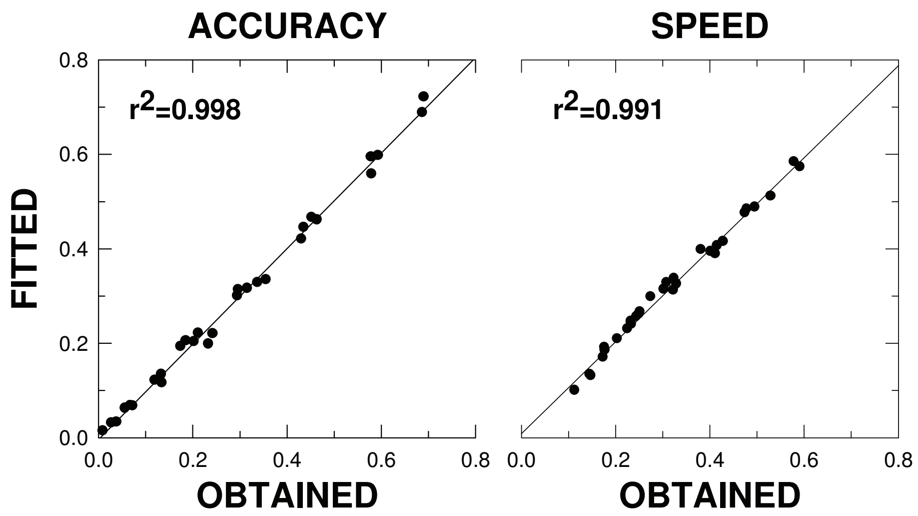

Figure 2. Fitted proportions, p''_{jg} 's plotted against the obtained proportions, p'_{jg} 's for the accuracy and speed stress conditions.

Summary and Conclusions

Application of Case D of Torgerson's (1958) Law of Categorical Judgment to the relative frequencies of confidence category usage in a line length discrimination task permitted estimates of the scale values for the confidence category boundaries. Upon normalization and rescaling, these scale values were linearly related to the objective confidence category boundaries in both the accuracy and speed stress conditions. The slope of each of these plots was essentially 1.0, thus showing that the confidence category scale values is equally spaced on the underlying subjective probability scale. Fitted proportions, based on the scale values for the category boundaries and the stimuli, corresponded very closely to the obtained, indicating that the Case D assumptions were justified.

References

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

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: Wiley.

Acknowledgement

This work was supported by a Natural Sciences and Engineering Research Council of Canada Discovery grant to William M. Petrusic.

THE TIME COURSE OF CONFIDENCE PROCESSING

Steven R. Carroll and William M. Petrusic
Department of Psychology, Carleton University, Ottawa, Canada
srcarrol@connect.carleton.ca; bill_petrusic@carleton.ca

Abstract

The time course of confidence processing is discussed in light of three decision-making experiments: two psychophysical tasks and one general knowledge task. Each experiment consisted of a confidence block of trials, wherein participants expressed a confidence rating following each rendered decision, a no confidence block, where confidence was never expressed, and a stop-confidence block, where participants were told to expect to have to express confidence but were instructed not to do so if a tone sounded. Stop-confidence tone onset delays were varied systematically. Decisional response time data reveal how confidence processing unfolds linearly throughout the primary decision-making process.

Baranski and Petrusic (1998) initiated the study of the locus and the time course of confidence processing in comparative judgements through investigation of the properties of the time to determine confidence. The time to determine confidence is defined by the time between the expression of the primary decision and the selection of a confidence category corresponding to the subjective probability that the decision rendered was correct. Baranski and Petrusic concluded that when the primary decision is made under speed stress confidence is computed post-decisionally. On the other hand, under accuracy stress, there was a strong suggestion that confidence is initiated and perhaps even completed during the primary decision process. Indeed, Petrusic and Baranski (2000, 2003) provided direct evidence that confidence is processed during the primary decision process under a stress for accuracy at the expense of speed. Using a between participants design in a line length discrimination task, one group of participants rendered confidence following the primary decision and the other simply made the comparative judgement. The requirement of confidence judgements substantially increased primary decisional response times (RTs). Moreover, confidence times varied systematically with confidence category, thereby implicating post-decisional confidence processing.

The Baranski and Petrusic (2003) and Petrusic and Baranski (2000, 2003) work demonstrate how rendering confidence increases primary decisional RTs but fails to specify the time course of the computation of confidence. For example, the balance of evidence may be computed during the final stages of decisional processing. On the other hand, the balance of evidence may evolve, continuously, over the course of decision processing. The presented experiments were designed to further examine both the loci and the time course of confidence processing by using a variant of Logan's (1983) "stop-signal" paradigm.

Experiments 1 and 2

Method

Participants. Forty-eight Carleton University undergraduate students participated in return for course credit. Twenty-four were assigned to the signal detection experiment, twenty-four to the line-length discrimination experiment.

Apparatus. The study was conducted using a desktop computer with a standard colour monitor. The computer was equipped with a Pentium-class processor, a Soundblaster soundcard, and a Windows 98 operating system. Stimulus presentation and response data collection was controlled via Superlab Pro v. 2.0. Participant responses were made via a control panel with two primary response buttons (labelled 'yes/no' or 'left/right' depending on the task) and seven confidence response buttons (labelled 'X/50/60/70/80/90/100').

Stimuli. Signal detection stimuli were 16, 100 x 100 pixel squares arranged to form a larger 4 x 4 square. Each of the 16 squares contained 2500 2 pixel x 2 pixel 'dots'. Each dot was either inactive (coloured white) or active (coloured black). The density level of the dots in the four centre squares varied from trial to trial, but on any given trial all of the centre squares were homogenously dense, with either 50% active dots (the 'noise' condition), 52%, 54%, 56%, or 58% active dots. Only 50% of the dots in the background squares were active on any given trial.

Line-length stimuli consisted of single, horizontal lines bisected by a vertical line 10 pixels in length. The vertical line was positioned so one of the two horizontal line-segments was always 100 pixels long, while the remaining segment was always longer being either 102, 104, 106, or 108 pixels in length. The lines were offset horizontally by 24 pixels to either the left or right of the centre of the screen, and were positioned 178 pixels above an instruction ('shorter' or 'longer').

Procedure. Participants in the signal detection task were asked, on each trial, to decide whether there was a greater density of active dots in the centre of the display than there were around the edges of the display. Participants in the line-length discrimination group were asked, on each trial, to decide which of the two line-lengths was either shorter or longer, depending on the instruction. Half of the trials in each block were noise trials and the other half signal trials, with each of the four signal strengths occurring equally often for a total 336 trials in each of three blocks for the signal detection group. Each of the 4 line length pairs appeared with each of the two instructions in each of the two left-right orders and this factorial combination was replicated 24 times for a total of 384 trials in each of four blocks for the line-length discrimination group.

Each participant received each of the three conditions, with condition presentation order varying between participants. In one condition, participants were asked, following each primary decision, to rate their confidence in the accuracy of the decision they had just made: a rating of '50' was indicative of a guess, a rating of '100' indicated certainty, and the ratings 60-90 were to be used accordingly. Participants were further instructed to select a confidence rating of 'X' if they were certain they had made a mistake.

In a second condition, participants never had to render confidence. In the third condition, participants were told to expect to have to rate their confidence, but were told they would not have to do so if they heard a short tone. The tone was presented on one-half of the trials, and its onset following stimulus presentation varied, in 100 ms increments, from 100 ms to 500 ms. As well the tone was presented simultaneously with the onset of the stimulus pair.

Results and Discussion

RTs three standard deviations above the mean were censored, accounting for 2.29 % of the 24192 trials in the detection task and 1.98 % of the 27648 trials in the line length task. Throughout, the primary dependent variables for each cell in the design are the mean of each participant's overall primary decisional RTs, mean confidence time, mean proportion correct

responses, and mean confidence rating. Only those results relevant to the discussion of the locus and time course of confidence processing are presented here. In each analysis of variance (ANOVA), Huynh-Feldt, epsilon adjustment of degrees of freedom was used. However, the degrees of freedom associated with each value of *F* are defined by the design and the Mean Square Errors provided in the text are those given by the conventional degrees of freedom. Level of significance was set at 0.05 throughout.

Primary Decisional RT Analyses. The main effect of condition (No confidence, Tone, No-Tone, Confidence) on primary decisional RTs was reliable for both the signal detection task, F(3, 54)=9.09, partial η^2 = .335, and the line-length discrimination task, F(3, 54)=15.57, partial η^2 = .464. Participants took longest to make decisions whenever they were required to render confidence.

The Effect of Tone Onset Delay on Primary Decisional RTs. The plots in Figure 1 provide mean primary decisional RTs for each of the stop-confidence signal onset delays with the stop-confidence condition for each task. As is strikingly evident for each task, decisional RTs increase *linearly* with the delay in the stop signal; i.e., with the amount of confidence processing. An ANOVA showed the linear component of the overall sum of squares for the effect of tone onset delay to be marginally significant for the signal detection task, F(1, 23)=3.21, p>0.08, partial η^2=.123, and reliable the line-length discrimination task, F(1, 23)=7.74, partial η^2 = .252.

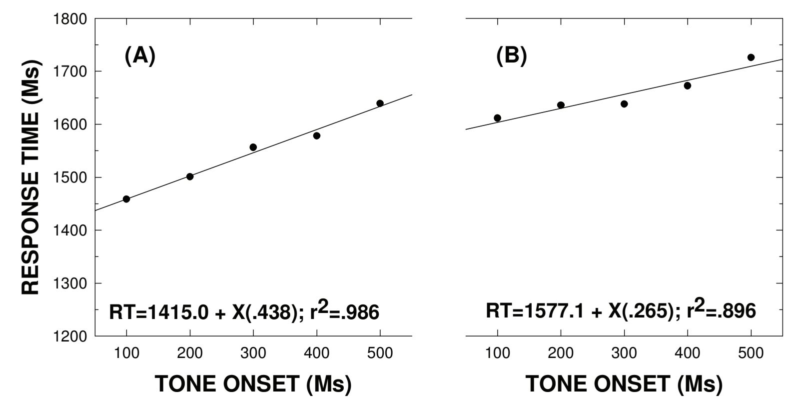

Figure 1. Primary decisional RTs with 95% confidence intervals during the Stop-Confidence condition for the signal detection (Panel A) and the line-length discrimination tasks (Panel B).

Time to Render Confidence. It took considerably more time to render confidence in the stop-confidence condition than in the confidence condition, with the main effects of condition evident for the both the detection, F(1,18)=70.808, partial η^2 = .797, and for the line-length discrimination tasks, F(1,18)=48.233, partial η^2 = .728. Replicating the findings of Petrusic and Baranski (2003), a main effect of decisional difficulty on time to render confidence was significant for the detection task, F(1,18)=6.475, partial η^2 = .265; confidence times decreased as the differences between the compared densities increased, indicating post-decisional confidence processing. In contrast, no effect of discriminative difficulty was found for the line-length discrimination task, F(1,18)=1.42, p>.25, partial η^2 = .073, suggesting little, if any, confidence processing occurred post-decisionally.

Experiment 3

These results beg the question of whether confidence continues to be processed at a constant rate throughout the entire decision-making process, or whether confidence processing asymptotes prior to the explicit expression of a decisional response. In Experiment 3 we attempted to answer this question by studying the effect of prolonging the stop confidence signal onset delay. Specifically, the delay was extended to a point where participants have had enough time to accrue most, if not all, of the evidence required to render a confidence judgement.

Method

Participants. Twenty-Four Carleton University undergraduate students participated in return for course credit.

Apparatus. Identical to that used in Experiment 2 (i.e., the line-length discrimination task).

Stimuli. Stimuli were 792 pairs of Canadian city names. Stimulus cities were selected at random from a list of 395 Canadian urban centres and were categorized according to the ratio of differences in size between the city populations. A comparison between two cities whose populations differed by a ratio greater than 2.99 was considered 'easy', a ratio between 2.98 and 1.6 was considered to be of 'moderate' difficulty, and a ratio less than 1.6 was considered 'hard'. 264 city pairs were selected for each difficulty category. Within each difficulty category, every effort was made to control for effects of city size. Canada's five most populous urban centres were excluded from the stimulus set.

Procedure. Participants were asked, on each trial, to decide which of the two presented cities was either largest or smallest, depending on the instruction. Instructions appeared centred on the screen for a period of two seconds before being replaced by the stimulus pair. For half of the trials the largest city was presented on the left, for the other half of the trials the largest city was presented on the right. Participants received each of the three conditions described in Experiments 1 and 2. For the stop-confidence condition, the tone was presented on one-half of the trials, and its onset following stimulus presentation varied, in 250ms increments, from 250ms to 2500ms. A 0ms tone onset delay was also included.

Results and Discussion

RTs three standard deviations above the mean were censored, accounting for 4.96% of the 19008 trials. Analytical methodology directly corresponded to that used in Experiments 1 and 2. Again, only those results relevant to the discussion of the locus and time course of confidence processing are presented here.

Primary Decisional RT Analyses. An effect of condition was found, $F(3, 54)=8.51$, $MSE=711049.4$, partial $\eta^2=.321$, but it is interesting to note that only the pure No Confidence condition differed reliably from the other conditions. This finding, along with an examination of Figure 2, suggests that confidence processing may indeed develop ballistic properties if the stop processing signal is delayed for a long enough time.

The Effect of Stop Confidence Signal Onset Delay on Primary Decisional RTs. Figure 2 presents the mean decisional RTs for stop confidence signal onset delays ranging from 250ms to 2500ms. There was a reliable effect of tone onset delay, $F(10, 180)=14.57$, $MSE=294912.98$, partial $\eta^2=.447$. The relationship between mean decisional RT and signal

delay was reliably quadratic, $F(1, 18)=29.74$, $MSE=282735.76$, partial $\eta^2=.623$. Increases in mean decisional RTs with onset delay were evident, though modest, up to the 1500ms delay, where the mean decisional response time was 3235ms: just 98ms less than the mean decisional RT for the confidence trials in this condition. Signal delays longer than 1500ms resulted in comparatively longer decisional RTs.

Segmental linear regression analysis confirmed that the 1500 ms delay represented the intercept between two linear segments. Neither of the two 95% confidence intervals constructed about the slopes of these segments contained 0, suggesting that both slopes described by Figure 2 are significantly positive. Further trend analyses conducted for the first five tone onset delays revealed how mean RT increased along with signal delay in a marginally significant and linear fashion, $F(1, 18)=4.32$, $MSE=178690.721$, partial $\eta^2=.194$, $p=.052$. Analyses for longer delays (1500ms - 2500ms) demonstrated a similarly linear relationship between mean RT and signal delay, $F(1, 18)=75.618$, $MSE=252611.299$, partial $\eta^2=.808$.

It would seem that long stop confidence processing signal delays (i.e., those that come after a participant has had enough time to completely process confidence) severely disrupted the decision-making process: a hypothesis supported by the finding that participants making hard decisions were correct significantly less often in the Tone condition compared to the other conditions.

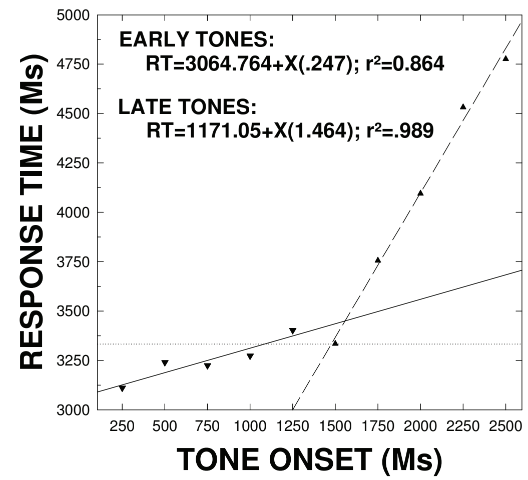

Figure 3. Primary decisional RTs as a function of stop confidence signal delay. The dotted line represents mean decisional response time for confidence trials in this condition.

Time to Render Confidence. No significant effect of decisional difficulty was found on time to render confidence, suggesting a largely decisional locus of confidence processing.

Decisional Accuracy. Importantly, there was an effect of condition on decisional difficulty, $F(3, 54)=3.75$, $MSE=.006$, partial $\eta^2=.173$ and a Condition x Decisional Difficulty interaction, $F(6, 108)=4.28$, $MSE=.007$, partial $\eta^2=.192$. Participants were correct least often when making hard decisions in the Stop Confidence condition, supporting the hypothesis that a stop confidence processing signal presented after participants have had enough time to entirely process confidence was quite disruptive to the overall decision-making process.

Summary and Conclusions

This study serves as an important extension to Petrusic and Baranski (2000, 2003) and Baranski and Petrusic (2003). While these earlier studies are clear in demonstrating a decisional locus for confidence processing, the present experiments illustrate how confidence processing begins immediately upon presentation of the stimuli to be compared. What is more, the present study sheds light on how confidence processing evolves throughout the decision making process.

Experiments 1, 2, and 3 demonstrate how confidence processing occurs in parallel to the decision-making process, how confidence processing is not ballistic, and how confidence processing evolves steadily throughout the course of decisional processing. Experiment 3 sets limits on these latter two findings by illustrating how the steady pace of confidence processing will continue only until such time as all required decisional evidence has been accrued (i.e., until the mean time required for participants in this experiment to make a decision and subsequently render confidence if allowed to do so). Once this threshold is passed, however, asking participants to discard confidence related evidence disrupts the decision-making process, resulting in significantly longer decisional response times as well as diminished decisional accuracy in certain cases.

References

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, **24**, 929-945.

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

Logan, G. D. (1983). On the ability to inhibit simple thoughts and actions: 1. Stop-signal studies of decision and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9**, 585-606.

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

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

Statistics Canada. *Population and dwelling counts, for Canada and census subdivisions (municipalities), 2006 and 2001 censuses*. Retrieved August 9, 2007, from http://www12.statcan.ca/english/census06/data/popdwell/Table.cfm?T=301&S=3&O=D

Acknowledgements

This work was supported by a Natural Sciences and Engineering Research Council grant to Carroll and a Natural Sciences and Engineering Research Council of Canada Discovery grant to Petrusic.

HOW WAS IT FOR YOU?
PSYCHOPHYSICS AND THE EVALUATION OF STUDENT EXPERIENCE OF E-LEARNING

Diana Eugenie Kornbrot, University of Hertfordshire
d.e.kornbrot@herts.ac.uk

Evaluating the student experience of Higher Education has become a matter of national importance in several countries. For example, in England & Wales the National Student Survey (NSS) is administered on line to all students in the final year of their undergraduate degree. The NSS uses 5-point Likert scales, giving extent of agreement or disagreement with positive statements. There are 22 questions, covering 6 aspects of student experience. This presentation considers how psychophysical methods based on signal detection theory or Luce's choice theory can be used to analyze such data. Such methods can determine how well the questions discriminate different aspects of experience, as well as how favourably the students experience these aspects of their education. Particular emphasis is given to exploring discipline differences together with the effects of recent technologies, such as managed learning environments and web 2.0 social software.

Higher education matters: to students, parents, communities, countries, and the planet. Consequently, everyone wants to evaluate the success of the higher education enterprise. This Ms. draws on both performance and satisfaction data from higher education (university) data on satisfaction and performance in England & Wales to illustrate the problem and show how psychophysics can make a contribution. There are two key reasons why evaluation remains so difficult. Namely, such evaluations are 'high stake' and have multiple stakeholders.

High stakes are manifest in the importance of a good degree. For the individual student, getting a 'good' degree is of paramount importance, preferably with as little effort and risk as possible. This poses a fundamental problem for people devising assessment procedures. Identical procedures do *not* measure the same underlying property over time. IQ tests provide an interesting example, with massive improvements over the C20th (Flynn, 1987). As achieving a high IQ is 'high stake', 'teaching to the test' became rife. Many cried 'foul', and tried: either to restrict such teaching, or to devise non-coachable items. Doomed to failure. If the results are high stake, then there is an inevitable 'arms race' between the assessors and the assesses. What the psychological instruments measure changes. Physicists don't have this problem. The ruler measures length, the scales measure weight, etc.

The problem is just as intractable when assessing student satisfaction. All the stakeholders would like to know how students experience features of education such as: teaching, assessment, learning resources, personal development, etc. Clearly here too, if it is in a student's interest to graduate from a high satisfaction institution, then student may modify their questionnaire responses in line with those interests.

Higher education evaluation has multiple stakeholders. Degree results are widely used to validate psychological properties that predict how people will perform on *future* tasks. This is equally true of the potential graduates who will be the actual task performers and of the employers who will profit from their efforts. It is also true of the communities, local, national and international that will benefit from a high quality graduate workforce. Furthermore, the paying clients of the higher education industry, be they students, parents, government or industry will want 'value for money'. In attempting to satisfy these needs higher education evaluation is actually fulfilling two, sometimes conflicting, purposes. Firstly, degree performance informs the world about what the student *already* knows.