On the Control of Confidence Processing in Sensory-Based Decision-Making Tasks

Steven R. Carroll

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the Master of Arts degree

Department of Psychology

Carleton University

Ottawa, Ontario

August, 2006

Your file Votre référence
ISBN: 978-0-494-18250-5
Our file Notre référence
ISBN: 978-0-494-18250-5

Abstract

Participants were asked to make a series of comparative judgements. Following half of these they were asked to render confidence (i.e., subjectively evaluate their surety of having made a correct choice). Participants either rendered confidence in alternating 'pure blocks' (i.e., confidence was always / never rendered following every trial in a block) or in 'mixed blocks' (i.e., confidence was rendered only on random trials). In one experiment, participants in the mixed block condition never knew when they would be expected to render confidence. Most, though not all, confidence rendering by these participants was done post-decisionally. In a second experiment, participants in the mixed block always expected to have to render confidence, but were told not to do so if they heard a tone. These participants began rendering confidence early and seemed able to stop processing confidence on demand. These findings imply an ability to control when confidence is processed.

Acknowledgements

I would first like to thank Dr. Joseph Baranski, my external committee member from Defence Research and Development Canada, not only for his feedback but also for taking the time to travel to Ottawa in order to participate in my defence. I would also like to thank Dr. Craig Leth-Steensen and Dr. Jo-Anne LeFevre for their insightful comments and recommendations.

Many thanks are owed to Dr. William M. Petrusic, my thesis supervisor, for being a fountain of knowledge, a patient guide, the provider of a great thesis idea (the stop-confidence paradigm), and for being willing to let me air my more contentious theories before the thesis committee. Though I still have a long row to hoe, and a great deal to learn, Bill has instilled within me a confidence that I can finally succeed in the endeavour. And that, I think, is the hallmark of a great teacher.

A debt of gratitude is owed to my participants, who collectively made 91,840 sensory-based decisions and 45,920 confidence judgements. Science owes much to the undergraduate's ability to endure tedium.

Finally, many many many thanks are owed to my colleague, best-friend, and wife Judith Godin. Everything I do, I do it for you.

Table of Contents

List of Figures

List of Tables

General Introduction

Recent investigations by Petrusic and Baranski (2003, 2000) and Baranski and Petrusic (2001) into the locus and time-course of confidence processing have created problems for two classes of model of sensory-based decision-making: post-decisional models and single-process models.

Post-decisional models typically make the assumption that a decision follows a series of discrete evidence accrual events. These models tend to refer to one's surety of having made a correct choice (i.e., confidence) as a function of the difference between accrued levels of evidence supporting each possible decisional outcome. Examples of post-decisional models include The Accumulator Model (Vickers, 1979), Slow and Fast Guessing Theory (Petrusic & Baranski, 1989), The Sequential Sampling Model (Juslin & Olsson (1997), and The Race Model (Van Zandt, 2000).

Single-process models tend to describe confidence as a by-product of the primary decision itself. Confidence in these models is assumed to reflect the difference between the observed strength of a signal and a criterion strength level subjectively established by the observer to help differentiate between actual signals and mere fluctuations in background noise. The most famous example of a single-process model is Signal detection Theory (Tanner & Swets, 1954).

Petrusic and Baranski (2003, 2000) and Baranski and Petrusic (2001) found that participants asked to render confidence following each of a series of sensory-based decisions took significantly longer to make their decisions than did participants who were never asked to render confidence. This finding implies confidence processing interferes with primary decision processing and, as such, confidence processing must begin *before*

the primary decision is made. Interestingly, these researchers also noted that post-decisional times to render confidence varied significantly, and that time to render confidence seemed to be related to decision difficulty. This finding implies confidence is also processed *after* the primary decision has been made.

The challenges posed by these findings to the aforementioned classes of decision-making model are these: post-decisional models cannot explain how confidence generation could begin before all evidence has been accrued, and single process models cannot explain why confidence is not known to the observer immediately once the primary decision has been made.

In this light, the question of the extent to which a decision-maker can exert control over confidence processing has become an important one. If, for instance, it can be demonstrated that confidence processing is unavoidable, unstoppable, automatically activated pre-decisionally in a sensory-based decision-making task, and that the process continues after the primary decision has been made, then neither post-decisional nor single-process models can justifiably ignore confidence effects. Confidence determination, if this were the case, would seem to be an inextricable part of the primary decision-making process. If, on the other hand, it can be shown that decision-makers exercise deliberate control over whether and when they initiate confidence processing, neither model type would have to be altered *per se*, though the ability of these models to explain all instances of sensory-based decision-making would still be diminished. If decision-makers can demonstrate the ability to postpone confidence processing until after the primary decision has been made, then post-decisional models could be said to predict the outcome of this limited instance of decision-making. If decision-makers can

demonstrate the ability to process confidence entirely pre-decisionally and without cost, then single-process models could be said to predict this limited instance of decision-making.

The present experiments offer a demonstration of the extent to which decision-makers can exert control over confidence processing. Such a demonstration might allow researchers to narrow their range of focus and concentrate more completely on the one class of models that most accurately describes sensory-based decision-making. This would be an important step towards the ultimate goal of decision-making research, namely the development of a single, complete, and accurate model of decision-making. Once psychophysicists can agree on such a model, we will be able to shift our attention away from merely trying to describe the process of decision-making to trying to discover a means of making this process more accurate.

Before a review of experimental methodology and results can be reported, however, it is first important to more specifically state the tenets of the aforementioned models of decision-making, and to briefly note how a demonstrable control over confidence, or a lack thereof, would support or hinder their author's claims.

Single Process Models

Signal Detection Theory

In offering researchers a veritable tool-chest of analytical methods for psychophysical experiments, few theories have had the same impact as Tanner and Swets' (1954) Signal detection Theory (SDT) on cognitive psychology. The task of a decision-maker, under SDT, is to decide whether a particular sample of perceptual evidence was obtained from a distribution of noise or from a distribution of signal-plus-

noise (Figure 1). In the absence of observer bias, assuming both of these distributions are normal with equal variance, a criterion point β is established by the observer at the point where these two distributions intersect. If the sensory observation, X, exceeds the criterion β then the observer responds "Yes". If $X<\beta$ the observer responds "No". If $X>\beta$, and the observation came from the signal plus noise distribution, a 'hit' is recorded. If $X>\beta$ and the observation came from the noise distribution, a 'false alarm' is recorded. Similarly, whenever $X<\beta$, the recorded results will indicate either a 'correct rejection' or a 'miss'.

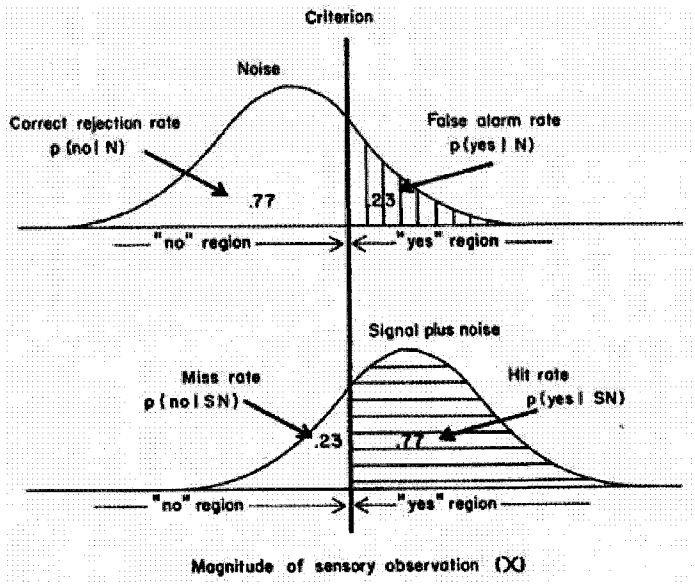

Figure 1. Signal Detection Theory's theoretical distributions (from Gescheider, 1997)

The likelihood of an observer making an error (i.e., a false-alarm or a miss) is a function of the degree of overlap of the noise and signal-plus-noise distributions. Where the assumptions of normality and equal variance hold in these distributions, the index of signal detectability becomes:

$$d' = \frac{\mu_{sn} - \mu_n}{\sigma_n} \quad (1)$$

(Modified from Gescheider, 1997)

Where μ_{sn} is the mean of the signal-plus-noise distribution, μ_n is the mean of the noise distribution, and σ_n is the common variance of the two distributions. Where the distributions cannot be assumed to have equal variance, the index of signal detectability becomes a function of the root-mean-square of the standard deviations of both distributions:

$$d_a = \frac{|a|}{\sqrt{.5(1+b^2)}} \quad (2)$$

(Modified from Gescheider, 1997)

where $|a|$ is the absolute value of the intercept when the observer's false-alarm scores are standardized (Z_N) and regressed on the observer's standardized hit scores (Z_{SN}), and b is the slope of this regression function.

Observer bias is often tested in SDT via the introduction of a pay-off matrix which alters the observers placement of the criterion β thusly:

$$\beta = \frac{p(N)}{p(SN)}\left[\frac{V[No \mid N] + V[Yes \mid N]}{V[Yes \mid SN] + V[No \mid SN]}\right] \quad (3)$$

(modified from Green & Swets, 1966)

Where $p(N)$ is the probability that the observation originated in the noise distribution, $p(SN)$ is the probability that the observation originated in the signal-plus-noise distribution, $V[No|N]$ is the value to the observer of responding 'no, there is no signal present' when the sample actually originated in the distribution of noise (this is usually a financial incentive established *a priori* by the researcher and made known to the

observer), *V[Yes|N]* is the value to the observer of responding 'yes, there is a signal present', when the sample actually originated in the distribution of noise, and *V[Yes|SN]* and *V[No|SN]* are similarly defined values where samples originate in the signal-plus-noise distribution. Apart, then, from the implicit assumption of (3) that every observer makes a cost-benefit analysis prior to making each decision in order to maximize gain and minimize loss, there is nothing in SDT that can explain non-optimal decision-making behaviour which does not benefit the observer in some quantifiable way. In other words, given two identical samples and two identical payoff matrices, an observer should reach two identical conclusions regarding whether or not a signal is present. Nothing in the theory, for example, would allow that participant behaviour can differ when a decision-maker makes the same choice twice: once with the requirement of rendering confidence after the decision is made, and once without.

SDT does not specifically discuss the origin of confidence, but it has been argued that confidence in SDT is a function of the distance between where in the sampled-from distribution an observer believes the sample to have originated and the position of criterion β (Petrusic & Baranski, 2003). Under SDT the difference between criterion β and the origin of the signal must be computed regardless of whether confidence is to be processed since it is through this comparison that the primary decision is made. Confidence rendering, therefore, should not create any additional burden for the decision-maker and a confidence rating should be available to the decider as soon as they have made a choice.

It has already been noted that the findings of Petrusic and Baranski (2003, 2000) and Baranski and Petrusic (2001) are problematic for SDT, since this theory cannot

explain why post-decisional times to render confidence vary. Though the point is not stressed by the aforementioned researchers, SDT should also be unable to explain why mean primary decision response times increase when observers are asked to render confidence since, under SDT, confidence processing does not alter the computational requirements of the primary decision.

The point of the present study is to demonstrate the extent to which decision-makers can exert control over confidence rendering. If confidence processes are subject to executive control, SDT could be thought of as describing a limited instance of sensory-based decision-making wherein the decision-maker has opted to render confidence entirely pre-decisionally. If, on the other hand, participants in the present study demonstrate an inability to stop processing confidence post-decisionally, the tenets of SDT would have to be viewed with scepticism for all of the reasons suggested by Petrusic and Baranski (2003, 2000) and Baranski and Petrusic (2001).

It should be noted that a secondary goal of Experiment 1 is to determine whether the cognitive mechanisms involved in confidence processing are different from those utilized during primary decision processing. If this turns out to be the case, single-process models by definition will be unable to explain any interaction between confidence rendering and decision-making, since single process models postulate a single underlying mechanism which governs both the primary decision and the generation of confidence.

Post-decisional Models

The Accumulator Model and the Balance of Evidence Hypothesis

Vickers (1979) has proposed what he has called the Accumulator Model of

decision-making (Figure 2). According to this model, an observer charged with making a sensory-based two-alternative forced-choice decision will gradually accrue evidence supporting both of the choice alternatives. When a criterion level of evidence has been accumulated in support of one of the alternatives, the observer makes a decision (Figure 2, Panels C and D). If the observer is faced with a time constraint and is prevented from gathering enough evidence to let either accumulator reach criterion, the decision will favour the alternative for which the most evidence has been accrued at the time the decision is made (Figure 2, Panels A and B).

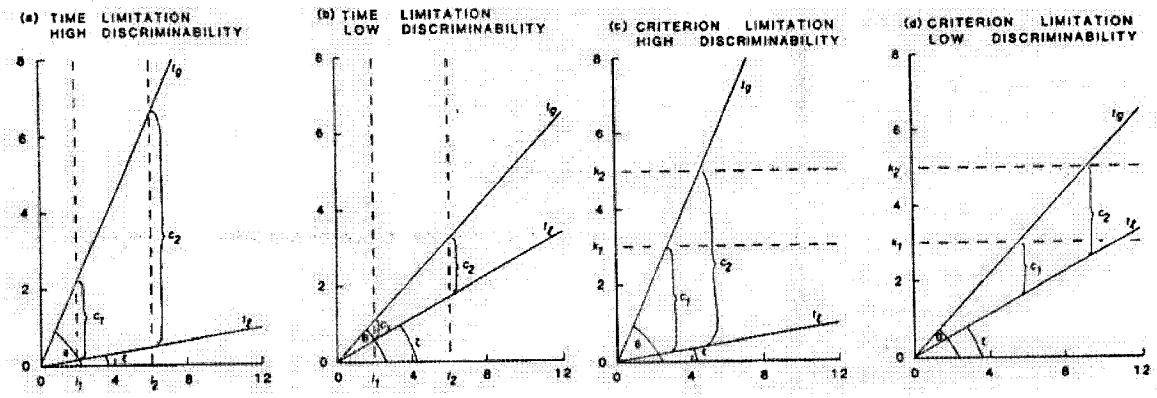

Figure 2. Hypothetical evidence accumulation plots under the Accumulator Model. The abscissa represents the number of observations and the ordinal represents the magnitude of accumulated evidence (from Vickers & Packer, 1982).

Confidence in the Accumulator Model is a function of the differences between accumulated levels of evidence (Figure 2, labels c1 and c2). As such, the model is able to make some clear predictions regarding the degree of confidence expressed by decision-makers under time constraints as well as by those under criterion constraints. For example, the model suggests at which time constraint level an observer will give the same confidence ratings as an observer allowed to accrue evidence at leisure.

The challenge posed by Petrusic and Baranski (2003, 2000) and Baranski and

Petrusic (2001) to this, as well as to other post-decisional models of decision-making, is that these models cannot explain how the requirement of rendering confidence causes increases in primary decision response times. All of these models consider confidence rendering to be a post-decisional computation of the difference between levels of accumulated evidence and, as such, cannot explain why a decision-maker would begin to compute this difference before enough evidence has been accrued to allow a decision to be made.

A component of a different Vickers model, Vickers' Adaptive Accumulator Module, does allow that confidence rendering can cause increases in primary-decision response times, but only in the long-run (1979). Following each decision made within a series of decisions, an observer compares her/his expressed confidence level to a subjectively determined target confidence level. If the expressed level of confidence is less than the target level of confidence, an 'under-confidence' accumulator is incremented. Alternatively, if expressed confidence is greater than target confidence an 'over-confidence' accumulator is incremented. If either of these accumulators reaches a threshold level, the primary decision-making criterion level is changed in order to maximize the efficiency of the decision-making process. In other words: if the observer has set a very high decision-making criterion and is repeatedly expressing absolute confidence in the correctness of their choices, the decision-making criterion will be lowered in order to allow the observer to spend less energy making what are, apparently, relatively easy decisions. Conversely, an observer consistently expressing low confidence following each of a series of decisions will raise the decision-making criterion in order to allow more evidence to be accrued before each future choice is made (Figure

2, Panels C and D show how this will help improve accuracy and increase confidence). As such, the module predicts that any increase in the overall contextual difficulty of a series of decisions will trigger an eventual increase in the primary decisional criteria, while any overall decrease in contextual difficulty will result in a comparable decrease in primary decisional criteria.

This module is unique among the evidence accrual models in that it allows for a post-decisionally-generated-confidence-based increase in decisional response times. Unfortunately, the model cannot explain the empirical finding that primary decision response times increase on trials where confidence rendering is required while the mean difficulty of the decisions being made is held constant over all trials.

The present study will demonstrate whether decision-makers can exert enough control over confidence processing to allow the suggestion that, occasionally, confidence is rendered entirely post-decisionally. As previously noted, such a finding would allow all post-decisional models a definite, though limited, utility. Further, if the cognitive mechanisms involved in confidence processing can be shown to be distinct from those involved in primary decision processing, then support will have been garnered for all post-decisional models since these models predict just such a separation.

Slow and Fast Guessing Theory

Petrusic and Baranski (1989; see also Petrusic & Jamieson, 1978; Petrusic, 1992; and Petrusic & Baranski, 2006) proposed a model of decision-making which allows for the accumulation of doubt in response to an evidence accrual event. Slow and Fast Guessing Theory (SFGT) holds that when an observer is attempting to make a two-alternative forced-choice decision, the observer establishes a set of two criterion points,

C_1 and C_2, along a decisional axis, d. Following evidence accrual event i, evidence accumulates in support of decision R_1 if $d_i < C_1$, and in support of decision R_2 if $d_i > C_2$. If $C_1 \leq d_i \leq C_2$, doubt evidence accrues (i.e., the evidence sampled does not conclusively support either alternative). Each of these three accrual event outcomes has a unique threshold: α_1, α_2, and α_3 respectively. Assuming the decision-maker does not have the option of responding 'I do not know', once doubt evidence has accumulated as far as the α_3 threshold the decision-maker will guess and respond R_1 with probability *g*, and R_2 with probability *1 – g* (Petrusic, 1992; see also Petrusic & Baranski 2006, Baranski & Petrusic 2003). Under SFGT assumptions, the probability of an observer deciding R_1 is:

$$p(R_1) = \sum_{s=0}^{\alpha_2-1}\sum_{t=0}^{\alpha_1-1}(\alpha_1+s+t-1)!\frac{p_1^{\alpha_1}p_2^s p_3^t}{(\alpha_1-1)!s!t!} + g\sum_{s=0}^{\alpha_1-1}\sum_{t=0}^{\alpha 2-1}(\alpha_3+s+t-1)!\frac{p_1^s p_2^t p_3^{\alpha_3}}{(\alpha_3-1)!s!t!} \quad (4)$$

Petrusic (1992)

where p_1, p_2, and p_3 are the probabilities that a sampled piece of evidence will support decision R_1, R_2, or doubt respectively. Further, the expected number of evidence accrual events required before decision R_1 is made is:

$$E(N \mid R_1) = \frac{\sum_{s=0}^{\alpha_2-1}\sum_{t=0}^{\alpha_1-1}(\alpha_1+s+t)!\frac{p_1^{\alpha_1}p_2^s p_3^t}{(\alpha_1-1)!s!t!} + g\sum_{s=0}^{\alpha_1-1}\sum_{t=0}^{\alpha 2-1}(\alpha_3+s+t)!\frac{p_1^s p_2^t p_3^{\alpha_3}}{(\alpha_3-1)!s!t!}}{p(R_1)} \quad (5)$$

Petrusic (1992)

SFGT is specifically concerned with modelling results obtained through empirical observations of the differences between decision-makers operating under accuracy stress and those operating under speed stress. The first class of observers are asked to, where necessary, sacrifice decision-making speed in favour of decision-making accuracy. The second class of observers are asked to do the reverse. According to the model, accuracy-

stress observers are assumed to set a high α_3 threshold relative to thresholds α_1 and α_2. Speed-stress observers set a relatively low α_3. Given these assumptions, and the use of Formulae 4 and 5, the model can accurately predict the results of a number of laboratory findings. For example, it correctly predicts that accuracy-stress observers will demonstrate greater primary decision response times (i.e., greater $E(N|R_1)$) when making incorrect responses compared to when they make correct responses. It also correctly predicts that speed-stress observers will behave oppositely, exhibiting faster response times when making incorrect decisions. (See Petrusic, 1992; Petrusic & Baranski, 1989).

SFGT also makes several unique claims about the nature of confidence processing in a sensory-based decision-making task, since the theory allows for two distinct sources of 'doubt' generation. First, the theory allows that doubt can accumulate following each evidence accrual event. Second, similar to the Accumulator Model, the difference between magnitudes of evidence accumulated supporting R_1 and R_2 will play a role in confidence generation. Specifically, if the doubt evidence accrual count is small, and the difference between accrued R_1 and R_2 evidence is great, the observer will express a high level of confidence. Conversely, if the doubt count is high and there are equal amounts of evidence gathered in support of R_1 and R_2, the observer will admit to having guessed.

These two assertions make this theory unique among the post-decisional class of models in that, in part, it allows for confidence to be generated both during the primary decision-making process as well as afterwards. Confidence is generated during the primary decision-making process via the accumulation of doubt evidence, and afterwards via a retrospective analysis of the differences between R_1 and R_2 levels of evidence. The first assertion limits SFGT, however, by preventing it from explaining how the

requirement of rendering confidence causes an increase in primary decision response times: SFGT proposes doubt evidence is gathered during the evidence accrual process whether confidence rendering is required or not.

The Sequential Sampling Model

Juslin and Olsson (1997) provide another model of confidence in a two-alternative forced-choice sensory-based decision-making task. This model, the Sequential Sampling Model (SESAM), describes the evidence accrual in terms of an observer's subjective 'mean sensation': after N sensory sampling differences are accrued, the mean sensory difference is calculated. If the absolute value of the mean sensory difference, $|\overline{X}|$, exceeds a criterion, C, an overt decision is rendered. If, on the other hand, $|\overline{X}| < C$ an additional sample observation is obtained. This process continues until either a decision is made or until $N = N_{max}$, at which point time has run out and the observer guesses:

$$p(\text{choosing alternative A}) = p(\text{choosing alternative B}) = .5.$$

Confidence ratings in SESAM are functions of the relative frequency of the number of discrete sensations that support the decision made compared to the total number of sensations stored in working memory. The exception to this rule is the instance where the observer has been forced to guess, in which case the observer reports the lowest possible confidence level regardless of how much evidence supporting either alternative is stored in working memory.

While in some ways similar to Vicker's Accumulator Model, SESAM makes some very different predictions and, as such, is included here as a separate model. Unlike the Accumulator Model, for instance, SESAM predicts a decision-maker who has run out

of time will have no more than a 50% chance of making a correct decision and will never report a confidence level greater than 'guess'. Furthermore, SESAM predicts that decision-making will be more sensitive/accurate than will confidence rendering: decision-making in SESAM is based on the sampling distribution of the mean sensation, which has a standard error $\sigma/\sqrt{N}$, while confidence is based on the distribution of individual sensations stored in working memory, which has a standard deviation σ. As such, it is rare for SESAM decision-makers under intense time constraints to exhibit over-confidence.

It should be noted that Vickers and Pietsch (2001) have thoroughly tested the theoretical assumptions of SESAM, paying particular attention to the assumption that observers have a limited memory window within which they can store their accumulated decision sensations. These researchers show that, among other things, SESAM predicts response time distributions that are incompatible with empirically obtained data, SESAM cannot correctly predict the accuracy results of decision-making studies where participants are faced with time constraints, and that a theoretical SESAM decision-maker will make more mistakes if they are allowed to gather more evidence over a longer period of time. This last finding is not only counter-intuitive but, according to Vickers and Pietsch, it also contradicts empirical observation. The researchers conclude: "despite an exhaustive search we were unable to find any region in the landscape of possible parameter values where the behaviour of the model adequately reflected the totality of empirical findings" (Vickers & Pietsch, 2001, p. 801). And while Juslin and Olsson (1997) claim to have found empirical support for their predictions, it should further be noted that these findings have not been found to be universally replicable (see, for

example, Baranski & Petrusic (1994), Baranski & Petrusic (1995), Baranski & Petrusic (1999), Petrusic & Baranski (1997), Petrusic & Baranski (2001), Petrusic & Baranski (2003), Petrusic (2003)).

Van Zandt's Race Model

Van Zandt (2000) presents a model which is, conceptually, quite similar to Vickers' model. It is here included as a distinct model because it makes some unique predictions about the distribution of the confidence ratings that should be obtained by decision-makers rendering confidence in a two-alternative forced-choice experiment. Van Zandt's Race Model describes decision-making as a race between opposing Poisson-process based, independent, evidence accumulation counters. As was the case with Vickers' model, however, each counter in the Race Model has its own separately adjustable threshold level. Also like the Vickers' model, confidence in the Race Model is a function of the differences between the two alternative evidence counters at the time that the decision is made. Specifically:

$$C = X_A(RT_{final}) - X_B(RT_{final}) \quad (6)$$

where C is the confidence rating, $X_A(RT_{final})$ is the number of evidence accrual events that have been accumulated in support of decision A at time RT, and $X_B(RT_{final})$ is the number of evidence accrual events that have been accumulated in support of decision B at time RT_{final} (modified from Van Zandt, 2000). Van Zandt goes on to suggest that, so long as the evidence accumulation rates for each alternative remain constant, the probability distribution of the C can be determined. The stability of this distribution depends on a constant evidence accrual rate of $\lambda_A + \lambda_B$ (i.e., the accrual rate of evidence supporting A plus the accrual rate of evidence supporting B) with the probability that any particular

evidence accrual event supporting A being $p_A = \lambda_A / (\lambda_A + \lambda_B)$ and the probability that the event will support alternative B is $p_B = \lambda_B / (\lambda_A + \lambda_B)$.

Van Zandt and Maldonado-Molina (2004) have presented a modified version of the Race Model that can account for the findings of Baranski and Petrusic (1998). It most notably predicts the finding that confidence is determined entirely post-decisionally by observers under speed-stress. The modified Race Model holds that four, rather than two, subjective decision-related thresholds are established by a decision-maker. The first two 'low' thresholds each correspond to one of the two possible alternative choices, with a 'very low' threshold suggesting the decision-maker is under a time constraint. The second two 'higher' thresholds are used to evaluate confidence. In essence, evidence accumulates until enough has been gathered to meet the requirements of one of the lower thresholds, at which point the decision-maker makes a corresponding choice. Evidence then continues to accumulate until the decision-maker has gathered enough to enable them to render confidence (*i.e.,* one of the higher evidence thresholds has been reached). While it should be noted that observers under accuracy stress only employ a single set of decisional criteria, confidence in either stress condition is still a function of the difference between the total accumulated levels of evidence supporting each of the two alternative possible choices.

This modified Race Model is important because it bridges a gap between Van Zandt's theory of decision-making and SFGT. Specifically, the modified Race Model suggests why, under speed constraints, a decision-maker is quick to make an initial decision but is relatively slow to render confidence. The first thresholds, in this case, are thought to be set considerably lower than the second set of thresholds. With the

thresholds set thusly, relatively little evidence would be required before a decision-maker could make a primary decision, yet much more additional evidence would be required before confidence could be rendered. Both versions of the Race Model, however, maintain a post-decisional locus of confidence generation and, as such, are subject to the same criticisms as are all post-decisional models: namely, that they cannot explain increases in primary decision response times when confidence rendering is required.

Introduction to Experiment 1

The Automaticity of Confidence

Motivating this experiment was the question of whether or not the initiation of confidence processing is a function of the larger context within which individual decisions are made. The experiment tested whether, in blocks of trials where the requirement of rendering confidence randomly changed from trial to trial, decision-makers would initiate confidence rendering processes on trials where they were not required to do so. In other words, if a decision-maker was required to render confidence on a random 50% of trials, would they adopt the strategy of processing confidence regardless of whether or not it was required rather than constantly 'turning the system on and off'?

Assuming a decision-maker has control over when they might decide to start processing confidence, there are three possible strategies a decision-maker might adopt in a random, or 'mixed', block of trials ('mixed block' is the terminology used by Los, 1999a: see below). First, the decision-maker might only render confidence on those trials where they are required to do so. Results, if this strategy were adopted, would be identical to those reported in other experiments that have used 'pure blocks' of trials

(terminology, again, is Los', 1999a) instead of mixed blocks of trials: participants would begin processing confidence before they have made their decision, and would complete the process post-decisionally (e.g., Baranski & Petrusic, 2001; Petrusic & Baranski, 2000, 2003).

A second possible strategy would involve simplifying the demands of the mixed block by always rendering confidence, but only reporting confidence when it was appropriate to do so. With this strategy, a decision-maker could avoid 'switching gears' from trial to trial and, regardless of whether confidence rendering was required on any given mixed block trial, primary decision response times would be identical to those found in a pure block of confidence trials. Mean times to render confidence, under this strategy, would also be completely parallel to those found in pure confidence blocks of trials.

A third strategy would be for a decision-maker to simplify the requirements of the mixed block task by choosing to never render confidence before making a primary decision, and only beginning to process confidence on trials where they are subsequently prompted to select a confidence category. Given this strategy, primary decision response times in the mixed block of trials should be identical to those found in a pure block of no confidence trials, and mean times to render confidence would be much larger than those found for a pure confidence block of trials.

If participants adopt the second strategy and consistently render confidence on trials where they are not required to do so, it could be argued that it is simply easier than not to go ahead and render confidence. As such, confidence could be considered the 'when-in-doubt default setting' of the decision-making process. If this were the case, no

author of a post-decisional nor single-process model could reasonably ignore the effects or time-course of confidence processing. On the other hand, if confidence is rendered only as required (i.e., participants are able to adopt either strategy one or three), modellers of decision-making could safely argue that, where their models cannot explain the effects of confidence rendering, they have instead modeled the limited instance where either a decision-maker has chosen to compute confidence entirely pre-decisionally (single process models) or the decision-maker is rendering confidence entirely post-decisionally (post-decisional models).

Mixing Costs: the Autonomy of Confidence?

Los (1999a, 1999b, 1996) describes mixing costs as systematic increases in participant response time when different levels of an independent variable are presented to participants during a single block of trials (i.e., the mixed block) as compared to participant response times when levels of the variable are presented each during a separate block of trials (i.e., the pure block). A mixing benefit is said to occur when response times decrease during mixed blocks of trials compared to pure blocks.

Los (1999a), in exploring the nature of mixing costs, asked participants to identify members of a set of digits, ranging from 2 to 5, that were obscured either by noise, in the form of extra black dots in close proximity to the digit presented, or by having parts of the digits 'whited-out'. He found that when, within a block of trials, the task randomly switched between identifying noise-obscured and white-out obscured digits, mixing costs occurred. He also found that when only noise-obscured digits were used, and the task switched between identifying digits obscured by either eight or twelve dots, no mixing costs were found.

Though Los' intent was to identify those conditions under which mixing costs do or do not emerge, he has inadvertently suggested a means by which one can identify whether two tasks involve distinct cognitive processes. Los (1999a, p. 15) postulated that "different computational processing demands of the stimuli constitute *a necessary condition for the occurrence of mixing costs*" (italics added). In other words, mixing costs and benefits will occur if and only if the two tasks being mixed utilize different cognitive processes. With regards to the current study, Los' findings have suggested several interesting, albeit secondary, hypotheses. If participants are presented with a mixed confidence block of trials (i.e., rendering confidence only following a randomly selected 50% of the trials), and these primary decision response times are compared to a pure confidence block of trials (i.e., 100% confidence trials), we would not expect to see mixing costs since, for both block types, the computational requirements of the primary decision do not differ. On the other hand, if the computational demands of confidence rendering differ from those of primary decision making, mixing costs might become manifest in post-decisional times to render confidence.

If the assertion that mixing costs become manifest only when the tasks being mixed involve separate cognitive mechanisms is a correct one, the discovery of mixing costs in times to render confidence would bolster the claims of post-decisional models and would be harmful to single-process models. Post-decisional models of decision-making explicitly predict that primary decision making and confidence rendering utilize separate cognitive mechanisms, while single process models claim both confidence and the primary decision arise from a single decision-making process.

Decision Types

Both experiments presented in this paper were conducted twice: once on a group engaged in a detection task, and once on a group performing a discrimination task. A detection task requires a decision-maker to indicate whether they believe a signal is being transmitted through a field of background of noise, and requires a 'yes/no' response. A discrimination task requires a decision-maker to decide which of two stimuli contain more/less of some specified characteristic, such as length. These are, arguably, two different types of sensory-based decisions, and both were examined in order to determine whether the findings of the present studies would generalize over both cases.

Experiment 1

Method

Participants

Forty first-year psychology students from Carleton University each participated in one 90-minute session in return for course credit. Six participants failed to complete the session, and were replaced. A computer malfunction resulted in the loss of 55 of 1000 of one participant's trials, but since this loss accounted for only 5.5% of the data collected, this participant's data were included.

Apparatus

The study was conducted using a desktop computer with a standard colour monitor. The computer was equipped with a Pentium-class processor, a Soundblaster soundcard, and a Windows 98 operating system. Stimulus presentation and response data collection was controlled via Superlab Pro v. 2.0. Participant responses were made via a control panel with two primary response buttons, labelled 'YES/NO' or 'LEFT/RIGHT'

depending on the task, and seven confidence response buttons, labelled 'X/50/60/70/80/90/100'.

Stimuli

Participants were assigned to either a line-length discrimination task or a signal detection task.

Line-length stimuli each consisted of a single, horizontal line bisected by a shorter 10 pixel high vertical line. The vertical line was positioned so that the length of one of the two line-segments making up the larger horizontal line was always 100 pixels in length, while the remaining segment was slightly larger (being either 102, 104, 106, or 108 pixels in length). Line-length stimuli were horizontally offset from centre of the computer monitor display, either 25 pixels to the right of centre or 25 pixels to the left to try to prevent participants from using the edge of the screen as the basis for their decision.

Accompanying each line-length pair was a text instruction, centred horizontally on the computer monitor, and positioned 125 pixels beneath the lines. The instruction read either 'LONGER' or 'SHORTER' and was printed using a 12 point, Arial Baltic Bold font face.

Each of the four line-length pairs in each of the left-right orders was paired factorally with each of the two instructions giving rise to 16 cells in the design. Over the course of the experiment each of the cells in the design was replicated 62 times for a total of 992 trials. 8 additional replications of randomly selected stimulus pairs were added to this total, allowing for an even 10 blocks of 100 trials.

The signal detection stimuli were 'dot-density displays'. Dot-density stimuli

consisted of 16 100 x 100 pixel squares. Each of the 16 squares contained 2500 'dots', represented by 2 x 2 pixel squares. Each dot was either inactive (coloured white) or active (coloured black). The density level of the dots in the four centre squares varied from trial to trial, but on any given trial all of the centre squares were homogenously dense (+/- .5%), and these densities were either 50% active dots (the 'noise' condition), 52% active dots, 54% active dots, 56% active dots, or 58% active dots. 50% of the dots (+/- .5%) in the surrounding squares were active on each trial.

Two-hundred and fifty of the noise condition stimuli, 62 of the 52% active stimuli, 62 of the 54% active stimuli, 62 of the 56% active stimuli, and 62 of the 58% active stimuli were each presented twice over the course of the experiment for a total 996 trials. 4 stimuli were randomly selected from each of the active categories and repeated, allowing for an even 10 blocks of 100 trials.

All other text displayed on the monitor during the course of the experiment, including the signals 'READY', 'CONFIDENCE', 'NO CONFIDENCE', and the confidence ratings '50/60/70/80/90/100' (see the Procedure section for more information) appeared centred horizontally and vertically, in a 12 point, Arial Baltic Bold font face.

Procedure

Participants were seated before the computer monitor, within comfortable reach of the response panel. Participants received verbal instructions from the experimenter and were presented with on-screen examples of the stimuli involved in the task. They were told that they would be presented with 10 blocks of stimuli, with a corresponding nine short breaks between each block.

Twenty participants were assigned to the line-length discrimination condition.

These participants were told that a stimulus and an instruction would appear simultaneously on the screen. If the instruction read 'LONGER', the participant's task was to decide which line-segment was longer, and to press the 'LEFT' or 'RIGHT' button on the response panel as was appropriate. Alternatively, if the instruction read 'SHORTER' participants were told to select the shorter line segment.

The remaining 20 participants were assigned to the signal detection condition. These participants were told that their task, following each presentation of a dot-density display, was to decide whether there was a greater density of dots in the centre of the display than there was around the edges of the display. If the densities varied, they were told to press the 'YES' button. Otherwise, they were told to push the 'NO' button.

All participants were asked to render confidence following one-half of the decisions they made. They were told that when it was time to render confidence the numbers '50/60/70/80/90/100' would appear on the display. To render confidence, they were asked to subjectively evaluate their certainty of having just made a correct decision, and were told to press the button that most corresponded to this evaluation: '50' was indicative of a guess, '100' indicated certainty, and the intermediate buttons (60-90) were to be used as the participant felt was appropriate. The 'X' button was to be pressed instead of a confidence rating if the participant was certain they had made a mistake.

Half of the participants in each decision-type condition, ten in the line-length discrimination group and ten in the signal detection group, were assigned to the pure block condition. These participants were asked to render confidence following each stimulus presentation only during alternating blocks of trials. Half of these participants, five for the line-length discrimination group and five for the signal detection group,

rendered confidence during odd numbered blocks, including block number 1, and the remaining participants, five per decision group, rendered confidence during even numbered blocks.

The remaining participants, ten per decision group, were assigned to the mixed block condition. These participants were asked to render confidence only in those situations where the word 'CONFIDENCE' preceded stimulus presentation. If the words 'NO CONFIDENCE' preceded stimulus presentation participants were told that they did not have to render confidence on that trial. Note that participants in the pure block condition saw the word 'READY' preceding each stimulus presentation, instead of the words 'CONFIDENCE' or 'NO CONFIDENCE'.

Results

Primary decision response times less than 200 ms in length were removed from analyses. Response times more than 3 standard deviations above each participant's mean response time were considered outliers and were also removed from analyses, as were 'mistakes' here defined as a participant pressing a confidence button while being required to press a decision button, or vice versa. Together, these cases accounted for 3.31 % of the 19945 trials within the signal detection group, and 2.9 % of the 20000 trials within the line-length discrimination group.

Response Time Distribution Assumptions

Since response time distributions are always positively skewed, Van Zandt, of the Race Model described in the General Introduction, has raised concerns about using traditional methods of null-hypothesis testing on a dependent variable which is never normally distributed (2002). This violation, however, was not of concern in this

experiment: nor, I would argue, should it be of concern to any researcher doing response time analysis research for two reasons. First, as is well known, according to the Central Limit Theorem the sampling distribution of the mean approaches a Gaussian distribution as the sample size approaches 30. In the case of the present studies, participant means were typically derived from samples of approximately 1000 observations. Second, though platykurtosis in a distribution can attenuate power (Stevens, 2002), response time distributions are rarely platykurtic and are typically quite peaked. This was certainly the case in the two experiments here presented, where the response time distributions were positively skewed and fairly leptokurtic.

Levels of Significance for the Analyses of Variance in Experiment 1

All analyses of variance (ANOVA) were within participant analyses, $\alpha_{critical}$ was set at .05, and only those main effects and interactions with a partial η^2 of .1 or greater were reported as being practically significant. Huyndt-Feldt degrees of freedom were used to measure significance, but the degrees of freedom here reported are those defined by the design. All graphs presented display 95% confidence intervals (CI).

Since p(confidence trial) = p(no confidence trial) = .5 throughout the mixed blocks of trials, the probability of achieving a run of increasingly longer length became increasingly small (e.g., p(run of 1)=.5, p(run of 2)=.25,...,p(run of 7)=.0078). As such, there were far more runs of size 0 or 1 in the analyses than there were, for example, runs of size 5 or 6. Given the random nature of the mixed block of trials, and the uncertainty of being able to have every participant involved in 'confidence/no confidence trial runs' of any given length when other independent variables were included in the design, ANOVAs involving the effect of run-length were analyzed using only run-length as a

within subjects variable, and only runs of up to length 4 were considered in the analyses. Run-length results should be interpreted with this consideration in mind.

Primary Decision Response Time Analyses

An ANOVA of primary decision response time within the signal detection group used block number (5 levels: blocks 1 and 2 were collapsed, blocks 3 and 4 were collapsed, blocks 5 and 6 were collapsed, blocks 7 and 8 were collapsed, and blocks 9 and 10 were collapsed), dot density (5 levels), and trial type (2 levels: confidence, no confidence) as within participant independent variables (IVs), and block type (2 levels: pure blocks, mixed blocks) as a between participant IV.

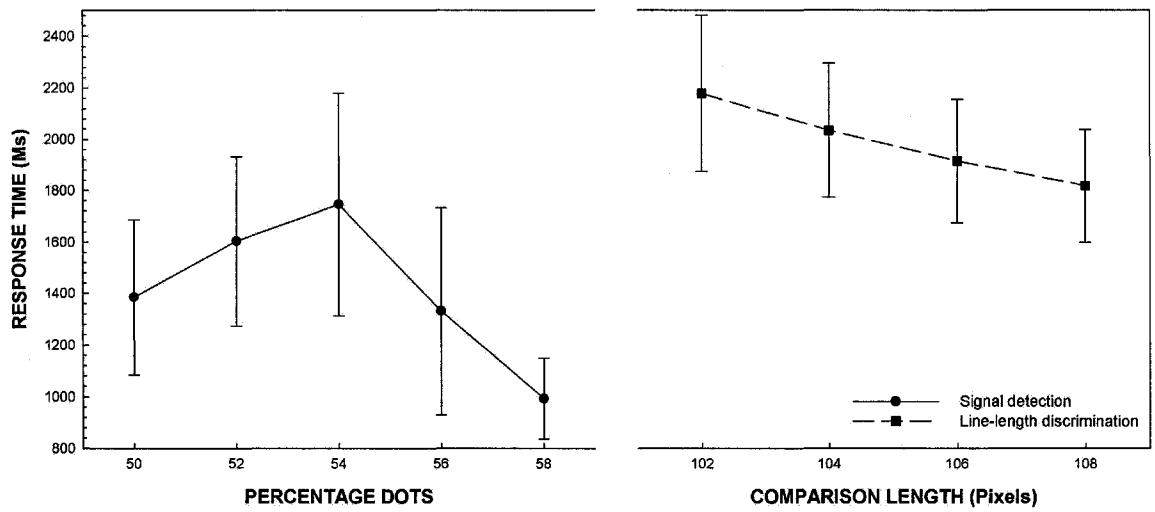

Figure 3. Effects of changes in decisional difficulty on primary decision response times for the signal detection group (Panel A) and the line-length discrimination group (Panel B).

A significant main effect of dot density was found, $F(4, 72) = 7.96$, partial $\eta^2 = .307$. On signal trials (the 52%, 54%, 56%, and 58% density levels), participants tended to respond more quickly as dot density increased, though an examination of Figure 3, Panel A and a review of the signal detectability and calibration analyses which follow suggests that participants often confused the 52% stimulus level with the 50% stimulus

level. A main effect of block number was found, $F(4, 72) = 31.19$, partial $\eta^2 = .634$, with participants responding more quickly as time wore on (a practice effect). An interaction between these variables (dot density and block number) was also found, $F(16, 288) = 7.39$, partial $\eta^2 = .291$: while in the first block of trials the difference in mean primary decision response time between a slight and obvious signal was relatively large, this difference grew less pronounced as block number increased.

A main effect of trial type was detected, $F(1, 18) = 10.04$, partial $\eta^2 = .358$, as was a significant interaction trial type and block type, $F(1, 18) = 17.64$, partial $\eta^2 = .495$. To further explain this interaction, analyses of simple effects were conducted. A simple effect of trial type within pure blocks was found, $F(1, 9) = 16.67$, partial $\eta^2 = .649$, where no such effect was found within the mixed blocks of trials. An examination of Figure 4, Panel A clarifies these findings by showing how participants took reliably more time to make decisions within pure confidence trials compared to pure no confidence trials, but were remarkably homogenous in their primary decision response times across trial types within mixed blocks of trials. Interestingly, no significant simple effects of block type were found within either type of trial.

A comparable ANOVA was conducted for the line-length group, with block number (5 levels), comparison line-length (4 levels), and trial type (2 levels) as within participant IVs, and block type (2 levels) as the between participant IV. The results of this ANOVA were similar to those found for the signal detection group: a main effect of comparative line-length was found, $F(3, 54) = 45.58$, partial $\eta^2 = .717$ (Figure 3, Panel B), and a practice effect was evidenced by a main effect of block number, $F(4, 72) = 34.84$, partial $\eta^2 = .659$. An interaction between these variables was also

found to be significant, $F(12, 216) = 7.742$, partial $\eta^2 = .301$: again, as block number increased, the differences in mean primary decision response time for the different comparison line-lengths decreased.

A significant effect of trial type was found, $F(1, 18) = 10.717$, partial $\eta^2 = .373$, as was a significant interaction between trial type and block type, $F(1, 18) = 10.956$, partial $\eta^2 = .378$. Simple effect analyses revealed a simple effect of trial type within pure blocks, $F(1, 9) = 14.557$, partial $\eta^2 = .618$, but no such effect was found within mixed blocks. No simple effects of block type within either trial type were found to be significant.

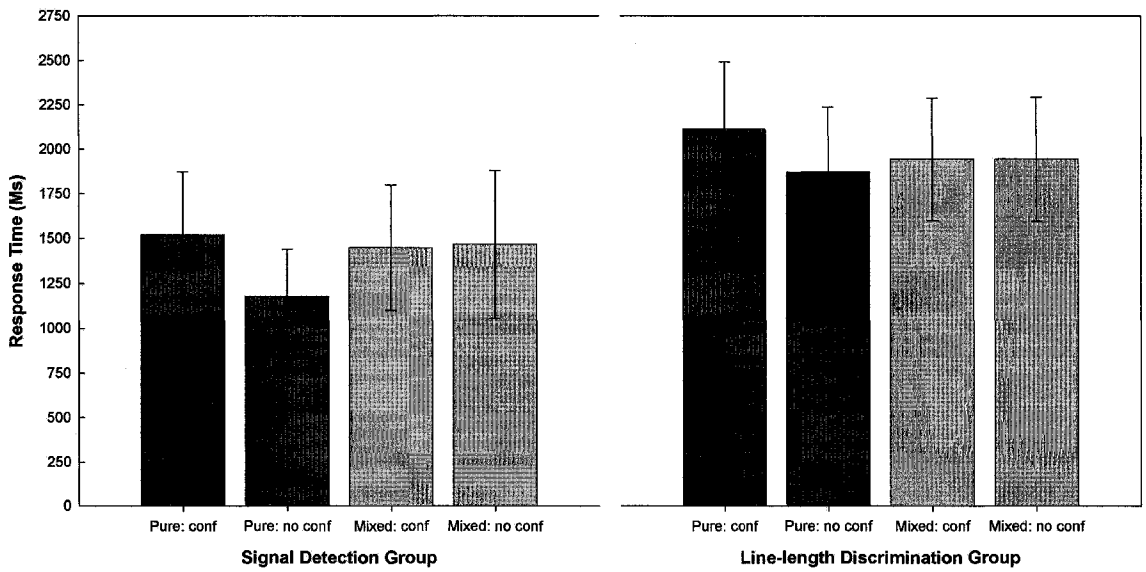

Figure 4. Changes in primary decisional response times as a function of trial type for both block types for the signal detection group (Panel A) and the line-length discrimination group (Panel B).

The finding, for both groups, that primary decision response times increased as the relative difficulty of the decision being made increased affirmed the effectiveness of the *a priori* manipulations of dot-density and differences in line-length. As well, the findings replicate the well documented dependence of response times on decisional

difficulty (e.g., Muensterberg, 1894). The finding that participants responded more quickly as time passed was indicative, perhaps, of their becoming experts at the task over the course of 1000 trials - though a more likely explanation is a gradual speed-accuracy trade-off in the face of a long experimental session. This later hypothesis seems borne out by the discriminative accuracy and detective sensitivity analyses which follow.

Importantly, homogenous mean primary decision response times were found within both decision-making groups in the mixed block of trials. Mixed-block participants were dedicating a constant amount of time to pre-decisional confidence processing regardless of whether or not confidence rendering was required. The signal detection mixed block group took almost, but not quite, as much time to process the primary decision as did participants in the signal detection pure confidence block of trials. This suggests that, within this decision-making group, a great deal of confidence processing was occurring before the primary decision was made (Figure 4, Panel A): though it should be noted that mixed block participants did not take reliably more time to make their primary decisions than did their pure no confidence block counterparts. Interestingly, the line-length discrimination mixed block group took homogenously less time to make each primary decision than did the pure confidence block line-length group, and more time to make each primary decision than their pure no-confidence counterparts; though, again, these differences were not reliable.

Response time analyses, therefore, seem to suggest some confidence processing was taking place in mixed-blocks of trials regardless of whether confidence rendering was required on any given trial.

The effects of run-length on primary decision response time.

Within the mixed-confidence blocks, of interest in this study was whether primary decision response times would reflect mixing costs or mixing benefits. If manifest, these would present themselves as systematic increases/decreases in mean primary decision response time with increases/decreases in either the successive number of confidence trials preceding a given trial (a 'confidence run') or the number of successive 'no confidence' trials preceding a trial (a 'no confidence run').[1]

No significant effect of either confidence trial or no-confidence trial run length was found for either type of trial within the signal detection group, neither was an effect of no-confidence trial run length found for either type of trial for the line-length discrimination group. Contrary to expectation, within the line-length discrimination group a significant effect of confidence trial run length on primary decision response time was found for both confidence trials, $F(4, 36)= 5.890$, partial $\eta^2 = .396$, and for no-confidence trials, $F(4, 36)= 9.869$, partial $\eta^2 = .523$. Primary decision response times in each of theses cases neither decreased nor increased systematically with increases in run-length. As such, it seems fair to conclude that these effects represent neither mixing costs nor mixing benefits.

Analyses of Time to Render Confidence

Two ANOVA's, one for each decision group, were used to examine the effects of the between participant factor block type (2 levels: mixed, pure), and the within

[1] It should be noted that, with regards to the literature on the effects of run-length, mixing costs have alternately been called 'switch costs'. Los (1999b) assigns equivalence to mixing costs and switch costs, and in so doing is able to hypothesize regarding the origins of the former as they are manifested in a run via a discussion of the literature on the latter.

participant factors block number (5 levels), and dot density / line-length (5 levels / 4 levels) on mean times to render confidence.

Importantly, significant main effects of block type were found for the signal detection group, with times to render confidence for the mixed blocks being consistently greater than times to render confidence in the pure blocks, $F(1, 18) = 33.88$, partial $\eta^2 = .653$ (see Figure 5, Panel A). A similar effect was found to be significant in the line-length discrimination $F(1, 18) = 5.84$, partial $\eta^2 = .245$ (see Figure 5, Panel B).

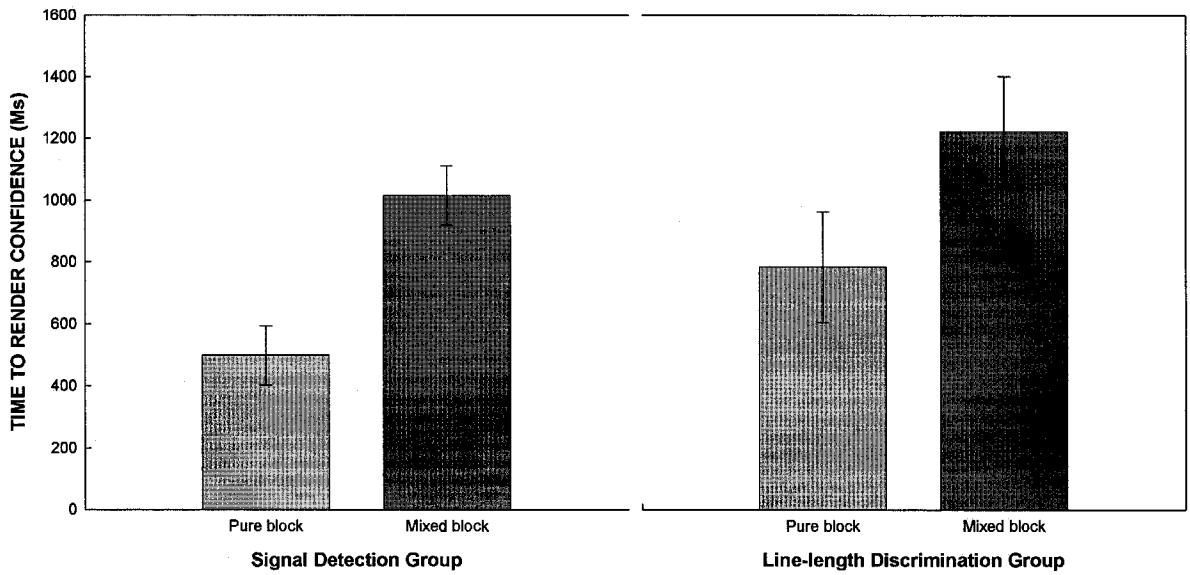

Figure 5. Changes in times to render confidence by block type.

A main effect of dot density on time to render confidence was found, $F(4, 72) = 6.96$, partial $\eta^2 = .279$, with times to render confidence tending to decrease with increases in signal detectability. No similar effect of comparative line-length was found (see Figure 6). Also of note in Figure 6 is the finding of a significant comparative line-length x block type interaction, $F(3, 54) = 5.24$, partial $\eta^2 = .111$, where no similar interaction was found for the signal detection group.

Practice effects were again evidenced by significant main effects of block number: signal detection group, $F(4, 72) = 50.91$, partial $\eta^2 = .739$; line-length discrimination group, $F(4, 72) = 22.94$, partial $\eta^2 = .56$; as well as a significant dot density x block number interaction, $F(16, 288) = 3.939$, partial $\eta^2 = .180$.

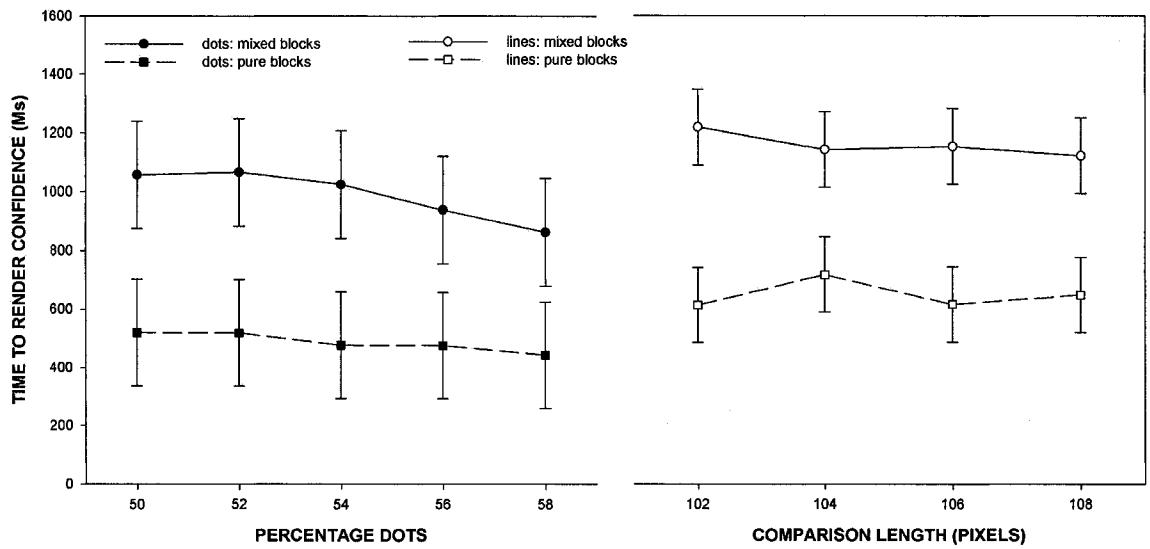

Figure 6. Changes in times to render confidence by block type as a function of dot density (Panel A) and line-length (Panel B).

It has already been noted that mean primary decision response times in the mixed block of trials were remarkably homogenous regardless of whether confidence rendering was required, and always greater than the mean response time for the pure no-confidence block of trials (Figure 4). This finding suggests that a consistent amount of confidence processing was always occurring pre-decisionally in the mixed block of trials. Coupled, however, with the finding that participant's times to render confidence were much slower in the mixed blocks of trials than in the pure confidence blocks of trials, an argument can be made that participants were adapting to mixed-block uncertainty by beginning to processes confidence before the primary decision was made, but that the bulk of

confidence processing was occurring almost entirely post-decisionally.

ANOVAs on time to render confidence with confidence category (6 levels) as a within participant IV and block type (2 levels) as a between participant IV found a main effect of confidence rating for the signal detection group, $F(5, 65) = 7.03$, partial $\eta^2 = .351$, but not for the line-length discrimination group (Figure 7). The former finding replicates Petrusic and Baranski (2003, 2000) and Baranski and Petrusic (2001), who found systematic decreases in time to render confidence with increases in expressed level of confidence. No reliable systematic decrease in time to render confidence was found for the line-length discrimination group.

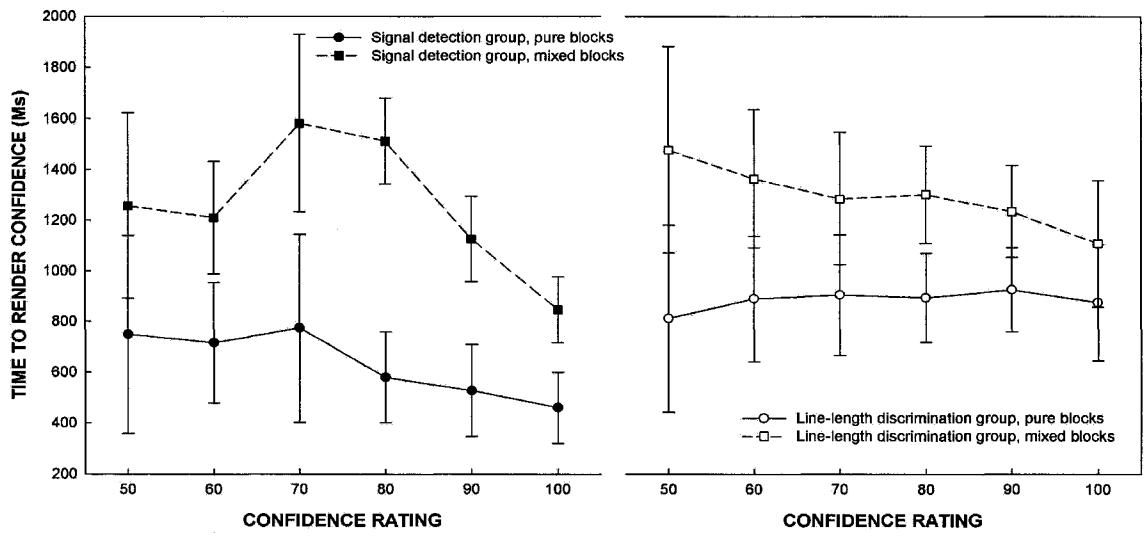

Figure 7. Mean time to render confidence as a function of confidence rating for pure blocks and mixed blocks for the signal detection group (Panel A) and the line-length discrimination group (Panel B).

The effects of run-length on time to render confidence.

Two ANOVAs were conducted for each decision-making group in order to test the effects of confidence trial run length and no-confidence trial run length on mean times to render confidence. Interestingly, significant main effects of both no-confidence trial

run length, $F(4, 36) = 5.682$, partial $\eta^2 = .387$, and confidence trial run length, $F(4, 36) = 16.842$, partial $\eta^2 = .652$, on time to render confidence were found for the signal detection group. This, following a review of Figure 8 (Panel A), suggests the presence of mixing costs and mixing benefits for this group. Similar effects were found for the line-length discrimination group (Figure 8, Panel B): confidence trial run length, $F(4, 36) = 9.227$, partial $\eta^2 = .506$; no-confidence trial run length, $F(4, 36) = 5.439$, partial $\eta^2 = .377$.

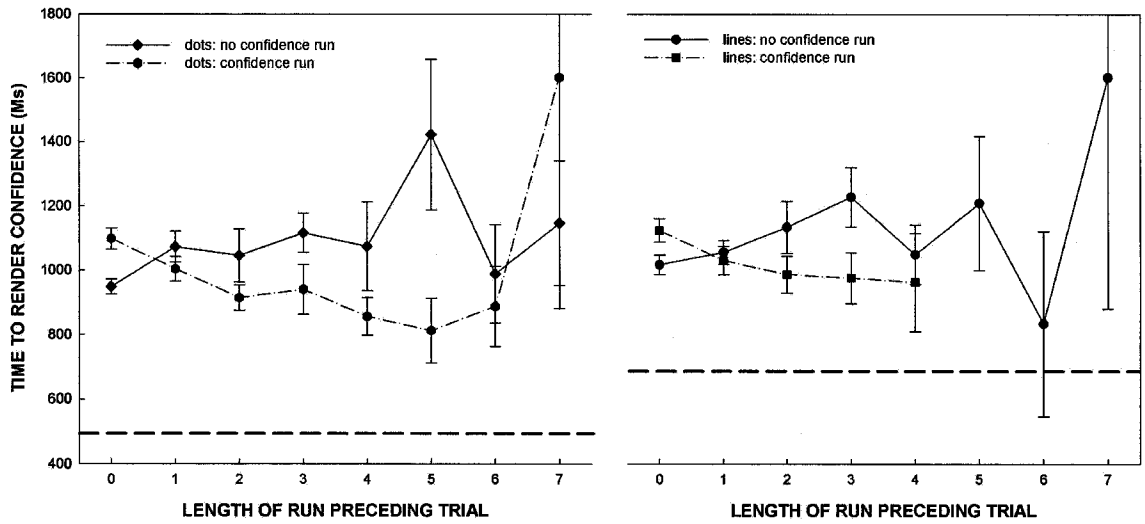

Figure 8. Effect of no confidence trial run length and confidence trial run length on time to render confidence for the signal detection group (Panel A) and the line-length discrimination group (Panel B). The dashed lines represent the mean time to render confidence collapsed over of both pure confidence blocks.

Generally, times to render confidence increased with the length of no-confidence trial run, and decreased with the length of confidence trial run: a run of no-confidence trials seemed to hinder confidence rendering and a run of confidence trials seemed to assist confidence rendering. According to the logic of Los (1999a), the presence of these mixing costs and benefits on mean time to render confidence suggest that confidence processing and primary decision-making are engaging different cognitive mechanisms.

Confidence Analyses

ANOVA structure for the analyses of mean confidence rating directly paralleled the ANOVA structure for the analyses of times to render confidence: two ANOVAs were conducted with block type (2 levels) as a between participant IV and with block number (5 levels) and dot density/line-length (5 levels/4 levels) as within participant IVs.

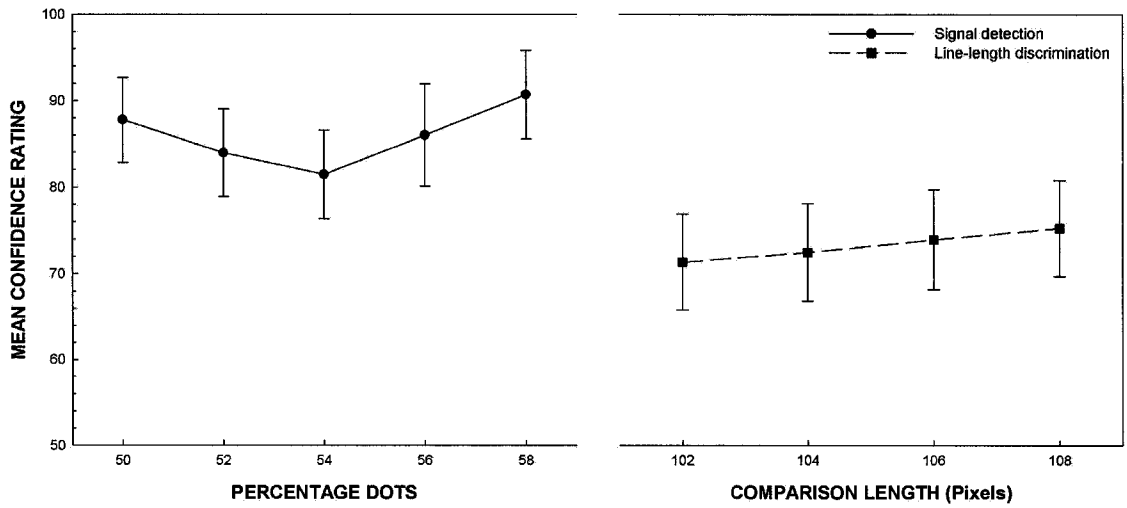

Figure 9. Mean confidence rating as a function of dot-density for the signal detection group (Panel A) and as a function of comparison length for the line-length discrimination group (Panel B).

Predictably, for the signal detection group, a main effect of dot density on expressed level of confidence was found, $F(4, 72) = 8.22$, partial $\eta^2 = .313$. When a signal was present, participant confidence generally increased with signal strength (Figure 9, Panel A). There were no significant differences in confidence ratings between the mixed and pure blocks of trials for the signal detection group.

Within the line-length discrimination group a significant main effect of comparative line-length on mean confidence was found, $F(3, 54) = 22.49$, partial $\eta^2 = .555$. As suggested in Figure 9 (Panel B), participants were (slightly) more confident

when presented with two lines that differed by a length of 8 pixels than they were when presented with two lines that differed in length by only 2 pixels.

Practice effects were again evidenced by a significant main effect of block number on mean confidence, $F(4, 72) = 3.08$, partial $\eta^2 = .146$, as well as significant block number x decisional difficulty interactions: signal detection group, $F(16, 288) = 5.83$, partial $\eta^2 = .245$; line-length discrimination group, $F(12, 216) = 1.974$, partial $\eta^2 = .099$.

Detectability and Discriminative Accuracy Analyses

ANOVAs on p(correct) used decision difficulty (signal detection: 5 levels; line-length discrimination: 4 levels) and block number (5 levels) as within participant IVs and block type (2 levels) as a between participant IV. Because of the dichotomous nature of the dependent variables in these analyses, p(correct) values obtained from participants were adjusted first using Berkson's correction (1953), which was used to attenuate extreme values of 0.0 and 1.0, and then all values were subjected to an arcsine transformation.

Main effects of dot density on detectability were found, $F(4, 72) = 100.63$, partial $\eta^2 = .848$, as was a main effect of line-length, $F(3, 54) = 111.53$, partial $\eta^2 = .861$. Predictably, in either case, participants became more accurate as the difficulty of the decision being made decreased.

Interestingly, participants in both groups were reliably, if only slightly, more accurate on trials where they were asked to render confidence: signal detection group, $F(1, 18) = 6.953$, partial $\eta^2 = .279$; line-length discrimination group, $F(1, 18) = 4.49$, partial $\eta^2 = .200$.

Main effects of block number were also found for the signal detection group $F(4, 72) = 5.64$, partial $\eta^2 = .239$, and the line-length discrimination group, $F(4, 72) = 11.61$, partial $\eta^2 = .392$. Participants in both groups became significantly less accurate as time passed. Along with the other practice effects previously noted, this finding suggests that participants gradually began to sacrifice accuracy in favour of speed.

It should also be noted that a significant interaction between block number and dot density was also found for the signal detection group $F(16, 288) = 5.30$, partial $\eta^2 = .227$.

The effect of run length on detectability and discriminative accuracy

Four ANOVAs were conducted for each decision-making group, two to examine the effect of confidence trial run length on accuracy for both trials where confidence was rendered (confidence trials) and trials where it was not (no-confidence trials), and two to examine the effects of no-confidence trial run length on accuracy on confidence and no-confidence trials. Accuracy DVs were adjusted as noted in the analyses preceding.

Within the signal detection group, a main effect of confidence trial run length was found for no-confidence trials, $F(4, 36) = 70.51$, partial $\eta^2 = .887$, and for confidence trials, $F(4, 36) = 10.35$, partial $\eta^2 = .535$. While there was no apparent pattern to the effect of confidence trial run length for no-confidence trials, accuracy seemed to diminish as confidence trial length increased for the confidence trials. Similarly, accuracy diminished with length of no-confidence trial run for no-confidence trials, $F(4, 36) = 21.45$, partial $\eta^2 = .704$, but there was no pattern to the significant effect of no-confidence trial run length for confidence trials, $F(4, 36) = 21.82$, partial $\eta^2 = .708$. Interestingly, within the signal detection group, even a small run of one sort of trial

caused accuracy to diminish when the run was followed by a comparable trial. For signal detectors, familiarity, it would seem, breeds contempt.

Within the line-length discrimination group no effect of no-confidence trial run length was found on accuracy for confidence trials, but a significant effect was found for no-confidence trials, $F(4, 36) = 5.44$, partial $\eta^2 = .377$. An effect of confidence trial run length on accuracy was found for both confidence trials, $F(4, 36) = 4.68$, partial $\eta^2 = .342$, and no-confidence trials, $F(4, 36) = 5.24$, partial $\eta^2 = .368$. There were no apparent patterns to any of these effects, with accuracy neither steadily increasing or decreasing with the length of trial run.

Analyses of Calibration and Resolution

Baranski and Petrusic (1994) suggested using measures of calibration and resolution in confidence-based analyses as a means of determining the extent to which participants' expressions of confidence reflect their actual levels of performance.

Calibration can be defined as the extent to which expressed confidence matches performance. For example, a participant with perfect calibration will make a correct response only 50% of the time when they express 50% confidence, will be correct 60% of the time when they express 60% confidence, etc... Resolution can be defined as a participant's ability to use confidence categories to differentiate right answers from wrong. A participant with perfect resolution will express 50% confidence every time they make an incorrect choice, and will express 100% confidence every time they make a correct choice. Two measures of resolution are here presented: resolution and standardized resolution (η^2). Standardized resolution is directly comparable to the familiar measure of effect size η^2, and can be thought of as the proportion of total

variability in participant p(correct choice) that is accounted for by the use of the confidence categories. A fourth measure of each participant's general bias is over/under-confidence, which is typically given by the mean percent confidence minus mean percent correct, with negative values indicating under-confidence and positive values indicating over-confidence.

Table 1. Mean Calibration, Over/Under Confidence, Resolution, and η^2 for the different levels of decisional difficulty for both groups in Experiment 1.

Stimulus Type	Level	Calibration	Over/Under-Confidence (expressed as a percent)	Resolution	η^2
Dot-density	50% dense	.025	-6.83	.009	.207
	52% dense	.492	59.07	.052	.498
	54% dense	.145	17.01	.037	.208
	56% dense	.028	-4.53	.009	.131
	58% dense	.016	-4.96	.005	.145
Line-lengths	100 vs 102	.065	4.30	.009	.041
	100 vs 104	.031	-5.84	.008	.049
	100 vs 106	.038	-11.89	.006	.050
	100 vs 108	.037	-11.94	.007	.074

For each participant indices of calibration, resolution, standardized resolution, and bias were obtained. Table 1 provides the mean of each of the individual participant indices as a function of decisional difficulty.

For each decision-making group, separate one-way ANOVAs were conducted for each performance index, separately using dot density/comparative line-length and block number as IVs. No significant main effects of decisional difficulty or block number were found for either resolution or η^2. While significant main effects on calibration were found for block number, (signal-detection, $F(9, 140) = 2.94$, partial $\eta^2 = .174$; line-length

discrimination, $F(9, 140) = 7.06$, partial $\eta^2 = .312$), the only easily interpretable effect was that of decisional difficulty on calibration: dot-density, $F(4, 95) = 40.48$, partial $\eta^2 = .630$; comparative line-length, $F(3, 76) = 4.57$, partial $\eta^2 = .153$. As can be seen in Table 1., replicating the findings of Baranski and Petrusic (1994), calibration for both groups improved as decisional difficulty decreased.

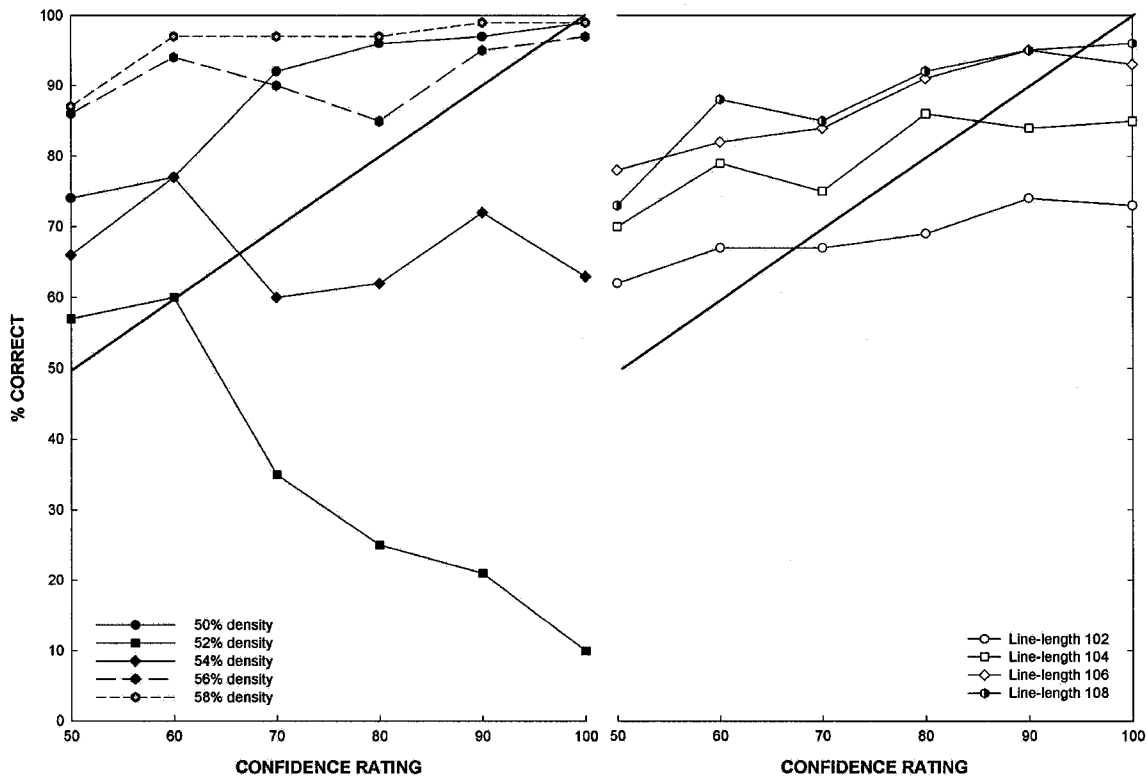

Figure 10. Mean percent correct as a function of confidence category for the signal detection group (Panel A) and the line-length discrimination group (Panel B) for each level of decisional difficulty in Experiment 1.

Also apparent in Table 1 is the significant effect of decisional difficulty on over/under-confidence: dot-density, $F(4, 95) = 48.90$, partial $\eta^2 = .673$; comparative line-length, $F(3, 76.323) = 7.0$, partial $\eta^2 = .224$. Participants were generally over-confident

when making difficult choices, and under-confident when making easy choices. Figure 10 further illustrates this phenomenon.

Discussion of Experiment 1

A review of Figures 3, 4, 9, and 10 will suggest that participants in the signal detection task occasionally behaved differently than did participants in the line-length discrimination task, with the two sets of participants diverging most particularly within indices of primary decision-making performance. For example, within the signal detection group the mean primary decision response times for the mixed blocks of trials were only slightly less than the means for the pure confidence blocks of trials, yet within the line-length discrimination group the mean primary decision response times for the mixed blocks of trials more closely paralleled the means for the pure no-confidence blocks. Though some minor differences between these two decision-making groups would have been expected in any case, the extent of some of these differences is striking and bears discussion.

These differences have two potential origins: one methodological, and one theoretical. First, it is possible that the differences between the two groups were amplified by the use of block-type as a between group variable: pure-block participants never made mixed-block decisions and *vice versa*. This concern was addressed with some success in Experiment 2, wherein block-type was treated as a within participant variable (see Experiment 2: Results).

The second possible source of divergence between these two groups originates in nature of each type of decision being made. While line-length discrimination participants knew that one of the two stimuli presented had to contain more of the quality being

probed, namely 'shortness' or 'longness', participants in the signal detection task had the option of rendering a more speeded though less accurate decision of "no, there is no difference between the stimuli" when presented with a faint signal. This suggestion is supported by the findings that signal detection group participants often confused the 52% signal with the 50% signal, and that they took less time to make decisions when presented with a 52% signal than when presented with a 54% signal. One could test this theory empirically by turning the signal detection task into a two-alternative forced-choice task, by asking participants which dot-density field contains "more" or "less" dots. In doing so, one could directly test whether dot-density is analyzed in a fashion comparable to line-length. Alternatively, the detection task might have been run with but a single signal strength and a noise condition within each block, with signal strength varying over blocks.

In any case, the confidence related indices for both decision-making groups in this experiment were quite comparable, and this was expected. Regardless of the model being considered and whether the nature or processing requirements of the decisions themselves might differ, the theoretical basis for confidence processing in each case remains consistent. As such, it seems reasonable to generalize the findings of this experiment across decision types and draw several, overarching conclusions.

Though participants in the mixed blocks of trials always engaged in some pre-decisional confidence processing, these participants were able to put off the bulk of their confidence processing until after the primary decision had been made. So while these participants were not able to fully embrace the third proposed strategy (postponing confidence processing until after a choice had been made) this was the direction they, at

least, tended towards in the mixed block of trials. This implies an impressive, albeit most likely unconscious, degree of control over when a decision-maker can choose to begin confidence rendering. With this finding in mind, the authors of post-decisional models of decision-making can make the principled claim that their works describe a limited instance of decision-making where, quite literally, decision-makers are choosing to render confidence post-decisionally. These modellers must acknowledge, however, that confidence processes are likely always initially engaged before any primary decision is made.

Mixing costs and benefits manifested themselves in times to render confidence during the mixed blocks of trials. If Los (1999a) is correct in his assessment that mixing costs only appear when the two tasks being mixed rely on different cognitive processes, one must allow that confidence rendering processes and decision-making processes are independent of one another. The discovery of mixing costs/benefits in times to render confidence are problematic for SDT and other single process models because these models suggest confidence and the primary decision are part-and-parcel of a single process. On the other hand, post-decisional models that suggest confidence and decision-making involve two distinct processes are bolstered by this finding.

Introduction to Experiment 2

The Stop-Signal Procedure

Band, van der Molen, and Logan (2003) describe a means by which one can evaluate a participant's ability to stop a cognitive process once it has been initiated in a mixed block of trials. They call this process the 'stop-signal procedure'. In a typical stop-signal experiment, participants are presented with a visual stimulus about which they

must make an evaluation. On random trials participants hear a tone, presented either simultaneously with or later than the presentation of the visual stimulus, and this tone signals them to stop evaluating the stimuli. By varying the temporal difference between the onset of the visual stimuli and the onset of the stop processing signal, and subsequently calculating the probability that a participant will respond regardless of the presentation of that signal, one can measure the 'ballistic' properties of a given cognitive process. For instance, if a participant is only able to hold back a response when the visual stimuli and the stop processing signal are presented simultaneously, one can conclude that the cognitive processes involved in generating the response are relatively unstoppable once initiated. On the other hand, if a participant can refrain from giving a response even at long stop Signal Onset Asynchronies (SOAs), it is reasonable to conclude that the response processes involved in the task are quite solidly under executive control.

This stop-signal paradigm can be used to ascertain when in a sensory-based decision-making task participants can start, and whether they can stop, rendering confidence. To do this, it must first be understood that the requirement of rendering confidence increases primary decision response times for participants in a pure confidence block of trials compared to those in a pure no confidence block (for one of many replications of this phenomenon, see Petrusic & Baranski, 2000; or simply note the differences in the mean primary decision response times for the pure blocks in Experiment 1). If participants in a sensory-based decision-making experiment were told to always expect to have to render confidence, and were told to refrain from doing so only if they heard a tone, one might expect the differences in primary decision response

times at various SOAs to reveal when confidence rendering is being initiated as well whether this process can be stopped. This hypothesis was the motivation behind Experiment 2.

It should be noted that this experiment is not a typical stop-signal experiment *per se* in that, due to limitations of the software used to manage the experiment, no direct record was kept of the number of times participants actually rendered confidence when they were signalled not to. Participant ability to stop rendering confidence was deduced by a comparison of their mean response times at various SOAs to their mean response times in trials where confidence rendering was required.

Experiment 2

Method

Forty-eight first-year psychology students from Carleton University each participated in one 90-minute session in return for course credit. Seven participants failed to complete the session, and were replaced.

Materials

Apparatus and visual stimuli were identical to those used in Experiment 1. The audio tone used in this experiment was presented through a standard pair of desktop computer speakers, had a frequency of 700 Hz, and was played for 185 ms at an amplitude/volume which varied for each participant, but which was set at a level the participant could hear without the tone being uncomfortably loud.

Procedure

As was the case in Experiment 1, half of the participants (n=24) were assigned to the line-length discrimination task. Each of the four line-length pairs used in this task

appeared with each of the two instructions for each of the two left-right orders, once offset to the left and once to the right, and this factorial combination was repeated 12 times for a total of 384 trials in each of three blocks.

The remaining participants (n=24) were assigned to the signal detection task. Within this task, half of the trials in each block were noise trials and the other half signal trials, with each of the four signal strengths occurring equally often. There were a total 336 trials in each of three blocks presented to participants in this group.

All subjects participated in three blocks of trials, though the order of these blocks was varied between participants, giving 6 block orders with four participants assigned to each order (hence twenty-four participants per task). Two of these blocks were identical to the pure blocks of trials described in Experiment 1: in one block participants never rendered confidence, and in a second block participants always rendered confidence following each primary decision. In a third block of trials, participants were told to always expect to have to render confidence but were instructed not to do so if they heard a tone. The tone was presented randomly on one-half of the trials within this mixed block, and it sounded equally often at each of the 0ms, 100ms, 200ms, 300ms, 400ms, and 500ms SOAs following presentation of the visual stimulus (32 trials at each SOA within the mixed block for the line-length discrimination group, 28 trials at each SOA for the signal detection group).

Results

As in Experiment 1, primary decision response times less than 200ms in length were removed from analyses, and response times more than 3 standard deviations above each participant's mean response time were considered outliers and were also removed

from analyses. Mistakes, again defined as a participant pressing a confidence button while being required to press a decision button, or vice versa, were also removed from analyses. Together these cases accounted for 2.41 % of the 24,192 trials within the signal detection group, and 2.13 % of the 27,648 within the line-length discrimination group.

Levels of Significance for the Analyses of Variance in Experiment 2

All analyses for the second experiment were within participant ANOVAs. Huyndt-Feldt degrees of freedom were used, but the degrees of freedom reported are those defined by the design. The level of significance was set at .05 throughout, but only those effects with a partial η^2 of .1 were reported as being practically significant. All graphs display 95% CIs.

The Effects of Block Order

As noted in the method section above, all participants took part in three blocks of trials (a pure confidence block, a pure no confidence block, and mixed stop-confidence block). The order of block presentation varied for each participant, with 4 participants from each decision group participating in each of the 6 possible block order presentations (e.g., 1. confidence, no confidence, stop-confidence; 2. confidence, stop-confidence, no confidence; etc...). Throughout the analyses conducted, significant interactions involving block order were found: though the effects manifested were disorderly and seemed to defy interpretation. These effects are, possibly, an artefact of the relatively small number of participants assigned to each order. No further detailed discussion of these between participant effects will follow.

Primary Decision Response Time Analyses

The ANOVA of primary decision response time within the signal detection group

used block type (pure confidence, pure no confidence, and the two stop-confidence sub-blocks of trials: tone and no tone) and dot density (5 levels) as within participant IVs, and block order (6 levels) as a between participant IV. A similar ANOVA, conducted for the line-length discrimination group, used comparison line-length (4 levels) as a within participant IV.

A main effect of dot density on primary decision response time was found for the signal detection group, $F(4, 72)=25.04$, partial $\eta^2 = .582$, and a similar main effect of line length was found for the line-length discrimination group, $F(3, 54)=27.03$, partial $\eta^2 = .600$. In either case, as was found in Experiment 1, response times decreased as detective sensitivity and discriminative accuracy increased.

Table 2. Mean primary decision response times (Ms) for each condition in the signal detection and the line-length discrimination tasks.

	Condition			
		Stop-Confidence		
Task	No Confidence	Tone	No-Tone	Confidence
Signal Detection	1421.16	1548.21	1810.31	1963.57
Line-length	1956.03	1653.36	1845.83	2301.99

A significant main effect of block type was found for both the signal detection group, $F(3, 54) = 9.09$, partial $\eta^2 = .335$, and the line-length discrimination group, $F(3, 54) = 15.57$, partial $\eta^2 = .464$. Comparing the two pure blocks of trials, participants were always slowest making their primary decisions during the pure confidence block of trials. Similarly, within the stop-confidence block of trials, participants were always slowest on trials when confidence rendering was required (i.e., the no-tone trials). These findings replicate Baranski and Petrusic (2001) and Petrusic and Baranski (2000, 2003) by suggesting that, for either block type, whenever confidence rendering was required,

processing began before the primary decision was made. As already noted, this finding is problematic for strictly post-decisional models of decision-making.

Curiously, the mean primary decision response time for the pure no confidence block of trials was less than either stop-confidence mean for the signal detection group, but was greater than either stop-confidence means for the line-length discrimination group (see Table 2).

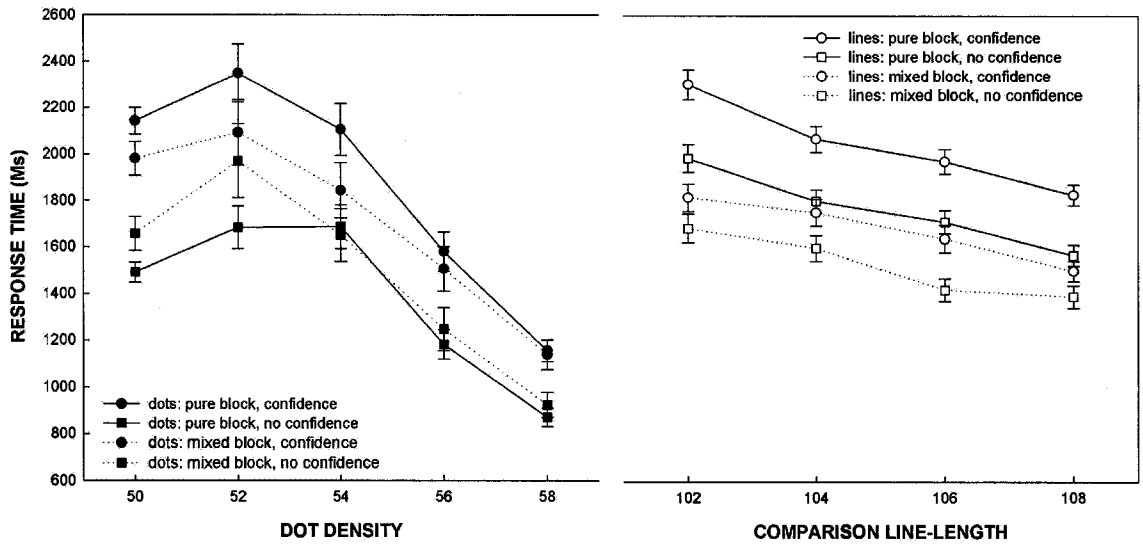

Figure 11. Mean primary decision response times for each block type at each level of decision difficulty for both the signal detection group (Panel A) and the line-length discrimination group (Panel B).

Interestingly, an interaction between block type and decisional difficulty was found for both groups: signal detection, $F(12, 216) = 1.93$, p=.055, partial $\eta^2 = .097$; line-length discrimination, $F(9, 162) = 3.68$, partial $\eta^2 = .168$. Within the signal detection group the differences in mean primary decision response time between blocks appear to decrease along with the degree of decision difficulty (Figure 11, Panel A). Of particular note is the way in which mean response times for confidence trials in mixed and pure blocks begin to converge as decision difficulty decreases, and the similar way in which

mean response times for the no confidence trials converge. Illustrated in Figure 11, Panel B is the pattern of interaction for the line-length discrimination group, though this pattern is not as easily interpreted as was that of the signal detection group.

These findings do not bode well for single process models. Recall that response time under SDT is a function of the difference between an arbitrarily established decisional criterion point and the point in the distribution of signal-plus-noise (or noise) that the decider believes a signal to have originated (i.e., larger differences equal shorter response times). The only way SDT could explain the observed converging patterns of mean decisional response times for pairs of otherwise identical trials (i.e., between confidence trials in the mixed vs. pure blocks, and between no confidence trials in mixed vs. pure blocks) is to allow that decision-makers shift decisional criterion for reasons having nothing to do with either signal strength or material gain. In this instance, SDT would have to allow that the criterion is shifting due to differences in the larger 'block context' within which each individual choice is being made. It is not immediately clear how SDT could describe these effects, nor how SDT could justifiably model decision-making without confidence, given two patterns of 'no-confidence trial' response times that seemingly depend on whether confidence rendering might be required on a future, adjacent trial.

The Effect of Tone Stimulus Onset Asynchrony on Primary Decision Response Times

To examine the effects of varying tone SOA on response time, ANOVAs were conducted for both decision groups for the stop-confidence block with SOA (7 levels) as a within participant IV.

A significant main effect of tone onset delay on primary decision response time was found for the signal detection group, $F(6, 138) = 3.43$, partial η^2=.130, as was a main effect of tone onset delay for the line-length discrimination group, $F(6, 138) = 7.21$, partial $\eta^2 = .239$. As suggested by Figure 12, a linear contrast was also found to be significant for both groups: signal detection, $F(1, 23) = 12.52$, partial $\eta^2 = .353$; line-length discrimination, $F(1, 23) = 19.18$, partial $\eta^2 = .455$.[2]

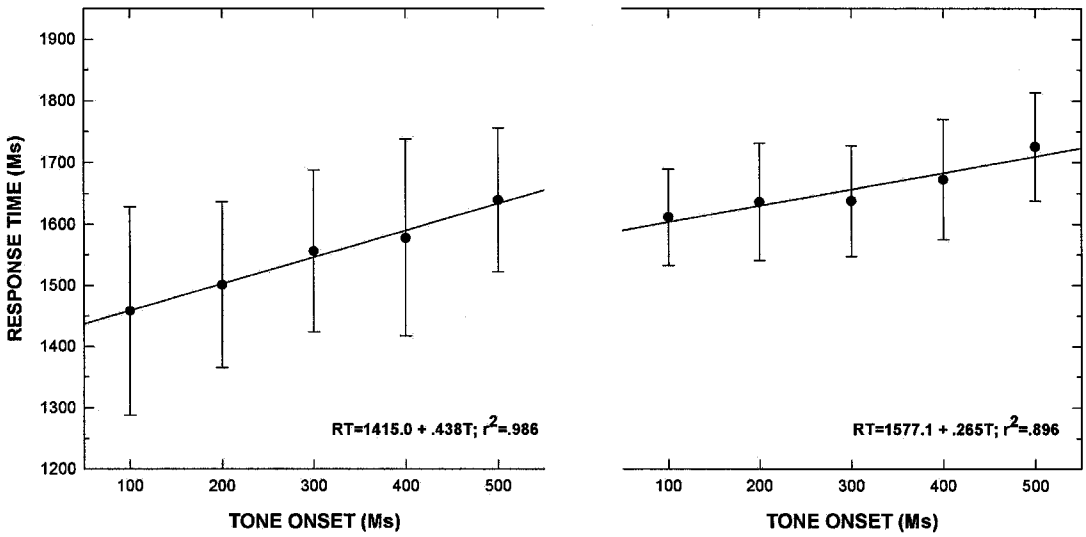

Figure 12. Mean primary decision response times for stop-confidence trials for the signal detection group (Panel A) and the line-length discrimination group (Panel B).

These findings were of paramount concern to the present study. First, they demonstrate that primary decision response times increase linearly with increases in the stop-confidence SOA (Figure 12). Apparently, participants in the stop-confidence block began to process confidence immediately following the presentation of the stimuli to be compared and, unless interrupted, confidence steadily evolved for at least the duration of the longest SOA used in this study (500 ms). Secondly, and this second point is implied

[2] As is evident in Figure 12, the linear analyses were restricted to tone onsets varying between 100 and 500 ms. Stop-signals also occurred at a 0 ms onset where the mean response time was 1555.83 ms for the signal detection group and 1637.38 ms for the line-length discrimination group.

by the first, participants were clearly able to stop rendering confidence on demand. This latter finding suggests, rather strongly, that confidence processes are not ballistic in nature: once initiated, the processes can be stopped, though at a cost in terms of primary decision response time.

Analyses of Time to Render Confidence

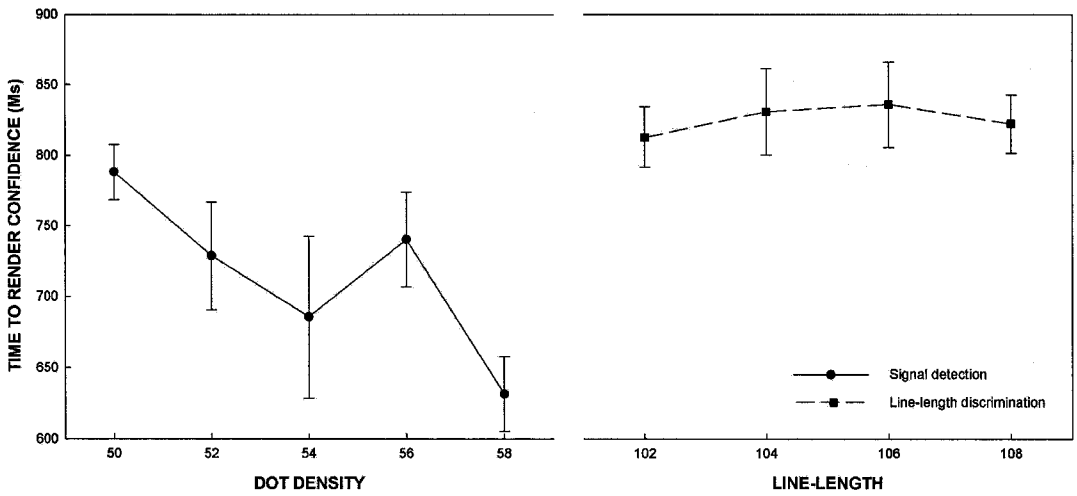

Figure 13. Times to render confidence by dot density for the signal detection group (Panel A) and by line-length for the line-length discrimination group (Panel B).

The ANOVA of time to render confidence within the signal detection group used block type (pure confidence and stop-confidence) and dot density (5 levels) as within participant IVs, and block order (6 levels) as a between participant IV. A similar ANOVA, conducted for the line-length discrimination group, used comparison line-length (4 levels) as a within participant IV.

A main effect of dot density on time to render confidence was found for the signal detection group, $F(4, 72) = 6.48$, partial $\eta^2 = .265$, where confidence times decreased with increases in dot density (Figure 12, Panel A). Oddly, no effect of discriminative difficulty was found for the line-length discrimination group (Figure 13, Panel B).

While the former finding is a replication of Petrusic and Baranski (2003), the later finding is curious. It would seem to indicate that, in general, little confidence processing was occurring post-decisionally within the line-length discrimination group. And yet, as was the case in Experiment 1, participants took significantly longer to render confidence in the mixed, stop-confidence block than they did in the pure confidence block for both groups (Figure 14): signal detection group, $F(1,18) = 70.81$, partial $\eta^2 = .797$; line-length discrimination group, $F(1,18) = 48.23$, partial $\eta^2 = .728$. This finding suggests post-decisional confidence processing is occurring within the stop-confidence block for both groups.

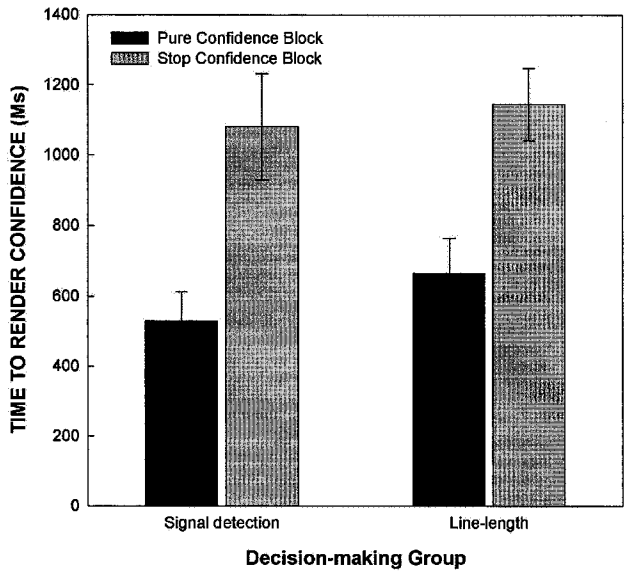

Figure 14. Times to render confidence as a function of block type and decision-making group

A significant block type by line-length interaction found for the line-length discrimination group, $F(3, 54)=3.97$, though partial η^2 only reached .019, where no similar interaction was found for the signal detection group, partially explains this paradox (Figures 15 and 16). As can be determined via an examination of Figure 16, while no mixing costs resulted from the length of 'no confidence trial run length', a

definite mixing-benefit was being derived as the length of 'confidence trial run' increased.

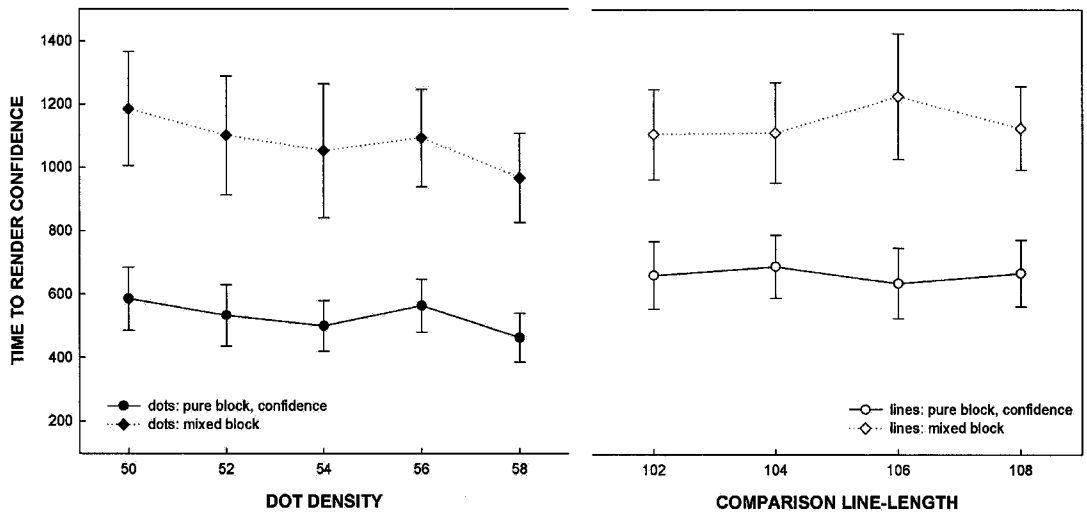

Figure 15. Times to render confidence as a function of decision difficulty for the signal detection group (Panel A) and line-length discrimination group (Panel B).

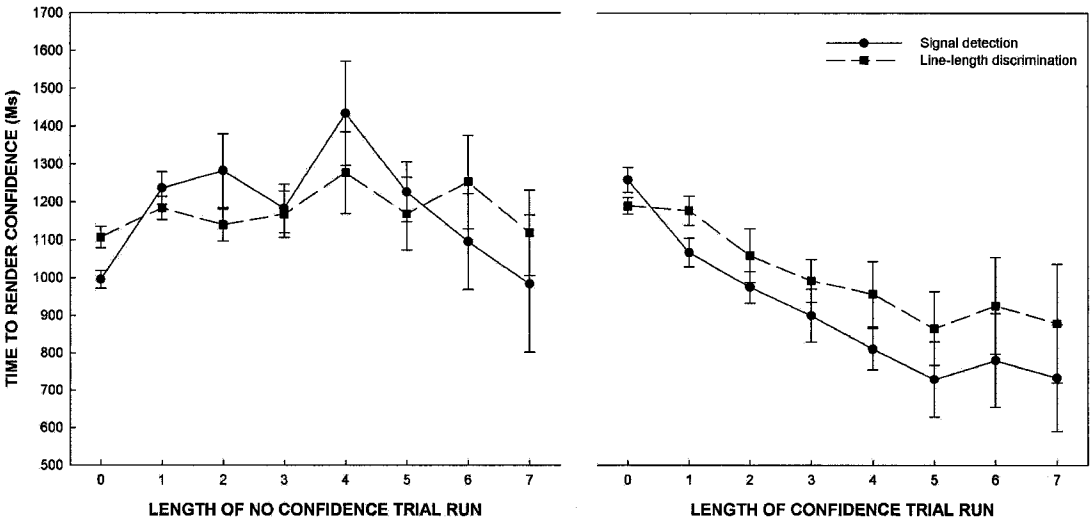

Figure 16. Times to render confidence as a function of no confidence trial run length (Panel A) and confidence trial run length (Panel B).

It would seem that a post-decisional processing of confidence was occurring less often as the length of confidence trial run increased, and that mean times to render

confidence following a run of seven confidence trials began to approach the mean for the pure block of confidence trials. It is in the pure block of confidence trials where, one can assume, very little post-decisional confidence processing was taking place within the line-length discrimination group.

After averaging over dot-density levels and stimulus strength pairs, an ANOVA on time to render confidence with block type (2 levels) and confidence rating (6 levels) as within participant IVs and order (6 levels) as a between participant IV found a significant effect of confidence rating for the signal detection group, $F(5, 40) = 3.42$, partial $\eta^2 = .299$, but not for the line-length discrimination group. As was the case in Experiment 1, participants rendered confidence slightly more quickly when they were more confident, though this effect was somewhat muted in this experiment (Figure 17).

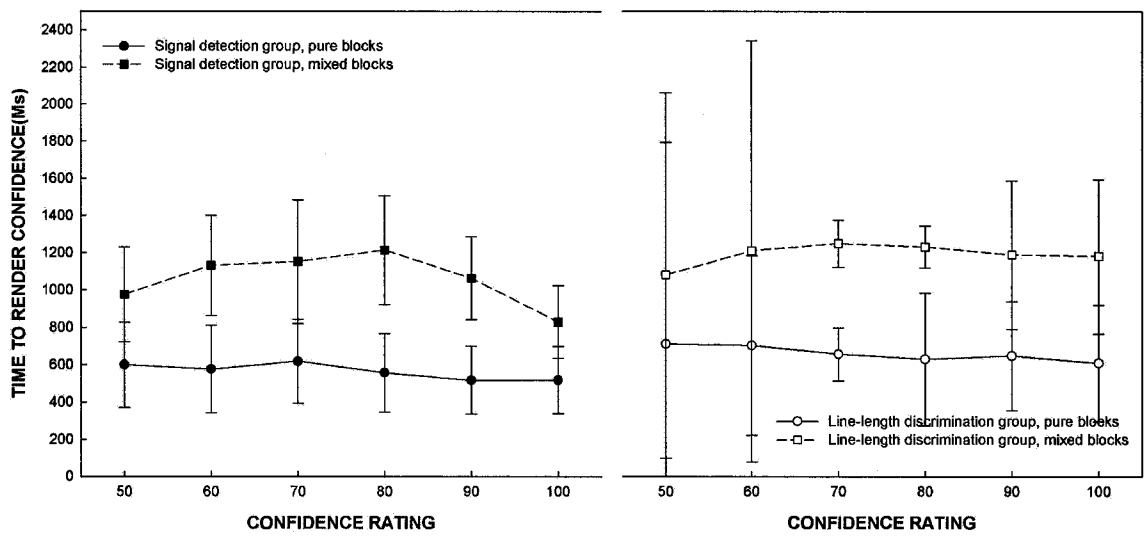

Figure 17. Mean overall time to render confidence as a function of confidence rating for the signal detection group (Panel A) and the line-length discrimination group (Panel B).

Detectability and Discriminative Accuracy Analyses

ANOVAs on p(correct) used decision difficulty (signal detection: 5 levels; line-length discrimination: 4 levels) and block type (3 levels) as within participant IVs, and

block order (6 levels) as a between participant IV. As was the case in experiment 1, p(correct) values obtained from participants were adjusted first using Berkson's correction and then were subjected to an arcsine transformation.

As illustrated in Figure 18, detective sensitivity increased as a function of dot density, $F(4, 72) = 67.96$, partial $\eta^2 = .791$ (Panel A), and discriminative accuracy increased as a function of comparative line-length, $F(3, 54) = 198.54$, partial $\eta^2 = .917$ (Panel B). Importantly, as can also be seen in Figure 18, no significant main effects of block type were found for either decision-making group. Participants in the stop-confidence condition were not sacrificing accuracy in favour of speed and, in all cases, it would seem the only factor affecting accuracy was ease of choice.

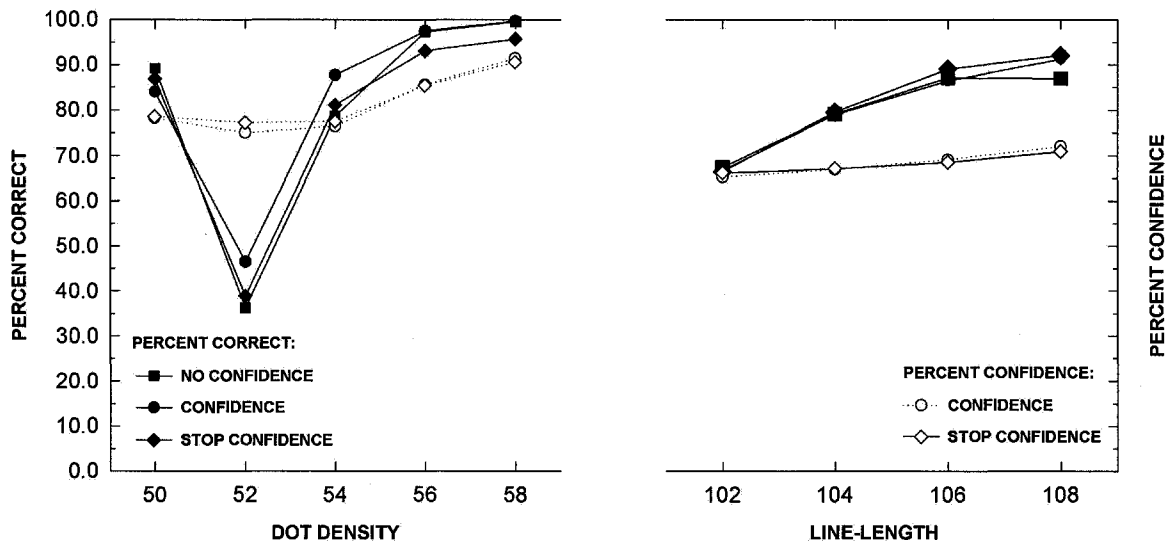

Figure 18. Percent correct and percent confidence as a function of dot density (Panel A) and line-length (Panel B).

Confidence Analyses

Again, as evidenced by Figure 18, a main effect of dot density on mean

confidence rating was found for the signal detection group, $F(4, 72)=36.28$, partial $\eta^2 = .668$ (Panel A), just as a main effect of line length on mean confidence rating was found for the line-length discrimination group, $F(3, 54) = 48.68$, partial $\eta^2 = .730$ (Panel B). All participants expressed greater confidence as the differences between the stimuli they were comparing became more obvious. As no main effects of block type were found, participants seem to have expressed comparable levels of confidence regardless of whether they were rendering confidence in a mixed block or a pure block of trials.

Analyses of Calibration and Resolution

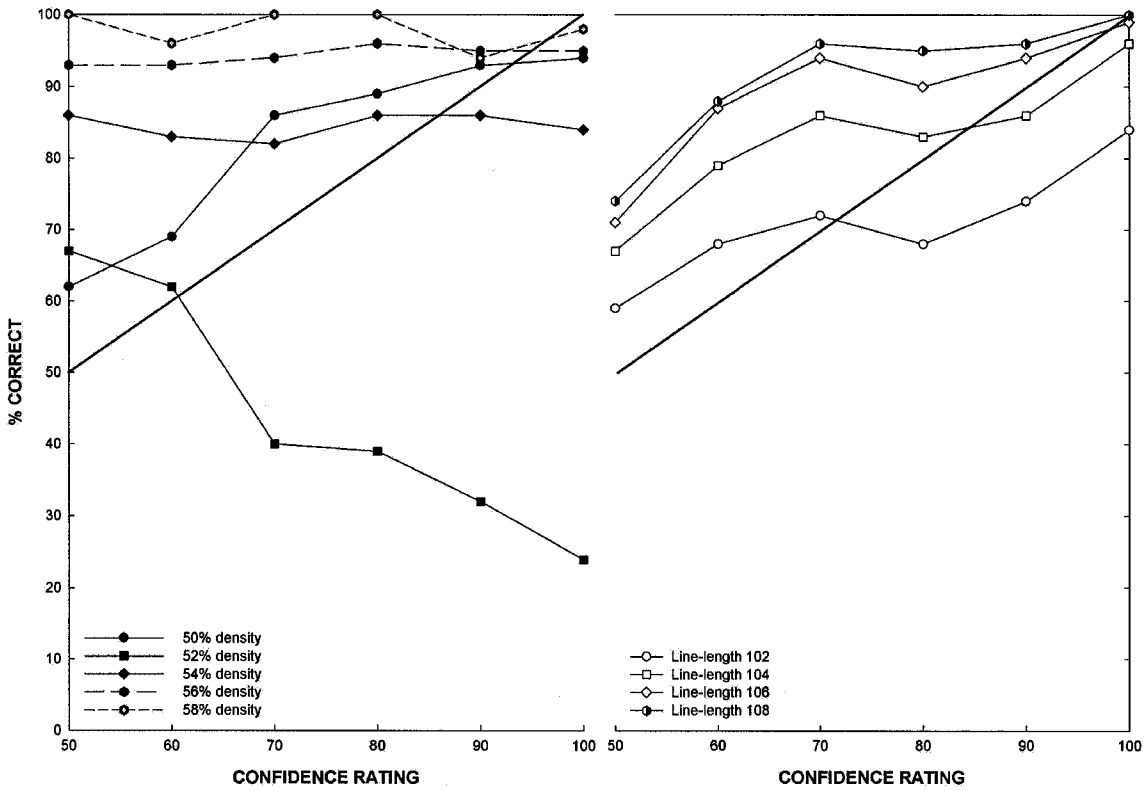

Figure 19. Mean percent correct as a function of confidence category for the signal detection group (Panel A) and the line-length discrimination group (Panel B) for each level of decisional difficulty in Experiment 2.

Calibration analyses for Experiment 2 were very similar to those carried out in Experiment 1, except that decisional difficulty and block type were used as the IVs in the ANOVAs.

Table 3. Significant main effects of decisional difficulty on Calibration, Over/Under-confidence, Resolution, and η^2.

Group	Index	$F_{obtained}$	η^2
Signal detection	Calibration	$F(4, 115) = 24.12$	.456
	Over/Under	$F(4, 115) = 25.51$	.470
	Resolution	$F(4, 115) = 6.53$	.185
	η^2	$F(4, 115) = 3.25$	.116
Line-length discrimination	Calibration	$F(3, 92) = 4.60$	.130
	Over/Under	$F(3, 92) = 14.72$	.325
	η^2	$F(3, 92) = 3.45$	.101

Table 4. Calibration, Over/Under Confidence,Resolution, and η^2 for the different levels of decisional difficulty for both groups in Experiment 2.

Stimulus Type	Level	Calibration	Over/Under-Confidence (expressed as a percent)	Resolution	η^2
dot-density	50% dense	.044	-6.82	.038	.314
	52% dense	.260	32.76	.100	.343
	54% dense	.065	-5.80	.028	.210
	56% dense	.035	-6.54	.012	.138
	58% dense	.023	-3.40	.051	.185
line-lengths	100 vs 102	.075	-1.45	.025	.119
	100 vs 104	.036	-12.15	.028	.217
	100 vs 106	.053	-18.43	.031	.522
	100 vs 108	.054	-19.06	.037	.805

While some of the analyses involving block type approached statistical significance, none of these were held to be practically significant. On the other hand, significant decisional difficulty effects were found for almost all confidence related

indices of performance for both decision-making groups (see Tables 3 and 4). For both groups, calibration improved as the choices became easier to make. As was the case in Experiment 1, under-confidence increased as the decisions became easier, but the most difficult decisions showed over-confidence; i.e., the ubiquitous hard-easy effect (Figure 19). Strangely, within the signal-detection group, resolution decreased with discriminative accuracy while, for the line-length discrimination group, resolution monotonically increased as discriminative accuracy increased (c.f., Baranski & Petrusic, 1994; Petrusic & Baranski, 1997).

Discussion of Experiment 2

Experiment 2 has demonstrated how participants in a stop-confidence, mixed block of trials, who always expected to have to render confidence, began to process confidence as soon as the trial was initiated. Confidence appeared to have then evolved steadily, in linear fashion, until participants were told to stop processing. Importantly, participants in this study also demonstrated the ability to stop processing confidence at will.

As noted in the results section, an interaction between block type and decision difficulty on primary decision response time was found (Figure 11). This interaction makes it difficult to salvage the single process models since it suggests confidence rendering processes can systematically affect primary decision making processes on trials where confidence rendering is not required. Single process models cannot, in theory, explain why a no-confidence trial in a mixed block would differ from an identical no-confidence trial in a pure block.

Conclusion

The goal of the present experiments was to determine the extent to which decision-makers can control whether and when confidence processing is initiated in a sensory-based decision-making task. Experiment 1 suggests that participants can delay a large part of confidence processing until after the primary decision is made. Experiment 2 suggests participants can begin to process confidence immediately upon presentation of the stimuli to be compared and, importantly, has also shown that participants can choose whether and when to stop processing confidence. Considering these findings, decision-makers would seem to have a surprising ability to control confidence: more than an automatically generated by-product of primary decision-making processes, confidence seems to be a much more malleable thing.

The presented findings have implications for both single-process and post-decisional models of decision-making. Signal detection theory, and similar single-process models, cannot describe the totality of the findings of either experiment. Post-decisional models, on the other hand, seem capable of describing a limited range of decision-making behaviours of the type revealed by the mixed block condition of Experiment 1.

Presently absent from the literature is a mathematical model of sensory-based decision-making that is flexible enough to account for the effects of confidence rendering regardless of whether or when such processes are initiated. While the development of such a model is beyond the scope of this paper, it is nevertheless possible to list the mandatory tenets of just such a model given the results heretofore described.

First: the model must allow that, when confidence processing is initiated, the cost

of computing confidence grows at a constant rate. If confidence processing were not based on a constant rate of 'confidence related evidence accrual', one could not have expected to see the strikingly linear plots detailed in Figure 12.

Second: it should be noted that, though confidence processing seems to require a constant rate of evidence accrual, it does not necessarily follow that primary decision-making relies on a similarly linear process of evidence accrual. Yet, since participants in these experiments tended to respond relatively quickly when given an easy choice and more slowly when given a difficult choice, a model of decision-making based entirely on evidence accrual seems reasonable. To wit:

Given a pool of possible evidence events supporting either choice A or B, where the threshold number of A evidence events that must be accrued before A is selected equals the threshold number of B evidence events that must be accrued before B is selected; if p(A evidence event) > p(B evidence event), A is the choice more likely to be made, and the speed with which it is possible to make that decision should be directly proportional to the difference between p(A evidence event) and p(B evidence event). This idea is supported by the present study which, as already noted, found that participants generally responded more quickly and accurately when the two stimuli being compared differed more obviously.

Third: confidence processing must be included in the model, but in a manner that would to allow it to begin at any point prior to or following the primary decision. One way to do this would be to allow that, periodically and consistently throughout the evidence accrual process, from any arbitrarily chosen moment of process initiation, stock is taken of the relative frequency with which evidence accrual events support either

alternative choice. A comparison of these relative frequencies would lead to high confidence in instances where they differ greatly, and low confidence in instances where they barely differ at all.

Fourth: since, from Experiment 1, it would seem that confidence processing can take place largely post-decisionally, a definitive model must allow that confidence processing could continue after the primary decision is made, as suggested by Van Zandt and Maldonado-Molina (2004). It should be noted, however, that comparison stimuli in the presented experiments were always removed from a participant's view once a primary decision was made. Post-decisional confidence processing must have thereafter relied on a memory of the stimuli that had been compared. A truly comprehensive model, therefore, would have to move beyond mere linear concepts of evidence accrual and into the larger realm of memory psychophysics, where confidence processing is based on an active meta-cognitive analysis occurring over the time-course of decision processing as well as a memory-based post-decisional review of the competing evidence accrual totals.

References

Band, G. P. H., van der Molen, M. W., & Logan, G. B. (2003). Horse-Race Model simulations of the stop-confidence procedure. *Acta Psychologica*, 112, 105-142.

Baranski, J. V., & Petrusic, W. M. (2003). Adaptive decision processes in perceptual comparisons: effects of changes in the global difficulty context. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 658-674.

Baranski, J. V., & Petrusic, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, 55(3), 195-206.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics*, 61(7), 1369-1383.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgements: experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 929-945.

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology – Revue canadienne de psychologie experimentale*, 49(3), 397-407.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. *Perception and Psychophysics*, 55(4), 412-428.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay and quantal response, based on the logistic function. *Journal of the American Statistical Association*, 48, 565-599.

Gescheider, G. A. (1997). *Psychophysics: the fundamentals*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons, Inc.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgement: a sampling model of confidence in sensory discrimination. *Psychological Review*, 104(2), 344-366.

Los, S. A. (1999a). Identifying stimuli of different perceptual categories in mixed blocks of trials: evidence for cost in switching between computational processes. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1), 3-23.

Los, S. A. (1999b). Identifying stimuli of different perceptual categories in pure and mixed blocks of trials: evidence for stimulus-driven switch costs. *Acta Psychologica*, 103, 173-205.

Los, S. A. (1996). On the origin of mixing costs: exploring information processing in pure and mixed blocks of trials. *Acta Psychologica*, 94, 145-188.

Muensterberg, H. (1894). Studies from the Harvard psychological laboratory (I). C. A psychometric investigation into psycho-physic law. *Psychological Review*, 1, 45-51.

Petrusic, W. M. (2003). Calibration of response times and confidence in perception and cognition. In B. Berglund & E. Borg. (Eds.). *Fechner Day 2003. Proceedings of the Ninetenth Annual Meeting of the International Society for Psychophysics*. Pp. 235-240. Larnaca Bay, Cyprus: The International Society for Psychophysics.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Human Perception and Performance*, 18(4), 962-986.

Petrusic, W. M., & Baranski, J. V. (2006). Contextual control of accrual thresholds. In B. Berglund & E. Borg. (Eds.). *Fechner Day 2006. Proceedings of the Twenty-Second Annual Meeting of the International Society for Psychophysics*. Larnaca Bay, Cyprus: The International Society for Psychophysics.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgements. *Psychonomic Bulletin & Review*, 10(1), 177-183.

Petrusic, W.M. & Baranski, J.V. (2000). Effects of expressing confidence on decision processing: Implications for theories of RT and confidence. In C. Bonnet (Ed.). *Fechner Day 2000. Proceedings of the Sixteenth Annual Meeting of the International Society for Psychophysics*. (Pp. 103-108). Strasbourg, France: The International Society for Psychophysics.

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

Petrusic, W. M., & Baranski, J. V. (1989). Context, context shifts, and semantic congruity effects in comparative judgements. In D. Vickers and P. L. Smith (Eds.), *Human Information Processing: Measures, Mechanisms, and Models* (pp. 231-246). North-Holland: Elsevier Science Publishers B.V.

Petrusic, W. M., & Jamieson, D. G. (1978). The relation between probability of preferential choice and the time to choose changes with practice. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 471-482.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 68(5), 301-340.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401-409.

Van Zandt, T. (2002). Analysis of response time distributions. In J. Wixted (Vol. Ed) and H. Pashler (Ed.), *Steven's Handbook of Experimental Psychology, 3rd Ed.* New York: John Wiley & Sons.

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

Van Zandt, T. & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6), 1147-1166.

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

Vickers, D. & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50, 179-197.

Vickers, D. & Pietsch, A. (2001). Decision making and memory: a critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. *Psychological Review*, 108(4), 789-804.