# NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was scanned as received.

66

This reproduction is the best copy available.

UMI®

Running head: CONFIDENCE PROCESSING IN COMPARATIVE JUDGEMENTS

# Confidence Processing in Comparative Judgements: Speed Versus Accuracy Stress

Joel Lucas

A thesis submitted to the Faculty of Graduate Studies

and Research in partial fulfilment of the requirements for

the degree of Master of Arts

Department of Psychology

Carleton University

Ottawa, Ontario

September, 2005

*Your file Votre référence*
*ISBN: 0-494-10060-5*
*Our file Notre référence*
*ISBN: 0-494-10060-5*

**Abstract**

In Experiment 1, participants compared the length of pairs of lines, with payoffs emphasizing accuracy at the expense of speed in one condition, and payoffs emphasizing meeting a 500 ms deadline, at the expense of accuracy, in another. In both the accuracy and the speed stress conditions, comparisons were made for an initial block of trials without the expression of confidence, and in subsequent blocks, on each trial, participants indicated how confident they were. In the accuracy but not the speed-stress condition, primary RT significantly increased with the addition of confidence assessments. In the speed-stress condition confidence RT was much longer than the accuracy condition. Our results suggest that under accuracy-stress confidence is processed both decisionally and post-decisionally but entirely post-decisionally under speed-stress.

Experiment 2 attempted to rule out the possibility that increased decision times observed during cognitive tasks accompanied by confidence ratings, are simply due to additional processing caused by anticipation of a making a choice after each discrimination. Participants compared population sizes of Canadian cities in conditions which were followed by confidence ratings, a choice response task, or no additional task. Primary response times were much longer in the confidence group than both the choice response task no-confidence group, indicating that the effects of confidence on primary decision RT is not caused by the simple process of making a choice after the discrimination task.

**Acknowledgements**

Well then...first of all I would like to acknowledge that this research was funded by a Discovery grant administered to W. M. Petrusic from the National Science and Engineering Research Council of Canada. Additionally, NSERC supported me financially with a personal scholarship, and I certainly appreciated it. I would also like to acknowledge P.D. McCormack for his financial support.

A big thanks to my thesis committee: John Zelenski, Craig Leth-Steenson, and Joe Baranski for their words of encouragement, their attention to detail, helpful feedback, and their kindness toward my mistakes. They all happen to be nice people as well. I would like to thank my SUPERvisor, Dr. Petrusic, who is the nicest person I've met since moving to the big city. He dared to take me on as a student despite my reputation as a trouble-maker. "The boy seems to try hard, but he's just out of control!!!" was a familiar lament that could be heard throughout the land.

Most of all I would like to thank my wonderful wife, Jennifer Love-Hewitt, for her undying love and support. I just couldn't have done it without you. Thank-you to my lab mates, David McGill, Steve Carroll, and Shmulik Shaki for helping me with my work, and just for being good guys. I would like to express my appreciation towards the Running Gods for blessing me with swift feet, a clear mind and a noble heart. I would like to give special thanks to Jimmy from "Jimmy-Z" custom guitars. Big thanks to Kate Bush, Dave Matthews Band and In Flames, and may the ghost of Antonio Vivaldi haunt my soul. Joel uses Laney amps exclusively, and is endorsed by Korg, Yamaha, Saucony and Kraft Foods.

## Table of Contents

## List of Figures

Page

**Confidence Processing in Comparative Judgements: Speed Versus Accuracy Stress**

Psychophysicists have long used experiments requiring stimulus discrimination as a means to understand cognitive processes involved in human judgement. Accompanying stimulus comparisons, participants are often asked to give a rating of confidence, in order to give further insight into cognitive processes. Studying subjective confidence judgements in relation to the actual accuracy of participants' performances may reveal several interesting properties. For instance, the question arises: Is our confidence accurate? Can we trust it? Measures of confidence may reveal the conditions and contexts in which we are most and/or least accurate. The discrepancy between confidence and true probability gives a baseline to measure the effects of procedures designed to improve the accuracy of our judgments. The study of confidence has generated various models to explain and predict various aspects of the discrimination process. General findings of confidence, which must be incorporated into such a model are: the fact that confidence is a function of discriminability; confidence varies inversely with response time; and confidence is a function of accuracy. Additionally, conditions emphasizing speed or accuracy and the occurrence of response biases affect these relationships. More recently, studies have begun to determine the locus of confidence occurrence and the basis for the computation. Results from these studies suggest that the additional task of giving confidence ratings after a perceptual discrimination task requires additional processing time. However it is unclear whether or not it is the processing of confidence per se that requires additional time, since it could be that the addition of *any* task would increase RT for the perceptual discrimination. This thesis is an attempt to determine whether or not confidence processing is a separate process, in relation to perceptual discrimination,

which requires additional mental resources and therefore processing time, or instead if it is integral to the perceptual decision and requires no additional processing. Also we investigate if confidence is processed at the same time as the perceptual discrimination, afterwards or partially both.

*Discriminability and Confidence*

Peirce and Jastrow (1884), in an effort to test if perceptual discriminations can be made below the threshold of a just-noticeable-difference (jnd), performed the earliest experiment to examine the dependence of confidence ratings on difficulty of task. They argued that it is possible to make perceptual discriminations above the level of chance, even when participants are not aware of any difference in the magnitude of the compared stimuli. The researchers used themselves as participants in an experiment designed to test the ability to detect changes in tactile pressure. The participant held a finger against a scale as weights were alternated by the experimenter, shifting from a lighter weight to heavier, and then lighter again, or the reverse, starting with a heavier weight. The task was to determine which sequence of weight change had occurred. As well, confidence ratings for their decision were elicited after each trial, based on a four-point scale. Several important findings from this study emerged. First, it was determined that confidence varies as a function of discriminability. Confidence increased with the difference in magnitude of the comparative stimuli (confidence increased as the task became easier and participants were therefore more accurate). Second, participants may underestimate their ability to discriminate under some circumstances, especially when discriminative accuracy is relatively high.

The latter finding revealed itself in participants' tendency to indicate the "absence of any preference for one answer over its opposite..." even though they actually scored well above chance.

*Confidence and the Speed/Accuracy Trade-off*

Replicating and extending the work of Peirce and Jastrow, Henmon (1911) also examined the relationships among discriminative accuracy, confidence and the response time (RT) of the primary decision. Henmon timed participants on a line-length discrimination task, in which the level of difficulty was held constant. Participants were required to give a rating of confidence for their decision, immediately after each trial. The four-point confidence scale was as follows: a –"perfectly confident"; b –"fairly confident"; c - "with little confidence"; and d – "doubtful. An important finding was that, as RT increased, accuracy and confidence decreased. This is a general characteristic of discrimination tasks, which consistently shows an inverse relationship between RT and accuracy (e.g. Baranski and Petrusic, 1998) known as the speed-accuracy trade-off. This is an example of the relationship between speed and accuracy from trial to trial *within* a condition, which more specifically, is known as the *micro* trade-off (Vickers et al., 1985). The *macro* trade-off deals with the speed-accuracy relationship *between* conditions(Vickers et al., 1985), which emphasize either speed or accuracy. The distinction is an important one since the relationship generally reverses from micro (less accurate as RT increases) to macro (more accurate under the condition allowing more RT). This relationship is confusing and has been a hindrance for discrimination models, trying to capture the dynamics between RT and accuracy across conditions, emphasizing speed or accuracy.

Vickers (et al., 1985) has made the argument that a distinction needs to be made, regarding the relationship between RT and accuracy/confidence in experimenter-controlled and subject-controlled conditions. He argues that in participant-controlled experiments (e.g., accuracy stress), participants are able to take as many observations as they desire to meet the decisional criteria. In this particular context, longer RTs are associated with a reduced *quality* of information accumulation due to the difficulty of discrimination or a "noisy" trial (the rate of information accumulation slows down), and results in less confidence and accuracy, compared to responses made in less time. In contrast, during experimenter-controlled conditions, with a limitation on the number of observations a participant may make (e.g., speed-stress), participants are typically unable to meet their decisional criteria, and therefore longer RTs are associated with a greater accumulation of information towards decisional criteria, resulting in more accurate and confident decisions. Relatively difficult discriminations and "noisy" trials will still cause lower confidence and accuracy, however, all other things being equal, as RT increases, accuracy and confidence will increase.

Following on Henmon, Johnson (1939) tested participants on a line-length discrimination task, which included varying levels of discriminative difficulty, rather than just one. A standard line length of 50 mm was used, with varying line-lengths of 40, 42, 44, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 58 and 60 mm in the method of constant stimuli. Additionally, participants were instructed to emphasize either speed or accuracy when making the line-length discrimination. Subsequent to each line length discrimination, participants also gave a confidence rating, from 1 to 100, by marking off a point on a linear scale. This was the first case in which confidence was measured on a perceptual

task with varying levels of difficulty, as well as including explicit instructions to alter the response time of the primary decision. As one might expect, within each condition, the comparisons of stimuli with a smaller difference in size (more difficult to discriminate) elicited longer RTs, lower accuracy and lower confidence than comparisons of stimuli with a larger difference. Surprisingly though, levels of confidence for the accuracy condition were not generally different than the speeded condition, a result consistent with an earlier report from Garrett (1922) and a later report by Festinger (1943), who also failed to find a difference in overall confidence levels between speed and accuracy. This was perplexing since it did not parallel the inverse relationship found between RT and confidence (Henmon, 1911), which would lead one to believe that the decrease in RT in the speed condition would cause a loss in confidence. It suggested though, that participants could willfully shift RT without sacrificing confidence. Johnson's study was important because it established that accuracy was not just the result of stimulus differences, but the context of speed or accuracy stress needed to be incorporated into analysis of the RT/confidence/accuracy relationship as well.

Vickers and Packer (1982) have since introduced an explanation for the seemingly perplexing results produced by Henmon (1911), Johnson (1939) and Festinger (1943). They proposed that there was a fundamental problem using between-subjects or between-sessions designs to compare confidence between speed and accuracy conditions. They hypothesized that within one session, it is possible participants scale confidence so that the easiest discrimination will automatically get a confidence rating of 100 per cent, while the most difficult, 0 per cent, whether it is a speed or accuracy condition. That is, confidence ratings are not given in absolute values but values relative to the difficulty of

the other levels of discrimination difficulty within that particular session. If a session in which participants made comparisons under speed-stress, was followed the next day with comparisons made under accuracy-stress, confidence ratings would not change, even though participants probably were more confident in absolute terms. They tested this hypothesis, by varying speed and accuracy blocks within a single session in an effort to elicit single scaling criteria. Participants compared line lengths in a discrimination task and gave a confidence rating after each trial, as blocks were alternated from speed emphasis to accuracy. The results showed greater confidence for the accuracy condition compared to the speed condition, which contradicts the lack of a difference found in studies from Garrett (1922), Johnson (1939) and Festinger (1943). The key difference was that in past studies, the researchers had varied the conditions from one session to the next, whereas Vickers alternated conditions within a single session.

Baranski and Petrusic (1998) replicated and extended Vickers and Packer's (1982) study. In a discrimination task, participants alternated between blocks, emphasizing either speed or accuracy, within single sessions. The results showed that during the first session, participants displayed more confidence in the accuracy condition than in the speed condition, supporting the results of Vickers and Packer (1982).

To summarize the relationship between speed, accuracy, discriminability and confidence:

1. Conditions in which an explicit instruction is given, to emphasize speed, elicit faster RTs than conditions emphasizing accuracy. Generally though, speed conditions are less accurate than conditions emphasizing accuracy. Earlier research indicates that confidence does not decrease in speeded conditions,

compared to accuracy conditions, although more recent evidence suggests that it does, although this is still equivocal.

2. Generally, *within* an accuracy condition (subject-controlled), as RT increases, accuracy and confidence decrease. Longer RTs are associated with either discriminations which are more difficult or a "noisy" trial.
3. Within conditions which are experimenter controlled (e.g., speed-stress) and internal criteria are not met, accuracy and confidence increase with increased RT.
4. As discriminability increases (becomes easier) accuracy and confidence increase and RT decreases.

*The Locus of Confidence*

Baranski and Petrusic (1998) initiated the investigation of when confidence is determined in relation to the primary discrimination, also known as the "locus of confidence" in a series of three experiments. This has become an area of interest for those trying to understand the basis of confidence since any model developed to account for confidence must incorporate the locus of confidence, and is therefore one more way in which to provide tests of such a model.

Discrimination models developed thus far make explicit or implicit predictions for the locus of confidence and generally fall into one of two categories. Decisional-locus models, including signal-detection theory and probabilistic mental models, assume that confidence processing is an integral part of the primary decision and requires no additional processing either during or after the primary decision. In contrast, post decisional-locus models, including sequential sampling models, assume that determining confidence is a process, distinct from the primary decision, which requires additional

mental processing *after* the primary decision. Therefore, it is not clear whether confidence is processed during or after the primary decision, and also if it is an integral part and incidental result of the primary decision, or a separate process, which requires additional mental capacities. One assumption of studies focussed on confidence processing is that an increase in RT associated with experimental manipulation indicates an increase in mental processing. Therefore, the general approach taken up to this point to determine the locus of confidence, has been to try to observe changes in both decisional and confidence RT, while manipulating experimental parameters.

*Baranski and Petrusic (1998)*

As indicated, Baranski and Petrusic (1998) attempted to uncover the locus of confidence in a series of three experiments. The major focus was on the confidence RT. In the first experiment, they tested participants in experiments utilizing a single stimulus pair, as in Henmon (1911). Participants made line-length discriminations and indicated their confidence rating on each trial immediately after, by pressing one of four response keys corresponding to their subjective level of confidence (guess, low moderate, certain). The results showed that the time to determine confidence was a function of the level of confidence selected. Subjects were slowest to respond in the "low" category and became progressively quicker to respond, as confidence increased. The "guess" category was the slowest for some participants, while for others it was much faster. The fact that confidence RT varies as a function of confidence category, indicates that confidence is at least partially determined after the primary decision. If it had been entirely processed during the primary decision, one would expect confidence RT to be a constant, to account for merely generating a report after the fact.

Baranski and Petrusic's second experiment revisited Johnson's (1939) classic experiment, and extended the study of confidence RTs by introducing additional levels of discrimination difficulty to the design of the experiment. As in Experiment 1, participants were required to respond as accurately and as quickly as possible. Precisely as in Johnson's experiment, primary decisional RTs decreased and confidence increased as the stimulus pairs became easier to discriminate. Finally, also precisely as in Experiment 1, confidence RTs varied systematically with confidence category; again exhibiting the "fast-guessing" property.

In a third experiment Baranski and Petrusic revisited Vickers and Packers' (1982), study in which participants were tested on a line-length discrimination task and rendered confidence judgements after every line-length comparison, under alternating blocks of speed and accuracy stress, within a single session. As mentioned above, Vickers and Packer had found confidence judgements were higher under accuracy stress compared to speed stress, a finding that was inconsistent with, yet more intuitive than past studies (Garrett, 1922; Johnson, 1939; Festinger, 1943). Consistent with the first two experiments, Baranski and Petrusic found that confidence RT was dependent on confidence category, indicating post-decisional processing, but only under speed stress. After sufficient practice under accuracy-stress, confidence RT was constant in all categories, and as expected, the time of the primary decision was longer under accuracy-stress compared to speed-stress. This would suggest that under accuracy stress, confidence was processed while the primary decision was made, supporting a decisional model of confidence, in this circumstance. Importantly though, confidence RT was much longer in the speed condition compared to the accuracy condition. This suggests that

confidence processing may have been processed during the primary decision in the accuracy-stress condition, but was delayed during speed-stress, until after the primary decision had been made. They hypothesized that during sped-stress confidence may in fact be processed entirely post-decisionally. Generally, this study suggests that regardless of its locus, confidence is a distinct process and not simply an incidental result of the processing of the primary discrimination task, given that it seems to have a post-decisional locus under speed stress and both a decisional and a post-decisional locus under accuracy stress.

*Petrusic and Baranski (2000, 2003) and Baranski and Petrusic (2001)*

Further experimentation focused on decisional RT, in an effort to more decisively implicate a decisional locus, since an increase in RT of decisions accompanied by confidence would indicate additional processing during the primary decision. Petrusic and Baranski (2000, 2003) tested participants in a discrimination task involving square-size, in which changes in decisional RT were observed, as some participants gave confidence responses immediately after the primary decision, and others did not. Both a decisional and post-decisional hypothesis would predict that the additional task of making confidence judgements should have no effect on primary RT. Petrusic and Baranski reasoned though, that if the requirement of confidence did increase decisional RT, it would call into question the validity of the present models of confidence, at least as currently formulated. The results showed that the RTs for the primary decision in the confidence condition were much longer compared to the no-confidence group indicating that confidence in this condition was processed during the primary decision.

This result was replicated in a later study, using line-length discrimination in one experiment and a task comparing population sizes of various Canadian cities in a second experiment (Baranski & Petrusic, 2001). A longer decisional RT with the requirement of confidence suggests that confidence processing has at least a partial decisional locus. Current models which support a decisional locus though, have hypothesized that confidence would not involve more processing time, a hypothesis that is not supported from these results.

In a second experiment, using a signal detection task, reported in Petrusic and Baranski (2000), participants were tested to see if altering the number of categories of confidence would affect the primary decision RT. Based on previous evidence, which shows that RT increases as the number of response categories increases (Merkel, 1885), they hypothesized that if confidence has a decisional locus, then increasing the number of confidence categories should increase primary decision RT. Participants were divided into four different groups, in which they either did not give confidence ratings, or they gave a confidence ratings based on a scale of two (certain, not certain), four (guess, low moderate and certain) or six (50, 60, 70, 80, 90, 100) categories. The results show that RT progressively increased as the number of categories increased.

Taken together, the experiments from Baranski and Petrusic (1998, 2001) and Petrusic and Baranski (2000, 2003) show strong evidence that confidence is processed at least partially during the primary decision. Additionally, in each experiment it was shown that the time to determine confidence was a function of the confidence category, indicating post-decisional processing. Collectively, the results also indicate that discrimination and confidence processing are distinct processes, which may occur simultaneously or

sequentially, depending on the demands of the task. It seems as though during accuracy stress, participants may take advantage of having the luxury of initiating confidence processing before explicitly making a decision for the discrimination task. Results from Baranski and Petrusic (1998) suggest that confidence processing may have an entirely post-decisional locus under speed-stress.

*The Basis of Confidence*

Several models have been developed which generally attempt to account for the process of perceptual discrimination between stimuli on some physical dimension. Peirce and Jastrow (1884) offered perhaps the first formula to account for the relationship between discrimination difficulty and confidence: $m = c \log (p / 1- p)$, where $m$ is confidence, p is the probability of the answer being right, and c is a constant. Later, Volkmann offered a formula with the hypothesis that confidence had an inverse relationship with RT: $(t = a / 2v-1) + b$, where v is the level of confidence, t is reaction time, and a and b are constants. Although confidence does indeed have an inverse relationship with RT, RT by itself has been shown not to be the basis for confidence (Petrusic and Baranski, 2005).

Since then, models of confidence have become more elaborate as the need to incorporate macro and micro speed/accuracy trade-offs became realized, as well as bias effects; differences in subject vs. experimenter controlled studies (see Vickers, 1979); and the difference in response times for errors and correct responses under accuracy stress. As described above, models of confidence may be divided into two main categories based on the locus of confidence that they assume: decisional models and post-decisional models.

*Decisional-Locus Models*

*Signal Detection Theory.* A signal detection theory of discrimination states that each physical stimulus evokes a response in the perceptual system or a strength value corresponding to a numeric value, which represents the actual physical dimension of the stimulus, as well as a degree of random error. Based on this value the participant makes discriminations easily when the strength values are different by a large magnitude, and with increasing difficulty as the stimuli become more similar, since the degree of error creates a distribution curve, which may overlap with distribution of other perceptions creating confusion.

When a decision is made in a binary forced-choice task, SDT predicts that perceived stimuli are compared to a set criterion along an axis of evidence, to determine which stimulus exceeds the other, on some dimension of interest (longer, shorter, etc.). Confidence is a direct result of this decision, determined by the distance from the criterion to the chosen stimulus. This distance will fall onto one of several boundaries, determined by the participant, each corresponding to a confidence rating (the greater the distance, the higher the confidence rating). Therefore the confidence rating is in effect, determined as soon as a stimulus is perceived as the correct choice (see also Egan, Schulman, & Greenberg, 1959).

*Probabilistic Mental Models (PMM) Theory.* Gigerenzer, Hoffrage, & Kleinbolting, (1991) propose a second model that assumes confidence is processed at the same time as the primary decision. Although the theory was initially developed to explain confidence associated with general knowledge tasks, it is general enough in principle to be applied to tasks with perceptual stimuli, as well. The model suggests that when a discrimination is

made, a reference class is conjured up which includes both stimuli, for example if you were comparing Thunder Bay and Ottawa to determine which city had a higher population, the reference class could be "cities in Canada". The reference class is important because it defines the set of available cues that may be used to compare the stimuli. To determine which city has the higher population we rely on these cues, which have a particular validity. Possessing an NHL team or not, could be one cue, which may be used to try to determine if one city has a higher population. For instance, we may remember that Ottawa hosts an NHL team, while Thunder Bay does not. In the true state of nature, there is in fact, an actual probability that a city with an NHL team will be more populous, than a city without one. The model assumes participants have some interpretation of this probability and therefore can rank this cue along with all other cues into a hierarchy, based on the highest probability. Participants then use the most valid cue available to try to determine if one stimulus exceeds the other on some property. Confidence therefore is the result of the cue validity or how much we believe one stimulus exceeds another on some property based on what we know about the two stimuli, relative to one another and the reference class.

*Post-decisional Models*

Post-decisional models all share several common characteristics. They assume that the physical difference between stimuli is represented as a normal distribution of probability, with the actual difference represented as the mean of the distribution and the variance of the distribution is attributed to neural noise distorting perception. The variability of this distribution arises from any number of influencing factors such as environmental distractions and slight fluctuations in physiological processes resulting in varying

attention or transduction errors, and fatigue. Some points in the distribution fall above zero supporting one decision, and some fall below zero, supporting the alternative choice. More difficult comparisons will have a more equal division of positive and negative points compared to an easier comparison with a much larger portion of the distribution attributed to one choice. Samples of information regarding the difference are taken in "snapshots", perhaps ten per second (Vickers, 1979). An explicit decision is made following an accumulation of observations which meets a criterion for one choice or the other. The assumption that information accumulates in discrete events nicely accounts for the relationship between RT, and discrimination difficulty. Bias towards one stimulus or another is represented by increasing or decreasing decisional criteria.

The post-decisional models to be described include: the "runs" model, which was the first discrimination model to account for confidence and served as the basis for other models to develop, including the random-walk model, the accumulator model, moving-window model, and the doubt-scaling model.

*The "Runs" Model* was proposed by Audley (1960) in an attempt to account for the inverse relationship of RT and accuracy found in discrimination tasks (Henmon, 1911). The model proposes that participants collect information implicitly, in small units of information, which supports one stimulus or another in a discrimination task on some property (A>B or B>A). In order to arrive at an explicit decision, one must collect an unbroken sequence of information supporting the choice of one stimulus over another. If the two stimuli were vastly different, a sequence of information would occur quickly since there was no contrasting information to interfere with the "chain". For instance, if A represents information in favour or one response and B, the other, a criterion of 4 may

result in a "fast" sequence of AAAA, with no information to support B. In contrast, if two stimuli were very similar, the information collected may sway to support one stimulus and then the other, and therefore the decision takes longer. A criterion of 4 in this instance may take more time since there may be vacillations until a sequence of 4 is obtained: AABAAABAABAAAA.

Audley defined confidence with the equation: C=1/(V+1). C is confidence, the 1's are constants and V represents the vacillations, or the number of times the sequence of information switches to support one stimulus or the other. In practice, the runs model seemed to be flawed as a model of discrimination since it predicts equal RT for both correct responses and errors (Vickers 1979), although it is well established that under accuracy stress errors in fact, take longer. Additionally, since confidence is directly related to RT, the runs model implies that confidence will be equal for correct responses and errors, when in fact confidence for errors has been found to be lower in several experiments (Petrusic, 1992).

*Balance-of-Evidence Hypothesis.* Vickers (1970, 1979) presented a model to account for perceptual discrimination in which magnitudes of information of differences between two stimuli are collected and accumulate in one of two bins, or accumulators (A > B or B > A). Stimuli are perceived with a degree of error, represented as overlapping distribution curves. The difference between the two stimuli can be represented as a distribution curve as well with a certain proportion corresponding to an explicit decision A > B and the other proportion B > A. Vickers hypothesized that the perceived absolute magnitudes of positive and negative responses are collected and summed in individual bins. When one bin reaches a certain criterion, the decision is made that, that stimulus is greater on some

particular property (e.g., line length). In the accumulator model, confidence is a function of the "balance of evidence" d or the magnitude difference of information between the chosen response (ta) and the amount accumulated in the alternate response (tb), at the time the decision is made. Using computer simulations Vickers demonstrated that the accumulator model correctly predicts all the basic properties of binary discrimination and the model provides a series of detailed predictions of the form of the response time distribution (e.g., mean, variance, skew, and kurtosis).

*The Random Walk Model.* Heath (1984) presented the random-walk model originally developed by Link (1975) and Link and Heath (1975) to account for confidence associated with discrimination tasks. Again differences in stimuli are observed sequentially and summed to meet set decisional criteria for each stimulus, but unlike the runs model, observations are not required to occur in succession to reach the decisional criteria. Heath (1984, p.57) describes the random-walk model as follows:

Two response thresholds are set by the participant (along a continuum representing stimulus differences), each corresponding to an explicit decision towards one stimulus. One threshold is set at "A" and the other, "–A", with "0" (no difference between stimuli) in between. The thresholds are influenced by response caution (larger for accuracy stress, compared to speed stress). When the accumulated difference reaches "A", an answer of "V > S" is generated or if the accumulator falls to "–A" the answer "V < S" is the result. If two stimuli are near equal on the underlying dimension, the observation of negative information will be near equal to positive information and the "drift" will be prolonged, accounting for RT.

Confidence can be accounted for by the formula (A-C) θ. The symbol "θ" represents the discriminability of the stimuli and increases proportionately to the difference in line length. "C" represents the starting point for the random walk, and is positioned between "A" and "–A", corresponding to response bias. Confidence varies from trial to trial only if either the boundary, "A" or the starting point of the walk randomly varies from trial to trial. The model has the attractive properties that, as is evident in the data, as discriminability increases, i.e. θ increases, confidence increases. The model also predicts that as the boundary, "A", is set further out, accuracy increases, RT increases, and confidence increases, i.e. under increasing accuracy-stress.

Heath supports his model by applying data from an experiment by Ascher (1974). Vickers and Smith (1985) reject Heath's random-walk model claiming that it is unable to perform in the manner it was originally presented. The confidence model (A-C) θ predicts that when a participant is more biased towards a particular response, the response will be made with less confidence. This prediction is inconsistent with data cited by Vickers (Packer, 1984) in which participants who had indicated greater response bias on trials for a discrimination task had actually expressed greater confidence. As well, as noted, the model predicts higher confidence under accuracy stress than under speed stress and is obtained, but rarely. This model also predicts equal RT for correct decisions and errors. It also incorrectly predicts that confidence levels will remain unchanged across trial with a constant level of discrimination difficulty.

*Sensory Sampling Model.* Juslin and Olsson (1997) have more recently offered a model specifically to account for discrimination processes in the perceptual domain as opposed to the intellectual cognitive domain. The authors argue that a distinction must be

made between Thurstonian (sensory) and Brunswickian (contextual) based error. They believe perceptual discrimination tasks (Thurstonian) are subject to error in judgement arising from variability of internal representations of stimuli, and therefore should be viewed upon differently than intellectual cognitive tasks (Brunswikian), which are subject to error arising from external cues. They believe this distinction is important since they claim that under-confidence, not over-confidence, is obtained in the sensory domain while over-confidence invariably occurs in the intellectual domain. However it should be noted that Baranski and Petrusic (1994, 1995, 1999) and Petrusic and Baranski (1997, 2005) consistently report over-confidence in the perceptual domain.

Based on the sensory/contextual distinction, they propose a model specifically for perceptual tasks. They propose that information regarding the magnitude of differences of stimuli, during discrimination, is collected sequentially over time. The subjective differences between stimuli that are collected vary due to neural noise. Short-term memory acts as a buffer for samples, with a limited capacity that when full, eliminates the "oldest" sample to make room for the "newest" sample. A decision is reached when an upper or lower criterion, each corresponding to one stimulus, is reached by the summation of positive or negative difference means, in favour of one stimuli or the other. Confidence is determined after a decision is made, by the proportion of samples in the buffer, which indicate that the chosen stimuli exceed the other on the task-specific property dimension (e.g., longer line-length). For instance if 6 out of 10 samples indicate $a > b$, then confidence is 60%.

Juslin and Olsson have supported their model by applying it to data from two experiments. Since Juslin and Olssons' publication, Vickers and Pietsch (2001) have

explicitly rejected the sensory sampling model, as a model for sensory discrimination tasks. One criticism was that the model did not adequately account for subjective confidence, since it predicts that confidence decreases as the sampling window increases in size, which is inconsistent with previous research that shows increased confidence with the number of observations made in expanded judgement tasks (Irwin 1956; Vickers et al. 1985). Also their model generally shows that accuracy will decrease as the size of the window increases, which Vickers and Pietsch argue is untrue. Therefore, since the model was created to account for over and under-confidence and the window size is a major determinant of over and under-confidence in their model, Vickers and Pietsch argue that the components which determine over and under-confidence (confidence and accuracy) have a basis which is fundamentally flawed.

*Doubt-Scaling Model.* A more recent approach has been developed by Baranski (1991), based on Slow and Fast Guessing Theory (Petrusic, 1992). Like previous sequential sampling models, information is hypothesized to accumulate in a series of distinct events, each corresponding to a particular explicit decision. One of three decisions is possible: A > B, B > A, or A = B (indifference), determined by the accumulation of events corresponding to each decision, until a preset criterion is reached for one response. In the event that the criterion is met supporting A = B, then the participant is hypothesized to randomly guess between two stimuli.

Three models were derived from Slow and Fast Guessing Theory in an effort to account for measures of confidence during discrimination tasks. In the first model introduced by Baranski, "The Competing Information Scaling" model, confidence is negatively correlated to the amount of competing evidence between alternatives, that is

the total number of discrete units of information accumulated for all three possibilities. A small amount of competing information results in high confidence, while a larger amount results in lower confidence. This is similar to the balance of evidence model proposed by Vickers, except in this model rather than calculating the difference between accumulators, the events in the accumulators are counted to determine competing evidence. In a second model proposed by Baranski, an "RT Scaling Model", confidence is inversely related to RT of the primary decision. For both of these models, Baranski plotted the expected number of evidence accrual events, for both correct responses and errors, in each confidence category. The plots predict the same decisional RT for all levels of difficulty across the different confidence categories, an absence of the difficulty effect. They also predict equal confidence response times for both errors and correct responses. This led Baranski to conclude that both the Competing Information Scaling Model and the RT Scaling Model were not viable models of confidence processing. The third model proposed by Baranski proved to be more fruitful. Like the first model, confidence is scaled from the number of units of competing information, but more influence is given to the category of inconclusive information (A = B). Baranski tested the "doubt-scaling" model and concluded that it reasonably accounted for their data. Baranski attained further evidence for the doubt-scaling model by fitting calibration curves from data provided in Keren (1988), Lichtenstein and Fischhoff (1977). The next section provides an overview of the measures that characterize calibration curves.

*Confidence viewed as subjective probability: Probability assessment indices*

In the context of subjective probability assessment analyses, three indices have been extensively studied (for formal treatments of this topic, see Baranski & Petrusic, 1994;

Murphy, 1973; Yaniv, Yates, & Smith, 1991; Yates, 1990). *Bias*, equivalently the over/under confidence index, provides a global index of the confidence/accuracy relation in terms of a deviation between the proportion confidence and proportion correct, $\bar{p}$- $\bar{e}$, where $\bar{p}$ denotes the mean subjective probability (proportion confidence) and $\bar{e}$ denotes the mean occurrence of the event (proportion correct). Hence, a negative score denotes under confidence and a positive score denotes over confidence. The *calibration* index, $\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{p}_j-\bar{e}_j)^2$, where $\bar{p}_j$ and $\bar{e}_j$ are the mean proportion confidence and mean proportion correct in confidence interval $j$, respectively, $n_j$ the number of observations in confidence interval $j$, and $n$ the total number of observations, indicates how closely confidence ratings, viewed as subjective probabilities match performance accuracy. The calibration score ranges between 0 (optimal score) and 1 (the worst possible score), although calibration is rarely larger than .25. The *resolution* index, $\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_j - \bar{e})^2$, provides an index of how well people use their confidence ratings to differentiate correct from incorrect responses. Typically, raw resolution scores are normalized by the "knowledge index", $\bar{e}(1-\bar{e})$, and the resulting normalized resolution score, known as $\eta^2$, ranges between 0 (no resolution) and 1 (perfect discriminability between correct and incorrect responses).

In the sequel, confidence ratings, obtained in the various conditions of the proposed experiments, will be treated as subjective probabilities and calibration curves will be obtained by plotting the percentage of correct responses associated with each confidence category. Perfect or ideal calibration arises when the data points fall along the main diagonal. Under-confidence is denoted by points above the diagonal and over-confidence

by points below the main diagonal. Given the percentage of correct responses in each confidence category and the relative frequencies with which the various confidence categories were used, the bias, calibration, and resolution indices can be readily calculated.

## Current Thesis

For this thesis, the completed experiments were designed to further investigate the locus of confidence. There are several findings that pertain to our experiments. As previously mentioned, Baranski and Petrusic (1998) found that when participants made perceptual discriminations under a deadline, the time to make subsequent confidence judgements increased substantially, in comparison to confidence judgements made subsequent to discriminations without a deadline, implying a post-decisional locus under speed-stress. However, Petrusic and Baranski (2003) found that when participants made discriminations under conditions stressing accuracy, decisional RTs were substantially longer when followed by confidence judgements, compared to when they were not, suggesting that, at least part of confidence processing occurs during the primary decision, under accuracy-stress. No experiment up to this point has included both confidence and no confidence conditions under both speed and accuracy stress.

## EXPERIMENT 1

Our initial experiment aims to replicate and expand upon Petrusic and Baranski (2003). Participants were tested on a line-length discrimination task, in conditions with and without a decisional deadline and with and without subsequent confidence judgements. In an additional condition, participants were required to make both the primary decision and the confidence judgement under a speeded deadline. If confidence

processing has an entirely post-decisional locus, then decisional RT should remain unaffected by the requirement of confidence in all conditions. This experiment is also exploratory in nature. Typically, confidence judgements are not elicited under a deadline so the effect of this constraint on accuracy and calibration in this condition, has yet to be determined. This experiment also provides a replication of Baranski and Petrusic (1998) and Vickers and Packer (1982), which both showed that confidence decreases under speed-stress in comparison to accuracy-stress, a result inconsistent with earlier research (Garrett, 1922; Johnson, 1939; Festinger, 1943), which showed no change in confidence across conditions.

## Method

*Participants*

Twenty-eight Carleton University undergraduate students participated for one session lasting from 1.5-2 hours, in return for course credit in an introductory psychology class.

*Apparatus*

Participants were seated so that stimuli were presented directly in front of, and at approximately eye level. The only light was provided from a shaded desk lamp (60 watt light bulb) positioned behind and to the left of the monitor and was intended to increase the level of participants' acuity, while still providing enough light to see around the room. Participants faced an 18" computer monitor (Samsung SyncMaster 750s) and gave responses to stimuli generated by an IBM clone (Raven 486/50) PC, running DOS software. Timing was accurate to within ± 1ms.

Responses to stimuli were made on a keypad, positioned to the right of the monitor, with a faceplate (25 X 22 cm), slanted downward, with three rows of push buttons (2 X 2

cm each). The top row of six buttons were spaced horizontally to form a semi-circle and labelled from left to right (50, 60, 70, 80, 90, 100) to represent confidence categories. The middle row consisted of 2 buttons labelled "LEFT" and "RIGHT" which corresponded to left and right stimuli on the monitor which participants used to make discriminations of line-length during trials. The single button of the bottom row, was used to initiate each trial and was also used to indicate an error during elicitation of confidence, and was labelled "START / ERROR".

*Stimuli and Procedure*

Stimuli consisted of pairs of white horizontal lines, presented on a black background. The lines varied in length in terms of pixels. Stimulus pairs, denoted by combinations of pixel length (x, y) that were presented were: (100, 101), (100, 102), (100, 104), (100, 106), (100, 108), and (100, 110). Pairs were presented so that one line was to the left of centre and the other to the right. Each combination of line length was presented randomly an equal number of times and was counterbalanced so that (x, y) was presented half of the time and (y, x) the other half. All participants completed all 5 conditions, each of which consisted of 96 trials (4 blocks of 24 trials) so that all 6 line-length combinations (levels of difficulty) were presented a total of 16 times in each condition (4 times in each block). The purpose of dividing each condition into blocks was to allow participants to occasionally rest before continuing with more trials. A message would appear on the screen after each block, "Rest. Press any key to continue." Participants initiated the next block when they decided to. At the end of each condition a message appeared which said, "This part of the experiment is over." At which point the researcher gave additional instructions pertaining to the next condition.

The conditions were either under speed stress or accuracy stress and either required additional confidence ratings or did not. However, some aspects of the experiment were common to all of the conditions. In each condition participants initiated each trial by pressing "START" on the key-pad, at which time they were presented with either the instruction "Longer" or "Shorter". One second later, a pair of lines was presented on the screen. If the instruction was "Longer", participants were to choose the longer line and if the instruction was "Shorter" participants were to choose the shorter line by pressing either the "LEFT" or "RIGHT" button on the keypad, depending on the position of their chosen line, on the screen. The primary decision time, was measured by the time the lines appeared on the screen until key was pressed.

Participants were told to emphasize speed or accuracy, depending on the condition, but were told that both speed and accuracy are important. Accuracy and speed were measured for all responses, in all conditions. Additionally, participants were rewarded and penalized a monetary value depending on the emphasis of the condition, in order to encourage speed or accuracy responding. A computer program calculated the amount earned by each participant and they were paid accordingly at the end of the session.

In addition to the line discrimination task, some conditions required participants to give a rating of confidence, related to the accuracy of their answers, for each line-length discrimination. Immediately after participants indicated a decision for the perceptual discrimination, the word "CONFIDENCE" appeared on the screen probing participants to give a confidence rating by pressing a key on the keypad which best corresponded with their level of confidence (50-100). Participants were instructed that a rating of "50" indicated a pure guess or 50% chance of being correct, while a rating of "100"

represented absolute certainty or a 100% chance of being correct and ratings from 60-90 represented varying levels of confidence between a pure guess and absolute certainty. Additionally, participants were instructed to press the "ERROR" key instead of a confidence key, when they were sure that they had chosen the incorrect line (accidentally). Confidence response time was measured from the time the word "CONFIDENCE" appeared on the screen until participants gave an answer. Participants received feedback after each trial regarding the accuracy of their response and whether or not they met the deadline when it was required. After each response the message appeared "Primary Response Time Too Slow" or "Primary Response Time OK", and "Response Was Correct" or "Response Was Incorrect." When there was a deadline the confidence response, the feedback was similar to feedback for the primary response.

In the first condition (Acc), participants were instructed to emphasize accuracy during perceptual discriminations. In the second accuracy-emphasized condition (Acc/Conf), participants gave confidence ratings in addition to making perceptual discriminations. In both accuracy-emphasis conditions, participants were rewarded $.02 for each correct answer and penalized $.02 for each incorrect answer. In the third condition (Speed) participants were instructed to try to make line discriminations under a 500 millisecond (ms) deadline. In the fourth condition (Speed/Conf) participants were asked to give confidence ratings (not under a deadline), in addition to making the initial line discrimination under a 500 ms deadline. In the fifth condition (Speed/SpeedConf), again participants were required to indicate a decision under a 500 ms deadline, followed by ratings of confidence. However, participants were also required to give confidence ratings under a deadline (750 ms). In the speed-stress conditions, participants were

rewarded if they made the deadline and penalized if they did not. If participants made their primary decision under the deadline, and they gave a correct answer for the line discrimination task, they received $.02 each time. If they made the deadline but were incorrect, they received only $.01. If they did not meet the deadline but gave a correct answer they were penalized $.01. If they did not meet the deadline and were incorrect they were penalized $.02. In the fifth condition participants were able to make an additional $.01 each time they met the deadline for confidence but were penalized $.01 if they did not.

## Results and Discussion

Data from four participants were eliminated due to corrupted data files, so that data from twenty-four participants were analyzed. The results from Experiment 1 are presented in four main sections: accuracy conditions, speed conditions, confidence judgements, and probability assessment analyses. For the accuracy conditions (NoConf, Conf), the speed conditions (NoConf, Conf, SpeedConf), decisional RT analyses are presented first, analyses of discriminative accuracy next. Analyses of the properties of the confidence judgements are provided following the main analyses of RTs and accuracy in the speeded deadline conditions. Significance levels for analyses of variance (ANOVA) were set at $p < .05$ and were based on the Huyhn-Feldt adjusted degrees of freedom, although the degrees of freedom provided are those specified by the design.

*Accuracy Conditions*

*Decision Time Analyses*

---

Insert Figure 1 about here

---

Data for the analysis of RT in the accuracy conditions plotted in Figure 1 were not censored. In order to examine the effect of rendering confidence on primary decision time, an ANOVA was conducted with mean decision RT as the dependent variable, with the two conditions (no-confidence versus confidence), four blocks, two instructions, and six stimulus pairs as within subjects factors, and the two orders in which the experiment was run, as a between-subjects factor. The ANOVA revealed no main effect of session (generally, the RTs for the accuracy conditions were the same in both orders). As is clearly evident in Figure 1, rendering confidence had the effect of significantly increasing RT, $F(1, 22) = 8.82$, $MSE = 7913999.45$. The main effect of block was marginally reliable $F(3, 66) = 2.83$, $MSE = 2028481.40$, $p = .054$, since participants generally responded more quickly with increasing practice. As expected, the main effect of stimulus pair was highly significant $F(5, 110) = 47.63$, $MSE = 2769747.16$.

The effect of confidence on primary decisional RT did not differ across difficulty levels. The interaction between condition and stimulus pair failed to attain statistical significance $F(5, 110) = 1.81$, $p = .150$, $MSE = 1400804.66$. However, as the plots in Figure 2 show, the increase in decisional RTs is relatively large for the more difficult stimulus pairs and considerably less with the easiest comparisons.

Insert Figure 2 about here

It has long been known that when the instructions to emphasize accuracy are effectively implemented, the RT for errors will typically be noticeably longer than for correct responses. Importantly, as is evident in the top panels of Figure 3, error RTs are uniformly longer in each block and each condition, precisely as reported previously. On the other hand, as the bottom panels of Figure 3 show, for the speed conditions, both when confidence is required and when it is not, RTs associated with errors are faster than those associated with correct responses, as is typically found. Interestingly, when confidence is speeded and primary decisions slow, RTs for correct and error responses are the same. An ANOVA was not conducted to provide statistical support for the overall configuration of findings evident in Figure 3, since not all participants had data in each cell.

*Discriminative Accuracy*

Insert Figure 4 about here

To examine the effect of rendering confidence on discriminative accuracy, an ANOVA was conducted, with proportion of correct responses as the dependent variable; within-subjects factors were the two conditions, four blocks, two instructions, and six stimulus pairs and the two orders in which the experiment was run was the between subjects factor. As in the RT analysis, the main effect of order was not significant, $F < 1$.

Importantly, discriminative accuracy significantly increased when confidence was required, as is evident in Figure 4, $F(1, 22) = 16.56$, $MSE = 0.02$. Also, as is evident in Figure 4, discriminative accuracy increased as the pairs became, a priori, easier to discriminate $F(5, 18) = 162.32$, $MSE$=0.04. Generally, discriminative accuracy varied idiosyncratically, but significantly, across blocks $F(3, 66) = 3.53$, $MSE = 0.04$. The 4-way interaction of group x block x instruction x session was reliable, but seemed uninteresting for the purposes of our investigation and uninterpretable as well.

Although the condition by block interaction failed to attain statistical significance $F(3, 66) = 1.63$, $MSE = 0.04$, examination of the changes in discriminative accuracy over blocks when confidence was required compared to when it was not, proved informative. As is evident in the plots provided in Figure 5, generally performance deteriorates and RT decreases over blocks when confidence is not required, perhaps due to increasing boredom. However, the requirement to render confidence uniformly increases discriminative accuracy over each block, perhaps due to enhancing motivation to perform the task well.

---

Insert Figure 5 about here

---

*Speed Conditions*

*Decision Time Analyses*

---

Insert Figure 6 about here

---

For all three speed conditions (NoConf, Conf, SpeedConf) the overall mean RT was 445.14 ms, s = 272.03. Because the presence of outliers dramatically altered results, the data were censored. RTs longer than 3 standard deviations above the mean, were eliminated resulting in a cut-off of 1261.23 ms. Censoring resulted in a loss of 0.99% of the observations. To determine the effect of rendering confidence on RT under speed stress, an ANOVA was conducted with primary decision time as the dependant variable, the within-subjects factors were 3 conditions, four blocks, two instructions, and the order in which participants were run was the between-subjects variable. The main effect of condition was highly significant, $F(2, 44) = 13.49$, $MSE = 1197.83$, as is evident in Figure 6. The respective means were 408.83, 416.45, and 462.43 for NoConf, Conf, and SpeedConf conditions.

A priori orthogonal comparisons, which partitioned sums of squares for conditions, showed that the NoConf and the Conf conditions did not differ reliably ($F < 1$); the requirement of confidence increased mean RTs by merely 7.62 ms. However, the comparison of average of the NoConf and Conf conditions compared to the SpeedConf condition was highly significant $F(1, 44) = 15.80$ indicating that speeding the confidence judgement substantially slowed primary decision times. Participants may well have used primary decision time to compensate for the decrease in available time otherwise provided during the confidence decision, in an effort to process confidence. Moreover, there was no cost for this increase in primary decision time since participants were still within the deadline.

Insert Figure 7 about here

The plots in Figure 7 in panels A and B show primary decision RTs for the three conditions required confidence ratings. An ANOVA was conducted with mean decisional RT as the dependent variable, with the three conditions, four blocks, six stimulus pairs as within subjects factors and order in which participants were run as a between-subjects factor. Order was not significant, $F < 1$. The main effect of condition was highly significant, $F(2, 21) = 83.06$, $MSE = 3.02$, reflecting the fact that the accuracy condition was much longer than either of the two speed conditions. Mean response times for the accuracy condition, speed condition without a deadline on confidence, and the speeded deadline confidence conditions were: 2076.32, 417.05, and 462.48 ms, respectively. The main effect of stimulus pair was also significant, $F(5, 18) = 9.87$, $MSE = 444882.82$, indicating that participants took longer to respond as stimulus pairs became more difficult to discriminate. The condition by pair interaction was also significant $F(10, 13) = 5.83$, $MSE = 446704.74$, reflecting the fact that the speed conditions were affected by a much lesser degree by pair discrimination difficulty, compared to the accuracy condition, as is also evident in the plots in Figure 7.

*Discriminative Accuracy*

Discriminative accuracy for each of the three speed conditions (Speed, Speed/Conf, Speed/SpeedConf) with each stimulus pair is also plotted in Figure 4. As is especially clear, and evidence of the effectiveness of the speed versus accuracy stress manipulations, discriminative accuracy is uniformly poorer under speed stress than under

accuracy stress at each stimulus pair. Overall, under accuracy stress, discriminative accuracy was 86.28% but only 68.39 % under speed stress.

In order to determine the effect of rendering confidence on discriminative accuracy in the three speed conditions, an ANOVA was conducted, with the three speed conditions, the six stimulus pairs, and the two instructions as within-subjects factors and order of the speed versus accuracy conditions as a between-subjects factor. There was no main effect for the order in which participants completed the experiment, $F < 1$ and the three conditions did not differ in discriminative accuracy, $F < 1$; discriminative accuracy was 68.24%, 67.76%, and 69.20% for No-Conf, Conf, and Speed-Conf conditions, respectively. The main effect of stimulus pair was highly reliable $F(5, 110) = 67.67$, $MSE = 0.03$. As well, the main effect of instruction attained significance, $F(1, 22) = 4.94$, $MSE = 0.41$. Discriminative accuracy was 69.5% with the instruction "Longer" and 67.3 % with the instruction "Shorter". As well the 4-way interaction of condition, instruction, stimulus pair and order was reliable, but not readily interpretable.

**Confidence Judgements**

---

Insert Figure 8 about here

---

The plots in Figure 8 provide mean confidence ratings for the accuracy and the two speed conditions as a function of stimulus pair. An ANOVA was conducted with the three confidence conditions, two instructions, and six stimulus pairs as within-subjects factors and order as a between-subjects factor. Trials in which participants indicated that they made an error were removed from the data. The effect of order in which participants

were run in the experiment was not significant, $F(1, 22) = 1.16$, $MSE = 1870.76$. The main effect of condition was significant, $F(2, 44) = 5.79$, $MSE = 244.53$. The overall confidence levels were: 85.62%, 83.64%, and 88.07% for accuracy, speed stress without a deadline on confidence, and with a deadline of confidence conditions, respectively. A Newman-Keuls test showed that only the difference between the two speed conditions was reliable $q(44) = 4.79$, $p < .05$.

The main effect of stimulus pair was highly reliable, $F(5, 110) = 42.08$, $MSE = 70.69$, reflecting the fact that confidence monotonically increased as stimulus pair became easier to discriminate, as is also evident in Figure 8. The interaction of condition and stimulus pair was reliable $F(5, 110) = 11.40$, $MSE = 42.61$ and arises as a consequence of two effects: First, as is evident in Figure 8, confidence remains relatively insensitive to difficulty in the Speed/SpeedConf condition, while in the Speed/Conf and Acc/Conf conditions, confidence increases as stimulus pairs become easier to discriminate. Second, upon examining the speed deadline with no deadline on confidence and the accuracy conditions it is clear that for stimulus pairs that are difficult [eg. (100, 101), (100, 102), and (100, 104)] confidence levels are very close to one another, but as stimulus pair becomes easier to discriminate, confidence is higher in the accuracy condition than in speed.

An additional ANOVA was conducted with just two conditions (Speed/Conf and Acc/Conf), two instructions, six stimulus pairs as within-subjects factors and order as the between-subjects factor. The difference between the two conditions failed to reach conventional significance, $F(1, 22) = 3.18$, $MSE = 564.57$ $p \geq 0.088$ (however one-tailed $p = 0.044$). Not surprisingly, the main effect of stimulus pair was highly reliable, $F(5,$

110) = 50.86, *MSE* = 68.39. Importantly, the interaction between condition and stimulus pair was highly significant, $F(5, 110) = 8.42$, *MSE* = 39.24.

*Post-Decisional Error Reports*

On trials where confidence was rendered, participants indicated they had made on error on but 12 of the 2304 trials (0.52 %) under accuracy stress. Moreover, as the plots in Figure 9 show, these participants had little or no knowledge of when they had made an error. On the other hand, participants working under speed stress had excellent knowledge of when they had in fact made an error, as is evident from the plots in Figure 9. For example, for when participants indicated confidence without a deadline on the rendering of confidence, they used the error key on 203 of the 2304 trials (8.81 %) and they in fact made errors on 61.7, 79.0, and 91.4 % of the trials with the hard, medium, and easy comparisons, respectively. When confidence was rendered under a speeded deadline, the error key was used on 132 of the 2304 trials (5.73 %) and errors occurred more frequently than not at each difficulty level as can also be seen in Figure 9 although in this condition, overall detection of error was considerably poorer than when confidence was not under speeded deadline stress. In sum, these findings nicely replicate and extend the calibration and resolution of errors first reported in Baranski and Petrusic (1994) and the fact that under accuracy stress resolution is poor but under speed it is excellent, as found in Baranski and Petrusic (1998).

*Confidence RT*

The Panels C and D in Figure 7 provide confidence times for the three conditions requiring confidence ratings. These plots show that confidence RT increased substantially in the Speed/Conf condition compared to the Acc/Conf condition, suggesting that,

indeed, confidence processing was occurring during the confidence RT window, which may have otherwise been processed during the primary decision, had there not been a deadline. In the Speed/SpeedConf condition, confidence RTs were extremely fast and remained constant over levels of difficulty. This result suggests that no confidence processing occurred in this condition, during the confidence RT window, but rather was fully processed during the primary decision.

An ANOVA was conducted with mean confidence RT as the dependent variable, with the three conditions, four blocks, and six stimulus pairs as within-subjects factors, and order in which participants were run as a between-subjects factor. Order was not significant, $F(1, 22) = 2.31$, $MSE = 2.43$, $p \geq 0.14$. The main effect of condition was highly significant, $F(2, 21) = 64.74$, $MSE = 950304.24$. Mean confidence RTs were 690.13, 1010.99, and 327.58 ms for the Acc/Conf, Speed/Conf, Speed/SpeedConf, respectively. The main effect of block was significant, $F(3, 20) = 9.79$, $MSE = 72462.93$, reflecting the fact that participants progressively and monotonically provided ratings of confidence more quickly with practice. The main effect of stimulus pair also was significant $F(5, 18) = 4.89$, $MSE = 70000.63$. The plots in Figure 7 show that generally confidence RT decreased as pairs became easier to discriminate, however the reliable condition x pair interaction $F(10, 13) = 5.77$, $MSE = 59085.00$, reflects the fact that the Speed/SpeedConf condition remained at a very fast response time across all levels of difficulty, and did not show the gradual decrease in RT as discrimination increased, displayed by the conditions Speed/Conf and Acc/Conf.

## Probability Assessment Analyses

*Group Calibration Curves.*

Calibration curves were obtained by plotting the percentage correct at each level of confidence for the accuracy, speed, and speeded confidence conditions, separately at each level of difficulty. These curves are provided in Panel A of Figure 10 and the plots in Panel B provide the relative frequencies with which each confidence category was used.

Examination of these calibration curves reveals, not surprisingly, a hard easy-effect. Indeed, in each condition, the calibrations curves are below the main diagonal for the difficult comparisons and, generally, above the main diagonal for the relatively easy comparisons, revealing over-confidence for the difficult comparisons and under-confidence for the relatively easy comparisons.

Under speed-stress, when the expression of confidence is not under a deadline, appropriately low confidence accompanies choices that are at or near chance, while with high confidence, equally appropriately, choices are made with high accuracy. More importantly, in this condition, the confidence categories are used in a maximally informative way as is evident in the bottom panels of Figure 10. For example, the use of the certain category is maximal with the easy comparisons and minimal with the hard comparisons and the opposite pattern is observed with the guess category. Accordingly, this overall configuration with the calibration curves, and the relative frequencies, manifests itself as enhanced resolution, in striking replication of the findings reported in Baranski and Petrusic (1994, Experiment 2).

On the other hand in the other speed stress condition, when the expression of confidence is under speed stress, the calibration curves are largely in disarray, although still reflecting the ubiquitous hard-easy effect. Most notably, though, the relative frequencies of confidence category usage fail to reflect the difficulty of the comparison. As a consequence the enhanced resolution obtained under speed stress is no longer evident.

Finally, it is important to note that under speed-stress, when the expression of confidence is not speeded, discriminative accuracy is below chance in the guess category, and the probability of an error increases as the comparison becomes easier. In contrast, under accuracy stress, as is evident in Figure 10, discriminative accuracy does not fall below chance in any of the confidence categories. Precisely the same configuration of findings with speed versus accuracy stress on the form of the calibration curves was reported in Baranski and Petrusic (1998, Experiment 3, see Figure 7).

---

Insert Figure 10 about here

---

*Individual Participant Analyses.*

For each participant, calibration, bias, and normalized resolution ($\eta^2$) indices were computed at each difficulty level in each of the three conditions in order to obtain quantitative characterization of the level of probability assessment in each condition with each index. Figure 11 provides plots of the mean values of each of these indices as a function of the difficulty of the stimulus pair. The six discriminative pairs were combined into three groups for analysis and presentation, for reliability of results and ease of

comprehension. Pairs (100, 101) and (100, 102) became "Hard"; pairs (100, 104) and (100, 106) became "Medium"; and pairs (100, 108) and (100, 110) became "Easy". ANOVAs were then conducted with the three conditions and three difficulty levels as repeated measures with each of the calibration, bias and normalized resolution indices as dependent variables.

---

Insert Figure 11 about here

---

*Calibration.* The three conditions differed reliably $F(2, 54) = 20.37$, $MSE = 0.0016$ in calibration. As is evident in the top panel of Figure 11, calibration was best in the accuracy condition and worst when confidence was speeded. As is also evident, calibration depended on the difficulty of the comparison; calibration monotonically improves (the index becomes smaller) as the comparison becomes easier $F(2, 54) = 81.22$, $MSE = 0.00159$. Moreover, the interaction between condition and difficulty level was reliable $F(4, 108) = 7.07$, $MSE = 0.0012$; the differences in calibration among the three conditions are most evident with the most difficult comparisons and least evident with the easiest.

*Over/Under Confidence: Bias.* As is evident in the middle panels of Figure 11, mean over/under-confidence (bias) also vary systematically with condition, with the accuracy condition showing the least amount of overall over-confidence and the speed stress with a deadline on confidence the most over confidence $F(2, 54) = 64.21$, $MSE = 102.76$. As well, and entirely as expected, overconfidence was maximal with the most difficult comparisons and the least with the easiest comparisons $F(2, 54) = 165.22$, $MSE = 51.13$.

Also, precisely as with the calibration index, the interaction between condition and difficulty level was significant $F$ (4, 108) = 5.22, $MSE$ = 64.61; the differences in over-confidence are most evident with the most difficult comparisons. Finally, it is worth noting that in the accuracy-stress condition, a classic hard-easy effect is obtained; over-confidence occurs with the difficult comparisons but there is under-confidence for the other difficulty levels. In contrast, both speed-stress conditions are uniformly and massively over-confident.

*Normalized Resolution ($\eta^2$).* The pattern of findings with the normalized resolution index, ($\eta^2$), is strikingly different from that of the other two probability assessment indices. Notably, for the speed-stress condition, when confidence is not rendered under speed stress, normalized resolution improves dramatically as the comparisons become easier. On the other hand, normalized resolution remains approximately constant over difficulty levels for the other two conditions, as is evident in the bottom panel of Figure 11. As a consequence, the interaction between difficulty level and condition is statistically reliable $F$ (4, 108) = 12.98, $MSE$ = 0.02. Moreover, normalized resolution is substantially, and reliably better under speed stress when there is no deadline on the expression of confidence compared to when it is rendered under a speed deadline and under accuracy stress $F$ (2, 54) = 41.18, MSE = 0.03. Finally, the main effect of difficulty is also reliable, largely due to the monotonic, and large increases in normalized resolution as the comparisons become easier $F$ (2, 54) = 5.42, $MSE$ = 0.04.

## EXPERIMENT 2

In our second experiment we address a criticism aimed toward results from "locus of confidence" studies, which have shown an increase in RT for the primary perceptual decision. The finding that RT for perceptual judgements increases when they are followed by confidence judgements, compared to conditions with no confidence judgements, is somewhat difficult to interpret. The possibility exists that it is not the addition of confidence judgements per se, which increases perceptual decision RT, but in fact, *any* additional task following the perceptual judgement may increase RT. In this experiment we compared decisional RTs in three conditions, in which participants discriminated the size of populations of the top fifty largest Canadian cities. One condition simply consisted of city comparisons, without any additional task following each comparison. The second condition consisted of city comparisons, followed by a rating of confidence in which participants selected one of six categories of confidence by pressing one of six keys. The third condition consisted of city comparisons followed by the presentation of six squares on the screen, corresponding to the six confidence keys. One of the six squares was darkened, changing from trial to trial, and participants were simply required to select the key, which corresponded to the darkened square. If both the choice reaction time condition and the confidence condition increase decisional RT, it would suggest that it is not solely confidence processing which is responsible for the increase in RT, but rather a much more general effect.

### Method

*Participants.* Forty-nine Carleton University undergraduate students participated for one session lasting from 1-1.5 hours, in return for course credit in an introductory psychology

class. The data from four participants were eliminated from the study. Three completed the task at an approximate 50% level of accuracy, and one was generally too slow. Therefore there were a total of 45 participants' data included in the final data. The number of participants in each group were: NoConf = 17, Conf = 16 and CRT =15 conditions defined by the second session treatment.

*Apparatus.* Stimuli were generated by an IBM clone (Raven) PC and presented on an 18" computer monitor (Sony). The experiment ran on SuperLab software. All instructions and stimuli were black, presented on a white background. Participants were seated so that stimuli were presented directly in front of, and at approximately eye level. Room lights were left on. Participants responded to stimuli on a modified computer keyboard. A cardboard "face plate" was placed over the keyboard allowing only the response keys to show through, while all other keys were physically removed. The keys used to make LEFT/RIGHT decisions were the "C" key and the "M" key, and the six confidence keys were "S", "D", "F", "J", "K" and "L".

*Stimuli and Procedure.* Stimuli consisted of pairs of the 50 most populated census metropolitan areas (CMA) in Canada from Statistics Canada (2001). Cities that were eliminated were: Toronto, Montreal, Vancouver and Ottawa because of their known status as the largest cities in Canada, and the next highest 50 were used. Each city was paired with every other city to produce 1225 possible combinations. Three pairs were randomly selected from each difficulty category to make a total of twelve practice trials. Difficulty of discrimination was determined by the difference in magnitude of population between cities, by dividing the larger population of each pair by the smaller (e.g. the pair Kitchener 387,319/ Regina 178,000 = a difficulty ratio of 2.17). The cities were then rank

ordered according to their difficulty ratio and divided into four groups, so that an approximately equal number of cities appeared in each group. The four categories contained the following difficulty ratios, from most difficult to easiest: (1.01-1.35), (1.36-2.00), (2.01-3.85), (3.87-16.52). Thirty pairs were then randomly selected without replacement, from each category to make a set of 120 pairs. A second set was also created using the same manner, but none of the pairs from Set 1 appeared in Set 2. All participants completed both sets of 120 trials, half completed Set 1 first, while half began with Set 2. The pairs were constructed so that half of the time the larger city appeared on the left and half on the right, and half of the time the correct answer appeared on the left and half on the right. The instructions SMALLER and LARGER appeared an equal number of times for each level of difficulty in each set. At the end of the experiment all participants had completed 4 levels of difficulty X 30 pairs X 2 conditions for a total of 240 trials.

There were three different conditions in this experiment. In Condition 1, participants only made comparisons of CMA populations. In Condition 2, participants made population discriminations followed by confidence judgements about the accuracy of their judgements. In Condition 3 participants made population discriminations followed by a motor-reaction-time task. All participants completed Condition 1 first and then either Condition 1, 2 or 3 for the second session. In Experiment 2, like Experiment 1, the participant initiated each trial by pressing a key and the presentation of the instruction "LARGER" or "SMALLER" appeared for a duration of 1 second. After a one-second inter-stimulus interval (ISI), a pair of cities appeared. The pairs were presented so that one city was to the left of centre and the other to the right. Participants selected the city

on the right or left by pressing the corresponding left or right key on the keyboard. The primary decision time, was measured by the time the cities appeared on the screen until a button indicating "LEFT" or "RIGHT" was pressed. Participants were then prompted with the instruction, "START" to initiate the next trial when they were ready, and an ISI of 500 ms occurred before the presentation of the next set of cities. There were 120 trials and after every 30 trials a prompt appeared on the screen allowing participants to rest before continuing, similar to Experiment 1.

For Group 1 (NoConf/NoConf) the second set of stimulus pairs was completed in exactly the same way as the first. For Group 2 (NoConf/Conf), the second set again involved city comparisons, although after each city discrimination, participants were prompted with the question, "How confident are you?" after a 1000 ms ISI. Participants were required to select one of six keys in a horizontal row indicating their level of confidence for the accuracy of their answer. Participants were instructed that the keys indicated a progression in the level of confidence from the key furthest to the left, labelled "Guess" to the key furthest to the right labelled "Certain". After the selection of confidence, the trial began again.

The third group (NoConf/RT) had the same requirements as the NoConf/Conf group, except that instead of the confidence probe, a picture of 6 black-outlined squares appeared on the screen, in a horizontal row. One of the squares was darkened, while the other 5 remained white. Each square corresponded to one of the 6 "confidence" keys. The participants were instructed to merely press the key that corresponded to the darkened square. Each darkened square appeared on an equal number of trials, in a randomized order.

In all of the conditions, participants were told that both speed and accuracy for their decisions was important. Accuracy though was emphasized since participants were rewarded 5 cents for every correct decision they made. A computer program calculated the amount earned by each participant and they were paid accordingly at the end of the session.

## Results and Discussion

The data from three participants were discarded because they performed at, or near, chance levels in at least one of the sessions (average percent correct was 51.54). As well, data from one other participant were excluded because the RTs were excessively long (over half of the RTs were longer than 10 s). RTs were censored to remove outliers. The overall mean RT was 2877.9 ms and the standard deviation was 2008.4 ms and RTs more than three standard deviations above the mean were removed. As well, RTs less than 200 ms were censored. Overall, 2.07 % of the 11520 observations were removed. As in Experiment 1, significance levels for ANOVAs were set at $p < .05$ and were based on the Huyhn-Feldt adjusted degrees of freedom, although the degrees of freedom provided are those specified by the design.

*Response Time Analyses*

---

Insert Figure 12 about here

---

Figure 12 provides plots of mean overall RTs for each group for each block in each session. The plots in Figure 12 are clear. Although the three groups differ at the outset,

RTs systematically decline for each group, reflecting the effects of practice. However, by the fourth block the mean RTs of the three groups are nearly the same. In sum, RTs for the control condition and the choice reaction time (CRT) conditions continue to show the orderly and continuing effects of practice over blocks within each session. However, in contrast, the requirement of rendering confidence results in a substantial and sustained increase in decision time.

An ANOVA was conducted with the two sessions, four blocks, four levels of difficulty and two instructions as the within-subjects factors and the three groups as the between-subjects factor. Overall, RTs were faster in the second session than the first, reflecting the effects of continued practice $F(1, 45) = 18.35$, $MSE = 3991305.9$. As is evident from the plots in Figure 12, the groups differ from the outset, simply as a result of random assignment, and they continue to differ subsequently. Indeed, the main effect of group is highly reliable $F(2, 45) = 4.28$, $MSE = 8276051.10$. As is also clear, RTs declined systematically over blocks in both sessions, and the main effect of blocks is significant $F(3, 135) = 10.48$, $MSE = 1027640.20$. The comparisons also differed reliably in difficulty, as is evident in the plots in the left panel of Figure 12, $F(3, 135 = 51.79$, $MSE = 622987.20$.

---

Insert Figure 13 about here

---

The session by block interaction was significant $F(3, 135) = 5.67$, $MSE = 773866.44$. In the first session, RTs become progressively faster while in the second session, RT generally levelled out after the second block. There was a significant instruction by

difficulty interaction, $F$ (3, 135) = 2.72, $MSE$ = 414793.25. Generally, RT increased for both instructions as pair became more difficult; however the most difficult pair was slightly faster than the second most difficult pair with the instruction, "Larger". There was a significant interaction of block by difficulty, $F$ (9, 405), $MSE$= 521961.62. In blocks 3 and 4 the most difficult pair RT was slightly faster, than the second most difficult pair. The 3-way interaction involving session, difficulty, and group was significant, $F$ (6,135) = 2.17, $MSE$= 428056.21, although it too, proved difficult to interpret.

*Discriminative Accuracy Analyses*

The plots in the right panel of Figure 12 are clear in showing that discriminative accuracy was nearly identical for the three groups, $F < 1$; 68.3%, 68.9%, and 67.9% of the responses were correct for the control, confidence and CRT groups, respectively. The plots in Figure 12 are also clear in showing a very large difficulty effect $F$ (3, 135) = 305.90, $MSE$ = 0.06, for each group. Mean percentage correct was 51.8, 62.5, 71.7, and 87.3 for the four difficulty levels from hardest to easiest, thereby attesting to the effectiveness of the assignment of city pairs to difficulty levels on the basis of the ratios of population size. The main effect of block was also reliable $F$ (3, 135) = 10.80, $MSE$ = 0.04; accuracy was the highest in the second block (71.6%) and approximately constant over the other blocks (67.2%).

A number of two way interactions attained significance (e.g., session by instruction, block by difficulty, block by instruction, difficulty by instruction) as did the three way interaction involving block, difficulty, and instruction but these interactions are of little consequence and do not compromise the main conclusion that the RT findings are not a

consequence of speed accuracy tradeoffs, more oft than not reflecting idiosyncratic effects that are not easily if at all interpretable.

*Adjustment of Initial Between Group Differences: Analysis of Covariance*

---

Insert Figure 14 about here

---

To determine more precisely, if indeed group differences in RT were statistically different in the second session, when the treatment conditions were administered, an ANCOVA was conducted, using the mean of the entire first session, for each participant as the covariate. In this ANCOVA the within-subjects factors were the four levels of difficulty and the four blocks, with group as the between-subjects factor. Plots of covariance-adjusted means for each group over the four blocks in the second session are provided in Figure 14.

The three groups differed in the first session as attested by a significant effect for the covariate $F(1, 44) = 36.46$, $MSE = 2519041.24$. Moreover, the covariate was effective in adjusting for between group differences in session 1 and in reducing within cell variability, since the covariate was reliably, linearly related to the dependent variable; the estimate of the common slope over the three groups was 0.567, with a standard error of 0.094 and was significantly different from zero $F(91, 44) = 36.48$. Consequently, and importantly, the main effect of group was significant, $F(2, 44) = 6.79$, $MSE = 2519041.24$. A priori, planned comparisons, with Bonferroni adjustment of significance levels were conducted. The increase in decision time due to rendering of confidence,

relative to the control condition proved reliable $F(1, 44) = 7.16$, $p < 0.05$ but the CRT group did not differ from the control group $F(1, 44) = 1.15$.

As is also evident in the plots in Figure 14, RTs continue to decline over blocks and the main effect of block was reliable $F(3, 135) = 6.56$, $MSE = 283343.21$. As well, of course, the main effect of difficulty of the comparison was significant $F(3, 135) = 38.55$, $MSE = 197328.24$.

Therefore, there is strong evidence that indeed, the prolonged primary decision times observed during discrimination tasks, with the inclusion of confidence decisions, compared to without, as Experiment 1 shows, is not simply due to the inclusion of an extra task requiring the selection of a response after the primary task, but rather is reflective of more sophisticated cognitive processing, which would be associated with a task like confidence assessment.

## General Discussion

The goals of these two experiments were the following: to determine the locus of confidence processing during discrimination tasks under both speed and accuracy stress; and to determine if even the addition of a simple choice reaction time task, following a discrimination task could prolong primary decision times in the same manner as confidence decisions.

In Experiment 1, under accuracy-stress, primary decision times increased dramatically with the inclusion of confidence processing, compared to without, indicating that at least some confidence processing occurs during the primary decision. This result replicates findings from Baranski and Petrusic (2001), Petrusic and Baranski (2000, 2003). Evidence for post-decisional processing was also found under accuracy-stress.

Confidence RTs were reflective of discrimination difficulty of pairs, indicating that some level of processing occurred post-decisionally. In fact, *any* difference in confidence RTs would indicate that some processing must occur post-decisionally, otherwise a constant RT, reflective of a choice reaction time task would be expected.

Under speed-stress, the primary decision time was the same with or without the requirement of confidence, suggesting that no additional processing took place during the primary decision. Importantly though, confidence RT was much longer in the speeded condition, compared to the accuracy-stress condition, replicating Baranski and Petrusic (1998). This is strong evidence that indeed, confidence in the speed stress condition was delayed from being processed during the primary decision, and was rather processed entirely post-decisionally. Again, confidence RTs were reflective of discrimination pair difficulty, indicating post-decisional processing of confidence.

The condition with a deadline on both the primary and confidence decision was somewhat exploratory in nature but generated some interesting results. If indeed confidence was processed entirely post-decisionally in the Speed/Conf condition, then a deadline on the confidence RT would be expected to either increase primary decision time, or diminished confidence calibration and resolution. Both of these results occurred. The primary decision time for the Speed/SpeedConf condition was higher than the Speed/Conf condition, suggesting some degree of confidence processing occurred during the primary decision. Additionally calibration and resolution were poor in comparison to the Speed/Conf condition suggesting that the mental processes involved in calculating confidence were compromised. Confidence RT in the Speed/SpeedConf condition was not dependent on discrimination difficulty. Therefore it is our opinion that confidence in

the Speed/SpeedConf condition may have been processed entirely during the primary decision, at a diminished level. It is interesting that participants made their confidence response well under the deadline. Stemming from our view that confidence was processed entirely during the primary decision, in the Speed/SpeedConf condition, it becomes somewhat puzzling as to why participants would choose not to use any time during the confidence response time window to process confidence. A possible explanation is that 750 ms is simply not enough time to calculate confidence and select one of six keys from the keypad. Therefore participants may have "given up" trying to make full confidence calculations. A more simple "high" or "low" confidence rating may have been easily processed during the primary decision, and therefore no additional processing was needed during post-decision. This may also help to explain the lower frequency of confidence responses, which fell in the middle categories (60-80%), observed in the Speed/SpeedConf condition. There is no clear evidence for this view, and is simply a suggestion which remains to be corroborated in future studies.

Confidence levels were higher in the accuracy condition, compared to the speed conditions. The effect however was more pronounced for pairs that were easier to discriminate. The increase in confidence in the accuracy condition compared to speed is an important finding because it is consistent with the results of more recent tests concerning the relationship between confidence and speed/accuracy stress. Baranski and Petrusic (1998) and Vickers and Packer (1982) also found increased confidence during accuracy stress, but earlier studies found no difference between conditions (Garrett, 1922; Johnson, 1939; and Festinger, 1943). Vickers and Packer (1982) had reasoned that a

within-subjects, within-sessions design was necessary, so that confidence for items under speed and accuracy may be scaled relative to one another.

Another interesting finding was the increase in accuracy, which occurred in the accuracy stress condition with the inclusion of confidence, compared to without. This result however was confounded with practice. The confidence condition *always* followed the no confidence condition, and it is quite possible that the increase in accuracy was due to practice and nothing else. Previous results have been mixed. Baranski and Petrusic (2001), and Petrusic and Baranski (2003) found that accuracy did not increase with the inclusion of confidence. Petrusic and Baranski (2000) reported evidence that confidence may in fact impair accuracy. At this point, our finding of increased accuracy with confidence is interesting but unconvincing, and remains an area to be explored in future experiments.

Another effect, which has been inconsistent in past studies, is the effect of confidence processing on discrimination difficulty. In Petrusic and Baranski (2000, 2003) the inclusion of confidence processing prolonged primary decision times to a greater extent for easier pairs compared to more difficult pairs. In Experiment 1, we found that the effect of confidence did not significantly affect primary RT, differently across stimulus pairs, which is also consistent with the findings of Baranski and Petrusic (2001). However upon observing the graph, we felt that there was a tendency of the inclusion of confidence to slow the more difficult pairs to a greater extent than the easier pairs, which is the opposite effect of what Petrusic and Baranski (2000, 2003) found. This interpretation of the results makes sense from the perspective of Van Zandt's (2004) model of confidence processing, based on Vickers', "Balance of Evidence" model, which

theorizes that a decision in a binary discrimination task is the result of the accumulation of evidence in separate accumulators until one accumulator reaches a set criterion. In Van Zandt's model, accumulation of evidence continues throughout the confidence decision, and the criteria for a confidence decision, are reset. A confidence judgement is made in favour of the stimulus that meets its criterion first, corresponding to the magnitude of difference between the accumulators, at the time the decision is made. If indeed this is how we make confidence judgements, then it is reasonable to expect longer confidence RTs for more difficult pairs because it would take longer for one accumulator to reach its criterion, due fluctuation between the accumulators caused by discrimination difficulty. At this point, the interaction of confidence and discrimination difficulty is unclear and remains a topic for further investigation.

In Experiment 2 we found compelling results regarding the expense of confidence processing on mental faculties, in comparison to a choice reaction time task. First though, it should be noted that the design included a 1000 ms ISI after each city comparison, immediately preceding the presentation of the probe to make a confidence or choice RT, on the screen. Therefore participants could not make a response until at least one second after they had completed the discrimination task. In retrospect we think that ideally there should not have been an ISI before the probe for the secondary task, however we believe this in no way affects the primary response time of the decision and therefore does not alter our conclusions.

Experiment 2 demonstrated that the inclusion of a choice reaction time task following a discrimination task had no effect on primary decision times. This is strong evidence that the increase in primary RT that is observed with the inclusion of confidence assessments

is not simply due to the anticipation of making a choice and a motor response, but rather a more elaborate cognitive process, which would be expected from a self-evaluation. This result does not however eliminate the possibility that many other tasks following a discrimination task would also have the effect of prolonging primary RT. For instance, after each selection of a city as "Smaller" or "Larger" we could ask participants to rate on a scale of 50-100, their desire to live that city. This would be much more comparable to a confidence assessment, than a choice reaction time task, because a *self-assessment* is required based on information from the *preceding* task. Our choice reaction time task had neither of these qualities.

Recently Petrusic and Baranski (2005) tested to see if RT was the basis for confidence in cognitive discrimination tasks. They required participants to make binary discriminations, including a city-size comparison analogous to our experiment, but instead of making confidence judgements, they were asked to assess how long their response was by selecting a key corresponding to the time it took them to make the comparison. In the second phase, all participants completed the tasks followed by actual confidence judgements. They argued that if in fact RT was the basis for confidence, then the RT assessments should show the exact same calibration and resolution properties as actual confidence judgements. Response time assessments, like confidence assessments possess the qualities of self-evaluation, and are based on the decision immediately preceding the secondary task. Therefore, their experiment is also in fact an excellent way to test if confidence processing per se, causes prolonged primary decision times or if it is a more general effect. Their study did not include a "no confidence" group like ours, but what they found was, the primary RTs followed by actual confidence judgements were

significantly longer than primary RTs followed by RT scaling. This finding is quite remarkable because it suggests confidence processing may affect decision making differently than other self-assessments participants make, based on the discrimination task.

The results from our experiments generally do not support either decisional or post-decisional theories of confidence processing. During accuracy-stress and speed-stress with a deadline on confidence, we did observe confidence processing during the primary decision. However decisional models predict that confidence processing is integral to the primary decision, and requires no *additional* processing, which is totally inconsistent with the increase in RT we observed in these conditions. In the speed-stress condition without a deadline on confidence, we observed no processing of confidence during the decision, which is consistent with post-decisional models, however these models are unable to account for the RT increase we observed under accuracy-stress, and speed-stress with confidence deadline. Baranski and Petrusic (1998) established that confidence processing is indeed a time-consuming process. They also suggested that it is a process, which runs parallel to decision-making and may have the ability to temporally shift from a decisional locus to a post-decisional locus. They suggest that under accuracy stress, confidence interacts with the decision process so that decision-making may adapt based on momentary confidence assessments. They propose that confidence "evolves" throughout the decision process, suggesting that confidence may change based on new information gathered during decision-making, which Van Zandt and Maldonado-Molina (2004) have presented evidence for. Our results are reflective of the views of Baranski and Petrusic (1998). Van Zandt's alteration of Vickers' Balance of Evidence model

seems to account for our results most accurately in the context of a speed-stress condition.

## REFERENCES

Ascher, D. (1974). A model for confidence judgments in choice tasks. Unpublished thesis submitted for the degree of Ph.D., McMaster University.

Audley, R. J. (1960). A stochastic model for individual choice behaviour. *Psychological Review, 67,* 1-15.

Baranski, J. V. (1991). *Theories of confidence calibration and experiments on the time to determine confidence.* Unpublished doctoral, Carleton University, Ottawa. Ontario, Canada.

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

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

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. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics, 61* (7), 1369-1383.

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

Festinger, L. (1943). Studies in decision: 1. Decision-time, relative frequency of judgement and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 291-306.

Fischhoff, B., Slovic, P. & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552-564.

Garret, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology, 56,* 10105.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswickian theory of confidence. *Psychological Review, 98* (4), 506-28.

Heath, R. A. (1984). Random walk and accumulator models of psychophysical discrimination: A critical evaluation. *Perception, 13,* 57-65.

Henmon, V. A. C. (1911). The relation of the time of a judgement to its accuracy. *Psychological Review, 18,* 186-201.

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 51,* 261-268.

Johnson, D. M. (1939). Confidence and speed in the two-category judgement. *Archives of Psychology, 241,* 4-51.

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

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

Lichenstein, S., Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational*

*Behaviour and Human Performance, 20,* 159-183.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology, 12,* 114-135.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Journal of Mathematical Psychology, 40,* 77-105.

Merkel, J. (1885). Die zeitlichen Verhaltinisse der Willensthatigkeit. *Philosophische Studien, 2,* 73-127.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meterology, 12,* 595-600.

Packer, J. S. (1984). Performance changes in perceptual discrimination and identification. Unpublished thesis submitted for the degree of Ph.D., University of Adelaide.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the national academy of sciences, 3,* 75-83.

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

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

Petrusic, W. M. & Baranski, J. V. (2000). Effects of expressing confidence on decision processing: Implications for theories of RT and confidence. In Bonnet, C. (Ed.) (2000) *Fechner day 2000. Proceedings of the sixteenth annual meeting of the international society for psychophysics.* (pp. 102-109). Strasbourg, France. 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. (2005). Probability assessment with response times and confidence in perception and general knowledge. A paper read at the "Understanding biases in human judgement" symposium at the Joint Meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science and the British Experimental Society, Montreal, Quebec, July 2005.

Statistics Canada (2001). *2001 Census: Population and Dwelling Counts.* Retrieved from http://www12.statcan.ca/english/census01/products/standard/popdwell/Table-CMA-N.cfm

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

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13,* 37-58.

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

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

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

Vickers, D. & Smith, P. (1985). Accumulator and random-walk models of

psychophysical discrimination: a counter evaluation. *Perception, 14* (4), 471-497.

Vickers, D., Smith, P. L., Burt, J., & Brown, M. (1985). Experimental paradigms emphasizing state or process limitations: II. Effects on confidence. *Acta Psychologica, 59,* 163-193.

Vickers, D. & Pietsch, A. (2001). Decision making and memory: A critique of Juslin and Olsson's (1997) Sampling model of sensory discrimination. *Psychological Review, 104* (4).

Volkman, J. (1934). The relation of time of judgement to certainty of judgement. *Psychological Bulletin, 31,* 672-673.

Yaniv, I., Yates, J. F. & Smith, J. E. K. (1991). External correspondence: Decompositions of the mean probability score. *Psychological Bulletin, 110,* 611-617.

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