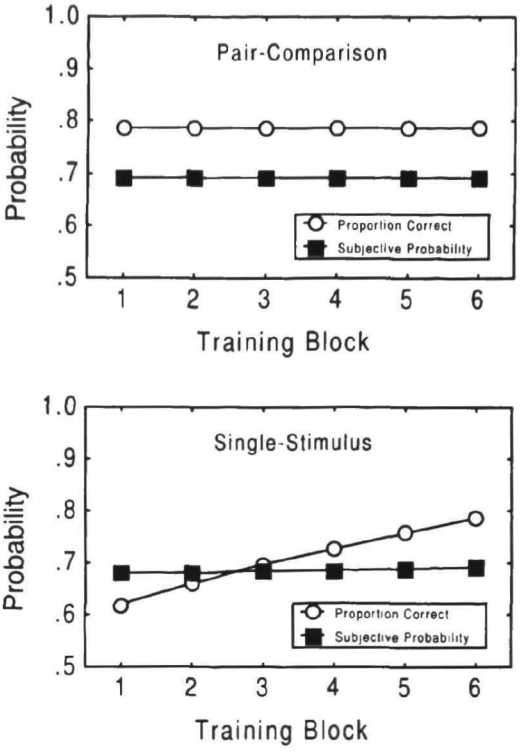

Figure 2: The predicted interaction between provision of outcome feedback and task. Top panel: Expected proportion correct and subjective probability as a function of feedback for pair-comparison. Lower panel: Simulated proportion correct and subjective probability as a function of feedback (or bias b) for single-stimulus. The values of b was: -.15, -.12, -.09, -.06, -.03, and 0. The response threshold $\phi_{\bar{x}}$ was .35, the stimulus difficulty μ was .25, and the sample window parameter n was 7 for all simulations.

The Experiment In one condition participants were to decide which of two lines that was the longer. In the second condition participants had to decide whether an observed line was longer or shorter than a given but not seen reference length, the length of a Swedish twenty-kronor note. A pilot test indicated that the participants did have a biased representation of this reference length (i.e., they guess that a presented length equal to a Swedish twenty-kronor note is longer than the note in 74% of the trials). In pair-comparisons there will be initial underconfidence that prevails in the face of feedback. With single stimulus, the provision of outcome feedback should make the participants underconfident. To the extent that there is no strong initial

over- or underconfidence bias, outcome feedback will make these participants *more poorly calibrated.*

Method

Participants

Twenty-four participants, aged between 19 and 32, attended. Thirteen were males and eleven were females. Most were undergraduate psychology students at Uppsala University who participated in exchange for credit for a course requirement. All participants had normal vision or corrected to normal vision.

Apparatus and Stimuli

All participants responded with a mouse. In the pair-comparison condition the stimulus display consisted of two vertical lines. The standard stimulus that was 130 mm (the same length as a Swedish 20 kronor note—the reference in the single-stimulus condition). There were three levels of difficulty, L1 (hardest), L2, and L3 (easiest). The difficulty of the standard-variable combination is expressed in terms of the ratio r of the longer line to the shorter line in the pair: 1.01, 1.02, and 1.025. Each stimulus unit consisted of two pairs of stimuli; in one pair the left line was longer; in the other pair the right line was the longer. Thus, in half of the presentations the standard was on the left, in the other half the standard was on the right. The order of the standard-variable pairs was determined randomly with a new permutation for each participant.

The lines were black, 1 mm wide and appeared on a white background. All of the standard-variable combinations were centered both horizontally and vertically. To ascertain that the stimuli combinations were of correct length and properly centered, all combinations were measured directly on the computer screen. This procedure was repeated on several occasions.

In the single-stimulus condition, the participant was presented with a single line. Half of the lines with stimulus difficulties L1, L2, and L3 were longer than the standard reference and the other half were shorter. The stimuli were the same as the variable stimulus in the pair comparison condition. The line appeared in equal proportion in the right and the left position with a random order of stimulus presentation.

Design and Procedure

Two independent variables were investigated; pair-comparison versus single-stimulus tasks (between-subjects), and performance before and after training with outcome feedback (within-subjects). Dependent measures were decisions and confidence, refined into measures of over/underconfidence. The effect of feedback was studied by first presenting participants with a pretest block without feedback. Then in four blocks, outcome feedback was given after each trial. Finally, a posttest without feedback was administered.

Each block consisted of 120 judgments. In a block the participant made 40 judgments of 3 different stimulus units. In the pair-comparison condition, a standard-variable pair of lines was shown on the screen and the question "Which line is the longest?" appeared below the stimulus unit. Beneath the question there were two small buttons, one labeled "Left" and one labeled "Right". When participants had decided which of the lines they thought was the longest, they clicked on the appropriate button with the mouse. Then the screen was cleared and the participants assessed how certain they were that they had made the correct decision. The subjective probability scale appeared in the middle of the screen and consisted of six buttons labeled "50%, "60%", "70%", "80%", "90%", and "100%". The written instructions briefly introduced the notion of calibration and the scale was anchored with "random choice" at "50%" and "certainty" at "100%". After the participants had selected a subjective probability level with the mouse, the screen was cleared and the next trial began.

In the single-stimulus condition, the question was "Is this line longer or shorter than a 20-kronor note?" and the two buttons were labeled "shorter" and "longer". In other respects the conditions were identical. The session was interrupted by a 20-minutes break and the whole session took between 2.5 and 3.5 hours.

Results and Discussion

In Figure 3 proportion correct, mean subjective probability and over/underconfidence scores are plotted for the pretest, the four training blocks, and the posttest for pair-comparison and single stimuli.

In the pair-comparison condition training had no (positive) effect on proportion correct. In fact, the proportion correct was higher in pretest than in posttest. The proportions correct were .79 (95 % confidence interval, CI, $\pm$ .05)[1] and .76 (95 % CI, $\pm$ .05) respectively). The prediction was that feedback would have no (positive) effect and this is the case. The confidence intervals overlap, so we can not be sure that the slight decrease in proportion correct constitutes a real effect.

In the single-stimulus condition, training had a positive effect on proportion correct as predicted. In Pretest the proportion correct was .64 (95% CI, $\pm$ .05). When feedback was given this figure grew steadily with every block until it peaked at .76 (95 % CI, $\pm$ .05) in training block 4. When feedback was withdrawn the proportion correct fell to .72 (95 % CI, $\pm$ .05) in the posttest. The drop in proportion correct from training block 4 to the posttest (.76 to .72) in the single-stimulus condition could be an effect of fatigue. One additional hypothesis is that the memory representation of the reference length may not be very stable and once the feedback is withdrawn, the older and more biased memory representation once again comes to dominate the responses. This latter hypothesis is not supported by the data, however. The bias towards responding "longer" was .74 in the pretest of the single-stimulus condition, decreased to .57 in the training block 4, but was still no more than .56 in the post-test. Fatigue seems to be a more viable alternative.

[1] All within group comparisons have a standard error contrast based on the condition x subject in groups mean square; see Estes, 1997, for details.

Therefore, we decided to test the effects of training on training block 4 rather than the posttest. This can be motivated on two grounds: First, the proportion correct in training block 4 of the single-stimulus condition (.76) is similar to the proportion correct in training block 4 for the pair-comparison condition (.77), and not much lower than the pretest result of .79 for pair comparisons. The small difference between the two conditions in training block 4 indicates that nearly all of the bias is eliminated. The structure of the single-stimulus condition is thus almost equivalent to the pair-comparison task, both of which can be modeled by the bias-free version of *SESAM*. Second, the signs of fatigue in the posttest of both conditions indicate that these data may not be representative of the true performance level of the participants. It is important to note, though, that all qualitative results are the same if the posttest data are used (see Figure 3).

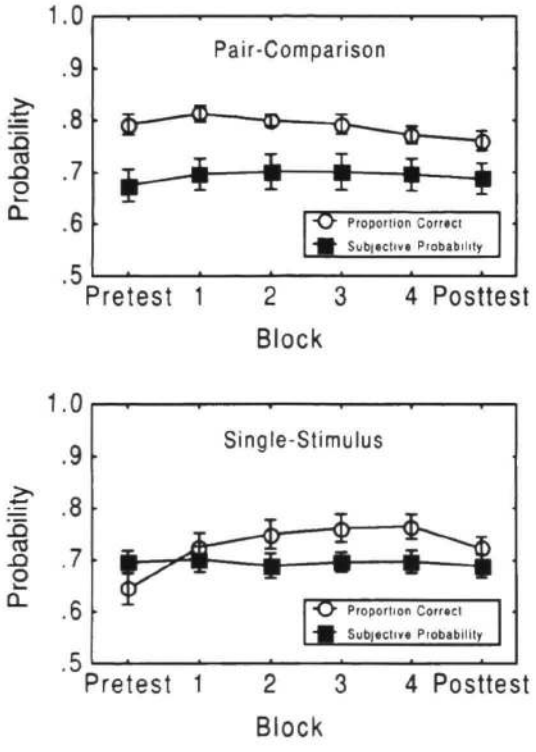

Figure 3: Proportion correct and mean subjective probability for pair-comparisons and single-stimulus as a function of training-block. The error bars are standard errors.

In Figure 3 we see that mean subjective probability is roughly the same in both conditions and constant across training blocks. In the pair-comparison condition, the pretest mean confidence is .67 (95 % CI, ± .03) and in training block 4 .70 (95 % CI, ± .03). In the single-stimulus condition, mean confidence is .70 (95 % CI, ± .03) both in the pretest and in training block 4. This means that there was underconfidence -.12 (95 % CI, ± .06) in the pretest of the pair-comparison condition. In the pretest of the single-stimulus condition there was a moderate overconfidence of .05 (95 % CI, ± .06). For the pair-comparison condition there was a clear underconfidence bias also in training block 4, -.08 (95 % CI, ± .06). As is evident from Figure 3, the minor decrease in underconfidence in the pair-comparison condition, is wholly explained by a decrease in the proportion correct, perhaps due to fatigue, rather than to accommodation of subjective probability judgments to feedback. In the single-stimulus condition there was also underconfidence in training block 4, -.07 (95 % CI, ± .06). this confirms two predictions. First, it was predicted that the reduction of bias should lead to a change in the direction of underconfidence and, second, that the single-stimulus condition should give the same result as the pair-comparison condition after training.

Note, finally, that the proportions correct in all blocks except the pretest of the single-stimulus condition (.64) are between .7 and .8, whereas mean subjective probability never exceeds .7. Across all blocks in the pair-comparison condition, underconfidence is -.10 at a proportion correct of .78. This pattern of uniform underconfidence for proportions correct between .7 and .8 deviates from the pattern in cognitive or inferential tasks, where there is generally close to zero over/underconfidence at this level of difficulty (Juslin, Olsson, & Björkman, 1997).

Conclusion

As predicted by *SESAM*, feedback produced different results in a pair-comparison and a single-stimulus task with no improvement in the former and deteriorating realism in the latter. These results illustrate that to predict the effect of outcome feedback on realism of confidence one needs to take into account how the specific task relates to the cognitive processes and representations that underlie performance.

The results from the conditions where the role of bias was minimized, pretest and training block 4 of the pair-comparison condition, and training block 4 of the single-stimulus condition, replicates the finding of underconfidence in sensory discrimination (for a review, see Juslin & Olsson, 1997). The results also indicate that bias is an important limiting condition for underconfidence to occur, a conclusion that is consistent with the data reported by Baranski and Petrusic (1994).

Acknowledgments

The research reported in this paper was supported by the Swedish Council for Research in Humanities and Social Sciences. We are indebted to Tomas Foucard for running the experiment. We thank Magnus Persson, Pia Wennerholm, and Anders Winman for helpful comments.

References

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

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfi-

dence phenomenon. *Perception & Psychophysics, 54*, 75-81.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review, 4*, 330-341.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for under-confidence in sensory discrimination. *Perception & Psychophysics, 57*, 246-254.

Fullerton, G. S., Cattell, J. M. (1892). *On the perception of small differences*. Philadelphia: University of Pennsylvania Press.

Johnson, D. M. (1939). Confidence and speed in two-category judgment. *Archives of Psychology, 34*, 1-53.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: a sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344-366.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes, 73*, 3-26.

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.

Kihlstrom, J. F., Barnhardt, T. M., & Tataryn, D. J. (1992). Implicit Perception. In R. F. Bornstein & T. D. Pittman (Eds.), *Perception without awareness* (pp. 17-54). New York: Guilford Press.

Olsson, H. (1999). *Explorations of the sensory sampling model: Sampling mechanisms, response time distributions, and bias*. Manuscript in preparation.

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

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology, 110*, 543-572.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgment: Two different accounts. *Scandinavian Journal of Psychology, 34*, 135-148.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.