National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

**THEORIES OF CONFIDENCE CALIBRATION AND
EXPERIMENTS ON THE TIME TO DETERMINE CONFIDENCE**

by

JOSEPH VICTOR BARANSKI

A dissertation submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
Department of Psychology
Carleton University
Ottawa, Ontario
July, 1991

© copyright
1991, Joseph V. Baranski

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-70922-7

Canada

The undersigned hereby recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis,

**THEORIES OF CONFIDENCE CALIBRATION AND
EXPERIMENTS ON THE TIME TO DETERMINE CONFIDENCE**

submitted by

JOSEPH VICTOR BARANSKI, B.Sc., M.A.

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Carleton University
September, 1991

.....
Chairman, Department of Psychology

.....
External Examiner

.....
Advisor

Abstract

Three models of confidence calibration in human judgment and discrimination are proposed and empirically evaluated. A model based on a categorical scaling of the amount of inconclusive information present at the completion of the judgment is shown to provide impressive fits to the calibration data of perceptual (Keren, 1988; Baranski & Petrusic, 1991a) and intellectual knowledge (Lichtenstein & Fischhoff, 1977) tasks and is more general than existing calibration theory (e.g., Ferrell & McGoey, 1980) since it provides quantitative predictions for the decision time associated with each confidence level.

As well, three experiments investigated the properties of the time to determine confidence with the goal of determining the processing locus of confidence judgments. The results suggest that confidence is determined post-decisionally when people are under demands to respond quickly; perhaps to free the primary comparison process of processing time and/or attentional resources. Under conditions emphasising accurate responding, however, the determination of confidence seems to be initiated, or can even be completed, during the decision process; perhaps because confidence is used in the adaptive regulation of decisional parameters. The latter finding also implies that the collection of confidence data may result in primary decision latencies that include time to determine confidence, rendering them difficult, if not impossible, to model empirically.

Acknowledgements

I would like to thank the members of my doctoral committee for their careful consideration of this dissertation: Dr. Bill Jones (chair), Dr. A.A.J. Marley (external examiner), Dr. Ann Laubstein (internal examiner), Dr. Robert Hoffmann, and Dr. Chris Herdman. I would like to give special thanks to Dr. William Petrusic, my advisor, mentor, and friend. Thank you for all the guidance, financial support, and for teaching me to be a scientist. Finally, I would like to thank my dear wife, Mary Anne, for her encouragement and love ("bears" forever).

This research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Dr. William Petrusic. Financial support for the author was provided by an Ontario Graduate Scholarship award, a Carleton University Dean's scholarship award for doctoral studies and research, teaching opportunities provided by chairman Bill Jones, and a research assistantship provided by Dr. William Petrusic.

Table of Contents

	Page
Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
 General Introduction	 1
 Chapter I: Confidence in Human Judgment and Discrimination	 3
Properties of Confidence Judgments	3
Decision Difficulty	3
Demands for Speed vs. Accuracy	4
E- vs. S-Controlled Situations	5
Semantic Congruity Effects	8
Theories of Decision Confidence	9
Early Theories	9
The 'Runs' Model	10
The Balance of Evidence Hypothesis	13
A Random-walk Theory of Confidence	14
 Chapter II: The Calibration of Confidence Judgments	 17
Introduction	17
Properties of Confidence Calibration	20
Overwhelming Overconfidence	20
Effects of Decision Difficulty	21
Effects of Training or Expertise	22
Factual Knowledge vs. Perception	23
Calibration and the Time to Choose	26
Models of Confidence Calibration	29
Heuristics and Biases	29
May's Calibration Model	30
The Calibration Model of Ferrell and McGoev	31
 Chapter III: Information Scaling Theories of Calibration	 35
Introduction	35
The Slow and Fast Guessing Theory	36
Information Scaling Theories of Calibration	38
The Scaling of Competing Information	39
The Scaling of Decision Time	52
The Scaling of Doubt	53
A Doubt Scaling Analysis of Keren (1988)	59
A Doubt Scaling Analysis of Lichtenstein and Fischhoff (1977)	 61

Summary	63
Chapter IV:	
The Locus of the Determination of Confidence	64
Introduction	64
An Internal Event Monitor Hypothesis	65
A Counting and Matching Hypothesis	66
Chapter V:	
Some Experiments on the Time to Determine Confidence .	68
Introduction	68
Apparatus	68
Experimental Conditions	69
Stimuli and Basic Procedure	69
Experiment 1: Henmon (1911) Revisited	70
Introduction	70
Subjects	71
Stimuli and Design	71
Results and Discussion	72
Experiment 2: Variations of Judgmental Difficulty .	80
Subjects	80
Stimuli and Design	80
Results and Discussion	81
Experiment 3: Difficulty, Speed vs. Accuracy, and the development of Automaticity	91
Subjects	91
Stimuli and Design	91
Results and Discussion	93
General Discussion	150
References	159
Footnotes	174
Appendix	178

LIST OF TABLES

TABLE	PAGE
1. Sequences of response terminating evidence accrual configurations and parameter estimates for the Competing Information Scaling Model fit to the data of Baranski & Petrusic (1991a): $\alpha_1=\alpha_2=\alpha_3=5$ and $\delta=.60, .70, .76$, and $.78$	45
2. Proportion correct (P) and response frequencies (F) for the four confidence categories as a function of sessions and speed (S) versus accuracy (A) in Experiment 3	97
3. Proportion correct and confidence rating as a function of speed versus accuracy and sessions in Experiment 3. Confidence ratings (1-4) are based on Guess=1, Low=2, Moderate=3, and Certain=4	106
4. Mean confidence rating for correct and error responses under speed and accuracy in Experiment 3. Confidence ratings (1-4) are based on Guess=1, Low=2, Moderate=3, and Certain=4	109
5. Mean decisional RT (ms) for each pair with the instructions "Shorter" and "Longer" under speed and accuracy in Experiment 3. The RT difference between the instructions shorter and longer are provided in parentheses	138
6. Mean confidence rating for each pair with the instructions "Shorter" and "Longer" under speed and accuracy in Experiment 3. The confidence difference between the instructions shorter and longer and the mean confidence are also provided	140
7. Mean confidence response time (ms) for each pair with the instructions "Shorter" and "Longer" under speed and accuracy in Experiment 3. The RT difference between the instructions shorter and longer and the mean confidence RT are also provided	141
8. Mean motor response time (ms) for each pair with the instructions "Shorter" and "Longer" under speed and accuracy in Experiment 3. The RT difference between the instructions shorter and longer and the mean motor RT are also provided	143

LIST OF FIGURES

FIGURE	PAGE
1. Hypothetical calibration curves illustrating the phenomena of overconfidence, under-confidence and perfect calibration	19
2. Calibration curves and relative response frequencies for the four levels of comparative difficulty for the data of Baranski and Petrusic (1991a): Level 1 ($r=1.10$) circles, Level 2 ($r=1.08$) diamonds, Level 3 ($r=1.05$) inverted triangles, and Level 4 ($r=1.03$) triangles (r is the ratio of the longer to the shorter extent of the comparison pair)	25
3. Mean correct (filled circles) and error (open circles) response times plotted as a function of the four levels of comparative difficulty for the data of Baranski and Petrusic (1991a): Level 1 ($r=1.10$) circles, Level 2 ($r=1.08$) diamonds, Level 3 ($r=1.05$) inverted triangles, and Level 4 ($r=1.03$) triangles (r is the ratio of the longer to the shorter extent of the comparison pair)	28
4. Calibration curves and relative response frequencies predicted by the Competing Information Scaling Model under accuracy for the 4 levels of comparative difficulty in the data of Baranski and Petrusic (1991a): Level 1 ($\delta=.78$, easiest) circles, Level 2 ($\delta=.76$) diamonds, Level 3 ($\delta=.70$) inverted triangles, and Level 4 ($\delta=.60$, most difficult) triangles (δ is the discriminability parameter)	44
5. Expected number of evidence accrual events (response time) for correct and error responses for the four levels difficulty as predicted by the Competing Information Scaling model. Parameter values used to derive these plots are provided in the text	51
6. Calibration curves and relative response frequencies predicted by the Doubt Scaling model under accuracy for the 4 levels of comparative difficulty: Level 1 ($\delta=.72$, easiest) circles, Level 2 ($\delta=.70$) diamonds, Level 3 ($\delta=.65$) inverted triangles, and Level 4 ($\delta=.59$, most difficult) triangles (δ is the discriminability parameter)	55

7. Calibration curves and relative response frequencies predicted by the Doubt Scaling model under speed for the 4 levels of comparative difficulty: Level 1 ($\delta=.72$, easiest) circles, Level 2 ($\delta=.70$) diamonds, Level 3 ($\delta=.65$) inverted triangles, and Level 4 ($\delta=.59$, most difficult) triangles (δ is the discriminability parameter)	56
8. Expected number of evidence accrual events (response time) for correct (filled points) and error (open points) responses for the four levels of difficulty as predicted by the Doubt Scaling model under accuracy and speeded conditions. Model parameter values are provided in the text	58
9. Doubt Scaling model fit to the data of Keren (1988, Experiment 1). Response frequency data points were provided in Keren (1988). However, calibration data points were approximated by eye. Model parameter values are provided in the text	60
10. Doubt Scaling model fit to the data of Lichtenstein and Fischhoff (1977, Experiment 5). Calibration and relative response frequency data points were approximated by eye from Ferrell and McGoev (1980). Model parameter values are provided in the text	62
11. Mean decisional response time (open circles), percent correct (open squares), and the time to determine confidence (filled circles) in Experiment 1. Each point is based on the mean of individual subject means at each confidence level	73
12. The time to determine confidence for subjects who display slow and fast guess response times ($N=9$ for each group) in Experiment 1	76
13. Mean correct (open circles) and error (filled circles) decisional response time and confidence response time for each confidence level in Experiment 1	79
14. Calibration difficulty effect: Mean percent correct and relative response frequency for each level of difficulty in Experiment 2 (r is ratio of the standard and variable stimuli)	82

15. Response time difficulty effect: Mean decisional response time and confidence response times for each level of difficulty in Experiment 2 (r is the ratio of the standard and variable stimuli	84
16. Mean correct (open circles) and error (filled circles) decisional response times and confidence response times in Experiment 2	86
17. Latency-Probability Functions (LPFs) for decision time and confidence time in Experiment 2 (p values less than .5 indicate error responses while p values greater than .5 indicate correct responses)	88
18. Speed/Accuracy calibration curves for each subject in Experiment 3	95
19. Probability correct as a function of difficulty level and set for speed versus accuracy for each subject in Experiment 3	102
20. Confidence rating as a function of difficulty level and set for speed versus accuracy for each subject in Experiment 3	103
21. Confidence rating as a function of stimulus difficulty and speed versus accuracy for the data of Vickers and Packer (1982) [re-plotted by eye] and the combined data in Experiment 3	105
22. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 1 in Experiment 3	113
23. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 2 in Experiment 3	118
24. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 3 in Experiment 3	122
25. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 4 in Experiment 3	127

26. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 1 in Experiment 3	131
27. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 2 in Experiment 3	132
28. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 3 in Experiment 3	133
29. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 4 in Experiment 3	134
30. Confidence response time distributions for subject 1 in Experiment 3	146
31. Confidence response time distributions for subject 2 in Experiment 3	147
32. Confidence response time distributions for subject 3 in Experiment 3	148
33. Confidence response time distributions for subject 4 in Experiment 3	149

GENERAL INTRODUCTION

Since the pioneering work of Fechner and Donders, decision psychophysicists have come to accept without question the utility of response accuracy and response time as tools for measuring elementary mental operations. However, for virtually as long as researchers have measured the speed and precision of decisions they have also measured the confidence people associate with the decisions they make. For early researchers, confidence provided quantified introspection and thus direct access to the elusive mind/body relationship; the measurable connection between the sensory experience of the elements in our world and our conscious understanding of them (e.g., Pierce & Jastrow, 1884; George, 1917; Lund, 1926; Williamson, 1915; Henmon, 1911). For contemporary researchers, on the other hand, confidence is studied by some because of its inherent utility in applied psychophysical contexts [e.g., the confidence based receiver operating characteristic (ROC) of Signal Detection Theory (see Egan, 1975; Green & Swets, 1966)]. For still others, confidence constitutes an elaborate 'meta-cognitive' ability that internally monitors performance and thus provides a basis for adaptive regulation of the decision making process (Vickers, 1978,1979). Independent of the motivation for research, confidence, as will be shown, is a meaningful and systematic aspect of our cognitive behaviour. It thus demands both empirical measurement and full recognition in any comprehensive decision theory.

Chapter 1 of this dissertation reviews the properties and models of confidence in human judgment and discrimination.

Chapter 2 reviews the research and theory concerned with the confidence 'calibration' problem (i.e., the relationship between subjective and overt performance). In Chapter 3, I outline three new quantitative models of confidence calibration. The behaviour of these models will be described and then evaluated through application to existing calibration data. In Chapter 4, I propose a new problem that extends the domain of confidence research. In particular, I wish to ascertain whether confidence is determined in conjunction with the primary decision or through a post-decisional evaluation of the information that motivated the primary judgment. Finally, in Chapter 5, I review the results of three experiments that address this new research. The overall objective is to theoretically address the stages and mechanisms involved in the comparative judgment task; from stimulus to response, through to the basis and locus for confidence in human judgments.

(I) CONFIDENCE IN HUMAN JUDGMENT AND DISCRIMINATION

1.1 Properties of Confidence Judgments

1.1.1 Decision Difficulty. Apparently the first experimental study of decision confidence was conducted by Pierce and Jastrow (1884). In a task involving successive comparisons of lifted weights, with only themselves as observers, Pierce and Jastrow noted that the confidence in the correctness of their judgments increased as the physical magnitude difference between the weights (and thus overt discrimination accuracy) increased. This result was soon confirmed in experiments with lifted weights and brightness discriminations by Fullerton and Cattell (1892) and with lifted weights and line-length comparisons by Garrett (1922).

Since these pioneering studies, the dependence of confidence on judgmental difficulty (e.g., stimulus differences) has been demonstrated on numerous occasions and with a variety of tasks (e.g., Trow, 1923; Seward, 1928; Johnson, 1939; Festinger, 1943; Pierrel & Murray, 1963; Vickers & Packer, 1982; Vickers, Smith, Burt, & Brown, 1985; Nicolson, 1982; Baranski & Petrusic, 1991a; Petrusic & Baranski, 1989a). Moreover, the result has proved to be consistent across a variety of confidence scales. For example, Henmon (1911) and Trow (1923) required their subjects to report confidence on a four-point scale: (a) certain, (b) fairly confident, (c) little confidence, and (d) guess. Seward (1928), in a recognition memory task, had subjects indicate confidence by 'marking-off' a scale ranging from -100 (absolutely sure the

stimulus was not presented) to +100 (absolutely sure the stimulus was presented). In this case, 0 indicated complete doubt. Alternatively, Vickers and Packer (1982), in a line-length comparison task, had subjects report confidence on a scale from 1-5 while Petrusic and Baranski (1989a, 1989b) and Baranski and Petrusic (1991a) had their subjects express confidence in terms of a percentage; from 0 (absolute certainty in an error) to 100 (absolute certainty in a correct response). In this case, 50 represented a guess response.

In sum, confidence varies systematically with the difficulty of judgments; independent of the number of confidence categories available or the metric of the scale employed. Furthermore, as noted by Audley (1964), subjects can quite readily report the confidence in their judgments under any of these conditions.

1.1.2 Demands for Speed Versus Accuracy. It is perhaps obvious that confidence and accuracy should vary systematically when discriminability is directly varied. However, it has long been known that discriminative accuracy, for a specific level of difficulty, can also be effectively varied indirectly simply by restricting the time people are allotted to make their judgment. Typically, accuracy decreases as the demands of the experiment require more rapid judgments (e.g., Garrett, 1922; Johnson, 1939; Festinger, 1943).

In view of this, it was surprising when confidence was found to show no such parallel in response to demands for speeded

responding. Rather, the degree of confidence for a given stimulus difference was consistent across both speed and accuracy emphasised conditions (Garrett, 1922; Johnson, 1939; Festinger, 1943).

Recently, however, Vickers and Packer (1982) suggested a resolution to this apparent paradox. They first noted that in the studies not finding variations in confidence under speed and accuracy conditions, the manipulations were conducted either between-groups or within-subjects but across sessions. They suggested that confidence may have been scaled differently under the speed and accuracy emphasized conditions:

For example, the most confident response in both conditions might have been assigned a rating of 100 per cent, and the least confident a rating of 0 per cent, irrespective of differences in the subjective impression experienced in the two conditions.... It corresponds to an unwanted range effect, or stimulus equalizing bias, discussed by Poulton (1973, 1979) ... (Vickers & Packer, 1982, pp. 186).

They argued that a more appropriate way to examine confidence under speed and accuracy conditions would be to ensure a common scaling and thus conduct the manipulation within-subjects and within-sessions. Their now landmark result was that subjects were significantly more confident on blocks emphasizing accurate responding (see Petrusic, 1991, for a recent replication).

1.1.3 Experimenter (E) versus Subject (S) Controlled Situations.

In 1956, Irwin, Smith, and Mayfield (1956) presented subjects with cards on which positive and negative numbers were printed.

The subjects were asked to judge, after each card, whether the mean of the whole pack of 500 cards was greater or less than zero, and to express their confidence in this judgment by a number between 0 and 100. They found that confidence was higher after 20 cards had been inspected than after only 10 were seen; suggesting that confidence increases with information provided and thus decision time allowed.

A similar result was reported by Geller and Pitz (1968) in the context of subjective probability revision. In their task, subjects were shown a sampling of poker chips from one of two bags. One bag contained 60 red chips and 40 white chips and the other contained 40 red and 60 white. Subjects were told that one of the two bags would be chosen at random and were shown a sequence of 20 chips drawn with replacement from the chosen bag. Prior to each chip sampled, subjects predicted the outcome (red or white). After each successive event, the subject chose a bag and indicated a confidence in the choice. As with Irwin et al. (1954), confidence increased as more information was provided.

At first, these results seemed to contradict the well-established finding in psychophysical comparison studies that longer decision times are typically associated with more difficult judgments and thus lower accuracy and lower confidence (Henmon, 1911; Johnson, 1939; Festinger, 1943; Pierrel & Murray, 1963). However, once again, this seemingly anomalous result was explained by Vickers and his associates (Vickers, Smith, Burt, & Brown, 1985).

Vickers et al. (1985) noted that in the traditional psychophysical comparison studies, showing an inverse relationship between confidence and decision time, subjects controlled the duration of the trial; terminating the stimulus display when the judgment was made. In the latter experiments where confidence seemed to be directly related to decision time, on the other hand, the duration of the trial, and thus the amount of information allotted to the subject, was controlled by the experimenter.

Vickers et al. (1985) distinguished these two paradigms as Subject (S) and Experimenter (E) Controlled paradigms, respectively, and suggested that the former more closely reflects an effect of information quality on confidence, while in the latter information quantity is systematically varied. As such, two completely different sets of predictions arise concerning decisional confidence. Namely, in the S-Controlled paradigm, quick responses will typically be associated with easier judgments and thus higher accuracy and confidence. In the E-Controlled paradigm, on the other hand, confidence and accuracy will typically increase as more information is provided and thus be a direct function of decision time.

Vickers et al. (1985a,b) subsequently performed a perceptual discrimination task that confirmed these predictions. In their experiment, subjects were presented with a successive sequence of horizontal lines extending varying amounts to the left and right of a vertical mid-line (representing zero difference). The task

required subjects to determine if the total sum of the extents on each trial was either positive (right) or negative (left). In the E-Controlled condition, the sequence was terminated after 6, 9, 12, 15, or 18 presentations, at which time subjects made a response and provided a confidence report (0-5) in the accuracy of their choice. In this condition, confidence and accuracy increased with number of observations (decision time). Alternatively, in the S-Controlled condition, the sequence of lines was terminated by the subject. Following the completion of the experiment, the number of observations was partitioned in order to correspond to the number of observations given in the E-Controlled condition. In this case, both confidence and accuracy steadily declined with number of observations. When taken together, the results clearly illustrated the critical roles that information quantity and quality play in the determination of confidence.

1.1.4 Semantic Congruity Effects. The inverse relationship between decision time and confidence in S-controlled tasks is also evident in a robust property of perceptual and symbolic comparison termed the "semantic congruity" effect. The semantic congruity effect, or "crossover" effect as it was originally dubbed by Audley and Wallis (1964), describes a response time interaction between the magnitude of the stimulus pair to be compared and the direction of the instruction (see Petrusic, 1991, for a recent review).

Recently, Petrusic & Baranski (1989a,b) showed that large semantic congruity effects can also be obtained with the confidence measure. For example, the "heavier" of two relatively heavy weights is selected more quickly and with greater confidence than the "lighter" of the same pair. Conversely, the "lighter" of two light weights is selected faster and with more confidence than the "heavier" of the pair (Petrusic & Baranski, 1989a). Since the Petrusic and Baranski (1989a) findings, semantic congruity effects have also been demonstrated with the confidence measure for visual proximity judgments (Petrusic, 1991; Petrusic & Baranski, 1989a,1989b) as well as with two-dimensional visual location judgments (Baranski & Petrusic, 1991b).

1.2 Theories of Confidence in Judgment and Discrimination

1.2.1 Early Theories. Early confidence theories were highly influenced by the fact that confidence consistently displayed an inverse relation to decision time. Rather than specifying the exact basis for this occurrence, however, several authors simply suggested and applied various unifying quantitative relations. For example, on the basis of his line-orientation discrimination data, Volkman (1937) proposed the simple inverse linear function,

$$t = a/(2c - 1) + b$$

where t is the decision time, a and b are constants, and c is the certainty with which the judgment is made.

Soon after, Johnson (1939), following upon his determination that the relation between decision time and confidence is seldom linear:

The functional relation then becomes: as doubt or uncertainty increases arithmetically, time of judgment increases geometrically (pp.45).

proposed the relation,

$$\log T = \log a + b(1-C)$$

where T is decision time, a and b are the slope and intercept constants, respectively, and C is confidence. After conditionalizing RT on various confidence levels and varying the parameters a and b, Johnson (1939) found relatively good fits to his line-length comparison data.

The major limitation of these early accounts, however, is that they are descriptive in nature; that is, they do not follow from any theoretical basis for the generation of decisional response time, let alone confidence. Nevertheless, the diligent research from which these properties were determined provided a comprehensive data-base for more elaborate quantitative theories.

1.2.2 The 'Runs' Model. In 1960, Audley proposed a stochastic counter model of human choice behaviour. The model was one of several decision theories that emerged in the early 1960's, attempting to expound the relations between decision time and accuracy [see also Stone (1960), Edwards (1965), and Laberge (1962)]. What distinguished Audley's model, however, was a clear recognition of the relationship between decision time and

confidence and, subsequently, the first attempt to formalize these properties within a quantitative process decision theory.

In the 'Runs' model, as in all modern decision theories relating decision time and accuracy, the overt (explicit) response is made only after a certain amount of 'implicit' evidence has been acquired. Audley (1960) assumed that this implicit evidence enters the sensory system over the course of the trial in the form of discrete units that favour the response $A > B$ or $B > A$; with each incoming unit of information consuming some small decision time component. Audley (1960) suggested that the overt response will be made when a criterion "run" (number) of implicit evidence favouring a particular response occurs in succession. This criterion is assumed to be internally set by the subject and is magnitude dependent on the experimental demands for speed versus accuracy. Hence, demands for speed generally lower the criterion, resulting in a more rapid achievement of the overt response but a greater likelihood of it being in error (i.e., the speed/accuracy tradeoff).

Audley (1960) proposed that confidence might be expressed by the relation,

$$C = 1/(V + 1)$$

where C is confidence expressed as a proportion and V denotes the number of "vacillations" of evidence accrual runs. Hence, if the criterion is 4, for example, the evidence accrual sequence AAAA (suggesting an easy comparison made quickly and with no competing evidence) has $V=0$ and thus $C=1$. The sequence ABABAAAA, on the

other hand (suggesting a longer, more difficult decision) has $V=4$ and thus $C=.20$. Thus, this simple view nicely captures the relation between decision time (in terms of judgment difficulty) and confidence.

The major limitation of Audley's (1960) model arises not directly from the formulation but rather as a consequence of the decisional theory in which it is embedded. Specifically, it has been shown that the 'runs' model will predict, under both speed and accuracy demands, equal mean correct and error response times (e.g., Vickers, Caudrey, & Willson, 1971; Audley & Pike, 1965; Vickers, 1979). The empirical evidence, on the other hand, indicates that error response times are typically longer than correct response times under conditions emphasizing accurate responding but comparable under speeded conditions (Henmon, 1911; Kellogg, 1931; Petrusic & Baranski, 1989b, Vickers, 1970; Laming, 1968; Link & Tindall, 1971; Swensson, 1972; Luce, 1986). Hence, since confidence is directly tied to response time, the runs model must also predict, for any level of discriminability, that confidence will be equal for correct and error responses (see Vickers, 1979). This contradicts considerable empirical evidence which has shown confidence for errors to be lower than for correct responses (e.g., Hertzman, 1937; Pierrel & Murray, 1963; Vickers & Packer, 1982; Vickers et al., 1985; Petrusic, 1991; Petrusic & Baranski, 1989a, 1989b).

Despite this shortcoming, however, the 'runs' stochastic choice model (as well as the others of it's time) set the stage

for more detailed and elaborate decision theories. It also uniquely illustrated the importance of modelling confidence in conjunction with decision time and accuracy in order to provide a full account of decision performance. In practice, however, the reporting of confidence data in psychophysical judgment tasks declined severely from the mid-sixties through to the mid-seventies; probably because the task of modelling the many complex interactions between response time and accuracy was formidable enough. However, this was soon to change quite dramatically.

1.2.3 The 'Balance of Evidence' Hypothesis. In 1970, Vickers proposed an Accumulator Theory of psychophysical judgment and discrimination. Unlike the 'runs' model (Audley, 1960), which assumed the accumulation of discrete unit increments of information over time, the accumulator theory assumes that the observer accumulates magnitudes of subjective stimulus differences (A-B)¹.

The accumulator theory (Vickers, 1970, 1979; Smith & Vickers, 1988) postulates two information stores. One store accumulates stimulus differences favoring the overt response $A > B$ while the other accumulates stimulus differences favouring $B > A$. A decision in favour of S_a or S_b occurs when one of the two stores first accumulates a criterion magnitude of stimulus difference information, k_a or k_b (where $k_a = k_b$ in the unbiased stimulus presentation case). As a basis for the generality and

applicability of the model, Vickers (1970, 1978, 1979, 1980) has shown, through extensive computer simulation, that the accumulator theory correctly predicts the various forms of the psychometric function with changes in discriminability (Vickers, 1979 fig. 32 p. 86) as well as the speed/accuracy trade-off relationship between error and correct response times under variations of k (Vickers, 1970, Fig.3, p. 45).

Vickers (1973, 1979) suggested that confidence is a function of the difference between the evidence total accumulated in favour of the successful response (t_a) and that accumulated in favour of the alternative response (t_b),

$$C(R_a) = t_a - t_b, \text{ where } t_a = k \text{ (criterion),}$$

and referred to this as the "balance of evidence hypothesis" for confidence. Subsequently, Vickers and Packer (1982) and Vickers et al. (1985) have provided an impressive array of empirical support for the computer simulation based predictions of the 'balance of evidence' hypothesis under variations of stimulus discriminability and relative demands for speed versus accuracy.

1.2.4 A Random-Walk Theory of Confidence. In 1984, Heath proposed a quantitative theory of confidence that was developed in the context of Link's (1975, 1990; Link & Heath, 1975) random-walk Relative Judgment Theory (RJT) of psychophysical discrimination. The RJT is similar to the Vickers (1979) accumulator model in that they both assume the accumulation of continuous, rather than discrete, stimulus difference information. The major difference

is that RJT assumes a single information store that totals both positive ($a-b$) and negative ($b-a$) difference magnitudes. The overt response $A > B$ or $B > A$ occurs when the accumulated evidence total reaches or exceeds the positive ($+A$) or negative ($-B$) response boundary, respectively; where the values A and $-B$ are determined by the relative demands for speed versus accuracy. Two other important parameters of the model are C , the response bias parameter, reflecting the discrepancy in the starting position of the random-walk and θ , the discriminability parameter, proportional to the overt stimulus difference.

Heath (1984) suggested that confidence will be proportional to $\theta(A-C)$ (i.e., a function of stimulus discriminability and directly proportional to the distance travelled by the random-walk), and provided support for the model by successfully fitting the data of Asher (1974). Heath (1984) concluded that his formulation was a viable alternative to the 'balance of evidence' hypothesis because it was mathematically tractable and did not rely on computer simulation for its derivations.

Vickers (1984) and Vickers and Smith (1985) responded by citing clear limitations to Heath's (1984) formulation, especially under conditions of biased stimulus presentation. For example, on Heath's (1984) view, the effect of biasing a particular response alternative (i.e., increasing C) would be to reduce the required distance the walk must travel and thus reduce confidence for that response. However, Vickers (1984) showed that in fact the opposite result is obtained; biasing the probability

of a particular response leads to an increase in the confidence for that response. Importantly, this result confirmed the simulated predictions of the balance of evidence hypothesis upon increases in k_a relative to k_b .

Perhaps a more serious limitation of the Heath (1984) formulation is the inability of the model to account for trial-to-trial variations in confidence under a fixed level of discriminability. Specifically, since θ , C , and A are fixed model parameters, the model must predict only one level of confidence for each stimulus difference relation. However, as Henmon's (1911) classic result showed, the same stimulus difference can be judged quickly and with high confidence on some trials but slowly and with low confidence on others.

In sum, the Balance of Evidence hypothesis has proven to provide a comprehensive account of the major properties of confidence judgments (see Vickers, 1979; Vickers & Packer, 1982; Vickers et al., 1985). In addition, Smith and Vickers (1988) have recently overcome the criticism (Heath, 1984) of relying on computer simulation for the predictions of the model by developing a mathematically tractable version of the balance of evidence hypothesis.

(II) THE CALIBRATION OF CONFIDENCE JUDGMENTS

2.1 Introduction

The present chapter focuses on the relationship between confidence and decision accuracy. In particular, the strong correlation between these two measures has already been noted. However, in early psychophysical studies the exact correspondence between confidence and accuracy could only be determined at the extremes of the confidence continuum; that is at "guess" and at "certain". Nevertheless, two consistent findings were that, (1) people were more often right than wrong when they said they guessed and, (2) people were not always 100% correct when they thought they were (e.g., Pierce & Jastrow, 1884; Fullerton & Cattell, 1892; Henmon, 1911; Garrett, 1922; Johnson, 1939).

In 1956, Adams suggested that it would be desirable to obtain a more detailed analysis of the relationship between confidence and accuracy; a 'realism of confidence judgments' as he later put it (Adams & Adams, 1961). He proposed that such an analysis could be obtained if confidence were to be, like accuracy, expressed in terms of a percentage. Hence, by determining performance accuracy conditional on various confidence levels [e.g., 50% (guess), 60%, 70%, ..., 100% (certain)], the exact correspondence between the measures could be obtained for the full range of confidence values.

Although Adams (1956) did not define it so, the study of this exact relationship between confidence and accuracy has become known as the "calibration" problem in modern-day decision

research. Specifically, a person or group is considered 'well-calibrated' at a specific task if, in the long run, their subjective estimates of performance closely correspond to their overt performance. Alternatively, a person or group is said to be 'over-confident' if subjective performance exceeds actual accuracy and 'under-confident' if the reverse is true. Figure 1 provides hypothetical calibration curves that illustrate the three basic phenomena of over-confidence, under-confidence, and perfect calibration.

To date, three indices of the relationship between confidence and accuracy have typically been reported in the calibration literature. The first is simply the difference between overall percent correct and overall percent confidence on a given task; which provides a global index of over- or under-confidence (Lichtenstein & Fischhoff, 1977).

The second, referred to as the calibration score (Lichtenstein & Fischhoff, 1977), is obtained by taking the weighted sum of squared deviations between the proportion correct obtained at each confidence level and the proportion correct expected on the basis of perfect calibration. Hence the calibration score, for half-range (i.e., 50%-100% confidence range) calibration studies, will vary between 0 (perfect calibration) and 1 (no calibration).

Finally, the resolution score (Murphy, 1971) provides an index of how close the obtained calibration curve lies to the main diagonal. That is, it is a measure of how well subjects can

Figure 1. Hypothetical calibration curves illustrating the phenomena of overconfidence, under-confidence, and perfect calibration.

assign a maximally different percent correct to each confidence category. The resolution score is obtained by taking the weighted sum of squared deviations between the proportion correct for each confidence category and the overall mean proportion correct obtained in the task. Hence, the resolution score varies between 0 (no resolution) and 1 (unit slope calibration curve).

2.2 Properties of Confidence Calibration

2.2.1 Overwhelming Over-Confidence. As mentioned, the first 'calibration' study was reported by Adams (1956). In that task, subjects viewed 40 words at 10 varying intensities of illumination. Each presentation lasted .25 sec and following each trial subjects wrote down the word they thought they saw and provided a confidence measure from 0-100 (in decades). Zero confidence was to imply certainty of an error and 100 certainty of a correct response. Adams (1956) found that people were consistently under-confident on this perceptual word-recognition task.

Soon after, Adams and Adams (1961) investigated the confidence-accuracy relationship in a task involving two-choice general knowledge (almanac) questions spanning a wide range of topics. Adams and Adams (1961) found good calibration for a psychologist but severe over-confidence for a hospitalized schizophrenic claiming to be Jesus Christ. Rather than being a general principle of dementia, however, Adams and Adams (1961), in a second experiment, found marked over-confidence for a group

of under-graduate students performing a nonsense syllable recall task.

Since the Adams and Adams (1961) findings, the phenomenon of over-confidence has been reported in numerous tasks. Nickerson and McGoldrick (1963) found over-confidence for judgments of the size of the U.S. states. Oscamp (1965) found that psychologists were grossly over-confident in their assessments of clinical case studies. Howell (1971) found that people were extremely over-confident for both internal (dart-throwing ability) and external (roulette) subjective probabilities of success. Similarly, many recent calibration studies, focussing almost exclusively on the assessment of performance on intellectual knowledge questions, have consistently reported over-confidence (e.g., Fischhoff, Slovic, & Lichtenstein, 1977; Koriatic, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980a; see Lichtenstein, Fischhoff, & Phillips, 1982 for a comprehensive review).

2.2.2 Effect of Decision Difficulty. A consistent finding in the latter studies was that people seemed to display the most extreme over-confidence on very difficult tasks (e.g., Nickerson & McGoldrick, 1965; Lichtenstein & Fischhoff, 1977). For example, in a task requiring discrimination between drawings made by European or Asian children, university students were about 70% confident while their actual correct probabilities were near .5 (Lichtenstein & Fischhoff, 1977). Importantly, Lichtenstein and Fischhoff (1977) also showed that over-confidence is reduced as

tasks get easier. In their fifth experiment, they classified a large set of general knowledge questions, on the basis of post-hoc analyses, according to whether they were easy or difficult. In what is now a classic result, their calibration analysis revealed over-confidence on difficult questions ($P_c=.60$) but slight under-confidence on easy ones ($P_c=.80$).

2.2.3 Effects of Training or Expertise. Since the finding of over-confidence has historically been so predominant and robust, a wealth of recent research has been concerned with its reduction. Two foci have been the study of task specific expertise and the provision of various corrective techniques such as training with feedback (Lichtenstein & Fischhoff, 1980a; Sharpe et al., 1988) or induced change in decisional strategy (e.g. Koriat et al, 1980).

Expertise has been shown to provide better calibration performance in, for example, auditors (Tomassini, Solomon, Romney, & Krogstad, 1982), calibration experts (Lichtenstein & Fischhoff, 1980b), and bridge players (Keren, 1987). On the other hand, medical doctors (DeSmet, Fryback, & Thornbury, 1977; Christensen-Szalanski & Bushyhead, 1981), psychology students evaluating their own exam performance (Sieber, 1974), clinical psychologists (Oscamp, 1962), and statistical experts calibrating probability assessments (Wagenaar & Keren, 1985) have not fared as well. Of all the experts studied, weather forecasters are the best calibrated (Murphy & Winkler, 1977). Of course, this is

based on years of recorded feedback on weather system precipitation probability outcomes.

Importantly, however, feedback alone does not always facilitate calibration performance. Choo (1976; cited in Lichtenstein et al., 1982), for example, found no effect of feedback. Sharp, Cutler, and Penrod (1988) found that sessional feedback improved resolution but not calibration on questions of general knowledge. Conversely, Lichtenstein and Fischhoff (1980a) found marked improvement when extensive feedback about calibration was provided after each session and Arkes, Christensen, Lai, and Blumer (1987) showed that even limited, but contextually specific, feedback can effectively reduce overconfidence.

Despite these important expertise and feedback distinctions, researchers still consistently reported varying degrees of overconfidence on intellectual decisions. Strangely though, research into the calibration problem has until relatively recently been almost exclusively concerned with calibration in intellectual knowledge tasks.

2.2.4 Factual Knowledge versus Perception. In 1980, Dawes asked, "Where would we find the reverse? (p.328)" That is, where would we find under-confidence? Dawes (1974) had proposed earlier that mis-calibration, in the form of under-confidence, might be found for perceptual judgments². Subsequently, Dawes (1980) attempted to provide this distinction between over- and under-confidence in

a study of both intellectual and perceptual tasks. Although Dawes (1980) was able to find over-confidence on an intellectual knowledge task (Experiment 2), his attempts involving perceptual comparisons were inconclusive. For example, in his third and fourth experiments, he found under- and over-confidence, respectively, on the same area judgment task differing only in the chosen population (college professors for Experiment 3 and university students for Experiment 4).

Some years later, Keren (1988) re-addressed the issue and found over-confidence on an intellectual knowledge task but not in an experiment involving Landolt-C visual acuity judgments or two others involving letter detection. On the basis of these findings, Keren (1988) concluded that perceptual tasks are indeed less likely to show overconfidence than intellectual knowledge tasks.

More recently, Baranski and Petrusic (1988, 1991a) and Petrusic and Baranski (1991a) found that, as with general knowledge questions, perceptual judgments exhibit both under- and over-confidence depending on the difficulty of the judgments. In a relatively easy visual proximity comparison task, Baranski and Petrusic (1991a) found global under-confidence. However, the global difficulty of the task was comprised of 4 specific difficulty levels ranging from almost errorless performance ($P_C = 96\%$) to relatively difficult judgments ($P_C = 72\%$). As is evident in Figure 2, Baranski and Petrusic (1991a) found, under both speed and accuracy emphasized conditions, clear under-confidence

Figure 2. Calibration curves and response frequencies for the four levels of difficulty in Baranski and Petrusic (1991a): Level 1 ($r=1.10$) circles, Level 2 ($r=1.08$) diamonds, Level 3 ($r=1.05$) inverted triangles, and Level 4 ($r=1.03$) triangles (r is the ratio of the longer to the shorter extent of the comparison pair). The top panels provide the data for subjects under accuracy stress while the lower panels provide the data for subjects who were shifted from accuracy to speed stress.

for the easiest judgments and a systematic trend towards over-confidence as difficulty increased. This result also reconciled Dawes' (1980) anomalous result since a re-examination of his Experiments 3 and 4, will show that the college professors were simply more accurate than the college students on the same task.

Baranski and Petrusic (1991a) concluded that the usual finding of over-confidence in intellectual knowledge tasks was likely due to the consistent selection of relatively difficult questions (evident in the very low accuracy rate typically reported) and that perceptual and non-perceptual confidence calibration performance may be regulated by a similar (if not identical) process of converting a feeling of certainty into a numerical probability estimate.

2.2.5 Confidence Calibration and the Time to Choose. As previously mentioned, Henmon (1911), Volkmann (1934), and Johnson (1939) provided the classic demonstrations of an inverse relationship between decision time and confidence. Henmon's (1911) study was particularly interesting because the task involved comparison of but a single pair of line-lengths of intermediate difficulty.

Importantly, Baranski and Petrusic (unpublished data) found a negative correlation between decision time and confidence in a general knowledge calibration task in which questions and alternatives were presented simultaneously on an video screen and response times were collected without the subjects' knowledge.

More recently, Baranski and Petrusic (1991a), in their visual proximity comparison task, extended Henmon's (1911) investigation of the relationship between confidence and response time in three important ways (see Figure 3). First, they showed that, as with the calibration of probabilities, there is a clear "difficulty effect" on response time when response times are conditionalized on the various confidence categories. Second, they found that this difficulty effect disappears when subjects are required to respond under speed stress. Finally, they showed that response times conditionalized on confidence levels preserve the general relationship between error and correct response times under speed and accuracy conditions. Namely, under accuracy, error times exceed correct times for each level of confidence while under speeded responding error and correct times are comparable.

When taken together, the findings established important constraints on theories of confidence calibration while pointing out the need for theories that are general enough to permit predictions for confidence, accuracy, and the time to decide.

Figure 3. Mean correct (filled circles) and error (open circles) response times as a function of confidence level for the four levels of comparative difficulty in Baranski and Petrusic (1991a): Level 1 ($r=1.10$) circles, Level 2 ($r=1.08$) diamonds, Level 3 ($r=1.05$) inverted triangles, and Level 4 ($r=1.03$) triangles (r is the ratio of the longer to the shorter extent of the comparison pair).

2.3 Models of Confidence Calibration

2.3.1 Heuristics and Biases. To date, several accounts of confidence calibration have focused on the phenomenon of over-confidence in general knowledge tasks. Tversky and Kahneman (1974), for example, suggested that people adopt simplifying rules or heuristics when assessing probabilistic outcomes. One such heuristic, called "anchoring and adjustment", states that people will rely too heavily on a very narrow range of subjective probabilities (i.e., the anchor). Lichtenstein et al. (1982) suggested that over-confidence might be due to an insufficient adjustment of probability estimates from this anchor.

Alternatively, Koriat et al. (1980) suggested a three-stage model of intellectual knowledge probability assessment. In the first stage the person searches memory for relevant information. Next, the person assesses the information to arrive at a feeling of certainty or doubt. In the final stage the person translates that feeling into a number. They suggested that over-confidence occurs because people tend strategically to select the 'available' information in memory that supports their hypothesis (i.e., the "availability heuristic"). In support of their position, they showed that over-confidence can be reduced when subjects are explicitly instructed to "make-up" arguments both for and against their original choice.

The problem with these heuristic views of calibration is, as mentioned, that they were specifically designed to account for over-confidence and thus do not address the critical effect of

judgment difficulty on calibration and how this leads to over- or under-confidence. More recently, Poulton (1989) proposed that mis-calibration results from a more general and pervasive estimation bias, termed response contraction, which is a tendency to overestimate the difficulty of easy judgments and underestimate the difficulty of more difficult ones. Although this may be true, Poulton (1989) offers no quantitative process theory of how this might arise.

2.3.2. May's Calibration Model. May (1986) has recently proposed an interesting view of calibration that accounts for both over- and under-confidence. Working in the context of judgments of relative "northness" or "southness" of world cities, May (1986) suggested that confidence will be a monotone function of the subjective difference between the two response alternatives; much like an internal psychophysical comparison. Hence, good calibration is predicted when the difference between the subjective and objective, or actual, scales is small and mis-calibration results when these differences are large. For example, when the relative locations of cities are reversed in memory (e.g., Rome is thought to be south of New York), subjects will show over-confidence because the subjective estimate will be based on a positive subjective difference for the incorrect response.

There are many positive features to this model, including the notion of an internal comparison process for remembered

magnitudes (see Petrusic & Baranski, 1991b; Baranski & Petrusic, 1991c, for recent reviews of the memory comparison literature) and implications concerning the effects of memory distortions on subjective estimations. However, as of yet, the model is not developed in sufficient detail to permit quantitative analyses and parameter estimation. In addition, it is not immediately clear how serious mis-calibration, in the form of over- or under-confidence, can arise when the difference between the physical and subjective stimulus scales are not subject to memory distortions, as in the case of direct perceptual judgments.

2.3.3 The Calibration Model of Ferrell and McGoey (1980). The Ferrell and McGoey (1980) calibration model consists of two parts: (1) an attempt by the observer to detect whether or not the response was correct and, (2) an assignment of a numerical probability value on the basis of (1).

The detection part of the task is described by signal detection theory (e.g, Green & Swets, 1966; Tanner & Swets, 1954). Hence, in the two-alternative forced choice (2AFC) task, response alternatives A and B are assumed to be represented by normal distributions, with common probability densities, along a continuum of affective decision strength. On each trial, a value x , of a scalar decision variable, X , is generated from a joint consideration of the stimulus and response. The value X , which can be considered as "subjective certainty", is a function of the difference between the means of the distributions corresponding

to the two response alternatives ($X = Y_1 - Y_2$) and thus the overt probability of a correct response $p(C)$ or the discriminability parameter, d' , in the signal detection sense. As in traditional signal detection theory, a cutoff value on X determines whether the choice was correct or not.

The numerical probability estimate is determined by partitioning the range of the decision variable, X , into m intervals with a set of cutoff values $\{x_i\}$. The set of m allowable probability responses, $\{r_i\}$, determined by the experimental conditions, is mapped onto the intervals, with higher values of r going to higher values of X . The subject then gives the confidence response r that corresponds to the interval into which the observed value x falls.

Importantly, the stimulus distributional properties of Signal Detection theory are general enough to characterize both perceptual and non-perceptual stimulus difference relations. Hence, the property of discriminability (d') mediated calibration can be easily extended to perceptual stimuli in addition to the general knowledge based stimulus representation for which the theory was developed.

However, the inverse relation between decision time and level of confidence is not easily accounted for by the Ferrell and McGoeys (1980) model; at least as it currently formulated. Specifically, because the model is theoretically conceived in the context of single-sample signal detection theory, and thus assumes only one "ideal observation" per trial (Tanner & Swets,

1954), the model predicts that decision times will be the same for all confidence categories (see Pike, 1973; Vickers, 1979).

This general limitation of signal detection theory can be overcome by simply assuming the existence of a latency function (Pike, 1973; Thomas, 1971) for the detection process in which response times decrease as a function of the distance of the sampled observation from the decision criterion. This view, commonly referred to as strength theory (e.g., Norman & Wickelgren, 1969; Wickelgren, 1968; but see Coombs, 1964, pp.530 for an earlier statement), when applied in the context of the Ferrell and McGoey (1980) model, predicts that response times will monotonically decrease as discriminability, and thus confidence, increases. In addition, because errors are typically associated with more difficult judgments, the model also predicts that error times will be longer than for the corresponding correct responses. Hence, with these auxiliary assumptions, the strength theory version of the Ferrell and McGoey (1980) model provides a qualitative account of the response time data under accuracy stress.

However, strength theory is still likely to fail in the long run. First, Murdock and Dufty (1972) showed that, contrary to empirical results, strength theory predicts smaller response time variability for errors than for correct responses. In addition, strength theory predicts that error times will always be longer than correct times (e.g., Coombs, 1964; Pike, 1973; Petrusic & Jamieson, 1978; Vickers, 1979). However, as previously seen,

Baranski and Petrusic (1991a) showed that correct and error times will be identical under conditions emphasizing speeded responding. Hence, although the strength theory version of the Ferrell and McGoey (1980) model provides a good account "accuracy" calibration data, it seems that a decision model that permits error times to exceed correct times under accuracy but not under speed, is required in order to provide a full account of RTs conditional on various confidence levels.

(III). INFORMATION SCALING MODELS OF CONFIDENCE CALIBRATION

3.1 Introduction

The present chapter introduces three new quantitative models of confidence calibration in human judgment and discrimination. In keeping with the research discussed thus far, the models will deal exclusively with confidence for decisions that have already occurred³.

The models are developed in the context of mathematical decision theory for purposes of tractability and application. However, the basis for confidence is more generally rooted in the notion that confidence depends on both the quantity and quality of information available to the decision maker. Specifically, for each two-choice decision there may exist evidence or information that favors the overt response, the alternative response, or neither response. This information may enter the system implicitly (as is assumed in perceptual discrimination models), or through the selective retrieval of information in long-term memory (as in the case of judgments based on verbal memory). How confident we are will ultimately depend on the relative proportion of available information. I now discuss how this all develops in relation to the time it takes to make decisions in general.

The pre-requisite for a theory of confidence is, of course, a primary decision theory. Hence, as a theoretical basis for the calibration models, I chose the Slow and Fast Guess Evidence Accrual Theory of choice behaviour (Petrusic, 1991; Petrusic &

Baranski, 1989b; Petrusic & Jamieson, 1978) which has been shown to provide a comprehensive account of the relations between decision time and accuracy under speed and accuracy emphasised conditions (i.e., monotone and non-monotone LPFs; see Petrusic, 1991; Petrusic & Baranski, 1989b). The calibration models will be developed in this context and will be described in detail following a review of the primary decision theory.

3.2 The Slow and Fast Guessing Theory

According to the slow and fast guessing theory, information about a particular stimulus configuration is assumed to accumulate over time in the form of a series of discrete, implicit, pre-decisional events, formally identical to sampling in sequential statistical decision making. Specifically, three classes of pre-decisional events are postulated:

(1) a_{rb} , an event whose occurrence favours the overt response a_{rb} (where a_{rb} denotes, for example, the response that stimulus A has greater magnitude than stimulus B). Let p_1 denote the probability of the evidence accrual event, a_{rb} .

(2) b_{ra} , an event whose occurrence favours the overt response b_{ra} . Let this event have probability p_2 .

(3) a_{ib} , an event whose occurrence favours neither a_{rb} nor b_{ra} but, rather, denotes a failure to resolve the stimulus configuration. Typically, a_{ib} events favour the overt indifference response a_{ib} , and subjectively corresponds to a complete lack of confidence in the correctness of the response.

The event $a_{\underline{I}b}$ has probability $p_3 = 1 - p_1 - p_2$.

The overt response occurring on a particular trial depends on which of three preset counters first achieves a criterion number of pre-decisional events. Let α_1 , α_2 , and α_3 denote the criterion number of events required for the overt responses $a_{\underline{R}b}$, $b_{\underline{R}a}$, and $a_{\underline{I}b}$, respectively. When the overt indifference response, $a_{\underline{I}b}$, is denied (e.g., in the two-alternative forced choice paradigm), it is assumed that observers guess from the terminal state $a_{\underline{I}b}$, responding $a_{\underline{R}b}$ with probability g and $b_{\underline{R}a}$ with probability $1-g$. Hence, the probability of the overt response, $a_{\underline{R}b}$, is given by the equation:

$$(1) \Pr(a_{\underline{R}b}) = \sum_{s=0}^{\alpha_2-1} \sum_{t=0}^{\alpha_3-1} (\alpha_1 + s + t - 1)! \frac{p_1^s p_2^t p_3^{\alpha_1-1}}{(\alpha_1 - 1)! s! t!} \\ + g \sum_{s=0}^{\alpha_1-1} \sum_{t=0}^{\alpha_2-1} (\alpha_3 + s + t - 1)! \frac{p_1^s p_2^t p_3^{\alpha_3-1}}{(\alpha_3 - 1)! s! t!}$$

where s and t are the number of evidence accrual events in counters favouring responses $b_{\underline{R}a}$ and $a_{\underline{I}b}$, respectively.

The expected number of evidence accrual events preceding an overt $a_{\underline{R}b}$ response (i.e., the expected number of samples prior to an overt response), conditional on the occurrence of an $a_{\underline{R}b}$ response, is given by the equation:

$$(2) E(N|a_{\underline{R}b}) = \left[\sum_{s=0}^{\alpha_2-1} \sum_{t=0}^{\alpha_3-1} (\alpha_1 + s + t) \frac{p_1^s p_2^t p_3^{\alpha_1-1}}{(\alpha_1 - 1)! s! t!} \right. \\ \left. + g \sum_{s=0}^{\alpha_1-1} \sum_{t=0}^{\alpha_2-1} (\alpha_3 + s + t) \frac{p_1^s p_2^t p_3^{\alpha_3-1}}{(\alpha_3 - 1)! s! t!} \right] / \Pr(a_{\underline{R}b})$$

and it is assumed that the observed decisional response time,

$T(aRb)$, is monotonically related to the average number of evidence accrual events preceding a particular response; i.e.,

$$(3) \quad T(aRb) = A * E(N | aRb) + B,$$

where A and B are non-negative constants (A = rate of evidence accumulation, B = stimulus encoding and motor execution time).

All applications of the model will assume unbiased stimulus presentation and guessing probabilities. This will be theoretically denoted by equal settings for the discriminant decisional criterion parameters ($\alpha_1 = \alpha_2$) and guessing probability ($g = 1 - g = .5$), respectively. One final parameter requires elaboration. This is the discriminability parameter, $\delta = p_1 / (p_1 + p_2)$, which monotonically increases with stimulus discriminability and is constrained between .5 (no stimulus difference) and 1.0 (errorless comparison).

3.3 Information Scaling Theories of Confidence Calibration

Three models of confidence calibration were considered in the context of the slow and fast guessing theory. Although these models are based on the same decision theory they provide distinct interpretations of how confidence is determined as a consequence of alternative rules for categorically scaling the evidence accrual information into confidence categories.

3.3.1 Scaling Competing Information. The first view of confidence calibration is closely related to the successful 'balance of evidence' hypothesis of confidence developed by Vickers (1979) and his associates (Vickers & Packer, 1982; Vickers et al, 1985) and discussed earlier. Recall that on the balance of evidence view, confidence is proportional to the difference between the accumulated evidence favoring the dominant (correct) and non-dominant (incorrect) responses. When the accumulated difference is large, confidence will be high. Conversely, when the accumulated evidence difference is small, confidence will be low.

In the context of the slow and fast guessing theory, this notion translates to a categorical scaling of the amount of competing information. Psychologically, the model implies that judgments will be made with high confidence when there is little or no information competing with the dominant response. Lower confidence, on the other hand, results when there is more competing information, or conflict between alternatives.

For purposes of illustration, let the decisional criterion values ($\alpha_1=\alpha_2=\alpha_3$) equal 3. Hence, three instances of implicit evidence must be accrued, favoring a particular response, in order to elicit the overt response. Note that a variety of response terminating evidence accrual configurations are possible on the basis of these decisional parameters. That is, the decision process can terminate from

the discriminant state with evidence totals (x,y,z) equal to: 3,0,0; 3,0,1; 3,0,2; 3,1,0; 3,1,1; 3,1,2; 3,2,0; 3,2,1; and 3,2,2 (where x,y , and z denote the number of counts accrued in the dominant, non-dominant, and indiscriminant accumulators, respectively). When the decision process terminates from the indiscriminant state (i.e., the criterion α_3 is reached and the overt response is based on guessing), the above set of accrual configurations is obtained but with x and z interchanged.

More generally, the decisional criterion values $(\alpha_1, \alpha_2, \alpha_3)$ give rise to a finite set of response terminating evidence accrual configurations $\{x_i, y_i, z_i\}$ for the response aRb , where $x_i = \alpha_1$, $0 \leq y_i \leq (\alpha_2 - 1)$, and $0 \leq z_i \leq (\alpha_3 - 1)$. In addition, when the overt response is based on guessing the set of accrual configurations are denoted by $0 \leq x_i \leq (\alpha_1 - 1)$, $0 \leq y_i \leq (\alpha_2 - 1)$, and $z_i = \alpha_3$.

On the Competing Information Scaling model, confidence is determined by categorically scaling these evidence accrual configurations into the available confidence categories. Instances in which there is no competing evidence (e.g., 3,0,0) will be scaled into the highest permissible confidence category. On the other hand, instances where there is much competing evidence (e.g., 3,2,2) will be classified into a low confidence category. Of course, the number and type of categories required are determined by the experimental conditions and the partitions used to categorize these

configurations are subject to individual differences.

As mentioned, two types of information can reduce confidence: competing evidence (i.e., $y > 0$) and inconclusive evidence (i.e., $z > 0$). However, on this model the basis for partitioning, or scaling, evidence into confidence categories is assumed to follow a lexicographic ordering of the values y and z in which information that directly competes (y) with the overt choice will influence confidence to a greater degree than inconclusive information (z). More generally, the confidence (C) associated with evidence configuration values (x, y, z) and (x', y', z') will satisfy the relation: $C(x, y, z) \geq C'(x', y', z')$, where $\{x=x'=\alpha_1\}$, $\{0 \leq y \leq (\alpha_2-1)\}$, and $\{0 \leq z \leq (\alpha_3-1)\}$, if and only if $y \geq y'$ and $z \geq z'$.

This means that the configuration 3,2,1, for example, cannot result in a higher confidence than the configuration 3,1,1 and the configuration 3,1,2, for example, cannot result in a higher confidence value than the configuration 3,1,1.

In the first application of the model an attempt was made to simultaneously fit the calibration and response frequency curves associated with the four levels of difficulty in the "accuracy" data of Baranski and Petrusic (1991a) described earlier. This is possible because the closed form equations 1 and 2 (above) provide the expected frequency ($f(x, y, z)$) and proportion correct ($p(x, y, z)$) for every possible response terminating evidence accrual configuration.

The parameter space for the model is based on four decisional parameters: the discriminant state accrual parameters ($\alpha_1 = \alpha_2$), the indiscriminant state accrual parameter (α_3), and the discriminability parameter, δ . In addition, there are two guessing parameters: P_3 , the probability of obtaining an inconclusive evidence event on any given sample, and g , the guessing accuracy parameter, which is assumed to equal .5. Finally, there are $N-1$ scaling parameters for partitioning the evidence accrual configurations into N confidence categories. However, in modelling the calibration difficulty effect, two severe parameter constraints were imposed on the model. First, because the four difficulty levels were intermixed in the experiment, it was assumed that subjects maintained stable decisional criteria and guessing probabilities for each difficulty level (see Petrusic & Baranski, 1989b). That is, the parameters $\alpha_1, \alpha_2, \alpha_3$, and P_3 were fixed for all difficulty levels; only the discriminability parameter, δ , was permitted to vary. Second, the 5 scaling parameters used to partition the evidence accrual configurations into the 6 confidence categories were fixed over all difficulty levels. That is, it was assumed that the evidence accrual configuration x_i, y_j, z_k would be assigned the same confidence regardless of the difficulty of the judgment from which it derived. Hence, once the configuration of decisional, guessing, and scaling parameters are chosen, the model's predictions for

calibration, response frequency, and a priori variations in difficulty, are based on variations of only 1 parameter, δ .

A parameter search was conducted for a simultaneous best fit to the four calibration curves and their associated response frequency curves⁴. The result, with $\alpha_1=\alpha_2=\alpha_3=5$ and δ values equal to .60 (most difficult), .70, .76, and .78 (easiest), is presented in Figure 4. Considering the constrained parameter conditions, the competing information scaling model provides a very impressive modelling of the calibration difficulty effect. In addition, the model also correctly predicts the pattern of response frequencies associated with each confidence category. As discriminability increases there is a progressive increase in the frequency of certain responses and a decrease in guess responses (see Appendix A and Table 1 for a detailed view of how the model's calibration and response frequency curves were obtained).

Several aspects of the model and data require additional discussion. Notably, an examination of Table 1 will show that the model predicts overconfidence for each difficulty level at 100% confidence because of an incorporation of competing and uncertain evidence accrual configurations into that category. That is, evidence configurations other than 5,0,0 (true subjective certainty) are tolerated under the classification of certainty. This results in severe overconfidence for the most difficult level of comparison ($\delta=.60$) because the same evidence configurations have lower

Figure 4. Calibration curves and relative response frequencies predicted by the Competing Information Scaling Model under accuracy for the 4 levels of comparative difficulty in the data of Baranski and Petrusic (1991a): Level 1 ($\delta=.78$, easiest) circles, Level 2 ($\delta=.76$) diamonds, Level 3 ($\delta=.70$) inverted triangles, and Level 4 ($\delta=.60$, most difficult) triangles (δ is the discriminability parameter).

Table 1

Parameter Estimates for the Competing Information Scaling Model Fit to the Data of Baranski & Petrusic (1991):
 $\alpha_1=\alpha_2=\alpha_3=5$, $\delta=.60, .70, .76$, and $.78$

$\delta=.60$	C_C	y	z	$f(x,y,z)$	$p(x,y,z)$	F_C	P_C
	100	0	0	5.20E-2	.883636	.268	.842
		0	1	2.60E-2	.883636		
		0	2	7.79E-3	.883636		
		0	3	1.82E-3	.883636		
		0	4	3.64E-4	.883636		
		1	0	9.90E-2	.835052		
		1	1	5.94E-2	.835052		
		1	2	2.08E-2	.835052		
	90	1	3	5.54E-3	.835052	.122	.778
		1	4	1.25E-3	.835052		
		2	0	.11571	.771429		
	80	2	1	8.10E-2	.771429	.081	.771
	70	2	2	3.24E-2	.771429	.152	.719
		2	3	9.72E-3	.771429		
		2	4	2.43E-3	.771429		
		3	0	.10830	.692308		
	60	3	1	8.66E-2	.692308	.142	.692
		3	2	3.90E-2	.692308		
		3	3	1.30E-2	.692308		
		3	4	3.57E-3	.692308		
	50	4	0	8.99E-2	.600000	.234	.599
		4	1	8.10E-2	.600000		
		4	2	4.05E-2	.600000		
		4	3	1.48E-2	.600000		
		4	4	4.45E-3	.600000		
	pure guessing			3.64E-3	.500000		

Table 1 continued

$\delta=.70$	C_C	y	z	$f(x,y,z)$	$p(x,y,z)$	F_C	P_C
	100	0	0	.100679	.985748	.421	.974
		0	1	5.03E-2	.985748		
		0	2	1.51E-2	.985748		
		0	3	3.52E-3	.985748		
		0	4	7.04E-4	.985748		
		1	0	.13850	.967365		
		1	1	8.31E-2	.967365		
		1	2	2.91E-2	.967365		
<hr/>							
	90	1	3	7.76E-3	.967365	.127	.930
		1	4	1.75E-3	.967365		
		2	0	.11707	.927027		
<hr/>							
	80	2	1	8.19E-2	.927027	.082	.927
<hr/>							
	70	2	2	3.28E-2	.927027	.126	.874
		2	3	9.83E-3	.927027		
		2	4	2.46E-3	.927027		
		3	0	8.09E-2	.844828		
<hr/>							
	60	3	1	6.47E-2	.844828	.106	.845
		3	2	2.91E-2	.844828		
		3	3	9.71E-3	.844828		
		3	4	2.67E-3	.844828		
<hr/>							
	50	4	0	5.27E-2	.700000	.138	.696
		4	1	4.75E-2	.700000		
		4	2	2.37E-2	.700000		
		4	3	8.70E-3	.700000		
		4	4	2.61E-3	.700000		
	pure guessing			2.94E-3	.500000		

Table 1 continued

$\delta=.76$	C_C	Y	Z	$f(x,y,z)$	$p(x,y,z)$	F_C	P_C
100	0	0	0	.15019	.996869	.550	.993
	0	1		7.51E-2	.996869		
	0	2		2.25E-2	.996869		
	0	3		5.26E-3	.996869		
	0	4		1.05E-3	.996869		
	1	0		.16331	.990153		
	1	1		9.80E-2	.990153		
	1	2		3.43E-2	.990153		
90	1	3		9.15E-3	.990153	.120	.974
	1	4		2.06E-3	.990153		
	2	0		.10808	.969470		
80	2	1		7.57E-2	.969470	.076	.969
70	2	2		3.03E-2	.969470	.100	.935
	2	3		9.08E-3	.969470		
	2	4		2.27E-3	.969470		
	3	0		5.81E-2	.909320		
60	3	1		4.65E-2	.909320	.076	.909
	3	2		2.09E-2	.909320		
	3	3		6.97E-3	.909320		
	3	4		1.92E-3	.909320		
50	4	0		3.00E-2	.760000	.079	.754
	4	1		2.70E-2	.760000		
	4	2		1.35E-2	.760000		
	4	3		4.95E-3	.760000		
	4	4		1.49E-3	.760000		
pure guessing				2.44E-3	.500000		

Table 1 continued

<u>$\delta=.78$</u>	<u>C_C</u>	<u>y</u>	<u>z</u>	<u>$f(x,y,z)$</u>	<u>$p(x,y,z)$</u>	<u>F_C</u>	<u>P_C</u>
	100	0	0	.17079	.998218	.596	.996
		0	1	8.54E-2	.998218		
		0	2	2.56E-2	.998218		
		0	3	5.98E-3	.998218		
		0	4	1.20E-3	.998218		
		1	0	.16985	.993711		
		1	1	.10191	.993711		
		1	2	3.57E-2	.993711		
<hr/>							
	90	1	3	9.51E-3	.993711	.114	.982
		1	4	2.14E-3	.993711		
		2	0	.10251	.978054		
<hr/>							
	80	2	1	7.18E-2	.978054	.072	.978
<hr/>							
	70	2	2	2.87E-2	.978054	.090	.948
		2	3	8.61E-3	.978054		
		2	4	2.15E-3	.978054		
		3	0	5.00E-2	.926309		
<hr/>							
	60	3	1	4.00E-2	.926309	.066	.926
		3	2	1.80E-2	.926309		
		3	3	6.00E-3	.926309		
		3	4	1.65E-3	.926309		
<hr/>							
	50	4	0	2.35E-2	.780000	.062	.771
		4	1	2.12E-2	.780000		
		4	2	1.06E-2	.780000		
		4	3	3.88E-3	.780000		
		4	4	1.16E-3	.780000		
	pure guessing			2.26E-3	.500000		

correct probabilities. By contrast, the under-confidence evident for the guessing category arises from an incorporation of discriminant information in addition to pure guessing. This results in severe under-confidence for the higher levels of discriminability because the same evidence configurations assume higher correct probabilities.

Hence, although the categorical scaling procedure of human subjects is orderly and systematic with overt performance the correspondence is not precise enough to render perfect calibration. On this view, then, mis-calibration results as a natural consequence of an orderly but imprecise scaling of information and the effect of decision difficulty on over/under-confidence is a natural consequence of a scaling process that remains fixed across decisional difficulty levels. It should be noted that this latter property of the model actually provides a quantitative description of Poulton's (1989) notion of response contraction bias, discussed in Chapter 2.

Modelling decision time and confidence. As previously stated, the basis for confidence resides in the categorical scaling of discrete evidence accrual configurations. Note that each configuration (x_i, y_i, z_i) gives rise to an expected number of evidence accrual events $(E(N|x_i, y_i, z_i) = x_i + y_i + z_i)$ which is assumed to be monotone with the observed decisional response time (see equation 3). Hence, the expected number of evidence accrual events $E(N|c_i)$ in

confidence category c_i is obtained by the weighted (by $f(x,y,z)$) sum of evidence accrual configurations assigned to category c_i , divided by the overall frequency of responding in that category (F_C):

$$E(N|c_i) = \frac{\sum[(x_i + y_i + z_i) * f(x_i, y_i, z_i)]}{F_C}$$

Figure 5 plots the expected number of evidence accrual events associated with each confidence category for the parameter values used to model the calibration data in Figure 4. Although the model correctly predicts monotone decreasing response times with increasing confidence it clearly fails to account for the "difficulty effect" on response time reported by Baranski and Petrusic (1991a). In addition, the model also incorrectly predicts equal correct and error times when these times are conditioned on the various confidence categories. As previously shown, Baranski and Petrusic (1991a) found error response times to be longer than correct response times at each difficulty level under accuracy.

A very large range of parameter values were explored with the hope that the model's predictions were a consequence of relatively low decisional criterion settings. Unfortunately, the response time difficulty effect and longer error times could not be produced at any criterion levels. For these reasons the competing information scaling model was rejected as a model of confidence calibration.⁵

Figure 5. Expected number of evidence accrual events (response time) for correct and error responses for the four levels of difficulty as predicted by the Competing Information Scaling model. Parameter values used to derive these plots are provided in the text.

3.3.2 An RT Scaling Model of Calibration. A second model of confidence calibration is based on the intuitively compelling notion that confidence is determined by simply partitioning the distribution of decision times into the number of confidence categories required. That is, trials involving long decision times will typically be associated with lower confidence and lower accuracy while quick decisions will be more accurate and more confident. Thus, this model loosely resembles the "Runs" model of confidence proposed by Audley (1960) and discussed in Chapter 1.

To illustrate this view in the context of the slow and fast guessing theory, assume again the decisional parameters $\alpha_1 = \alpha_2 = \alpha_3 = 5$. Hence, the set of response terminating evidence accrual configurations $\{x_i, y_i, z_i\}$ will range over: 5,0,0; 5,0,1, ... , 5,4,4 (plus, as before, the configurations based on guessing, where x and z are reversed). On this model the sum of these configurations, which denote the time taken on a particular judgment, determine the confidence ratings. Note that the sum of evidence accrual events can vary in integer values from 5 (5,0,0) to 13 (5,4,4). The confidence categories are obtained by partitioning these times - and thus their associated frequencies of occurrence and proportion correct - into the 6 confidence categories. The method and assumptions used to fit the data, as well as the number and meaning of the parameters, are the same as for the competing information scaling model. The exception is that

the lexicographic ordering on the values y and z is dropped. Here, y and z have equal weighting since they are both viewed as simply components of decision time.

The model was examined through many criterion and δ values for a fit to the calibration and frequency data reported above. As with the competing information scaling model, the RT scaling model provided an excellent account of these data. However, as with the competing information scaling model, the RT scaling model failed to produce a difficulty effect on response time and also predicted equal correct and error times. Hence, despite its simplicity and intuitive appeal, the RT scaling notion was also rejected as a model of confidence calibration.

3.3.3 The Scaling of Doubt (Inconclusive Information). The final model that was considered is identical to the competing information scaling model except that the lexicographic ordering on the values y and z is reversed. Hence, this model also assumes that competing and inconclusive evidence will influence confidence but this time the amount of inconclusive information, or doubt, assumes the more influential role.

The method for applying the doubt scaling model follows the same procedure as for the competing information scaling model except, as indicated above, the lexicographic ordering on the values y and z is reversed. A simultaneous best fit to the Baranski and Petrusic (1991a) "accuracy" data was

obtained with $\alpha_1=\alpha_2=\alpha_3=12$, $P_3=.25$, and $\delta = .59, .65, .70$, and $.72$ for the difficulty levels ranging from most to least difficult.⁶ The results, plotted in Figure 6, show that the doubt scaling model provides a good account of the calibration difficulty effect. Although the model captures the general trends in the response frequency data, the actual fit is decisively poorer than for the competing information scaling model in that there is less dispersion at the high and low ends of the confidence scale. Nevertheless, the account was reasonable enough to pursue an investigation of the response time predictions and the effects of speeded responding on calibration.

In order to model the effects of speed it was assumed that the decisional criterion parameters would be considerably lower in this case than for those under accuracy (see Petrusic & Baranski, 1989b). However, the four δ parameters, and the P_3 parameter (0.25), were assumed the same as for the accuracy data. A best fit to the speed data was obtained with $\alpha_1=\alpha_2=4$ and $\alpha_3=3$ and the results are plotted in Figure 7. The fit of the difficulty effect on calibration is impressive because the model captures the overall decrease in intercept for each calibration curve as well as the convergence of each curve to 50% correct at the guessing confidence level. This latter result occurs because the lower criterion for guessing (α_3) under speed results in more pure guesses. Once again,

Figure 6. Calibration curves and relative response frequencies predicted by the Doubt Scaling model under accuracy for the 4 levels of comparative difficulty: Level 1 ($\delta=.72$, easiest) circles, Level 2 ($\delta=.70$) diamonds, Level 3 ($\delta=.65$) inverted triangles, and Level 4 ($\delta=.59$, most difficult) triangles (δ is the discriminability parameter).

Figure 7. Calibration curves and relative response frequencies predicted by the Doubt Scaling model under speed for the 4 levels of comparative difficulty: Level 1 ($\delta=.72$, easiest) circles, Level 2 ($\delta=.70$) diamonds, Level 3 ($\delta=.65$) inverted triangles, and Level 4 ($\delta=.59$, most difficult) triangles (δ is the discriminability parameter).

however, the fit to the response frequency curves is not so impressive. On the other hand, the model does capture the gross features of the data; especially the marked increase in the proportion of guess responses with each difficulty level.

Modelling decision time and confidence. Figure 8 plots the expected number of evidence accrual events for each confidence level for the 4 difficulty levels under the accuracy and speeded conditions based on the parameters used to model the calibration data. Under accuracy the doubt scaling model produces monotone decreasing response times, a clear response time difficulty effect, and correctly predicts that error times will be longer than correct times for each difficulty level. Conversely, under speed the model predicts that error and correct times will be comparable, that the response time difficulty effect will not occur, and that the slopes of the response time functions will be shallower than those under accuracy.⁷

In sum, although not able to provide a perfect account of some of the data, the doubt scaling model does provide a good account of all aspects of the Baranski and Petrusic (1991a) data. However, because their calibration curves were characteristically flat through the 60%-100% confidence range (i.e., poor resolution) it was important to determine if the model can, under different response frequency configurations, model calibration curves that show better resolution. A search of the recent literature revealed that Keren (1988)

Figure 8. Expected number of evidence accrual events (response time) for correct (filled points) and error (open points) responses for the four levels of difficulty as predicted by the Doubt Scaling model under accuracy and speeded conditions. Model parameter values are provided in the text.

obtained such data in a recent investigation of calibration for Landolt-C visual acuity judgments.

3.4 A Doubt Scaling analysis of Keren (1988)

As mentioned in Chapter 2, Keren (1988, Experiment 1) obtained calibration and response frequency data for two levels of discriminability in a Landolt-C visual acuity task. A best fit to the data was obtained with $\alpha_1=\alpha_2=\alpha_3=7$, $P_3=.25$, and $\delta=.58$ and $.64$ for the hard and easy discriminability levels, respectively. Again, all decisional parameters, guessing parameters, and scaling parameters were fixed for the two difficulty levels.

The results, plotted in Figure 9, show that the doubt scaling model provides a good account of the data. In fact, the fit of the calibration curves would have been quite impressive had the accuracy for the hard judgments at 70% confidence followed the trend in the data. In addition, although the response frequency fit misses at the low end of the confidence scale, the model does capture the global pattern of the data.

In sum, the doubt scaling model can predict different varieties of calibration curves, including better resolution, when the pattern of response frequencies induce such an outcome. In fact given the similarities between calibration in perceptual and intellectual knowledge tasks it was a natural extension to apply the doubt scaling model to a set

Figure 9. Doubt Scaling model fit to the data of Keren (1988, Experiment 1). Response frequency data points were provided in Keren (1988) and calibration data points were approximated by eye. Model parameter values are provided in the text.

of non-perceptual calibration data. The classic demonstration of the calibration difficulty effect by Lichtenstein and Fischhoff (1977) seemed to be a reasonable data set for such an attempt.

3.5 A Doubt Scaling analysis of Lichtenstein and Fischhoff (1977)

Recall from Chapter 2 that in their experiment, Lichtenstein and Fischhoff (1977, Experiment 5) classified a large set of general knowledge questions into a hard and an easy set on the basis of overall percent correct. A calibration analysis of these sets revealed that the hard set produced severe over-confidence while the easy set produced slight under-confidence. A doubt scaling analysis was performed on the data and a best fit to the calibration and response frequency curves was obtained with $\alpha_1=\alpha_2=\alpha_3=5$, $P_3=.25$, and $\delta=.55$ and $.64$ for the hard and easy questions, respectively. The results, plotted in Figure 10, show that the doubt scaling model again produced a good account of the calibration and response frequency data for both levels of discriminability. The model clearly missed the calibration at 100% confidence for the hard questions but this is true of other attempts to model this set of data (Ferrell & McGoey, 1980). In general, accuracy "jumps" of 20% between the 90% to the 100% confidence categories will not be predicted by the model. On the other hand, accuracy "dips" of 10% to 20%

Figure 10. Doubt Scaling model fit to the data of Lichtenstein and Fischhoff (1977, Experiment 5). Calibration and relative response frequency data points were approximated by eye from Ferrell and McGoe (1980). Model parameter values are provided in the text.

between the 60% and 50% confidence category can be predicted.

3.6 Summary

The present chapter has shown that alternative categorical scaling strategies can be applied to the output of a discrete evidence accrual decision theory in order to provide various models of how decision confidence is determined. Detailed evaluations of these models suggest that confidence is obtained by scaling the amount of inconclusive information (doubt) present at the completion of the judgment. Whether this model provides a full account of all calibration phenomena remains to be determined. However, it is likely that a model of this variety and level of complexity will be required in order to simultaneously account for confidence, accuracy, and the time to decide.

(IV). THE LOCUS OF THE DETERMINATION OF CONFIDENCE

But will all quantitative treatment of mental processes be out of the question then? By no means! An important factor seemed to be susceptible to measurement: I refer to the time required for simple mental processes (Donders, 1969/1868).

4.1 Introduction

Although the Doubt Scaling model provides a comprehensive account of confidence calibration data, an important and intriguing theoretical question remains. Namely, is confidence immediately available upon making a decision or does confidence have to be determined? To pose this question differently we might ask, is confidence determined in conjunction with the primary decision or does confidence have a post-decisional locus?

The empirical study of this problem can be directed by a very basic assumption. Namely, if confidence is indeed immediately available upon making a decision, then the time to report (determine) that confidence should be constant over any confidence level available to the observer. On the other hand, if confidence is determined post-decisionally, then there is reason to believe that the time to determine confidence will be influenced by factors known to effect the duration of the decision process and should thus resemble the response time properties of the primary decision (see below).

On the one hand, we would think that an efficient psychophysical discrimination system should be "finely tuned" to the nature and complexity of the incoming stimulus

information in order to permit adaptive regulation of the decisional parameters. On this basis, it would be curious and surprising if it took more or less time to report one level of certainty over the other. On the other hand, the required allocation of attentional resources needed to simultaneously monitor and to classify the incoming stimulus information may result in a less efficient (or at least slower) primary decision process. In either case, it is not difficult to suggest and test alternative hypotheses for the locus of confidence determination.

4.2 A Decisional Locus Model: 'Internal' Event Monitoring.

A general description of a decisional locus model of confidence is captured by the assumption that the evidence accrual events are systematically classified upon arrival and thus there is direct and immediate access to the available evidence (information) that forms the basis of the primary judgment. In addition, it is assumed that each response terminating evidence accrual configuration is associated with a specific confidence level that simply has to be 'read-off' an internal scale such as Table 1, for example. Hence, with direct access to the response terminating evidence accrual configuration, the subject has to simply select the appropriate confidence level. The immediate and clear prediction of this view is that independent of the difficulty of the decision, the relative speed demands imposed on the

observer, or the confidence category selected, the time to determine confidence will be a (relatively fast) constant (motor execution time).

4.3 A Post-Decisional Locus Model: Event Counting and Matching.

By contrast, a general description of a post-decisional locus model for the determination of confidence is based on the assumption that there is no direct and immediate access to the quantity or quality of evidence accrued at the completion of the comparative judgment and thus no immediate basis for the degree of confidence in the decision. Rather, the response terminating evidence accrual configuration must be post-decisionally evaluated, or interrogated, before the appropriate confidence level can be selected.

One possible view is that the number of evidence accrual events present in each accumulator (i.e., the response terminating evidence accrual configuration x_i, y_i, z_i) must first be post-decisionally counted⁸ and then matched to an internal array, like Table 1, which defines the confidence levels to be associated with each response terminating evidence accrual configuration.

Importantly, note that although the response terminating evidence accrual configurations 5,1,3 and 5,3,1, for example, may result in different confidence values, they must be determined in the same amount of time (i.e., the number of

counts is the same). However, as previously seen in the development of the calibration models, these two configurations will not occur with the same frequency. Hence, exactly the same basis for determining the decision time associated with each confidence category for the doubt scaling model is assumed under the counting and matching hypothesis for the determination of confidence response times. Since the mean number of evidence accrual events increases with decreasing confidence, the required number of counts must also increase. Thus, the clear prediction of this view is that the time to determine confidence will mirror the properties of decision time associated with each confidence category in terms of the relationships between correct and error judgments, decisional difficulty effects, and demands for speed versus accuracy.

Preliminary support for a post-decisional locus of confidence was provided almost a decade ago by Petrusic (1982). He found that both decision time and confidence time increased as the duration difference decreased in a successive duration-discrimination task. Hence, the present studies sought confirmation of these results in a simultaneous comparison task and extend the investigation of confidence response time to include the effects of confidence level, error versus correct responses, extensive practice, and speed versus accuracy stress.

(V). SOME EXPERIMENTS ON THE TIME TO DETERMINE CONFIDENCE

5.1 Introduction

The present chapter reviews the results of three experiments that investigated the time to determine confidence under various conditions. Because the same basic task was used in each study, for convenience, a general description of the methodology will be presented first and the unique features of each experiment will be described in the respective Design and Procedure sections.

5.1.1 Apparatus. Stimuli were presented on an Amdek-310A video monitor. High resolution graphics were available with a Hercules monochrome card and MetaWindows graphics under Turbo Pascal software control. Timing, accurate to within ± 1 msec, was possible with a Data Translation clock board and extensive software development. Graphics production, presentation of instructions and stimuli, event sequencing and timing, and the recording of responses and response times was controlled by an IBM-PC/XT clone computer.

Subjects responded on a 13 cm x 17 cm response-panel possessing two collinear rows of square (2 x 2 cm) buttons. Row 1 (nearer to the subject) contained three buttons labeled L, X, and R (from left to right). Row 2 contained four buttons labeled GUESS, LOW (confidence), MODERATE (confidence) and CERTAIN (from left to right). The two rows were separated by 3 cm and the buttons in each row were

separated by 2 cm.

5.1.2 Experimental Conditions. Subjects were comfortably seated in a 3 m x 5 m experimental room. The stimuli were presented at eye level and at a viewing distance of approximately 60 cm. A small desk lamp, positioned behind the video screen, provided sufficient light for responding without interfering with the visual display.

5.1.3 Stimuli and Basic Procedure. Subjects initiated each trial by depressing and holding down the "home" button marked 'X'. One second later, an instruction was presented (LONGER or SHORTER), centered horizontally near the top of the screen, and remained there for the duration of the trial. The stimulus display appeared 1 second after the presentation of the instruction. This consisted of a central marker, indicated by a 2 cm vertical line, and two horizontal line segments extending different lengths to the left and right from the vertical mid-point of the central marker. All lines were 1 mm wide and appeared in amber colour on a black background.

The subjects' task was to determine which horizontal extent (i.e., left or right) was either 'LONGER' or 'SHORTER'. Subjects made the response by releasing the 'X' button and depressing either the 'L' or 'R' button. Thus, the decision time, or primary judgment time, was measured from

stimulus onset to (L/R) response selection. Subjects provided their confidence rating immediately following the primary response by depressing one of the four confidence buttons. Thus, the time from the depression of the response key (L or R) to the depression of the confidence key (i.e., GUESS,..., CERTAIN) constituted the time to determine confidence. Importantly, because the two instructions appeared equally often with each stimulus pair, and because each pair was counter-balanced for the left-right presentation order within blocks, the mean distance from the response keys to each confidence key was effectively controlled.

Detailed instructions and procedural information were read to the subjects before commencing the session. An emphasis was placed on instructions that required subjects: (1) to make all responses with the index finger of the preferred hand, (2) to respond as accurately as possible, and (3) to employ the full range of confidence alternatives to the best of their ability⁹.

5.2 Experiment 1: The Single Stimulus-Pair Design (Henmon (1911) revisited)

5.2.1 Introduction. In order to systematically study the properties of the time to determine confidence it was important to begin with the simplest possible design and thereafter build in complexity. For this reason, and because of historical significance, I chose to first replicate Henmon's (1911) classic single-pair design (see Chapter 1)

but this time also measure the RT from the decisional response to the confidence response.

5.2.2 Subjects. Twenty Carleton University undergraduate students served for one experimental session of approximately one-hour duration in return for introductory course credits. Subjects had normal or corrected-to-normal vision and were naive with respect to the nature and aims of the experiment.

5.2.3 Stimuli and Design. As mentioned, one pair of linear extents was used. Throughout, I will define stimulus pairs by the notation (x,y) to describe a pair with an extent x pixels (3.13 pixels = 1 mm) to the left of the central marker and y to the right.

It was important to choose a pair that provided neither too difficult nor too simple a judgment in order to preclude the exclusive use of extreme confidence categories. For this reason I chose the pair (100,103) which provided an intermediate level of discrimination in pilot sessions conducted on myself and two other subjects.

Comparison was made under a balanced, within-block, randomization of two levels of presentation order (i.e., 100,103; 103,100) and instruction (LONGER, SHORTER). In all, 400 trials were obtained from each subject: 5 blocks x 80 trials/block. Thus, a block of trials was comprised of the 2 presentation orders x 2 instructions x 20 replications of

each type of comparison. Each block was separated by a break period; the duration of which was under the subjects' control. Finally, each session began with a preliminary block of 24 practice trials (2 orders x 2 instructions x 6 replications). In this and all other experiments, I remained with the subject during the first few practice trials in order to ensure compliance with the instructions and to answer any queries. As such, practice trials were never analyzed.

5.2.4 Results and Discussion. The data of two subjects, who provided unusually long decisional response times (i.e., > 5000 ms) and poor accuracy (i.e., $P_c < 60\%$), were excluded. In addition, in all experiments to be reported, three types of trials were excluded from the analyses: (1) trials on which decision time exceeded 10 seconds, (2) trials on which confidence time exceeded 5 seconds, and (3) trials on which no confidence was reported. This accounted for 145 of the 7,200 trials obtained in the present study.

5.2.4.1 Effects of Confidence Category. Figure 11 provides plots of the decision time (open circles), percent correct (open squares), and time to determine confidence (filled circles) associated with each confidence category in Experiment 1. Also provided are the proportion of times each confidence category was used. Each point in Figure 11 was

Figure 11. Mean decisional response time (open circles), percent correct (open squares), and the time to determine confidence (filled circles) in Experiment 1. Each point is based on the mean of individual subject means at each confidence level.

obtained by taking the mean of individual subject means at each confidence level.

Separate analyses of variance (ANOVAs) were performed on the three dependent measures with the four confidence levels as the within-subjects factor. Evident in the top panel of Figure 11 is a clear replication of Henmon's (1911) results; increasing confidence is associated with a decrease in decision time ($F(3,51) = 32.65, p < .001$)¹⁰ and an increase in accuracy ($F(3,51) = 47.90, p < .001$). Importantly, the lower panel of Figure 11 shows that confidence level has a significant effect on the time to determine confidence ($F(3, 51) = 4.20, p < .025$). Specifically, subjects were relatively quick to elicit the "certain" response and took progressively longer to make confidence judgments with less certainty.

This latter result clearly contradicts any pure decisional locus view for the determination of confidence and supports the view that confidence is (to some degree) determined following the primary decision. In addition, it is also important to note that if confidence response times were uniformly flat across confidence levels (i.e., consistent with a decisional locus view), then it would be unclear as to whether the decisional response times of every psychophysical experiment that required confidence judgments were actually pure decision times or some combination of decision times and time to determine confidence.

5.242 Evidence for Slow and Fast Guessing. The results presented above provide preliminary support for a post-decisional locus view of confidence determination. However, a closer examination of the individual subject data revealed that the pattern of confidence response times showed an interesting dichotomy. Specifically, although virtually all subjects displayed confidence response times that increased from the "certain" through to the "low" confidence category, exactly half of the subjects displayed faster confidence response times for the "guess" category than for the "low" category while the other half displayed the reverse. That is, all subjects were very similar in their pattern of confidence response times except for the time required to elicit the "guess" response. Figure 12 provides plots of these two patterns of responding, labelled "slow" and "fast" guessing responses, respectively (N=9 in each group).

Additional analyses were conducted in order to determine a possible basis for the difference in guess responding. ANOVAs were performed with "slow" and "fast" guessing subjects as a between-group factor and the four levels of confidence as a within-subjects factor. The dependent variables investigated were decisional response time, confidence response time, percent correct, and the relative response frequency for each confidence category. The latter was included to determine if the groups differed in the proportion of times the various confidence categories were

Figure 12. The time to determine confidence for subjects who displayed slow (filled points) and fast (open points) "guess" responses in Experiment 1 (N=9 for each group).

used. For example, if the fast-guessing subjects used the "guess" category more often than the slow-guess group, then the result evident in Figure 12 could be attributed to a simple consequence of response category familiarity.

The results of the ANOVAs revealed no main effects of group for any of the dependent variables ($F < 1.5$). In addition, the only reliable interaction between group and confidence category was for the confidence response time measure (evident in Figure 12). Hence, in the absence of alternative explanations for slow- or fast-guessing, it seems that all subjects determine confidence by a post-decisional mechanism but that some subjects are able to identify when they come out of the pure guessing state (as defined by the slow and fast guessing theory) and thus are able to report these instances almost as quickly as the "certain" response. Although this view is admittedly ad hoc, we should realize that the entire venture of modelling decision confidence is based on assumptions that apparently must be ad hoc to the primary decision process.

5.2.4.3 Correct versus Error Responses. The final analysis in Experiment 1 focussed on the relationship between error and correct response times for the primary response and the confidence response. The mean of individual subject mean correct and error response times for the primary response and the confidence response are plotted as a function of

confidence level in Figure 13 (there was no difference between the "slow" and "fast" guessing subjects in these analyses as well). Because several subjects made no errors when they reported "certainty", the desired ANOVA, with confidence level and error and correct responses as within-subjects factors, could not be performed.

However, overall, error response times were longer than correct response times for the primary judgment (2844 ms vs. 2486 ms) and this difference was reliable ($F(1,17)=19.01$, $p < .001$). By contrast, error confidence response times were only 4 ms longer than correct confidence response times (728 ms vs. 724 ms; $F < 1.0$), contrary to the predictions of the post-decisional locus view. However, it might be that the difference between error and correct decisional response times (376 ms), although statistically reliable, was not large enough to induce the desired effect on confidence response time.

Experiment 2 investigated additional properties of the time to determine confidence by introducing three additional levels of judgmental difficulty into the design. On the basis of the findings reported by Baranski and Petrusic (1991a), a "difficulty effect" should be observed for accuracy and decision time when these measures are conditionalized on the various confidence levels. The critical test for the post-decisional locus view, therefore, will be whether or not a similar difficulty effect can be obtained for the time to

Figure 13. Mean correct (open circles) and error (filled circles) decisional response time and confidence response time for each confidence level in Experiment 1.

determine confidence. Finally, Experiment 2 also provides the opportunity for a more elaborate study of the relationship between correct and error decision times and confidence response times under various levels of difficulty.

5.3 Experiment 2: Variations of Judgmental Difficulty

5.3.1 Subjects. 30 Carleton University undergraduate students served for one, 1-hour session in return for course credits. None of the subjects had participated in the previous experiment, all had normal or corrected-to-normal vision, and all were naive with respect to the nature and aims of the experiment.

5.3.2 Stimuli and Design. It is well known that the difficulty of comparative judgments can be effectively manipulated by varying either the ratio (x/y) or the difference ($x-y$) of the stimulus pair (e.g., Munsterberg, 1896; Petrusic & Jamieson, 1979). Consequently, as in Experiment 1, I selected a 100 pixel standard and varied the variable extent by 1 pixel increments: 100,103; 100,104; 100,105; and 100,106.

Each subject made 384 comparisons per session: 4 blocks x 96 trials/block. Thus, each block was comprised of a randomized presentation of 6 replications of the 4 stimulus pairs x 2 orders of presentation (left, right) x 2 instructions (LONGER or SHORTER). Each subject began the

session with a preliminary set of 32 practice trials (4 stimuli x 2 presentation orders x 2 instructions x 2 replications).

5.3.3 Results and Discussion. The data of two subjects, who used only two confidence levels, were discarded. In addition, 173 of the 10,752 trials obtained in the study were excluded from the analyses because they failed to meet one of the criteria listed in section 5.2.4.

5.3.3.1 Effects of Difficulty. The top panel of Figure 14 provides plots of the mean of individual subject mean percent correct for each level of difficulty at each level of confidence. The lower panel provides the proportion of times each confidence level was used for each level of difficulty. The plots reveal a clear replication of the effect of decisional difficulty on calibration reported by Baranski and Petrusic (1991a); there is a clear tendency towards increased "over-confidence" as difficulty increases and performance declines (technically, we should not use terms such as over- or under-confidence when confidence is not expressed as a proportion or percentage. However, the convention will be followed in the present studies because it adequately describes the relative calibration performance between discriminability levels). In addition, the lower panel shows that the confidence measure was highly sensitive to the

Figure 14. Calibration difficulty effect: Mean percent correct and relative response frequency for each level of difficulty in Experiment 2 (r is ratio of the standard and variable stimuli).

various difficulty levels; there is a systematic increase in the use of lower confidence levels as difficulty increases.

The top panel of Figure 15 provides plots of the mean of individual subject mean decisional response times at each level of difficulty, at each confidence level. The plot shows some evidence of a difficulty effect on RT (i.e., at "guess") but it is clearly not as robust as the effect reported by Baranski and Petrusic (1991a).

The lower panel of Figure 15 provides plots of the mean of individual subject RTs for the time to determine confidence. As in Experiment 1, some subjects exhibited slow-guessing responses (N=13) while others displayed fast-guesses (N=15). The combined effects of these two types of responses resulted in the faster guess responses evident for each difficulty level. Importantly, the plots also show no evidence of a difficulty effect on the time to determine confidence. For each difficulty level, confidence response times are approximately equal.

Separate analyses of variance were conducted with mean percent correct, decisional response time, and confidence response time as dependent measures. The four levels of the difficulty factor and the four levels of confidence served as within-subject factors¹¹. Confidence level had a highly reliable effect on accuracy ($F(3,81)=53.05$, $p < .0001$), decision time ($F(3,81)=45.42$, $p < .0001$), and confidence response time ($F(3,81)=5.12$, $p < .007$); the latter result

Figure 15. Response time difficulty effect: Mean decisional response time and confidence response times for each level of difficulty in Experiment 2 (r is the ratio of the standard and variable stimuli).

2 of/de 2

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

replicating the main finding in Experiment 1. Difficulty level had a highly reliable effect on accuracy ($F(3,81)=11.56$, $p < .0001$), but not on decision time ($F(3,81)=2.74$, $p > .05$) or the time to determine confidence ($F(3,81)=1.61$, $p > .10$). There were no interactions between difficulty level and confidence category for any of the three dependent measures ($p > .30$).

The failure to obtain a difficulty effect on confidence response time contradicts the predictions of the counting and matching hypothesis. However, although the difficulty effect was highly reliable for the accuracy measure, it was not reliable for decisional response time; which is the prerequisite for anticipating the effect on confidence response time.

5.3.3.2 Correct versus Error Response Times. Figure 16 plots the mean of individual subject mean correct and error decision times and confidence response times. In support of the post-decisional view, error response times are longer for both the primary judgment and the confidence judgment. Again, confidence level could not be used as a within-subjects factor because many subjects made no errors when they reported "certainty". However, overall, error decision times were much longer than correct decision times (2927 ms vs. 2312 ms) and this difference was highly reliable ($F(1,27)=30.88$, $p < .0001$). In addition, error confidence

Figure 16. Mean correct (open circles) and error (filled circles) decisional response times and confidence response times in Experiment 2.

response times were longer than correct confidence response times (650 ms vs. 582 ms) and this difference was also reliable ($F(1,27)=4.78$, $p < .04$).

It is important to understand why error confidence response times were longer than correct confidence times in the present experiment but not in Experiment 1. Figure 17 provides plots of response time versus response probability for the primary responses and the confidence responses. The top panel plot, commonly referred to as the Latency-Probability-Function (LPF; see Audley & Pike, 1965; Pike, 1968, 1971; Petrusic & Jamieson, 1978, 1989; Petrusic & Baranski, 1989b) depicts the relationship between correct and error decisional response times for the four pair-difficulty levels in the task. This decreasing LPF is typical of the performance of subjects working under an emphasis to respond as accurately as possible; that is, error response times are substantially longer than for the corresponding correct responses. Notice that pair 4, the a priori easiest to discriminate, shows the fastest correct response time but the slowest error response time and the difference between correct and error response times decreases as the pair difficulty levels increase.

Importantly, the LPF for confidence response time shows a similar decreasing relationship between correct and error response times. Note, in particular, that the difference between error and correct confidence response times is the

Figure 17. Latency-Probability Functions (LPFs) for decision time and confidence time in Experiment 2 (p values less than .5 indicate error responses while p values greater than .5 indicate correct responses).

smallest for pair 1, the a priori most difficult level of discrimination. Moreover, recall that pair 1 in the present experiment was the single pair used in Experiment 1. Hence, the difference between correct and error confidence response times was apparently not reliable in Experiment 1 because of where that pair places on the LPF for decision and confidence response times.

Experiment 3 was conducted with several purposes in mind. First, in order to extend the examination of the effects of decisional difficulty on decision time and confidence time a wider range of difficulty levels was used. Hence, in this experiment we expect to obtain a more robust difficulty effect on decision time. The critical question remains as to whether confidence response times can be sensitive to a wider range of decisional difficulty variations.

A second aim of Experiment 3 was to study the effects of the relative demands for speeded versus accurate responding on the time to determine confidence. Following Vickers and Packer (1982), I employed a paradigm in which subjects worked under alternating sets (blocks) emphasising speed versus accuracy. The prediction of the post-decisional locus view is straight-forward: Since confidence response time must mirror decisional response time, the time to determine confidence must be faster under speed than under accuracy.

The third goal of Experiment 3 was to examine possible

changes in the properties of confidence response time over a large number of sessions. For example, it is possible that subjects might originally display performance that is consistent with a post-decisional view. However, with considerable practice the hypothesized counting and matching process involved in the determination of confidence might become automatized; resulting in performance indices that are consistent with a decisional locus view. For example, one possible interpretation of "fast-guessing" confidence reports is that they reflect a transition from a post-decisional confidence determination strategy to one with a decisional locus. On this view, the subject gradually begins to identify (learn) which confidence levels are to be associated with each response terminating evidence accrual configurations. Aside from subjective certainty, no other confidence level is likely to be more inherently familiar to people than guessing.

The fourth goal of Experiment 3 was to provide a view of the distributional properties of confidence response time for the various confidence levels and under speed and accuracy emphasized conditions.

Finally, in what was initially considered to be a slight departure from the main experimental focus of this dissertation, Experiment 3 provides an examination of the properties of motor execution time over sessions, confidence levels, difficulty levels, and under speed and accuracy

emphasised conditions. The present paradigm readily permits such investigation since the time from the release of the "home" button to the depression of the primary (L/R) response button provides a simple and convenient method of measuring the motor execution component of the overall response latency. To date, decision psychophysicists have (cautiously) assumed that such "non-decisional" time factors (others include stimulus encoding and signal propagation, for example) are separable RT components of the overall response time, and are treated as a separate, constant, parameter in modelling comparative RTs (see Luce, 1986 or Townsend & Ashby, 1983 for more detailed discussions). To the extent that the present paradigm permits an adequate measure of the motor execution component of overall response time, Experiment 3 will provide preliminary support or refutation of these assumptions.

5.4 Experiment 3: Difficulty, Speed vs. Accuracy, and the Development of Automaticity

5.4.1 Subjects. Four Carleton University graduate students served as paid volunteers for 10 experimental sessions of approximately 1-hour in duration. All subjects had normal or corrected-to-normal vision and were naive with respect to the nature and aims of the experiment.

5.4.2 Stimuli and Design. As mentioned, a wider range of difficulty levels was used in order to induce a difficulty

effect on decision time and, perhaps, confidence time. Hence the 100 pixel standard was maintained but the variable stimuli were varied in 2 pixel increments: 100,102; 100,104; 100,106; 100,108.

The 4 blocks of trials in each session were performed under alternating sets for speed (S) versus accuracy (A). The demand for speeded responding on S-blocks was enforced with external feedback to a 500 ms response deadline (i.e., "Too Slow").

Two subjects began session 1 with an accuracy block (ASAS) and two began with the speed block (SASA). Thereafter, subjects alternated between a primary speed or accuracy block from session to session. Prior to commencing accuracy blocks, subjects were presented the following message on the screen:

ACCURACY BLOCK

PLEASE MAKE THE FOLLOWING SERIES OF JUDGMENTS
AS ACCURATELY AS YOU CAN

Prior to commencing speed blocks, subjects were presented the following message on the screen:

SPEED BLOCK

PLEASE MAKE THE FOLLOWING SERIES OF JUDGMENTS
AS QUICKLY AS YOU CAN (WITHOUT BEING CARELESS)

Subjects began each session with a preliminary set of 32 practice trials (4 pairs x 2 orders x 2 instructions x 2 replications). The practice trials were followed by 4 blocks of 96 trials which arose from the factorial combination of 4 pairs, two instructions, 2 presentation orders, and 6

replications of each type of trial. Hence, in all, 3840 experimental trials were obtained from each subject.

Subjects were paid a base rate of \$6.00 per session. Under accuracy blocks, subjects were rewarded 1 cent for each correct response and penalized 1 cent for each error. Under speed blocks, subjects were rewarded 1 cent for correct responses that beat the deadline and were rewarded 1/2 cent for incorrect responses that beat the deadline. On the other hand, subjects were penalized 1/2 cent for correct responses that were "too slow" and were penalized 1 cent for incorrect responses that failed to beat the deadline.

5.43 Results and Discussion.¹²

435 of the 15,360 trials obtained in the present study were excluded because they failed to meet one of the criteria listed in section 5.2.4: 88 trials from subject 1 (S1), 183 from subject 2 (S2), 130 from subject 3 (S3), and 34 from subject 4 (S4). For the most part, these trials arose from omissions of a confidence report during the initial speed blocks.

The results are presented in seven major sections. The first looks at the relationship between confidence and accuracy over sessions and under speed and accuracy stress. Sections two and three describe the effects of stimulus difficulty, practice, and set for speed versus accuracy on performance accuracy and confidence, respectively. Sections

four and five characterize the properties of response time (decision RT, confidence RT, and motor RT) conditional on the various confidence and difficulty levels, respectively. Section six investigates "semantic congruity effects". Finally, section seven concludes the study with a view of the distributional properties of confidence response time.

In all analyses involving sessions, performance in session 1 is presented separately in order to permit comparison with the results obtained in the previous experiments. Performance for the remaining nine sessions will be grouped as follows: 2-4, 5-7, and 8-10.

5.431 Effects of Speed/Accuracy and Practice on Calibration. Figure 18 provides a view of individual subject calibration under speed and accuracy stress. The data for S1, S2, and S4 show the "difficulty effect" (cf. Baranski & Petrusic, 1991a) on calibration that speed produces; this is especially evident for S1 and S4, whose data are strikingly similar. On the other hand, the calibration data for S3 are quite similar under speed and accuracy. This latter result should be noted because S3 will display additional similarities between speed and accuracy in forthcoming analyses. In addition, note that unlike S1, S2, and S4, who seem to rely on extreme confidence levels under speed (i.e., "guess" and "certain"), the pattern of response frequencies for S3 show a direct trade-off in the proportion of times the

Figure 18. Speed/Accuracy calibration curves for each subject in Experiment 3.

various confidence categories are used under speed and accuracy. This might imply that S3, unlike the others, is employing a similar confidence scaling strategy under speed and accuracy and is thus actually more sensitive to the objective increase in difficulty that speed stress produces. Additional evidence for this possibility will be provided in sections 5.4.3.3 and 5.4.3.4.

Table 2 provides a view of calibration and the proportion of times each confidence category was used across sessions and under speed and accuracy conditions. Note that all subjects are extremely accurate when they report the "certain" confidence category, over the 10 sessions and under both speed and accuracy stress. Conversely, the proportion of correct responses made with the "guess" confidence level show very large differences under speed and accuracy, especially for S1 and S4. Specifically, under accuracy, the proportion of correct responses is, on average, well above chance. Under speed, however, subjects seem to have special knowledge about error trials which, in turn, are assigned to the "guess" category (I will discuss this result in more detail in section 5.4.3.3). Hence, it seems that some different processes or strategies are being used to determine confidence under speed and accuracy. The one exception to this pattern is S3, who, again, provides highly similar data under speed and accuracy stress.

Table 2

Proportion Correct (P) and Response Frequencies (F) for the Four Confidence Categories as a Function of Sessions and Under Speed (S) and Accuracy (A) in Experiment 3.

	<u>P(A)</u>	<u>P(S)</u>	<u>F(A)</u>	<u>F(S)</u>
<u>SUBJECT 1:</u>				
<u>Session 1:</u>				
Guess :	.813	.219	.084	.214
Low :	.933	.759	.314	.281
Moderate:	.897	.915	.304	.369
Certain :	.982	.962	.298	.135
<u>Session 2-4:</u>				
Guess :	.797	.292	.239	.302
Low :	.945	.688	.349	.192
Moderate:	.981	.949	.269	.175
Certain :	.988	.957	.142	.332
<u>Session 5-7:</u>				
Guess :	.679	.294	.202	.300
Low :	.869	.903	.291	.174
Moderate:	.989	.987	.323	.296
Certain :	1.000	1.000	.184	.229
<u>Session 8-10:</u>				
Guess :	.768	.305	.271	.330
Low :	.911	.859	.312	.163
Moderate:	.975	.960	.283	.306
Certain :	1.000	.991	.134	.201

Table 2 (continued)

	<u>P(A)</u>	<u>P(S)</u>	<u>F(A)</u>	<u>F(S)</u>
<u>SUBJECT 2:</u>				
<u>Session 1:</u>				
Guess :	.610	.414	.219	.291
Low :	.468	.267	.251	.150
Moderate:	.694	.567	.262	.300
Certain :	.900	.808	.267	.260
<u>Session 2-4:</u>				
Guess :	.508	.262	.318	.301
Low :	.630	.333	.286	.278
Moderate:	.807	.776	.260	.234
Certain :	.936	.935	.135	.187
<u>Session 5-7:</u>				
Guess :	.492	.173	.333	.279
Low :	.500	.328	.276	.317
Moderate:	.898	.759	.277	.259
Certain :	.984	.975	.113	.145
<u>Session 8-10:</u>				
Guess :	.273	.105	.259	.278
Low :	.361	.272	.287	.317
Moderate:	.824	.688	.289	.255
Certain :	.967	.963	.165	.150

Table 2 (continued)

	<u>P(A)</u>	<u>P(S)</u>	<u>F(A)</u>	<u>F(S)</u>
<u>SUBJECT 3:</u>				
<u>Session 1:</u>				
Guess :	.594	.514	.167	.563
Low :	.849	.711	.276	.237
Moderate:	.914	1.000	.365	.142
Certain :	.892	1.000	.193	.058
<u>Session 2-4:</u>				
Guess :	.448	.259	.183	.339
Low :	.655	.570	.297	.315
Moderate:	.891	.939	.238	.231
Certain :	.969	.969	.282	.114
<u>Session 5-7:</u>				
Guess :	.409	.156	.171	.323
Low :	.595	.537	.290	.264
Moderate:	.725	.777	.261	.185
Certain :	.841	.937	.278	.228
<u>Session 8-10:</u>				
Guess :	.292	.133	.119	.237
Low :	.552	.392	.176	.231
Moderate:	.624	.702	.303	.193
Certain :	.748	.820	.401	.339

Table 2 (continued)

	<u>P(A)</u>	<u>P(S)</u>	<u>F(A)</u>	<u>F(S)</u>
<u>SUBJECT 4:</u>				
<u>Session 1:</u>				
Guess :	.778	.340	.148	.288
Low :	.969	.628	.352	.234
Moderate:	.986	.889	.407	.440
Certain :	1.000	1.000	.093	.038
<u>Session 2-4:</u>				
Guess :	.856	.149	.184	.281
Low :	.948	.605	.412	.225
Moderate:	.979	.919	.336	.368
Certain :	1.000	1.000	.069	.126
<u>Session 5-7:</u>				
Guess :	.801	.140	.306	.359
Low :	.931	.670	.329	.189
Moderate:	.958	.918	.290	.295
Certain :	1.000	1.000	.075	.156
<u>Session 8-10:</u>				
Guess :	.852	.133	.282	.353
Low :	.970	.524	.353	.216
Moderate:	.964	.881	.289	.235
Certain :	1.000	.982	.077	.197

5.4.3.2 Effects of Difficulty and Speed/Accuracy on Judgment Accuracy. Figure 19 provides plots of $P(\text{correct})$ conditional on the four difficulty levels and set for speed versus accuracy for each subject in Experiment 3 (each point is based on approximately 460 observations). These plots confirm that both the difficulty manipulation and the speed/accuracy manipulation were effective in inducing large variations in the response accuracy measure. Namely, each subject is much more accurate under accuracy than under speed and each subject displays an increase in accuracy as the stimulus ratio increases. The one exception is S2, who apparently experienced some difficulty under speed stress (the actual response probabilities for S2 were .49, .51, and .48 at stimulus ratios 1.02, 1.04, 1.06, respectively). Finally, note again that performance is very similar for S1 and S4 and their performance is, overall, superior to S2 and S3.

5.4.3.3 Effects of Difficulty, Practice, and Speed/Accuracy on Confidence. Figure 20 provides individual subject plots of mean confidence, conditional on the four difficulty levels and set for speed versus accuracy (again, each point is based on at least 460 observations). The confidence ratings were obtained by assigning the following values to the four confidence categories: Guess=1, Low=2, Moderate=3, and Certain=4.

Figure 19. Probability correct as a function of difficulty level and set for speed versus accuracy for each subject in Experiment 3.

Figure 20. Confidence rating as a function of difficulty level and set for speed versus accuracy for each subject in Experiment 3.

For each subject, confidence increases as discriminability increases under both speed and accuracy. However, surprisingly, three of the four subjects do not replicate the Vickers and Packer (1982) result; only S3 shows clearly higher confidence under accuracy than under speed.

Figure 21 provides a direct comparison of the present data (collapsed over subjects) to those of Vickers and Packer (1982). Their task employed 5 levels of confidence (1-5) and 5 levels of line-length discriminability. However, judging from their plot of percent correct versus difficulty levels (Vickers & Packer, 1982, Figure 2), the levels of discriminability used by Vickers and Packer (1982) were somewhat higher than those used here. Nevertheless, the pattern of results are quite similar to those reported here.

Aside from the difference in levels of discriminability, the other difference between the present study and that of Vickers and Packer (1982) is that subjects in the present study participated for 10 sessions while the data reported by Vickers and Packer (1982) are based on a single session. Table 3 provides a view of the present data as a function of speed/accuracy and sessions. Importantly, every subject in the present study shows higher confidence under accuracy than under speed in session 1. After session 1, S1, S2, and S4 show comparable confidence under speed and accuracy or even slightly higher confidence under speed. Even S3, who shows the largest difference in confidence under accuracy and speed

Figure 21. Confidence rating as a function of stimulus difficulty and speed versus accuracy for the data of Vickers and Packer (1982) [re-plotted by eye] and the combined data in Experiment 3.

Table 3

Proportion Correct and Confidence Rating as a Function of Speed Versus Accuracy and Sessions in Experiment 3. Confidence Ratings (1-4) Based on Guess=1, Low=2, Moderate=3, and Certain=4.

		<u>P(correct)</u>	<u>Confidence</u>
<u>SUBJECT 1:</u>			
<u>Session 1:</u>	Accuracy:	.927	2.82
	Speed :	.729	2.42
<u>Sessions 2-4:</u>	Accuracy:	.925	2.31
	Speed :	.704	2.54
<u>Sessions 5-7:</u>	Accuracy:	.894	2.49
	Speed :	.767	2.45
<u>Sessions 8-10:</u>	Accuracy:	.902	2.28
	Speed :	.733	2.37

		<u>P(correct)</u>	<u>Confidence</u>
<u>SUBJECT 2:</u>			
<u>Session 1:</u>	Accuracy:	.674	2.58
	Speed :	.540	2.53
<u>Sessions 2-4:</u>	Accuracy:	.679	2.21
	Speed :	.528	2.31
<u>Sessions 5-7:</u>	Accuracy:	.663	2.17
	Speed :	.490	2.27
<u>Sessions 8-10:</u>	Accuracy:	.572	2.36
	Speed :	.480	2.28

Table 3 (continued)

		<u>P(correct)</u>	<u>Confidence</u>
<u>SUBJECT 3:</u>			
<u>Session 1:</u>	Accuracy:	.839	2.58
	Speed :	.658	1.69
<u>Sessions 2-4:</u>	Accuracy:	.762	2.62
	Speed :	.595	2.12
<u>Sessions 5-7:</u>	Accuracy:	.665	2.65
	Speed :	.549	2.32
<u>Sessions 8-10:</u>	Accuracy:	.621	2.98
	Speed :	.535	2.63

		<u>P(correct)</u>	<u>Confidence</u>
<u>SUBJECT 4:</u>			
<u>Session 1:</u>	Accuracy:	.951	2.45
	Speed :	.674	2.22
<u>Sessions 2-4:</u>	Accuracy:	.945	2.29
	Speed :	.642	2.34
<u>Sessions 5-7:</u>	Accuracy:	.904	2.13
	Speed :	.604	2.25
<u>Sessions 8-10:</u>	Accuracy:	.937	2.16
	Speed :	.560	2.27

in session 1, displays a systematic decline in the difference between confidence under accuracy and speed over sessions.

In sum, when discriminability levels are relatively high and the data are based on a single session and collapsed over subjects, as was the case in the Vickers and Packer (1982) study ($N=20$), the results will likely show a main effect of speed versus accuracy as well as an interaction between speed/accuracy and difficulty level. On the other hand, it seems that, over time, subjects will adopt different confidence scaling strategies that result in similar overall confidence under speed and accuracy. Table 4 may provide some insight as to how and why this occurs.

Table 4 provides a view of the confidence for correct and error responses under speed and accuracy. For each subject, confidence is much lower for errors than for correct responses (c.f., Pierrel & Murray, 1963; Vickers & Packer, 1982; Petrusic, 1991; Petrusic & Baranski, 1989a), and this is especially evident under speed stress. The reason for the difference in confidence for correct and error responses under speed can be understood as follows. First, consider a purely post-decisional locus for the judgment of confidence under speed. In addition, we know that many (overt) errors occur under speed stress because subjects make fast-guess responses (i.e., the guessing criterion, α_3 , which is set very low, is achieved first) in order to beat the deadline. Now, when the response terminating evidence accrual

Table 4

Mean Confidence Rating for Correct and Error Responses Under Speed and Accuracy in Experiment 3. Confidence Ratings (1-4) Based on Guess=1, Low=2, Moderate=3, and Certain=4. Number of Observations in Parentheses.

		<u>ACCURACY</u>	<u>SPEED</u>
<u>SUBJECT 1:</u>	Corrects:	2.50 (1722)	2.88 (1363)
	Errors :	1.51 (172)	1.28 (495)
<u>SUBJECT 2:</u>	Corrects:	2.59 (1207)	2.91 (891)
	Errors :	1.72 (673)	1.67 (886)
<u>SUBJECT 3:</u>	Corrects:	2.95 (1299)	2.79 (1058)
	Errors :	2.23 (556)	1.61 (797)
<u>SUBJECT 4:</u>	Corrects:	2.27 (1767)	2.86 (1162)
	Errors :	1.53 (131)	1.38 (746)

configuration is interrogated post-decisionally it will be the case that, on some proportion of trials, the accumulator corresponding to the overt response has a higher count total than that for the alternative response (but not yet at criterion). These instances will subsequently be given a relatively high confidence rating (and be correct with $P(\min)=\delta$), even though the overt judgment was based on guessing. On the other hand, on some proportion of trials, the accumulator corresponding to the response not made will have the larger count and thus the subject will have a basis to suspect that an erroneous response had been made, even though the overt response was again based on fast-guessing. Thus, a post-decisional mechanism for confidence provides the opportunity to, in a sense, re-evaluate the primary judgment when that judgment was based on guessing. If this is true, then perhaps we have identified a limitation of the current methodology. It may be that not providing subjects with a confidence alternative in which they can report the subjective feeling of being in error results a situation in which they have no alternative but to use the "Guess" category (the lowest available confidence level) despite the subjective feeling of being in error (i.e., re-scale the levels of confidence available). In addition, this view provides a plausible explanation of why the actual proportion correct is so much below chance when subjects report the "guess" confidence level under speed stress.

5.4.3.4 Effects of Confidence Level, Speed/Accuracy, and Practice on RT. Figures 22-25 provide individual subjects plots of decision time, confidence time, and motor execution time at each confidence level under speed and accuracy stress. Accompanying these plots are individual subject ANOVAs with decision time, confidence time, and motor time as dependent measures and speed/accuracy, sessions (10), and confidence categories (4) treated as levels of between-factors. Unless specified otherwise, the reliable main effects and interactions to be reported are significant at the $p < .001$ level. In addition, each reliable main effect or interaction is followed by an Omega-squared (ω^2) value, which provides an estimate ($0 \leq \omega^2 \leq 1$) of the strength of association between the independent and dependent variables (see Tabachnick & Fidell, 1989, pp.55).

Subject 1

Decision Times. The top two panels in Figure 22 provide plots of the decision time data for S1. The left panel plots show the typical result that, under accuracy, confidence level has a large effect on decision time. The right panel plots show that the deadline manipulation successfully produced flat and fast decision times across the four confidence categories (cf. Figure 3). In addition, note that under speeded responding there is a small practice effect in which response times decrease with sessions. Under accuracy, however,

response times initially decrease from session 1 to sessions 2-4 but then increase substantially for sessions 5-7 only to show a slight decline for the remaining 3 sessions.

The main effect of speed versus accuracy was, of course, highly reliable ($F(1,3751)=5550.08$, $O=.48$), as were the main effects of sessions ($F(9,3751)=26.68$, $O=.02$), and confidence category ($F(3,3751)=218.19$, $O=.06$). The interaction between speed/accuracy and sessions was reliable ($F(9,3751)=26.59$, $O=.02$), as were the interactions between speed/accuracy and confidence category ($F(3,3751)=265.15$, $O=.07$) and sessions and confidence category ($F(27,3751)=4.03$, $O=.007$). The three-way interaction between speed/accuracy, sessions, and confidence category was also reliable ($F(27,3751)=4.84$, $O=.009$).

Confidence Response Times. The middle two panels of Figure 22 provide plots of the confidence response time data for subject 1. Of special and surprising interest is that under accuracy, confidence response times are extremely fast and show no effects of confidence category or sessions; i.e., consistent with a decisional-locus view of confidence determination. Under speeded responding, however, confidence response times are much longer and show large effects of confidence category and sessions.

Figure 22. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 1 in Experiment 3.

Evidently, for this subject, confidence cannot be determined in conjunction with the primary judgment under extreme speed demands; providing a view of relatively pure post-decisional confidence determination times. Under accuracy, however, this subject apparently determines confidence in parallel (or concurrent) with the primary judgment; perhaps accounting for the rather long decisional response times evident in the upper left-hand panel.

These results provide a new perspective on our understanding of how confidence is determined. Specifically, they suggest that the confidence response times observed under accuracy conditions in Experiments 1 and 2 were perhaps only a portion of the overall confidence determination processing times and that the determination of confidence is actually initiated during the decision process; at the cost of an increase in overall decisional response time. Under extreme speed stress, however, the more efficient strategy would seem to be to delay the determination of confidence until the completion of the trial; perhaps because the required allocation of attentional resources needed to simultaneously monitor and classify the incoming stimulus information would result in a less efficient primary comparison process and/or a failure to meet the response deadline.

These results suggest two important points. First, they obviously cast serious doubt on the purity of the decisional

response times collected in psychophysical comparison experiments in which confidence judgments were required. This is particularly important for researchers who have attempted to model the decision time data collected in such studies (e.g., Vickers & Packer, 1982; Vickers et al, 1985a, 1989; Petrusic & Baranski, 1989b; the present research)¹³. Second, until experimental conditions permit an unconfounding of decision time and confidence time under accuracy conditions, a direct comparison of the true confidence response times under speed and accuracy conditions cannot be made. That is, confidence response times may still be found to be longer under accuracy than under speed, and thus parallel decision times, but the question will not be resolved with the present data. I will propose some possible resolutions to this issue in the General Discussion section.

The ANOVA involving confidence response times confirmed all of the main effects and interactions evident in Figure 22. Namely, the main effect of speed/accuracy was highly reliable ($F(1,3751)=2140.47$; $O=.314$; overall, 326 ms under accuracy and 585 ms under speed), as were the main effects of sessions ($F(9,3751)=41.76$, $O=.05$) and confidence category ($F(3,3751)=49.93$, $O=.02$). The interaction between speed/accuracy and sessions was reliable ($F(9,3751)=23.59$, $O=.03$) as were the interactions between speed/accuracy and confidence category ($F(3,3751)=60.37$, $O=.03$) and sessions and confidence category ($F(27,3751)=4.49$, $O=.01$). The three-way

interaction between speed/accuracy, sessions, and confidence category was also reliable ($F(27,3751)=3.23$, $O=.009$).

Motor Response Time

The lower panels in Figure 22 provide plots of the motor execution time for S1. These plots display several points of interest. First, under accuracy, motor times show a clear practice effect with sessions. Second, note that the properties of confidence response time expected in the confidence RT measure are evident in the early sessions for the decisional motor execution time. With practice, however, the properties of confidence time disappear from the motor time; again, likely because the subject is using the decision time to determine confidence, and accounting for the increase in decision time between sessions 5-7. Under speeded responding, motor times are extremely fast and flat across confidence categories, again, because it would be inefficient to determine confidence during that phase of the trial.

Also of interest is that the actual duration of the motor response times under speed are faster than those observed for simple reaction times (which are typically between 200-250 ms; see Luce, 1986 for a review of a considerable portion of the simple RT data); suggesting that the ballistic motor response has been programmed during the evidence accrual (decision) stage. Taken together, it seems that the assumption of constancy and separability of the

motor component of overall response times may not be acceptable since these times display a large speed/accuracy difference as well as evidence of confidence processing in the accuracy condition.

The main effect of speed/accuracy on motor times was highly reliable ($F(1,3751)=1692.73$; $O=.29$; overall, 206 ms under accuracy and 81 ms under speed). The main effect of sessions was reliable ($F(9,3751)=11.41$, $O=.02$) and was qualified by an interaction between speed/accuracy and sessions ($F(9,3751)=13.96$, $O=.02$). The main effect of confidence category was reliable ($F(3,3751)=7.74$, $O=.002$) and was qualified by an interaction between confidence category and speed/accuracy ($F(3,3751)=8.06$, $O=.004$). There were no other reliable interactions for the motor RT measure.

Subject 2

Decisional Response Times. The top two panels in Figure 23 provide plots of the decisional response time data for Subject 2. For the most part, these data are quite similar to those obtained for S1. There are two exceptions. First, notice that S2 had considerable difficulty beating the deadline in the first session under speed. In fact, although the response times are faster than for any of the accuracy sessions, the pattern of results resembles the accuracy data more than the remainder of the speed data. After session 1, S2 responded extremely fast under the speeded blocks and, as

Figure 23. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 2 in Experiment 3.

a consequence, had the poorest accuracy performance of the four subjects (see Figure 19). The other difference between S2 and S1 is evident in the overall faster decision times under accuracy and the presence of a systematic practice effect under accuracy.

The main effect of speed/accuracy was highly reliable ($F(1,3656)=9443.61$, $O=.55$) as were the effects of sessions ($F(9,3656)=91.65$, $O=.05$) and confidence category ($F(3,3656)=414.92$, $O=.07$). Also reliable were the interactions between speed/accuracy and sessions ($F(9,3656)=32.26$, $O=.02$), speed/accuracy and confidence level ($F(3,3656)=346.98$, $O=.06$), and sessions and confidence level ($F(27,3656)=6.60$, $O=.009$). The three-way interaction between speed/accuracy, sessions, and confidence level was also reliable ($F(27,3656)=3.02$, $O=.002$).

Confidence Response Times. The middle panels of Figure 23 provide plots of the confidence response time data for S2. I would like to first discuss performance in session 1 where the pattern of confidence response times are very similar to those obtained in Experiments 1 and 2. Interestingly, confidence response times are faster under speed than under accuracy for session 1, after which confidence response times are faster under accuracy, as they were for S1. The interesting aspect of this data is that S2 had trouble beating the deadline in session 1; that is, the data are more

in line with an instructional manipulation to respond quickly, rather than the enforced deadline adhered to thereafter. The significance of this result is that it suggests that when the subject uses a similar strategy to determine confidence under speed and accuracy, confidence times may in fact be longer under accuracy.

After session 1, when S2 begins to beat the deadline, confidence times become much longer because confidence is being determined post-decisionally. Under accuracy, confidence times show a practice effect with sessions but are relatively flat, as they were for S1.

The main effect of speed/accuracy was highly reliable ($F(1,3656)=2067.77$; $O=.26$; overall, 765 ms under accuracy and 1107 ms under speed) as were the main effects of sessions ($F(9,3656)=122.52$, $O=.14$) and confidence level ($F(3,3656)=78.60$, $O=.03$). The interaction between speed/accuracy and sessions was reliable ($F(9,3656)=63.28$, $.07$) as were the interactions between speed/accuracy and confidence level ($F(3,3656)=59.16$, $O=.02$) and sessions and confidence level ($F(27,3656)=4.77$, $O=.01$). The three-way interaction was also reliable ($F(27,3656)=1.61$, $O=.002$).

Motor Response Times. The lower panels of Figure 23 provide the motor execution time data for S2. Note that S2 also displays clear evidence of confidence processing time in the motor time under accuracy, and this persists throughout the

10 sessions. This is also true for the first session of the speed data, after which motor times are extremely fast and exhibit no confidence time as with S1. Hence, after session 1, time to determine confidence is no longer evident in the decision time or the motor time, enabling S2 to beat the deadline, but confidence times increase dramatically.

The main effect of speed/accuracy was reliable ($F(1,3656)=3657.58$; $O=.47$; overall, 229 ms under accuracy and 61 ms under speed) as were the main effects of sessions ($F(9,3656)=9.96$, $O=.01$) and confidence level ($F(3,3656)=11.66$, $O=.004$). There were only two reliable interactions: speed/accuracy x sessions ($F(9,3656)=21.24$, $O=.02$) and speed/accuracy x confidence level ($F(3,3656)=13.27$, $O=.005$).

Subject 3

Decisional Response Times. The upper two panels in Figure 24 provide plots of the decision time data for S3. Overall, the pattern of results are very similar to those of S2. The only difference is that S3 displays an even more pronounced practice effect under accuracy. In fact, for the final six sessions, the decision times are not much slower than those under speed. The significance of this result will become evident when we review the confidence response time data.

The main effects of speed/accuracy ($F(1,3709)=2876.57$, $O=.25$), sessions ($F(9,3709)=218.04$, $O=.17$), and confidence

Figure 24. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 3 in Experiment 3.

category ($F(3,3709)=134.89$, $O=.14$) were all reliable. So too were the interactions between speed/accuracy and sessions ($F(9,3709)=143.26$, $O=.11$), speed/accuracy and confidence level ($F(3,3709)=173.79$, $O=.05$), sessions and confidence level ($F(27,3709)=8.43$, $O=.02$), and speed/accuracy by sessions by confidence level ($F(27,3709)=10.57$, $O=.02$).

Confidence Response Times. The middle two panels of Figure 24 provide the confidence response time data for S3. Notice that the pattern of confidence response times are very similar under speed and accuracy. As with S1 and S2, confidence response times are longer under speed than under accuracy but the difference is much smaller than for the previous subjects. This is likely due to the relatively fast decision times exhibited by S3 under accuracy and suggests that S3 is determining confidence post-decisionally in both conditions.

The main effect of speed/accuracy was slightly reliable ($F(1,3709)=16.05$; $O=.003$; 823 ms under accuracy and 880 ms under speed), as were the effects of sessions ($F(9,3709)=86.40$, $O=.15$), and confidence category ($F(3,3709)=56.93$, $O=.03$). The interactions between speed/accuracy and sessions ($F(9,3709)=7.58$, $O=.01$), sessions and confidence level ($F(3,3709)=3.21$, $O=.01$), speed/accuracy and confidence level ($F(3,3709)=2.89$, $p < .05$ $O=.001$), and speed/accuracy, sessions, and confidence level ($F(27,3709)=2.45$, $O=.008$), were all reliable.

Motor Response Times. The lower panels in Figure 24 provide plots of the motor execution time data for S3. Under speed, S3 provides the typical fast and flat motor times evident for S1 and S2. Under accuracy, however, S3 shows a large practice effect on confidence response times and, as with S2, clear evidence of confidence processing time in the motor time throughout the 10 sessions.

The main effect of speed/accuracy was reliable ($F(1,3709)=888.93$; $O=.15$; overall, 255 ms under accuracy and 114 ms under speed) as were the main effects of sessions ($F(9,3709)=53.49$, $O=.08$) and confidence level ($F(3,3709)=22.09$, $O=.01$). Also reliable were the interactions between speed/accuracy and sessions ($F(9,3709)=38.22$, $O=.06$), speed/accuracy and confidence level ($F(3,3709)=38.79$, $O=.02$), and sessions and confidence level ($F(27,3709)=2.50$, $O=.007$). The three-way interaction between speed/accuracy, sessions, and confidence level was also reliable ($F(27,3709)=4.89$, $O=.02$).

Subject 4

As is evident in Figure 25, the overall pattern of decision times, confidence times, and motor times for S4 were very similar to those reported for S1. Briefly then, under speed, decision times and motor times were fast and flat while confidence times were long and exhibited the typical

properties of post-decisional confidence determination. The only difference is that S4 seems to acquire knowledge of the "guess" response more gradually, and thus shows a gradual transition from slow to fast guessing over sessions. Under accuracy, as with S1, there is no evidence of confidence processing times in either the confidence response times or the motor response times but the decisional response times are extremely long; suggesting that confidence is being determined in conjunction with the primary response.

Decisional Response Times. The main effects of speed/accuracy ($F(1,3726)=6037.80$, $O=.57$), sessions ($F(9,3726)=21.35$, $O=.02$), and confidence level ($F(3,3726)=30.39$, $O=.008$) were all reliable, as were the interactions between speed/accuracy and sessions ($F(9,3726)=25.55$, $O=.02$), speed/accuracy and confidence level ($F(3,3726)=47.94$, $O=.01$), and speed/accuracy, sessions, and confidence level ($F(27,3805)=1.87$, $p < .005$, $O=.002$).

Confidence Response Times. All main effect and interactions were reliable: Speed/accuracy ($F(1,3726)=1099.62$; $O=.20$; overall, 366 ms under accuracy and 571 ms under speed), sessions ($F(9,3726)=40.19$, $O=.06$), confidence level ($F(3,3726)=15.35$, $O=.008$), speed/accuracy by sessions ($F(9,3726)=6.55$, $O=.009$), speed/accuracy by confidence level ($F(3,3726)=16.52$, $O=.008$), sessions by confidence level

($F(27,3726)=4.39$, $O=.02$), and speed/accuracy by sessions by confidence level ($F(27,3726)=3.35$, $O=.01$).

Motor Response Times. As is evident in Figure 25, the only reliable effects were the main effect of speed/accuracy ($F(1,3726)=370.46$; $O=.09$; overall, 160 ms under accuracy and 78 ms under speed), sessions ($F(9,3726)=5.95$, $O=.01$), and the interaction between speed/accuracy and sessions ($F(9,3726)=1.89$, $p < .05$, $O=.002$).

Figure 25. Decision time, confidence time, and motor execution time as a function of confidence level under speed and accuracy conditions for subject 4 in Experiment 3.

5.4.3.5 Effects of Difficulty, Speed/Accuracy, and Practice. Figures 26-29 provide plots of decision time, confidence time, and motor execution time for each difficulty level as a function of speed versus accuracy condition and practice for each subject. In a fundamental way these data convey similar information to those just reported. The difference is that we exchange an index of subjective difficulty, as measured by confidence level, for an index of a priori difficulty, as measured by the stimulus ratio. Interestingly, the plots show that primary judgment time and confidence time show more robust effects when conditionalized on the levels of subjective difficulty than when conditionalized on the levels of stimulus ratio (see, in particular, the data of S2 and S3).

Since the effects of speed/accuracy and sessions were just described, the following analyses focus only on main effects and interactions involving stimulus pair difficulty level. Unlike in Experiment 2, the wider range of difficulty levels used in the present experiment did produce the sought-after effect of difficulty on decision time and confidence time; at least for some subjects.

Subject 1

The data for S1 are presented in Figure 26. For the primary judgment time, the main effect of stimulus pair was reliable ($F(3,3751)=73.12$, $O=.02$) and was qualified by an

interaction between speed/accuracy and pair ($F(3,3751)=71.87$, $O=.02$) since difficulty effects were not obtained under speed (cf. Link & Tindall, 1971). For the confidence response time measure, only the main effect of pair was reliable ($F(3,3751)=3.17$, $p < .03$, $O=.001$). There were no main effect or interactions involving pair for the motor RT measure.

Subject 2

The data for S2 are presented in Figure 27. As is evident in the plots, there were no main effect or interactions involving pair for the primary judgment time or the motor response time. The main effect of pair was the only reliable effect for the confidence RT measure ($F(3,3656)=6.96$, $O=.003$).

Subject 3

The data for S3 are presented in Figure 28. For the primary judgment time, the main effect of pair was reliable ($F(3,3709)=4.39$, $p < .005$, $O=.001$) and also interacted with the speed/accuracy manipulation ($F(3,3709)=4.68$, $p < .004$, $O=.001$). There was no main effect nor interactions involving pair for the confidence response time measure but there was a reliable effect of Pair for the motor time measure ($F(3,3709)=4.17$, $p < .007$, $O=.002$) as well as an interaction between pair and sessions for the motor time measure ($F(27,3709)=1.52$, $p < .05$, $O=.003$).

Subject 4

The data for S4 are presented in Figure 29. For the primary judgment time, the main effect of pair was reliable ($F(3,3726)=11.47$, $O=.003$) and also interacted with the speed/accuracy manipulation ($F(3,3726)=12.08$, $O=.003$). For the confidence response time measure, the main effect of pair was not reliable but pair did interact with the speed/accuracy manipulation ($F(3,3726)=5.46$, $O=.003$) since there were effects of difficulty under speed but not under accuracy. Again, there were no main effects or interactions involving pair on motor response time.

Figure 26. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 1 in Experiment 3.

Figure 27. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 2 in Experiment 3.

Figure 28. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 3 in Experiment 3.

Figure 29. Decision time, confidence time, and motor execution time as a function of difficulty level under speed and accuracy conditions for subject 4 in Experiment 3.

Summary of sections 3 and 4. Under conditions emphasizing accuracy, only S3 provided clear evidence for post-decisional confidence determination. However, given S3's relatively fast primary judgment times, it is unlikely that S3 was operating under strict accuracy conditions. On the other hand, S1 and S4 displayed very long primary decision times and no evidence of post-decisional confidence determination; either in the confidence judgment or in the motor execution operation. This suggests that some subjects may slow their primary judgment times in order to determine confidence in conjunction with the primary decision. The data for S2 fell somewhere in between these two strategies. S2's decision times were of intermediate duration and although there was no clear evidence of confidence determination in the confidence judgment there was considerable and persistent evidence of confidence determination in the motor selection operation.

Under extreme speed demands, all four subjects adopted a purely post-decisional strategy for determining confidence. This was evident in the fast and flat decision times and motor execution times and by the fact that confidence response times were longer for each subject under speed than under accuracy.

Primary judgment times, confidence response times, and motor response times are more sensitive to subjective difficulty variations than to a priori variations in stimulus

difficulty.

Motor execution times display evidence of decisional processing and are much faster under speed than under accuracy; faster than even simple RT measures, suggesting a programmed ballistic motor execution.

5.4.3.6 Semantic Congruity Effects. Semantic congruity effects were discussed briefly in Chapter 1 (some readers may want to review that section at this point). At that time we saw that semantic congruity effects have been observed with primary decision times as well as with the confidence measure (but rarely, if ever, with the percent correct measure, see Petrusic, 1991; Petrusic & Baranski, 1989a, 1989b). The notion to include appropriate stimuli for the study of congruity effects (i.e., both short and long line-length pairs) was considered during the early planning stages of the current series of studies. However, Petrusic and Baranski (1989, Experiment 1) found that the inclusion of long line-length pairs into the design can undo difficulty effects on primary response times (that otherwise might be observed) because of the additional visual scanning times required. Hence, since a primary goal of the present studies was to investigate the effects of difficulty on decision time and confidence time, it was decided that the investigation of congruity effects for the confidence response time measure should be delayed until the relation between decisional RT and confidence RT is

fully understood. As such, stimulus pairs in the present studies were selected such as to span a very limited range of the length continuum. For example, Pair 1, the a priori shortest (and most difficult) pair in the present study (100,102 pixels) has a mean value of 101 pixels or 32.27 mm. On the other hand, Pair 4, the a priori longest (and easiest) pair of lines (100,108) has a mean value of 104 pixels or 33.22 mm. Hence, the difference between the longest and shortest pairs in the present study was just under 1 mm. This made the discovery of semantic congruity effects in the present study quite astonishing.

Table 5 provides the mean decisional RTs for each pair with the instructions "shorter" and "longer" under speed and accuracy. These results provide several points of interest. First, as mentioned, semantic congruity effects are evident in this extremely narrow range of line-length pairs. Specifically, RTs are faster with the instruction "shorter" for the a priori shortest pair of lines and then switch over to show faster RTs with the instruction "longer". Importantly, these results are not a consequence of averaging; all subjects showed evidence of a congruity effect - either over all sessions, or for a majority of the sessions. The second important result is that congruity effects are also present under speeded responding - but are much smaller in magnitude. This result provides a replication of the findings of Petrusic (1991) and Petrusic and Baranski

Table 5

Mean Decisional RT (ms) for each Pair with the Instructions "Shorter" and "Longer" under Speed and Accuracy in Experiment 3. The RT Difference Between the Instructions Shorter and Longer is Provided in Parentheses.

	<u>ACCURACY (S-L)</u>		<u>SPEED (S-L)</u>	
PAIR 1: SHORTER:	1812	(+54)	419	(+5)
LONGER :	1866		424	
PAIR 2: SHORTER:	1748	(-64)	422	(+7)
LONGER :	1684		429	
PAIR 3: SHORTER:	1614	(-24)	429	(-1)
LONGER :	1590		428	
PAIR 4: SHORTER:	1550	(-51)	428	(-7)
LONGER :	1499		421	

Note: Each of the 16 entries in Tables 5-8 is based on approximately 960 observations.

(1989a, 1989b) [in fact, the magnitude of the congruity effects observed in the present study - under speed and accuracy - are comparable to those observed in the above mentioned studies].

The finding of congruity effects in such a narrow range of stimulus pairs suggests that the congruity effect is much more pervasive than previously imagined. In fact, their presence here should leave no doubt as to how intimately they are tied to the microstructure of comparison.

As in Petrusic (1991) and Petrusic and Baranski (1989a, 1989b), semantic congruity effects were also evident in the confidence measure for these comparisons. Table 6 provides a view of mean confidence (expressed, as in previous analyses, from 1-4) for each pair with the two instructions.

Note that the congruity effect for the confidence measure is of the "funnel" variety (Audley & Wallis, 1964)- a weaker form of congruity effect which is denoted by, (a) overall more confident judgments with the instruction "longer" for all pairs and, (b) an increasing difference between the two instructions as the pairs become longer. Table 6 also shows, as was seen in previous analyses, that the mean confidence rating increases as discriminability increases and that confidence is higher under accuracy than under speed at the higher levels of discriminability.

Table 7 provides the semantic congruity analysis for the confidence response time measure under speed and accuracy.

Table 6

Mean Confidence Rating for each Pair with the Instructions "Shorter" and "Longer" under Speed and Accuracy in Experiment 3. The Confidence Difference Between the Instructions Shorter and Longer (S-L) and the Mean Confidence are also Provided.

	<u>ACCURACY</u>			<u>SPEED</u>		
	<u>CONF</u>	<u>(S-L)</u>	<u>(MEAN)</u>	<u>CONF</u>	<u>(S-L)</u>	
<u>(MEAN)</u>						
PAIR 1: SHORTER:	2.06	+0.03	2.07	2.03	+0.09	2.08
LONGER :	2.09			2.12		
PAIR 2: SHORTER:	2.19	+0.12	2.25	2.10	+0.30	2.25
LONGER :	2.31			2.40		
PAIR 3: SHORTER:	2.48	+0.08	2.52	2.28	+0.26	2.41
LONGER :	2.56			2.54		
PAIR 4: SHORTER:	2.70	+0.17	2.79	2.46	+0.26	2.59
LONGER :	2.87			2.72		

Table 7

Mean Confidence Response Times (ms) for each Pair with the Instructions "Shorter" and "Longer" under Speed and Accuracy in Experiment 3. The RT Difference Between the Instructions Shorter and Longer (S-L) and the Mean Confidence RT for each Pair, are also Provided.

		<u>ACCURACY</u>			<u>SPEED</u>		
		<u>RT</u>	<u>(S-L)</u>	<u>(MEAN)</u>	<u>RT</u>	<u>(S-L)</u>	<u>(MEAN)</u>
PAIR 1:	SHORTER:	577	-1	577	802	+7	806
	LONGER :	576			809		
PAIR 2:	SHORTER:	568	+6	571	780	+8	784
	LONGER :	574			788		
PAIR 3:	SHORTER:	562	+7	566	776	+10	781
	LONGER :	569			786		
PAIR 4:	SHORTER:	564	+3	566	772	-8	768
	LONGER :	567			764		

Several results are of interest here. First, as seen in previous analyses, mean confidence RTs are relatively fast and flat across pair difficulty levels under accuracy because confidence has been determined during the primary judgment for three of the four subjects. Importantly, there is also no evidence of a semantic congruity effect for the confidence RT measure under accuracy. Second, mean confidence RTs monotonically decrease with increasing discriminability under speed and, surprisingly, there is a small but orderly semantic congruity effect.

Finally, Table 8 completes this section with semantic congruity analyses for the motor response time measure. Note that congruity effects are not evident with the motor response measure under speed nor accuracy. Rather, response times are extremely fast and flat across difficulty levels under speed whereas under accuracy, there is a lexical markedness effect in that response times are uniformly faster with the instruction "Longer". In addition, response times decrease with increasing discriminability under accuracy; suggesting that decisional processing of the primary judgment, the confidence judgment, or both, is included in the motor time.

In sum, semantic congruity effects occurred in every condition that exhibited clear evidence of decisional processing; i.e., under speed and accuracy conditions for the primary judgment time and under speed for the confidence

Table 8

Mean Motor Response Times (ms) for each Pair with the Instructions "Shorter" and "Longer" under Speed and Accuracy in Experiment 3. The RT Difference Between the Instructions Shorter and Longer (S-L) and the Mean Motor RT for each Pair, are also Provided.

		<u>ACCURACY</u>			<u>SPEED</u>		
		<u>RT</u>	<u>(S-L)</u>	<u>(MEAN)</u>	<u>RT</u>	<u>(S-L)</u>	<u>(MEAN)</u>
PAIR 1:	SHORTER:	234	-32	218	82	+1	83
	LONGER :	202			83		
PAIR 2:	SHORTER:	219	-16	211	80	+3	82
	LONGER :	203			83		
PAIR 3:	SHORTER:	220	-16	212	87	0	87
	LONGER :	204			87		
PAIR 4:	SHORTER:	215	-12	209	83	+2	84
	LONGER :	203			85		

response time measure. In addition, congruity effects were also evident with the confidence measure, replicating recent findings. On the other hand, congruity effects were not evident in the motor time, for the accuracy measure, or in the confidence response time under accuracy stress. The latter result was perhaps expected since previous analyses revealed that there was no evidence of confidence processing under accuracy for three of the four subjects.

5.4.3.7 Distributions of Confidence Response Times.

Figures 30-33 provide a view of the distributional properties of confidence response time at each confidence level, under speed and accuracy emphasized conditions. The plots are based on the data obtained in sessions 8-10, where performance has stabilized (c.f. Figures 22-25; open diamonds). The response times were grouped into 100 ms intervals, starting from a 50 ms base RT.

Figure 30 provides the data for S1. Under accuracy, confidence times are fast and show very little variance in the four distributions (i.e., note the relative probability at 300 ms). Under speed, confidence response times are not only longer but more variable and longer in the tail. In addition, the plots clearly show the ordering of response times from the "low" through to the "certain" confidence level, as well as the phenomenon of fast guessing.

Figure 31 provides the data for S2. Under accuracy,

confidence times are again relatively fast and show little variability between confidence levels. Under speed, the RT distributions are less skewed and preserve the ordering of mean RTs for each confidence level; in this case with guess responses falling within the range of alternative confidence levels.

Figure 32 provides the data for S3. Recall that the mean decisional RTs for S3 were not very different under speed (378 ms) and accuracy (590 ms) during sessions 8-10. This led to virtually identical confidence RT distributions under speed and accuracy; consistent with a similar confidence determination strategy.

Figure 33 provides the data for S4 which are very similar to the data for S1: Low variance distributions peaking at about 300 ms for each confidence level under accuracy and high variance, long tailed, distributions that preserve the ordering of mean times for each confidence level under speed.

Figure 30. Confidence response time distributions for subject 1 in Experiment 3.

Figure 31. Confidence response time distributions for subject 2 in Experiment 3.

Figure 32. Confidence response time distributions for subject 3 in Experiment 3.

Figure 33. Confidence response time distributions for subject 4 in Experiment 3.

GENERAL DISCUSSION

Three models of confidence calibration were proposed and empirically evaluated. Of the three, only the doubt scaling model was able to capture the effects of decision difficulty and demands for speed versus accuracy on the relationship between confidence and accuracy as well as the pattern of decision times corresponding to the various confidence levels.

The doubt scaling model proposes that people arrive at a feeling of certainty by scaling, respectively, the amount of inconclusive and competing information acquired during the decision. The model was fitted to two sets of perceptual calibration data (i.e., Baranski & Petrusic, 1991a; Keren, 1988) as well as to the classic general knowledge calibration data of Lichtenstein and Fischhoff (1977).

Although the doubt scaling model provided a comprehensive account of confidence calibration it left unspecified the details of how and where in the information processing stream the assumed scaling process evaluates the hypothetical response terminating evidence accrual configurations. In Chapter 4, two general hypotheses for a processing locus and mechanism for the determination of confidence were proposed.

The first was that the incoming evidence accrual events are monitored and systematically classified upon arrival and that upon reaching the decisional criterion, or the guessing

criterion, the confidence to be associated with each evidence accrual configuration is simply "read off" an internal subjective scale. This decisional-locus view predicted that the time to determine confidence would be relatively fast and exhibit no decisional properties.

The second possibility considered assumed that there is no immediate scrutiny of the incoming information but, rather, confidence is something that must be determined post-decisionally. One hypothesis is that, upon reaching the criterion and eliciting the response, the evidence accrual events are interrogated (counted) and then the resulting configuration matched to a subjective scale which defines the confidence to be associated with each evidence accrual configuration. This view led to the prediction that the greater the number of evidence accrual events (i.e., the longer the decision time) the longer it would take, on average, to determine a confidence in the response.

The predictions of these two general views of confidence determination were evaluated in three experiments that examined the time to determine confidence under various experimental conditions. A major finding was that confidence will be determined post-decisionally when decision times are relatively fast. In fact, under extreme time pressure, confidence determination times are extremely long, relative to those typically found under accuracy demands for the primary judgment. However, evidence for post-decisional

confidence determination was also evident under demands for accurate responding if the subject (S3) elects, by choice, to respond quickly. The benefit of a post-decisional strategy under speed would seem to be to free the primary judgment process of additional processing time and/or attentional resources. In addition, it may also simultaneously provide the opportunity to identify errors on trials where the primary judgment was based on guessing.

The other major result was that under accuracy stress, the determination of confidence appears to be initiated during the decision process. If primary judgment times are of intermediate duration (Experiments 1 and 2), the properties of confidence determination time will be observed following the primary judgment; although it is unclear as to what proportion of the overall confidence determination time is being observed during the selection of confidence. On the other hand, if subjects are under incentive to be very accurate, then there may be no evidence of post-decisional confidence determination time; either in the selection of confidence or in the motor time for the primary judgment (e.g., S1 and S4 in Experiment 3). The benefit of such a strategy might be to allow confidence to provide adaptive feedback during the decision process, perhaps to adjust the decisional criteria during the course of the trial in order to obtain better accuracy performance.

As alluded to previously, the very real possibility that

primary decision times may include confidence determination times poses a problem for researchers attempting to model decision time in tasks in which confidence judgments are collected. For example, in the present research, following the development of the doubt scaling model, an attempt was made to fit the decisional RT data of Baranski and Petrusic (1991a) (see footnote #7). The attempt was unsuccessful because the model predicted a psychologically meaningless negative residual RT (i.e., the B parameter on the slow and fast guessing theory). Perhaps this result was a consequence of inflated decision times due to the confounding of confidence times.

Hence, one goal of future research should be to determine just how much time the requirement of confidence judgments adds to the primary judgment time. Research on this problem should proceed in several converging directions. First, an estimate can be obtained in a between-groups design in which both groups perform the same judgment task but confidence is required of only one. However, the "no confidence" group should also perform a secondary response selection operation in order to offset any motor preparation time included in the primary judgment time.

A second strategy should involve a within-subject and within-session design in which the requirement of confidence is varied in alternating blocks. Two control groups would be required; one performing confidence judgments all along and

the other not requiring confidence at all.

Finally, one can record primary judgment times without confidence judgments over a large number of sessions. After performance stabilizes, we can introduce the requirement of confidence and note the changes, if any, in overall decision time for the various difficulty levels.

The possibility that confidence judgments effect decision time is consistent with the view of Nicolson (1982); although for different reasons. Nicolson (1982) suggested that the requirement of a secondary task, such as the reporting of confidence, might be irritating or boring for the subject and thereby effect their performance on the primary task. Nicolson (1982) suggested that confidence judgments could be obtained implicitly in a secondary, simple RT, response in which the subject corroborates feedback about the primary judgment (i.e., correct response "left" or "right"). Nicolson (1982) found that the time to corroborate the feedback was highly correlated with confidence level and thus, according to Nicolson (1982), provides an alternative to the collection of overt confidence judgments. The problem with Nicolson's (1982) study, however, is that the corroborating stimulus always provided feedback to the location of the correct response. Hence, it is unclear as to whether the corroboration time is actually an implicit indication of confidence or a consequence of conflict on error trials, which also increase with decreasing confidence.

Moreover, such a technique clearly precludes calibration-type analyses, confidence determination time analyses, and is, in all likelihood, no less annoying to subjects than reporting a confidence following the primary judgment.

Of course, one option that is always available to decision researchers - especially those concerned with modelling only RT and accuracy data - is to not require confidence data from their subjects at all. However, such an "out of sight, out of mind" approach does not necessarily preclude the possibility that subjects still use confidence in order to adjust their decisional criterion levels in accordance with the nature and complexity of the task (see Vickers, 1978,1979). For example, Baranski and Petrusic (1991d) recently found - both with and without the reporting of confidence - evidence for adaptive criterion adjustment when subjects are switched from an easy to a hard global difficulty level of perceptual comparisons.

Besides the need for a more complete understanding of the effects of confidence judgments on the decision process, future research should also compare and contrast existing models of calibration as well as consider alternative possibilities for the judgment of confidence. Beginning with the pioneering work of Audley (1960), the only quantitative proposals for confidence to date have been those based on stochastic evidence accrual decision theories (Vickers, 1979; Heath, 1982; the present models). As was the case with

Audley's (1960) model, the ultimate success or failure of a model of confidence or confidence calibration may not only be determined by it's ability to account for confidence data but also by the success or failure of the underlying decisional model. For, if a particular decision theory is unable to model the primary decision data, then it cannot be seriously considered as a basis for confidence judgments.

At the present time, this appears to be the major concern for the doubt scaling model; whose primary decision theory (the slow and fast guessing theory) has yet to be fitted to primary decision data. On the other hand, the balance of evidence hypothesis, the only current alternative to the doubt scaling model, has the continuous accumulator theory as a basis, which has recently been successfully applied to primary decision data (Smith & Vickers, 1988, 1989).

Apparently, two research directions need to be explored. First, the slow and fast guessing theory needs to be more fully examined as a model of psychophysical comparison data. Given that the slow and fast guess theory possesses three more parameters than the continuous accumulator theory, it is likely that the model cannot be fitted to primary judgment data without establishing certain decisional parameters from confidence calibration data; as was attempted in the present research.

The second research focus must determine if the balance

of evidence hypothesis can provide as comprehensive an account of calibration data as the doubt scaling model. Although such model comparisons will be complex and time consuming they will hopefully lead to a better understanding of how confidence is determined and how confidence, accuracy, and decision time are related.

Finally, it is possible that confidence might best be accounted for by some other, different, view. For example, many researchers have recently begun to turn their focus away from mathematical or information processing models of cognitive performance to models based on parallel distributed processes or neural networks (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986; Martindale, 1991). However, to date, there have been no successful applications of a neural network approach to the modeling of perceptual discrimination data. Moreover, it is difficult to understand how such an approach would be able to account for the many known, and complex, interactions between response time and response accuracy (e.g., monotone and non-monotone LPF's). Hence, perhaps we should reserve enthusiasm for a connectionist view of confidence until such a modelling architecture can be shown to provide as comprehensive an account of psychophysical discrimination data as that currently provided by mathematical sequential sampling models.

Conclusion

The pioneering studies of confidence in psychophysical judgment and discrimination were inspired by the need to understand the relationship between the sensory elements in our world and our conscious understanding of them. That inspiration remains today. In fact, it is perhaps comforting to know that the relationship between confidence and accuracy, although clearly not perfect, is a good one. At a fundamental level, we would expect no less from a cognitive system designed to accurately evaluate the sensations and perceptions in our world and, if needed, provide us with feedback about our performance when none is externally available. Indeed, the study of human psychophysics would be a futile pursuit if the clarity and detail of the physical world were represented by but a vague, conscious impressionism.

REFERENCES

- Adams, J.K. (1956). A confidence scale defined in terms of expected percentages. American Journal of Psychology, 70, 432-436.
- Adams, J.K., & Adams, P.A. (1961). Realism of confidence judgments. Psychological Review, 68, 33-45.
- Arkes, H.R., Christensen, C., Lai, C. & Blumer, C. (1987). Two methods of reducing over confidence. Organizational Behaviour and Human Decision Processes, 39, 133-144.
- Asher, D. (1974). A model for confidence judgments in choice tasks. PhD thesis, McMaster University, Hamilton, Ontario, Canada.
- Audley, R.J. (1960). A stochastic model for individual choice behaviour. Psychological Review, 67, 1-15.
- Audley, R.J. (1964). Decision-making. British Medical Bulletin, 20, 27-31.
- Audley, R.J., & Pike, A.R. (1965). Some alternative stochastic models of choice. British Journal of Mathematical and Statistical Psychology, 18, 207-225.
- Audley, R.J., & Wallis, C.P. (1964). Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. British Journal of Psychology, 55, 59-73.
- Baranski, J.V. (1987). Investigating adaptive decisional processes: Effects of stimulus difficulty context. Unpublished M.A. thesis, Carleton University, Ottawa,

Ontario.

Baranski, J.V. & Petrusic, W.M. (1988). The calibration of perceptual comparisons. Presented at the 49th annual meeting of the Canadian Psychological Association, Montreal, Canada, June, 1988.

Baranski, J.V., & Petrusic, W.M. (1991a). The Calibration of perceptual judgments: Data and a theory based on the scaling of doubt. Journal of Experimental Psychology: Human Perception and Performance. Manuscript under revision.

Baranski, J.V., & Petrusic, W.M. (1991b). Comparative judgments in two-dimensional space. Manuscript in preparation.

Baranski, J.V., & Petrusic, W.M. (1991c). The discriminability of remembered magnitudes. Memory & Cognition. Manuscript under revision.

Baranski, J.V., & Petrusic, W.M. (1991d). Adaptive judgmental processes: Effects of shifts in contextual difficulty. Manuscript in preparation.

Choo, G.T.G. (1976). Training and generalization in assessing probabilities for discrete events (Technical Report 76-5). Uxbridge, England: Brunel Institute of Organizational and Social Studies.

Christensen-Szalanski, J.J.J., & Bushyhead, J.B. (1981). Physicians' use of probabilistic information in a real clinical setting. Journal of Experimental Psychology:

Human Perception and Performance, 7, 928-935.

- Coombs, C. (1964). A theory of data. Wiley and Sons: New York.
- Dawes, R.M. (1974). The role of the expert in constructing predictive systems. Proceedings of the 1974 International Conference on Systems, Man and Cybernetics of the IEEE Systems, Man and Cybernetics Society, 522-525.
- Dawes, R.M., (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E.D.Lantermann & H.Feger (Eds.), Similarity and Choice: Papers in Honour of Clyde Coombs. Bern: Hans Huber.
- DeSmet, A.A., Fryback, D.G., & Thornbury, J.R. (1979). A second look at the utility of radiographic skull examination for trauma. American Journal of Radiology, 132, 95-99.
- Donders, F.C. (1969/1868). On the speed of mental processes. Acta Psychologica, 30, 412-431. Translation by W.G.Koster of Die Schnelligkeit psychischer prozesse. Archiv fur Anatomei und Physiologie und wissenschaftliche Medizin, 657-681.
- Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. Journal of Mathematical Psychology, 2, 312-329.
- Egan, J.P. (1975). Signal Detection Theory and ROC Analysis.

New York: Academic Press.

Fechner, G.T. (1860). Elements of Psychophysics. New ed. published 1966 (H.E. Adler trans.). Holt, Reinhart, & Winston: New York.

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

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

Fischhoff, B. (1982). Debiassing. In D. Kahneman, P. Slovic, and A. Tversky (Eds.). Judgment under uncertainty: Heuristics and biases (Chapter 31). Cambridge University Press: Cambridge.

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.

Fullerton, G.S. & Cattell, J.M. (1892). On the perception of small differences. University of Pennsylvania Philosophy Series, No. 2.

Garrett, H.E. (1922). A study of the relation of accuracy to speed. Archives of Psychology, 56, 1-105.

Geller, E.S., & Pitz, G.F. (1968). Confidence and decision speed in the revision of opinion. Organizational

Behaviour and Human Performance, 3, 190-201.

George, S.S. (1917). Attitude in relation to the psychophysical judgment. American Journal of Psychology, 28, 1-37.

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

Halliday, D., & Resnick, R. (1978). Physics (3rd edition). New York: Wiley & Sons.

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 judgment to its accuracy. Psychological Review, 18, 186-201.

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. American Journal of Physics, 54, 791-798.

Hertzman, M. (1937). Confidence ratings as an index of difficulty. Journal of Experimental Psychology, 21, 113-119.

Howell, W.C. (1971). Uncertainty from internal and external sources: A clear case of over-confidence. Journal of Experimental Psychology, 89, 240-243.

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 judgment. Archives of Psychology, 34, 1-53.
- Kellogg, W.N. (1931). The time of judgment in psychometric measures. American Journal of Psychology, 49, 65-86.
- Keren, . (1987). Facing uncertainty in the game of bridge: A calibration study. Organizational Behaviour and Human Decision Processes, 39, 98-114.
- Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. Acta Psychologica, 67 , 95-119.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118.
- LaBerge, D. (1962). A recruitment theory of simple behaviour. Psychometrika, 27, 375-396.
- Laming, D. (1968). Information theory of choice reaction times. New York: Academic Press.
- Lichtenstein, 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.
- Lichtenstein, S. & Fischhoff, B. (1980a). Training for calibration. Organizational Behaviour and Human Performance, 26, 149-171.
- Lichtenstein, S. & Fischhoff, B. (1980b). How well do probability experts assess probability? (Decision

Research Report # 80-5). Eugene, Oregon: Decision Research.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgement under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

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

Link, S.W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. Journal of Mathematical Psychology, 34, 2-41.

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

Link, S.W., & Tindall, A.D. (1971). Speed and accuracy in comparative judgments of line length. Perception and Psychophysics, 9, 284-288.

Luce, R.D. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

Lund, F.H. (1926). The criteria of confidence. American Journal of Psychology, 37, 372-381.

Martindale, C. (1991). Cognitive psychology: A neural-network approach. Brooks/Cole: Pacific Grove, CA.

- May, R.S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B.Brehmer, H.Jungermann, and G.Sevon (Eds.) New directions in research on decision making (pp. 175-189). Elsevier: North Holland.
- McClelland, J.L., & Rumelhart, D.E. (Eds). (1986). Parallel distributed processing, Vol.2. Cambridge, MA: MIT Press.
- Munsterberg, H. (1984). Studies from the Harvard psychological laboratory. A psychometric investigation of the psychophysic law. Psychological Review, 1, 45-51.
- Murdock, B.B., Jr., & Dufty, P.O. (1972). Strength theory and recognition memory. Journal of Experimental Psychology, 94, 284-290.
- Murphy, A.H., & Winkler, R.L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? National Weather Digest, 2, 2-9.
- Myers, J.L., & Well, A.D. (1991). Research design and statistical analysis. New York: HarperCollins Publishers Inc.
- Nickerson, R.S. & McGoldrick, C.C., jr. (1965). Confidence ratings and level of performance on a judgmental task. Perceptual Motor Skills, 20, 311-316.
- Nicolson, R. (1982). The corroboration time measure: An implicit confidence judgment. Acta Psychologica, 51, 27-

- Norman, D.A., & Wickelgren, W.A. (1969). Strength theory of decision rules and latency in short-term memory. Journal of Mathematical Psychology, 6, 192-208.
- Oscamp, S. (1965). Overconfidence in case-study judgments. The Journal of Consulting Psychology, 29, 261-265.
- Petrusic, W.M. (1982). Psychophysical judgment and the time to determine confidence. A paper read at the annual meeting of the Psychonomic Society, Minneapolis, Minnesota, November, 1982.
- Petrusic, W.M. (1991). Semantic congruity effects and theories of the comparison process. Journal of Experimental Psychology: Human Perception and Performance. Under revision.
- Petrusic, W.M., & Baranski, J.V. (1989a). Context, context shifts, and semantic congruity effect in comparative judgments. In D.Vickers and P.Smith (Eds.) Human Information Processing: Measures, Mechanisms, and Models (pp. 231-251). North Holland Press: Amsterdam.
- Petrusic, W.M., & Baranski, J.V. (1989b). Semantic congruity effects in perceptual comparisons. Perception & Psychophysics, 45, 439-452.
- Petrusic, W.M., & Baranski, J.V. (1991a). Effects of feedback and contextual difficulty in the calibration of perceptual judgments. Acta Psychologica. Under revision.
- Petrusic, W.M., & Baranski, J.V. (1991b). On the properties

of long term memories for percepts: Binary and quaternary relational judgments with remembered stimuli. Journal of Experimental Psychology: Learning, Memory, & Cognition. Under revision.

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

Petrusic, W.M., & Jamieson, D.G. (1979). Resolution time and the coding of arithmetic relations on supraliminally different visual extents. Journal of Mathematical Psychology, 19, 89-107.

Petrusic, W.M., & Jamieson, D.G. (1989). Comparative judgments of the ease of sameness-difference judgment: Matching probabilistic structures. American Journal of Psychology, 102, 69-90.

Pierce, C.S., & Jastrow, J. (1885). On small differences in sensation. Proceedings of the National Academy of Sciences, 3, 75-83.

Pierrel, R., & Murray, C.S. (1963). Some relationships between comparative judgment, confidence and decision time in weight-lifting. American Journal of Psychology, 76, 28-38.

Pike, R. (1968). Latency and relative frequency of response in psychophysical discrimination. The British Journal of Mathematical and Statistical Psychology, 21, 161-182.

- Pike, R. (1971). The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on simple counting. Perception and Psychophysics, 9, 455-460.
- Pike, R. (1973). Response latency mechanisms for signal detection. Psychological Review, 80, 53-68.
- Poulton, E.C. (1973). Unwanted range effects from using within-subject experimental design. Psychological Bulletin, 80, 113-121.
- Poulton, E.C. (1979). Models for biases in judging sensory magnitude. Psychological Bulletin, 86, 777-803.
- Poulton, E.C. (1989). Bias in quality rating judgments. Lawrence Erlbaum Associates: Hillsdale.
- Rumelhart, D.E., & McClelland, J.L., (Eds). (1986). Parallel distributed processing, Vol.1. Cambridge, MA: MIT Press.
- Seward, G.H. (1928). Recognition time as a measure of confidence. Archives of Psychology, No.99, 1-54.
- Sharp, G.L., Cutler, B.L., & Penrod, S.D. (1988). Performance feedback improves the resolution of confidence judgments. Organizational Behaviour and Human Decision Making, 42, 271-283.
- Sieber, J.E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. Journal of Personality and Social Psychology, 30, 688-694.
- Smith, P.L. (1989). A deconvolutional approach to modelling

- response time distributions. In D.Vickers and P.Smith (Eds.) Human information processing: Measures, mechanisms, and models. Elsevier Science Publishers: North-Holland Press.
- Smith, P.L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. Journal of Mathematical Psychology, 32, 135-168.
- Smith, P.L., & Vickers, D. (1989). Modeling evidence accumulation with partial loss in expanded judgment. Journal of Experimental Psychology: Human Perception and Performance, 15, 797-815.
- Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.
- Stone, M. (1960). Models for choice reaction time. Psychometrika, 2, 251-260.
- Swensson, R.G. (1972). The elusive tradeoff: Speed versus accuracy in visual discrimination tasks. Perception and Psychophysics, 12, 16-32.
- Tabachnick, B.G., & Fidell, L.S. (1989). Using multivariate statistics (2nd edition). New York: Harper & Row Publishers.
- Tanner, W.P. & Swets, J.A. (1954). A decision making theory of visual detection. Psychological Review, 61, 401-409.
- Thomas, E.A.C. (1971). Sufficient conditions for monotone hazard rate: An application to latency-probability curves. Journal of Mathematical Psychology, 8, 303-332.

- Tomassini, L.A., Solomon, I., Romney, M.B., & Krogstad, J.L. (1982). Calibration of auditors probabilistic judgments: Some empirical evidence. Organizational Behaviour and Human Performance, 30, 391-406.
- Trow, W. C. (1923). The psychology of confidence. Archives of Psychology, 67, 1-47.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.
- Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. Ergonomics, 13, 37-58.
- Vickers, D. (1978). An adaptive module for simple judgments. In: Attention and Performance VII, J. Requin (Ed.) Lawrence Erlbaum, Hillsdale, N.J., 599-618.
- Vickers, D. (1979). Decision processes in visual perception. New York: Academic Press.
- Vickers, D. (1980). Discrimination. In A.T. Welford (Ed.) Reaction times. New York: Academic Press.
- Vickers, D. (1985). Antagonistic influences on performance change in detection and discrimination tasks. In G d'Ydewalle (Ed.), Cognition, information processing and motivation. Proceedings of the XXIIIrd International Congress of Psychology (Vol 3). Amsterdam: North Holland.
- Vickers, D., Burt, J., Smith, P., & Brown, M. (1985a). Experimental paradigms emphasising state or process

- limitations: I. Effects on speed-accuracy tradeoffs. Acta Psychologica, 59, 129-161.
- Vickers, D., Caudrey, D., & Willson, R.J. (1971). Discriminating between the frequency of occurrence of two alternative events. Acta Psychologica, 35, 151-172.
- Vickers, D., Foreman, E.A., Nicholls, M.E.R., Innes, N.J., & Gott, R.E. (1989). Some experimental tests of the application of an interval of uncertainty model to a time-limited expanded judgment task. In D.Vickers and P.Smith (Eds.) Human Information Processing: Measures, Mechanisms, and Models. Elsevier Science Publishers: North Holland Press.
- Vickers, D., & Packer, J.S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. Acta Psychologica, 50, 179-197.
- Vickers, D., & Smith, P. (1985). Accumulator and random-walk model of psychophysical discrimination: A counter evaluation. Perception, 14, 471-497.
- Vickers, D., Smith, P., Burt, J. & Brown, M. (1985b). Experimental paradigms emphasizing state or process limitations: II. Effects on confidence. Acta Psychologica, 59, 163-193.
- Volkmann, J. (1934). The relation of time of judgment to certainty of judgment. Psychological Bulletin, 31, 672-673.

- Wagenaar, W.A., & Keren, G.B. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. Organizational Behaviour and Human Decision Processes, 36, 406-414.
- Wickelgren, W.A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. Journal of Mathematical Psychology, 3, 316-347.
- Williamson, G.F. (1915). Individual differences in belief, measured and expressed by degrees of confidence. Journal of Philosophy, Psychology, and Scientific Methods, 12, 127-137.

FOOTNOTES

1. For this reason, the theory is often referred to as the continuous accumulator theory, in contrast to the discrete evidence accrual models (e.g., Audley, 1960; LaBerge, 1962; Petrusic, 1991).
2. Dawes (1980) made this statement apparently without knowledge of the original findings of Adams (1957) who, as previously mentioned, found under-confidence on a perceptual word-recognition task.
3. We can contrast this with confidence people may have concerning events that will occur in the future; commonly referred to as "forecasting" (see Fischhoff (1982) for a review).
4. The parameter search was conducted with BASIC programs for the Slow and Fast Guessing theory (to derive the response frequencies and probabilities at various parameter levels) and to derive the empirical estimates of the models. However, it should be noted that the chosen parameter values arise from a non-exhaustive search of the permissible parameter space. Therefore, the results presented may not be 'optimal.' It is a future goal of this research to develop an interactive computer routine that employs an 'optimising' (e.g., least-squares) principle to simultaneously fit the calibration and response frequency data.
5. An interesting aspect of the slow and fast guessing theory is that it provides the discrete analogue of the Vickers

(1979) continuous accumulator theory when $\alpha_3=0$ and $P_3=0$. With this modification, the discrete version of the balance of evidence hypothesis was examined as a model for confidence calibration. The results were virtually identical to those reported for the Competing Information Scaling model: the model produced a good account of calibration and response frequency but failed to predict the RT difficulty effect.

6. The doubt scaling model was fit to the combined data of the four subjects who participated in the accuracy and speeded conditions in order to assume the same δ and guessing parameters under speed and accuracy.

7. The determination of the number of evidence accrual events (N) for each confidence category permits, for the first time with a discrete evidence accumulator theory, estimation of the non-decisional time parameters A (rate of accrual) and B (encoding and motor execution). Following equation 3 (above), A and B correspond, respectively, to the slope and intercept of the plot of N against the observed decisional response time. Although the estimate of A approximated the value obtained with the continuous accumulator theory (cf. Smith & Vickers, 1989), the B parameter took on a psychologically meaningless negative value that resulted from a relatively large observed response time range between the guess and certain confidence categories. However, evidence will be presented in the experiments to be reported that this result may be a consequence of the paradigm rather than a failure of

the model.

8. 'Counting' can be assumed to take the form of a high-speed serial scan of items in working memory (e.g., Sternberg, 1966).

9. I decided not to employ trial-by-trial feedback in any of the experiments because previous research (Baranski, 1987; Petrusic & Baranski, 1989b), with a similar task, showed feedback to induce a speed-accuracy tradeoff; with the direction of this trade-off determined by the global difficulty of the task. Since an important goal of this research was to study the effects of speed versus accuracy stress, I opted to remove the potential for such effects until they were under direct control (Experiment 3).

10. Significance levels reported in Experiments 1 and 2 are with the Greenhouse-Geisser epsilon adjusted degrees of freedom. However, degrees of freedom reported in the text are not adjusted.

11. One subject did not provide any "certain" confidence reports with the most difficult level of comparison. This missing cell was estimated for the ANOVA according to the procedure recommended by Myers and Well (1991, pp.256-258).

12. The third block of the seventh session of subject 3 was interrupted by a local power failure. The third block was restarted following the delay.

13. This may also include studies by Smith (1989) and Smith and Vickers (1988, 1989). Although confidence data were not

reported in these studies, Vickers et al. (1989, p.257) note that "most experiments" in their laboratory require confidence judgments following each response.

Appendix A

Table 1 provides a breakdown of how the calibration and response frequencies were fit for the competing information scaling model (Figure 4). The first column provides the confidence categories, C_c . Columns 2 and 3 together provide the configuration of competing (y) and indiscriminant (z) evidence accrual events and illustrate the lexicographic ordering on the variables y and z . Column 4 provides the expected frequencies of these various evidence accrual configurations, $f(x,y,z)$, for each value of δ . Column 5 provides the probability correct for each of these configurations, $p(x,y,z)$. Column 6 provides the overall response frequency associated with each confidence category, F_c . These values were obtained by summing the frequencies of each configuration within a given category, C_c . Best fits to the group frequency data were obtained by optimally partitioning the range of evidence accrual configuration frequencies into the 6 permissible confidence categories. Finally, column 7 provides the expected proportion correct associated with confidence category, P_c . This value is obtained by the weighted (by $f(x,y,z)$) mean of the values of $p(x,y,z)$ in each confidence category.

END

28·0|5·92

FIN