EMPIRICAL FOUNDATIONS FOR A COMPREHENSIVE MODEL OF
n-ALTERNATIVE FORCED-CHOICE DECISION-MAKING

by

Steven R. Carroll

A thesis submitted to
the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Psychology

Carleton University
Ottawa, Canada

©2011 Steven R. Carroll
NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

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

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
Abstract

In a series of four experiments, participants were asked to evaluate $20 \times 20$ arrays of coloured squares, and decide which of two, three, or four presented colours were represented either "MOST" or "LEAST" often in the display, depending upon the instruction. Following each of these decisions, participants were asked to express how confident they were that they had just made a correct choice. Decisional response times, accuracy, mean expressed confidence levels, and post-decisional times required to express confidence were recorded. The data generated by these experiments represents the first time all four of these dependent variables have been recorded in a series of experiments wherein the number of viable response alternatives available to participants exceeded two. As such, these data provide a framework by which models of two-alternative forced-choice (2AFC) decision-making can be extended to the $n$-alternative case. A race model, called the EPIC model, governed by Stevens Power Law (1957) and Shannon's Information Theory (1948), is presented and its ability to replicate the results of these three experiments is tested against the abilities of other race models to do the same. The results of two additional experiments, which were designed to generate semantic congruity effects (SCEs) within a $n$-AFC decision-making paradigm, are also presented, as is the results of an attempt to have the EPIC model replicate these more advanced effects.
Acknowledgements

There are a number of people who deserve thanks for helping me push this boulder all the way up to the top of the hill.

First, thanks go to my thesis supervisor, Dr. William M. Petrusic, for providing me with an entire decade’s worth of instruction in the psychophysical arts.

I would also like to thank my external examiner, Dr. Denis Cousineau, and my internal examiner, Dr. Raj Singh, for their helpful advice and insightful questions.

Special thanks go to committee members Dr. Jo-Anne LeFevre, who provided a very useful and thorough review of an early draft of this paper, and Dr. Craig Leth-Steenschen, who helped me program the BIC calculator.

Many thanks go to Etelle Bourassa, the Psychology Department Graduate Studies Administrator. Thank you for doing everything that you do for your graduate students.

I would also like to thank Dr. Janet Mantler, not only for her continuous encouragement and support, but for helping me to become a full-time student.

Thanks to T-Rex and Locksley, who helped me organise my priorities.

Most of all, I’d like to thank my best-friend and wife Dr. Judith Godin. We fought the forces of the Ivory Tower together. Battered and bruised, we've emerged from the trenches, graduate degrees in hand. I had the time of my life, Judith, and it was all thanks to you.
# Table of Contents

Abstract .......................................................................................................................... ii
Acknowledgements ....................................................................................................... iii
Table of Contents ........................................................................................................... iv
List of Tables ................................................................................................................ xiii
List of Figures ................................................................................................................ x
List of Appendices ......................................................................................................... xiii
Chapter 1: Historical Overview .................................................................................... 1
  Decisional Complexity and Decisional RT ................................................................. 3
  Decisional Complexity, Decisional RT, and Confidence ............................................ 11
  Decisional Complexity, Decisional RT, and Decisional Accuracy ............................ 23
  Decisional Complexity and Post-Decisional Times to Render Confidence ............... 28
Chapter 2: General Method for Experiments 1 – 3 ...................................................... 29
  Apparatus .................................................................................................................. 29
  Materials / Stimuli .................................................................................................... 30
  Procedure .................................................................................................................. 33
  Participants ............................................................................................................... 36
  Analytical Method and Results ................................................................................. 36
Chapter 3: Experiments 1 – 3 ...................................................................................... 39
  Experiment 1 ............................................................................................................ 39
  Experiment 2 .......................................................................................................... 45
  Experiment 3 .......................................................................................................... 51
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>An example of Murdock's distinctiveness hypothesis</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>Statistical results from analyses of Experiments 1-3</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>Statistical results from analyses of Experiment 1</td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>Decisional difficulty and the emergence of nonlinear RT</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>Statistical results from analyses of Experiment 2</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>Decisional difficulty and the emergence of nonlinear RT × Blocks</td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>Statistical results from analyses of Experiment 3</td>
<td>58</td>
</tr>
<tr>
<td>8</td>
<td>Proportion of correct responses: Decisional difficulty and complexity</td>
<td>61</td>
</tr>
<tr>
<td>9</td>
<td>Multinomial probabilities of all possible n = 5 scenarios</td>
<td>68</td>
</tr>
<tr>
<td>10</td>
<td>The derivation of E(RT) for sample sizes ranging from 1 to 25</td>
<td>70</td>
</tr>
<tr>
<td>11</td>
<td>$C_{nAFC}$ vs. $C_{base}$</td>
<td>76</td>
</tr>
<tr>
<td>12</td>
<td>Entropy levels for Experiment 1-3</td>
<td>78</td>
</tr>
<tr>
<td>13</td>
<td>Observations within each of the conditions in Experiment 1</td>
<td>84</td>
</tr>
<tr>
<td>14</td>
<td>Observed vs. modelled accuracy for Experiment 1 data</td>
<td>86</td>
</tr>
<tr>
<td>15</td>
<td>Observed vs. modelled accuracy for Experiment 2 data</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>Observed vs. modelled accuracy for Experiment 3</td>
<td>91</td>
</tr>
<tr>
<td>17</td>
<td>Model families tested based on Equation 15</td>
<td>102</td>
</tr>
<tr>
<td>18</td>
<td>Best BIC values produced by the SIMPLEX tests</td>
<td>105</td>
</tr>
<tr>
<td>19</td>
<td>The probabilities of various decisional RTs where nAFC = 2</td>
<td>114</td>
</tr>
<tr>
<td>20</td>
<td>The probabilities of various decisional RTs where nAFC = 3</td>
<td>115</td>
</tr>
<tr>
<td>Page</td>
<td>Section</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>E(confidence): Proportion of mode evidence accrued</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Observed vs. mode-based modelled confidence for Experiments 1-3</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Statistical results from analyses of Experiment 4</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Observed results from Experiment 4</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>“MOST” vs. &quot;LEAST&quot; colour representation in a 4AFC stimulus display</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Observed vs. modelled accuracy for Experiment 4</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Observed vs. mode-based modelled confidence for Experiment 4</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Statistical results from analyses of Experiment 5</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Statistical results from analyses of Experiment 6</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Possible extent comparisons using the stimuli in Experiment 6</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Best-fitting C_base and discomfort exponent 'a' values for Experiment 6</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Modelled main effect of complexity on accuracy for Experiment 6 data</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Modelled main effect of complexity on confidence for Experiment 6</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Modelled interaction between set and complexity on accuracy</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Modelled interaction between complexity and instruction on confidence</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Modelled interaction between set and instruction on confidence</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Modelled interaction: condition, instruction, and complexity on accuracy</td>
<td></td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A simplified rendition of Vickers' (1979) APEP model</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>A simplified rendition of Usher and McClelland's (2001) LCAM</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>The layout of the response panel used in Experiments 1-4</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Stimulus examples</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>Within trial stimulus display presentation order</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>Experimental results from Experiment 1</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>Experimental results from Experiment 2</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>Results of Experiment 3: Confidence blocks</td>
<td>56</td>
</tr>
<tr>
<td>9</td>
<td>Results of Experiment 3: Non-confidence blocks</td>
<td>57</td>
</tr>
<tr>
<td>10</td>
<td>Results of Experiment 3: Both block types collapsed</td>
<td>57</td>
</tr>
<tr>
<td>11</td>
<td>Decisional RTs, collapsed over difficulty for Experiments 1-3</td>
<td>59</td>
</tr>
<tr>
<td>12</td>
<td>E(RTs) for correct decisions in Experiment 1: Equation 11</td>
<td>75</td>
</tr>
<tr>
<td>13</td>
<td>E(RTs) for correct decisions in Experiment 1: Equation 13</td>
<td>80</td>
</tr>
<tr>
<td>14</td>
<td>The flow of the basic EPIC model</td>
<td>82</td>
</tr>
<tr>
<td>15</td>
<td>Modelled Experiment 1 decisional RT and accuracy plots</td>
<td>85</td>
</tr>
<tr>
<td>16</td>
<td>Modelled Experiment 2 decisional RT and accuracy plots: Basic model</td>
<td>87</td>
</tr>
<tr>
<td>17</td>
<td>Modelled Experiment 2 decisional RT and accuracy plots</td>
<td>89</td>
</tr>
<tr>
<td>18</td>
<td>Modelled Experiment 3 decisional RT and accuracy: No confidence</td>
<td>93</td>
</tr>
<tr>
<td>19</td>
<td>Modelled Experiment 3 decisional RT and accuracy: Confidence</td>
<td>94</td>
</tr>
<tr>
<td>20</td>
<td>Modelled decisional RT and accuracy plots: Leite and Ratcliff model 1</td>
<td>106</td>
</tr>
</tbody>
</table>
Modelled decisional RT and accuracy plots: Leite and Ratcliff model.
Overview of the RT-scaling confidence module.
RT-scaling confidence module: Experiment 1.
RT-scaling confidence module: Experiment 2.
RT-scaling confidence module: Experiment 3.
Overview of the mode-based confidence module.
Mode-based confidence module: Experiment 1.
Mode-based confidence module: Experiment 2.
Mode-based confidence module: Experiment 3.
Observed data for Experiment 4, regular instruction blocks.
Observed data for Experiment 4, irregular instruction blocks.
Post-decisional times to render confidence ratings in Experiment 4.
A comparison of times to render confidence to confidence criteria 'K'.
Modelled Experiment 4 decisional RT and accuracy plots: 'Irregular'.
Modelled Experiment 4 decisional RT and accuracy plots: 'Regular'.
Modelled 'irregular' mean confidence and time to render confidence.
Modelled 'regular' mean confidence and time to render confidence.
The layout of the response panel used in Experiments 5-6.
Experiments 5-6: Stimulus examples.
Main effects of decisional complexity for Experiment 5.
Main effects of stimulus set for Experiment 5.
Main effects of instruction for Experiment 5.
<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Interactions between complexity and instruction for Experiment 5</td>
<td>167</td>
</tr>
<tr>
<td>44</td>
<td>Interactions between complexity and instruction for Experiment 5</td>
<td>168</td>
</tr>
<tr>
<td>45</td>
<td>Main effects of decisional complexity for Experiment 6</td>
<td>176</td>
</tr>
<tr>
<td>46</td>
<td>Main effects of stimulus set for Experiment 6</td>
<td>177</td>
</tr>
<tr>
<td>47</td>
<td>Main effects of instruction for Experiment 6</td>
<td>178</td>
</tr>
<tr>
<td>48</td>
<td>Main effects of condition for Experiment 6</td>
<td>179</td>
</tr>
<tr>
<td>49</td>
<td>Interactions between complexity and stimulus set for Experiment 6</td>
<td>180</td>
</tr>
<tr>
<td>50</td>
<td>Interactions between complexity and instruction for Experiment 6</td>
<td>181</td>
</tr>
<tr>
<td>51</td>
<td>Interactions between complexity and condition for Experiment 6</td>
<td>182</td>
</tr>
<tr>
<td>52</td>
<td>Interactions between stimulus set and instruction for Experiment 6</td>
<td>183</td>
</tr>
<tr>
<td>53</td>
<td>Interactions between stimulus set and condition for Experiment 6</td>
<td>184</td>
</tr>
<tr>
<td>54</td>
<td>Interactions between instruction and condition for Experiment 6</td>
<td>185</td>
</tr>
<tr>
<td>55</td>
<td>Interactions between complexity, instruction, and condition</td>
<td>186</td>
</tr>
<tr>
<td>56</td>
<td>Interactions between set, instruction, and condition for Experiment 6</td>
<td>187</td>
</tr>
<tr>
<td>57</td>
<td>Modelled main effects of decisional complexity for Experiment 6</td>
<td>203</td>
</tr>
<tr>
<td>58</td>
<td>Modelled interaction between complexity and set for Experiment 6</td>
<td>204</td>
</tr>
<tr>
<td>59</td>
<td>Modelled interaction between complexity and instruction</td>
<td>205</td>
</tr>
<tr>
<td>60</td>
<td>Modelled interaction between complexity and speed/accuracy condition</td>
<td>206</td>
</tr>
<tr>
<td>61</td>
<td>Modelled interaction between set and instruction for Experiment 6</td>
<td>207</td>
</tr>
<tr>
<td>62</td>
<td>Modelled interaction: complexity, instruction, and condition</td>
<td>208</td>
</tr>
<tr>
<td>63</td>
<td>Modelled interaction: set, instruction, and condition: Accuracy</td>
<td>209</td>
</tr>
<tr>
<td>64</td>
<td>Modelled interaction: set, instruction, and condition: Speed</td>
<td>210</td>
</tr>
</tbody>
</table>
List of Appendices

Appendix A: Consent Form and Debriefing Form ......................................................... 237
Appendix B: EPIC Model MATLAB Code (Experiment 2 Version) ......................... 239
Chapter 1: Historical Overview

Introduction

A century’s worth of attention has been paid to two-alternative forced-choice (2AFC) decision-making tasks (an early example can be found in Cattell, 1902). The study of 2AFC judgement has generated volumes of empirical data, dozens of predictive models, and numerous debates concerning the relationships between such variables as decisional response time (RT, studied as early as Martin and Müller, 1899), decisional accuracy (as early as Henmon, 1911), expressed confidence in the accuracy of one’s decisions (as early as Henmon, 1911), and the time it takes to render confidence judgements following a decision (Baranski & Petrusic, 1998).

Without doubt, the utility of an accurate 2AFC model is undeniable since there are many important and strictly binary decisions made by individuals every day: “Chunky or smooth?” “Will you marry me?” “To be or not to be?” The true value of such a model, however, would lie in its ability to explain more complex decision-making behaviours. A ‘complex decision’ is operationally defined herein as a nAFC decision, where $n \geq 2$. After all, human beings must be capable of selecting a particular element from a set containing more than two items without resorting to binary comparisons of all possible pairs of set elements. If this were not the case, a decision as simple as “picking the nicest tomato from the grocer’s tomato bin” would become dauntingly complicated. If there were as few as 20 tomatoes in the bin, a series of strictly binary comparisons would involve $(20 \ choose 2) = 190$ separate evaluations of the relative “niceness” of each
Complex decisions

tomato in all possible pairs of tomatoes. And yet, while most 2AFC models offer
important insight into the mechanics governing the simplest of decisions, their inability to
generalize will but leave shoppers stranded in the produce department.

Over the last century a fair number of researchers have generated mathematical
models predicting RTs for complex decision-making experiments, but only a handful of
these models simultaneously predict decisional accuracy (see literature review below).
Other related dependent variables have been modelled by substantially fewer researchers,
and none of these have attempted to gather the data required to thoroughly test the
predictions of their models. Hence, I found no examples of mean confidence or post-
decisional time to render confidence data in the complex decision-making literature. This
lack is curious, because several authors have noted how their models of complex
decision-making behaviour require the inclusion of a confidence rendering module in
order to allow for the generation of realistic RT patterns (Vickers, 1979; Vickers & Lee,
1998; Vickers & Lee, 2000; see also Usher & McClelland, 2001). What research has
been done will be presented in the literature review immediately following.

I will present the results of six sensory-based decision-making experiments where
decisional complexity was systematically manipulated. The data generated by these
experiments have, for the first time, set empirically derived constraints on complete
models of complex decision-making, where the term 'complete' refers to an ability to
account for mean decisional RT, accuracy, expressed confidence ratings, and post-
decisional time to render confidence data. Following my presentation of Experiments 1-
3, I will describe how I have used the data generated by these experiments to test several existing models of complex decision-making behaviour as well as a new, better performing model: the Evidence accrual based on the Power-law and Information theory by Carroll model (or the EPIC model for short). Experiments 4, 5, and 6 were designed to challenge the abilities of the new model, and I present the results of these challenges.

**Decisional Complexity and Decisional RT**

Hick (1952) was among the first researchers to suggest how simple binary decision-making processes could be used to generate complex decisions. According to Hick's serial elimination model, decision makers attack a complex decision by sequentially comparing pairs of groups of stimuli. Following each comparison an entire group is barred from further consideration. As such, the number of comparisons required to make a final decision is limited to $k_m = \log_2(m + 1)$ comparisons of increasing specificity, where $m$ represents the number of stimuli being compared, and the $+1$ represents an interval of uncertainty situated between each comparison wherein the decision maker has to decide whether or not the stimuli can yet be perceived so that when $m = 0$ the decision maker makes $k_m = 0$ sub-decisions. To return to the example cited in the Introduction, a produce shopper governed by Hick’s logic, and faced with a bin of $m = 20$ tomatoes, would on average only ever need to make 4.392 comparisons before finding his salad ingredient, presumably by first dividing the pile in half based on some desirable feature characteristic, then sorting this better half into two groups based on
some other feature, and continuing in this fashion until the perfect tomato is obtained.

The equation \( k_m = \log_2(m + 1) \), and the RT predictions that stem from it, have come to be known as “Hick’s Law”.

Hick suggested that the RT required for each of the \( k \) comparisons is a constant and, as such, RTs should increase logarithmically with \( m \). RT per comparison (\( RT_k \)) is a function of the information one expects to receive from an inspection of each pair of stimulus groups combined with the information one expects to lose due to imperfections of the sensory apparatus. Hick notes that, according to Shannon’s Information Theory (1948), the amount of information one can expect to receive from a source, otherwise known as entropy, is

\[
H = -\sum p_i \log_2 p_i \tag{1}
\]

where \( p_i \) represents the expected probability that event \( i \) will occur. Given this, Hick suggests that

\[
RT = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \tag{2}
\]

represents expected decisional response time, where \( H(X) \) is the entropy due to expected signal probability, \( H(Y) \) is the entropy due to expectations of sensory fallibility, and \( H(X,Y) \) is a joint entropy arising from the interaction between the stimulus and response mechanisms (note: A more thorough review of Shannon's work can be found in Chapter 5).

Hyman (1953) suggested a modification to Hick’s theory when he discovered that RT varied linearly as a function of entropy. Hyman’s manipulation of entropy was based
on Hick’s supposition that the amount of information conveyed by a stimulus is a function of the number of possible responses available to a decision maker \( (n) \), and that entropy changes when the \( p(n_i) \) are unequal. In each of three experiments Hyman found the formula \( RT = b + a(\log_2 n) \) was a better fit to his data than was Hick’s original \( RT = a(\log_2 n + 1) \).

The addition of an intercept to Hick’s model prompted considerable debate regarding the best way to formulate the intercept and coefficient of \( \log_2 n \) within the function. For instance, Rapoport (1959) demonstrated how the models suggested by both himself and by Christie and Luce (1956) can each be reduced to the form suggested by Hyman even though each has complicated and differing ideas concerning the formulation of values \( a \) and \( b \). Smith (1977), on the other hand, had his own ideas regarding these parameters as he defined the intercept and coefficient in terms of neuronal excitation. Hyman’s modification further inspired debate about whether the addition of an intercept to the model actually does provide a better description of all observed data. Welford (1960) suggested that Crossman’s (1955) data supported Hick rather than Hyman. Hick himself, through personal correspondence with Welford, questions the validity of Hyman’s model by noting Hyman uses \( n \) instead of \( n + 1 \) in his model, and pointed out how \( \log n \) will approach negative infinity as \( n \) approaches 0, which is clearly an impractical result.

In addition to the debate regarding whether an intercept should be added to Hick's model, some researchers have questioned the surface validity of the theory itself. Vickers
(1979), for example, noted that the theory cannot explain how some decisions are made more quickly and with greater ease than other entirely comparable decisions. Indeed, it has often been demonstrated that, in a 2AFC decision-making task, RTs increase as the differences between the stimuli being compared decrease (e.g., see 2AFC results for Experiments 2-4 following). Vickers also complains that the model does not predict changes in participants’ performance over time, and he questions the extent to which individuals can generate and utilize the probabilities associated with stimulus or sensory system fault occurrence, and the joint probabilities associated with these two event classes.

Although these debates call into question the validity of Hick's model and theory, the nonlinear increases in decisional RT which occur when decisional complexity increases have been observed by many others in the years following Hick's original publication (for examples see the remainder of this historical review, and the RT results in each of the six experiments following). It is also worth noting how Hick's suggestion that the rate of information processing is constant over all levels of decisional complexity, as well as his idea that Shannon's Information Theory can provide a measure of the amount of information requiring processing in a nAFC task, were successfully implemented in the EPIC model to be presented later in this thesis.

Crossman (1955), another early researcher into the effects of decisional complexity on decisional RT, had two findings which were of particular importance insofar as the present study is concerned. First, Crossman discovered via a card-sorting
task that stimulus magnitude, as represented by the number of dots on the cards, did not affect card-sorting RT so long as stimulus ratios were held constant. For example, it took participants just as long to sort cards into piles of 1 versus 2 dots as it did to sort them into piles of 5 versus 10 dots. This finding informed the generation of the stimuli used in the present study's Experiments 1-4.

Crossman’s second relevant experimental manipulation was found in his fourth described experiment (1955). In another card-sorting task, Crossman varied the ratios of dots on the cards to be sorted while simultaneously varying the number of sorting categories from two to five. By simultaneously manipulating these two variables, and observing how both manipulations reliably affected card-sorting RT, Crossman provided some of the first empirically derived data on the effects of increasing decisional complexity while simultaneously manipulating decisional difficulty. Comparable manipulations were made in the present study's Experiments 2-4.

Crossman executed these manipulations as a means to test the predictions of his 'confusion function': A model predicting a positive correlation between decisional RTs and the ease with which a correct response can be confused with other possible response alternatives. A decision maker under Crossman’s paradigm first identifies a signal as most likely belonging to a particular category, and then attempts to confirm this hypothesis by comparing it to each possible alternative category. RT, therefore, is a function of the number of comparisons to be made as well as the extent to which these categories resemble the correct category.
Without questioning the validity of Crossman’s experimental findings, Hughes (1964) points out two serious flaws with Crossman’s model. First, Crossman never explains how a decision maker can choose the correct category to serve as the standard to which the other categories are to be compared, yet Crossman nevertheless claims this decision is always error-free. In effect, Crossman has suggested decision makers always know the correct solution to every problem before they have had a chance to review any relevant evidence.

Second, Hughes demonstrates how, if contrary to Crossman’s claim decision makers occasionally choose an incorrect category as the standard, the result is absurd. Crossman proposes that

\[ RT_{\text{each category comparison}} = \frac{1}{\log X_{\text{correct category}} - \log X_{\text{compared category}}} \]  

provides an adequate description of nAFC decisional RT data, provided decision makers are able to flawlessly choose the correct stimulus to use as a comparative standard. If, perhaps more realistically, the decision maker randomly chooses a category to act as the standard but fails to choose the correct response category, she/he must inevitably make the following comparison:

\[ RT = \frac{1}{\log X_{\text{correct}} - \log X_{\text{compared}}} = \frac{1}{\log X_{\text{correct}} - \log X_{\text{correct}}} \to \infty \]

which represents a time consuming mistake indeed (Hughes, 1964)!

Though Crossman’s model is demonstrably flawed, his assertion that RTs should increase as a combined function of decisional difficulty and decisional complexity is
certainly testable and was replicated, though somewhat indirectly, by Kaswan and Young (1965).

Stimuli in the Kaswan and Young experiment consisted of arrays of 1/4-inch long vertical lines arranged horizontally on blank cards. The spaces between these lines were varied systematically so that the lines could be categorized as either evenly spaced, unevenly spaced, spaced in groups of 3, or spaced in a 1-2 pattern. While the primary focus of the experiment was to examine the relationship between length of exposure to a stimulus and a participant’s ability to correctly categorize it, the experiment nevertheless hinted at the validity of Crossman’s predictions. The experimenters demonstrated how participants were able to more rapidly categorize stimuli when only two categories of stimuli were presented in a block compared to when three category blocks (next fastest) or four category blocks (slowest) were presented. Furthermore, they showed participants always took the least time to categorize evenly spaced stimuli and always took the most time to categorize the unevenly spaced stimuli. While it is not immediately obvious why evenly spaced stimuli should be more rapidly recognizable than other ordered patterns, it might be argued that the unevenly spaced stimuli were more easily confused with the “3” or “1-2” patterned stimuli. As such, the differences between these groups might represent a manipulation of decisional difficulty. If so, this experiment would seem to confirm both of Crossman’s predictions.

More recently, Leite and Ratcliff (2010) provided the first published direct test of Crossman’s assertions. Stimuli in these published experiments were white letters...
Complex decisions displayed on a black background 90 pixels high by 90 pixels wide. Easy, medium, and hard difficulty levels were generated by “flipping” 28.5%, 32.1%, or 36.1% of the black/white pixels in each display, thereby adding noise to each stimulus presentation. Following a brief (160 ms) exposure to each stimulus, participants decided which member of a set of possible alternative letters was being displayed. Valid response sets contained either two, three, or four letters. Confirming Crossman’s predictions, these researchers found effects of both decisional complexity and decisional difficulty on RT, but found no significant interaction between these variables.

The present dissertation’s Experiments 2-4 also provide a direct test of the Crossman hypotheses by directly and unambiguously manipulating both decisional complexity and decisional difficulty.

Unlike Hick and Crossman, some early theorists speculated on the nature of complex decision-making processes without gathering the empirical data required to support their hypotheses. Christie and Luce, for example, did not gather any data to support either of the models they presented in their 1956 paper and, as a result, subsequent researchers who did test their models found them to be poor fits of observed RT data (Welford, 1960; see also Welford, 1971). Nevertheless, Christie and Luce deserve credit for generating two hypotheses which have informed the work of other psychophysicists.

First, having assumed that any complex decision requires a preceding set of more elementary decisions, they suggested that observed decisional RT is a member of a
Complex decisions

normally distributed set of possible decisional RTs. The shape of this distribution is itself affected by the RT distributions of the set of elementary decisions. In conceptualizing decisional RT not as something arrived at linearly via a continuous evidence accrual process but rather as an interplay between the probability distributions of potential RTs, these researchers paved the way for such models as Usher and McClelland’s leaky competing accumulator model (LCAM, 2001) which allow random noise to play a factor in the derivation of predicted decisional RTs (for more on this model, see Chapter 4).

Second, Christie and Luce suggest that elementary decisions might not be performed sequentially, but might instead be conducted in parallel to one another. Although Welford found no evidence of parallel processing (1960, 1971), Christie’s and Luce’s parallel processing hypothesis informs several more modern models of complex decision-making process, including those of Vickers (1971), Vickers and Lee (1998, 2000), and Usher and McClelland (2001; see descriptions following), as well as the EPIC model to be described shortly.

**Decisional Complexity, Decisional RT, and Confidence**

This is not to say that decisional RT was the only dependent variable of interest to early modellers of decision-making behaviours. Henmon (1911), for example, asked his participants to provide trial-by-trial confidence ratings, and it has since become fairly standard practice in 2AFC decision-making experiments to report the means of such
Complex decisions

ratings (e.g., Juslin & Olsson, 1997; Baranski & Petrusic, 1998; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004).

A post-decisionally expressed confidence rating is here defined as a subjective evaluation of the integrity of the decision-making process itself. Mean confidence ratings are robustly negatively correlated with decisional difficulty and decisional RT (e.g., Carroll, 2006; first noted by Garrett, 1922 and subsequently by Volkmann, 1934), and participants in decision-making experiments seem to have little difficulty reliably expressing mean confidence ratings using either an ordinal scale or an interval scale (e.g., Carroll, Petrusic, & Leth-Steensen, 2009; see also the initial study by Audley, 1964). In other words, confidence ratings are not randomly generated. They have been shown time and again to be reliable, and predictable, things.

It remains unclear which aspects of the decision-making process influence the generation of confidence ratings, though a number of theories have been suggested. One such theory suggests that confidence is simply a post hoc scaling of decisional RT (e.g., Volkmann, 1934; Reed, 1951). According to this theory, decisions which take a relatively long time to make result in low confidence ratings, while comparably short decisions result in high confidence ratings. Unfortunately, this theory is not supported by the totality of empirically derived data. Petrusic and Baranski (2009), for example, found that calibration and resolution measures derived from mean expressed confidence ratings were not entirely comparable to calibration and resolution measures derived from mean expressed RT ratings. These latter ratings were drawn from an experiment where, instead
Complex decisions

of asking participants to render a confidence judgement, participants were asked to rate 'how long the decision took'.

Another theory suggests that confidence is a function of some physical aspect of the judged stimuli themselves (e.g., Vickers, 1979). Garrett (1922), for example, reported that confidence ratings tend to be positively correlated with degree of difference between two compared stimuli. In other words, easy decisions generated high confidence ratings while difficult decisions generated low confidence ratings. This common sense notion has proven to be a very reliable and easily replicable finding (for one of many examples of this phenomenon see Carroll, 2006).

Regardless of what aspect of the decision-making process is actually being most accurately reflected by these confidence ratings, confidence judgements seem to be demonstrably principled evaluations. As such, a number of researchers have suggested that decision makers use confidence ratings themselves to regulate decision-making performance. The theory goes like this: If, having made a series of decisions, a decision maker notices that she/he has repeatedly expressed low confidence in the accuracy of their judgements, the decision maker will deliberately take more time to make subsequent decisions in an attempt to increase both decisional accuracy and confidence. Vickers’ (1979) adaptive parallel eventuation process (APEP), as illustrated in Figure 1, will serve as an example of a model built upon this assumption.
Complex decisions

Figure 1. A simplified rendition of Vickers' (1979) APEP model. A stimulus variable V is compared to various standards S, one for each viable response alternative, and evidence is accrued accordingly. When criterion C amount of evidence supporting one of the S is obtained, an overt decision is made. Subsequently, current level of confidence k is compared to ideal confidence level K, and C is adjusted whenever over- or under-confidence occurs.
APEP was proposed as an alternative to Hick’s and other similar models of complex decision-making. A decision maker under APEP constraints, charged with categorizing a given stimulus, simultaneously compares the presented stimulus variable, V, to each member of a series of relevant category standards stored in memory, denoted S₁, S₂, ..., Sₙ. Evidence accumulates via a series of discrete accrual events, with each event prompting a simultaneous and separate evaluation of the difference between the sampled evidence and the various Sᵢ. Evidence accrual continues until one evidence accrual counter reaches a criterion level of evidence C at which point the overt decision V = Sᵢ is made. If a particular evidence accrual event leads to several simultaneous conclusions V = Sᵢ, the decision maker decides (V = Sₓ) where Sₓ is selected at random from the set of Sᵢ for which a criterion amount of evidence has been obtained.

Following each decision made, Vickers hypothesizes that a human decision maker will reflect on how confident they are that they have made a correct choice, and will compare this current level of confidence to an ‘ideal’ level of confidence K. Though APEP does not provide a theoretical justification for the setting of parameter K, trial-by-trial confidence is operationally defined by Vickers as the numerical difference between the amount of accrued evidence supporting the decision V = S and the amount of evidence supporting the decision V ≠ S (as per Garrett, 1922). If this difference fails to exceed K (i.e., [Σ(evidence V = S) − Σ(evidence V ≠ S)] < K), then the decision maker is under-confident and the criterion amount of evidence C required to make future decisions V = S or V ≠ S is raised. As such, future decisions will require more time to
make and, consequently, the decision maker will tend to experience greater levels of confidence. Similarly, if the difference between accrued evidence counters exceeds K (i.e., \[\sum(\text{evidence } V = S) - \sum(\text{evidence } V \neq S) > K\]) then the decision maker is overconfident and the decisional criterion C is lowered. This post-decisional confidence evaluation, and the subsequent adjustment of the decisional criterion C, is fundamental to the APEP model. Vickers’ own computer-based simulations of APEP show that without this confidence-based adjustment the model will incorrectly predict faster RTs as decisional complexity increases. When the model is permitted to make confidence-based adjustments to decisional criteria, however, it begins to predict RT patterns that are consistent with Hick’s Law and with Vickers’ data. Thus the interplay between the primary decisional criterion C and the confidence criterion K are central to this model of complex decision-making.

Although Vickers’ hypotheses regarding the utility of confidence were APEP’s saving grace as far as RT predictions are concerned, empirically derived confidence data may be the model’s downfall. Past research has demonstrated that Vickers’ conceptualization of confidence as something evaluated post-decisionally is flawed. It seems confidence processing often begins before the primary decision is made (e.g., Carroll, 2006; Petusic & Baransi, 2003, 2000; Baransi & Petusic, 2001).

Vickers and Lee (1998) describe how the APEP model will, given unconstrained post hoc manipulation of the model parameters C and K, support almost any empirically derived RT or discriminative accuracy finding. As a result, APEP has almost no
predictive utility. The adaptive generalized accumulator module (AGAM, Vickers & Lee, 1998) remedies this problem by allowing confidence to constrain the manner in which these parameters are manipulated. What is more, AGAM represents a further departure from Vickers’ earlier model by allowing mean confidence ratings to be affected by the number of available response alternatives considered by the decision maker. For example, given a set of 3 possible response alternatives {A,B,C}, confidence that the correct response is A \( (\text{confidence}_A) \) is a function of the mean of the differences between the amounts of accrued evidence supporting A and the amounts of accrued evidence supporting the other response options. Specifically:

\[
\text{confidence}_A = \frac{(A_{\text{evidence}} - B_{\text{evidence}}) + (A_{\text{evidence}} - C_{\text{evidence}})}{2}
\]

(5).

As was the case with APEP, an AGAM decision maker will follow each decision with a comparison of observed confidence to an ideal level of confidence. Once again, over-confidence leads to a lessening of decisional criteria while under-confidence results in more stringent decisional criteria. An interesting departure from APEP, however, is the inclusion of a “coefficient of proportionality” that serves to regulate the control that confidence has over these criteria:

\[
x(t_{\text{under confidence}} - t_{\text{over confidence}})
\]

(6)

where \( x \) is the coefficient of proportionality and the \( t \) are accumulators keeping track of over/under confidence levels.

Although AGAM answers one of the objections raised regarding APEP, that confidence was being affected by only one of all available stores of evidence \( S_i \), the
structure of the model presents two interesting problems of its own. First, while confidence is allowed to regulate the primary decisional parameters of the model, it is unclear what mechanism or hypothesis is regulating the secondary parameters 'ideal confidence' and the 'coefficient of proportionality'. If these parameters are being set post hoc in an attempt to describe empirically derived data, then AGAM is no better at addressing the “will-describeanything-eventually” objection than was the APEP model. Vickers and Lee acknowledge this problem in their General Discussion.

Second, if Vickers and Lee have correctly deduced the origin of observed confidence in Equation 5 above, then they have predicted there will be no difference between mean expressed confidence levels between an observer making the judgement “MOST” and an observer making the judgement “2nd LEAST” in a 2AFC decision-making task, since these two instructions are semantically identical. This hypothesis was tested directly within the context of the present Experiment 4, and was rejected (see the Experiment 4 Results section following).

Although these problems call into doubt the overall validity of AGAM and APEP, it is worth noting that Vickers’ broader conceptualization of a series of n evidence accumulators acting in parallel greatly informed the EPIC model to be presented in the sections following. Unlike AGAM and APEP, however, this new model does not use confidence to establish optimum decisional criterion levels.

Usher and McClelland’s (2001) leaky competing accumulator model (LCAM) represents an attempt to merge the philosophies governing classical evidence accrual
models with a "neuronally inspired" connectionist network. The LCAM is governed by four basic principles: First, that information processing results from a series of evidence accrual events. Second, that noise results in differing quanta of information being generated for each different evidence accrual event. Third, that accrued information gathered in 'memory' stores is subject to leakage or decay. Fourth, that response alternatives “compete” with each other via a process of lateral inhibition. The first two of these premises stem from traditional models of evidence accrual. The first is inspired by models such as those presented by Vickers (1979), and the second by authors such as Christie and Luce (1956). The third premise, which posits that stored accrued evidence is subject to decay, stems from the finding that all measurable human performance must “peak” so that giving a decision maker an infinite amount of time to accrue decisional evidence will not insure p(correct) = 1.0 (see Vickers, 1979 for discussion; for an example see Carroll, 2006). The fourth premise comes from connectionist theory, which allows units in a parallel distributed processing network to either activate or inhibit one another (e.g., Rumelhart & McClelland, 1986).

The LCAM network, like many other connectionist networks, consists of a layer of input units and a layer of output units. Changes in levels of activation of the output units (dx_i) over a fixed period of time are described thusly:

\[ dx_i = \left[ I_i - \lambda x_i \right] \frac{dt}{\tau} + \varepsilon_i \sqrt{\frac{dt}{\tau}} \]  

(7)
Complex decisions

where \( I_i \) includes input from all other units to output unit \( x_i \), including sensory input units, inhibition from other units, and recurrent excitatory input from the output unit back to itself. \( \lambda x_i \) describes the amount of information leakage suffered by the unit, \( dt/\tau \) represents the time interval in question, and \( \epsilon \) is simply a Gaussian noise term that reflects noise both internal to and external to the network’s “nervous system”.

The general flow of the LCAM decision-making process is illustrated within Figure 2. This process begins when the network samples a pattern of sensory input. This event, in turn, causes the activation levels of the network’s output units to change. When an output unit \( x_i \)’s activation level is greater than zero, it will inhibit every other output unit to an extent positively correlated with its degree of activation. The system will continue to sample sensory information until an output unit \( x_i \) reaches a criterion level of activation, at which time the system makes an overt response. In the time interval between the initial sensory sampling event and the overt response, levels of activation of each output unit \( x_i \) are affected by activation ‘leakage’, which is the tendency for unit \( x_i \)’s activation level to continuously decrease. This process is sometimes referred to as ‘decay’ (e.g., Leite & Ratcliff, 2010).
**Figure 2.** A simplified rendition of Usher and McClelland’s (2001) LCAM. Signal strength is calculated for each input unit $S_i$ based on the degree to which the unit is sensitive to some aspect of the stimulus $V$, and evidence is accrued accordingly. When $C$ amount of evidence supporting one of the $S_i$ is obtained, an overt decision is made.
Usher and McClelland found that in order to add additional output units to the network, thereby allowing the model to make increasingly complex decisions, they were forced to alter the system's activation parameter. Failure to change this parameter resulted in unacceptably variable decisional accuracy. When this parameter was allowed to shift, however, a constant error rate was maintained and a logarithmic relationship between the number of output units in the network and mean decisional RT emerged.

If the LCAM must maintain a constant error rate in order to satisfy Hick's Law, the model provides only a limited description of complex decision-making behaviour. Consider an experiment where a decision maker is not given trial-by-trial performance feedback. In such a case, feedback would have to be self-generated, and it is unclear how she/he would compare observed and optimal accuracy without making reference to subjective feelings of trial-by-trial confidence. There is, however, no expressly stated basis for the simulated generation of confidence in Usher and McClelland's LCAM. At best the model could use inverse RT as a means of determining confidence, where fast decisions are made with high confidence and slow decisions with low confidence, but, as has been noted, such a model would not reflect the totality of empirically derived data (Petrusic & Baranski, 2009). Nevertheless, it is worth noting that Usher, Olami, and McClelland (2002) later considered confidence to be the solution to the problem of performance without feedback when they stated "...the use of metacognitive signals such as confidence may be used by participants when feedback about performance is unavailable" (p. 712).
Furthermore, consider an experiment wherein a decision maker is presented with a series of decisions of varying decisional complexity. If decisional accuracy can be shown to systematically decrease as complexity increases, and given that Hick’s Law predicts mean RTs will increase logarithmically along with decisional complexity, then the LCAM will either incorrectly predict a fixed level of accuracy or will incorrectly fail to generate nonlinear decisional RT patterns. Comparably, if decision makers in such an experiment were not presented with trial-by-trial feedback and were instead forced to use confidence to self-evaluate performance, then LCAM predicts that mean confidence levels will be unaffected by increases in decisional complexity. As will be noted in the data analyses following all six of the conducted experiments, Usher and McClelland’s conceptualization of a fixed error rate or, conversely, a fixed mean confidence rate is untenable.

The suggestion that information decays as time passes and that evidence accumulators compete with one another via inhibitory processes was the subject of a recent investigation by Leite and Ratcliff (2010, see description below) and were also tested within the context of the present study (see sections following).

**Decisional Complexity, Decisional RT, and Decisional Accuracy**

Usher and McClelland (2001) note that models which predict both error rates and RTs are a relatively new and rare phenomenon. Indeed, there are still a number of modern researchers who choose to model error-free complex decision-making tasks.
McMillen and Holmes (2006), for example, claim to be investigating the relationship between RT and decisional accuracy, yet caution that their proposed modification to Hick’s Law should only be applied when an observed accuracy rate could be held to 95% or above. A more extreme example is found in Roe, Busemeyer, and Townsend (2001), who extended a connectionist model of nAFC decision-making RTs to the realm of preferential choice wherein the concept of decisional accuracy cannot be said to apply in any measurable way since a preference is neither accurate nor inaccurate.

Despite the dearth of models that predict the effect of complexity on both RT and accuracy, a few researchers have expressed interest in the relationship between these variables, some as early as the mid-twentieth century. For example, in Hick’s (1952) second experiment participants responded under speed stress instead of accuracy stress. This manipulation was performed in order to demonstrate how the predictions of his RT model would hold even in cases where the amount of information gained through observation of the stimulus was limited by the imposition of a deadline. This deadline was set so that his participants made errors "without abandoning altogether the attempt to make correct responses" (p. 15). Though Hick proceeded to argue that his model provided a reasonable fit for this accuracy data, the objections to his model have already been discussed.

Usher, Olami, and McClelland (2002) revisited Hick’s hypothesis that entropic-degradation is the root cause of decisional inaccuracy within the context of a study in speed-accuracy tradeoffs. Rejecting Hick’s model in favour of a race model (see Chapter
4), these researchers make the additional assumption that decision makers habitually attempt to maintain a constant level of decisional accuracy throughout the course of an experiment. In order to maintain this level of accuracy in the face of increases in decisional complexity, these judges must logarithmically increase the amount of evidence they require before an overt decision is made. This is because the probability that any given sample of evidence will support the correct decision decreases nonlinearly as decisional complexity increases. Within this framework, these researchers provide what they feel is a reasonable explanation, not only for Hick-like RT distributions, but also for the oft reported finding that speeded decisions are less accurate than more cautious ones. Specifically, they claim that speed-related errors are caused by a decision maker's inability to completely process all of the information provided by the stimuli.

Unfortunately, the results sections of all six experiments presented in the chapters following cast doubt on the basic assertions of this model, as there is no evidence to suggest decision makers attempt to maintain constant levels of decisional accuracy or, in the absence of trial-by-trial feedback, a constant level of confidence in their decisions.

Murdock (1960) was another early researcher who suggested a model for predicting errors in complex decision-making tasks, and he was the first to suggest that decisional accuracy in these tasks was a subject worth studying in and of itself. He demonstrated how, though a series of stimuli arranged along an interval scale may differ from one another by an arithmetically constant amount, stimuli closer to the ends of the scale are nevertheless more distinctive than their counterparts and will therefore be
Complex decisions correctly identified more often. Murdock’s hypothesis is perhaps better understood with reference to Table 1. Distinctiveness measures are presented in the table's second column, and are a function of the summed differences in signal strength between the given stimuli and all the other stimuli to which these might be compared. This value will always be largest for the stimuli arranged near the ends of the stimulus strength spectrum, and smallest for stimuli arranged closer towards the centre.

Table 1

An example of Murdock’s distinctiveness hypothesis.

<table>
<thead>
<tr>
<th>Stimulus Strength</th>
<th>Sum of the differences between stimulus and other stimuli</th>
<th>% of total summed stimulus differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12 (i.e., (4-2)+(6-2)+(8-2)=12)</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Total summed stimulus differences: 40</td>
<td></td>
</tr>
</tbody>
</table>

Murdock went on to demonstrate how distinctiveness decreases as more alternatives are made available to decision makers such that with two alternatives distinctiveness has a maximum value of 50%, whereas with three alternatives distinctiveness has a maximum value of 37.5%. In the four alternative example described by Table 1, distinctiveness has a maximum value of 30%. In three experiments, wherein participants had to identify tones of various decibel intensities, Murdock found $p(\text{correct})$
values fairly close to those predicted by this model. This "distinctiveness hypothesis" served as inspiration for Experiment 4 though the results unfortunately cast doubt on Murdock’s assertions (see Chapter 7).

Unfortunately, neither Hick (1952), Usher et al. (2002), nor Murdock (1960) chose to study errors caused by changes in decisional difficulty. As such, none of the theories put forward by any of these researchers can explain errors that result from decisions which are, simply put, more difficult to make than other decisions of comparable complexity.

Leite and Ratcliff (2010) presented the first published example of a sensory-based decision-making experiment wherein decisional complexity and decisional difficulty were manipulated simultaneously and their effects on decisional accuracy and decisional RT were recorded (see discussion preceding). These researchers state that the observed RT distributions were consistent with those predicted by Hick, and they also discovered decreases in decisional accuracy whenever either decisional complexity or decisional difficulty was increased. Interestingly, these researchers decided that the model best able to describe their observed results is one which, contrary to LCAM predictions, neither “leaked” nor “competed” and, contrary to the predictions of Usher et al. 2002, required no alterations to the required levels of accrued evidence regardless of level of decisional complexity. Leite and Ratcliff’s modelling method will be discussed and replicated in Chapter 5.
Decisional Complexity and Post-Decisional Times to Render Confidence

As already stated, I have found no examples in the literature of a sensory-based, decision-making experiment wherein decisional complexity was manipulated and time to render confidence judgements were recorded. This is surprising, since such data would provide support for those who suggest confidence is processed post-decisionally, including Vickers (1979), Vickers and Lee (1998), and, arguably, Usher and McClelland (2001). If observed times to render confidence are flat and fast over all levels of decisional complexity, then, contrary to the predictions of these researchers, confidence is likely being processed prior to the expression of an overt decision since very little additional post-decisional time is required before confidence is expressed. As will be seen in the results sections of Experiments 1-6, post-decisional processing is not always observed, which suggests confidence is processed in parallel to the primary decision (note that this has been demonstrated within the context of a 2AFC paradigm: See Baranski & Petrusic, 1998).
Chapter 2: General Method for Experiments 1 – 3

General Purpose

In the present research, data obtained experimentally were used to provide a record of participant behaviour in a sensory-based decision-making task wherein decisional complexity was manipulated. An iterative program was subsequently used to determine which of a number of different models could provide the best description of this observed data. Leite and Ratcliff (2010) used a similar approach, however the present research diverged from Leite and Ratcliff's methodology in that no limits were placed on the length of time participants could choose to view each presented stimulus and, following each decision, participants were asked to indicate how confident they were that they had made a correct choice. As such, these experiments represent the first where decisional RT, decisional accuracy, mean confidence, and mean post-decisional time to render confidence judgements were all collected in a series of complexity experiments. The resulting dataset, therefore, provides a more comprehensive benchmark for those who wish to model complex decision-making behaviour than do those datasets which have preceded it.

Apparatus

All three experiments were conducted using a desktop computer with a standard colour monitor. The computer was equipped with a Pentium-class processor and a Windows 98 operating system. Stimulus presentation and response data collection was
controlled by Superlab Pro v.2.0. Participant responses were made via a control panel fashioned out of a standard desktop keyboard with all but 11 keys removed. Bristol board was used as a casing for the response panel, with the labels “BLACK”, “RED”, “GREEN”, and “BLUE” written in appropriately coloured ink beneath four keys arranged along the bottom of the response panel. The labels “X”, “50/60/70/80/90/100”, “33/46/60/73/86/100”, and “25/40/55/70/85/100” were written in black ink beneath the remaining seven keys aligned along the top of the response panel. The words “GUESS” and “CERTAIN” were written in black ink above the top row’s second and seventh buttons respectively. The layout of the response panel is illustrated in Figure 3.

![Response Panel Layout](image)

*Figure 3.* The layout of the response panel used in Experiments 1-4.

**Materials / Stimuli**

Stimuli were 20×20 arrays of 20 pixel×20 pixel squares. Each square was coloured either red (RGB value: 255, 0, 0), green (RGB value: 0, 255, 0), blue (RGB
value: 0, 0, 255), or black (RGB value: 0, 0, 0). All squares were surrounded by a 2 pixel wide white (RGB value: 255, 255, 255) border. In total, each array was 440 pixels wide by 440 pixels high. Stimulus examples are provided (see Figure 4).

**Figure 4.** Stimulus examples. Panel A is an example of a two alternative decision, Panel B is an example of a three alternative decision, and Panel C is an example of a four alternative decision. Stimuli illustrated here are not to scale.

Stimuli were generated via a program written by the author. Each stimulus was comprised of a matrix of squares each of which were either one of two colours (Figure 4, Panel A), one of three colours (Figure 4, Panel B), or one of four colours (Figure 4, Panel C). The number of colours in each stimulus corresponded to the number of possible valid responses available to a participant judging the array. As such, there were three levels of decisional complexity in each experiment: 2AFC, 3AFC, and 4AFC.

Since the task for participants in these studies was to decide, for each colour array, which colour was represented either most often or least often depending on instruction (see Procedure below), manipulations of decisional difficulty were achieved.
Complex decisions

by varying the ratios between groups of coloured squares. For example, a very easy
2AFC decision would be one where only 10% of the squares were red while 90% were
black: A ratio of 1:9. An impossibly difficult 2AFC decision would be one where exactly
50% of the squares were red while 50% were black: A ratio of 1:1.

For all stimuli in Experiment 1, and for all ‘medium’ difficulty stimuli in
Experiments 2 and 3, the ratio between groups of coloured squares was held as nearly as
possible to 1:1.08 for all stimuli. Within the two coloured stimuli there were 208 squares
of one colour and 192 squares of a different colour (208 + 192 = 400; 208 / 192 =
1.0833). Within the three coloured stimuli there were 144 squares of one colour, 133
squares of a second colour, and 123 squares of a third colour (144 + 133 + 123 = 400;
144 / 133 = 1.0827, 133 / 123 = 1.0813). Within the four coloured stimuli there were 108
squares of one colour, 100 squares of each of two other colours, and 92 squares of the
final colour (108 + 100 + 100 + 92 = 400; 108 / 100 = 1.08, 100 / 92 = 1.0869).

For the ‘easy’ difficulty stimuli in Experiments 2 and 3, the between colour ratio
was held as nearly as possible to 1:1.16. Within the two coloured stimuli there were 215
squares of one colour and 185 squares of a different colour (215 + 185 = 400; 215 / 185 =
1.1622). Within the three coloured stimuli there were 154 squares of one colour, 132
squares of a second colour, and 114 squares of a third colour (154 + 132 + 114 = 400;
154 / 132 = 1.1666, 132 / 114 = 1.1578). Within the four coloured stimuli there were 116
squares of one colour, 99 squares of each of two other colours, and 86 squares of the final
For the 'hard' difficulty stimuli in Experiments 2 and 3, the between colour ratio was held as nearly as possible to 1:1.04. Within the two coloured stimuli there were 204 squares of one colour and 196 squares of a different colour (204 + 196 = 400; 204 / 196 = 1.0408). Within the three coloured stimuli there were 139 squares of one colour, 133 squares of a second colour, and 128 squares of a third colour (139 + 133 + 128 = 400; 139 / 133 = 1.0451, 133 / 128 = 1.0391). Within the four coloured stimuli there were 104 squares of one colour, 100 squares of each of two other colours, and 96 squares of the final colour (104 + 100 + 100 + 96 = 400; 104 / 100 = 1.04, 100 / 96 = 1.0417).

Four practice session stimuli, one for each colour, were also generated. These practice arrays consisted of 20 x 20 arrays of squares comprised of homogenously coloured elements.

Procedure

Participants were seated before the computer monitor within comfortable reach of the response panel. In order to familiarize participants with the layout of the response panel, and in order to insure participants could distinguish between the colours used in these experiments, subjects were asked to make 20 'practice decisions' before beginning the experiment itself. These practice decisions were made using the four practice stimuli, each of which was presented five times in a random order. The task for participants making these decisions was to choose the colour that best described each practice stimulus. Participants were encouraged to emphasize accuracy over speed during these
practice blocks of trials. Practice blocks were repeated until participants were able to respond to an entire block of 20 trials without making any mistakes. All but two of the participants were able to do this on their first attempt and no participant required more than two practice blocks of trials.

**Figure 5.** Within trial stimulus display presentation order: Instruction precedes stimulus by 500 ms, and both are replaced by the confidence prompt once an overt decision is made.

Prior to the presentation of the experimental blocks of trials participants were shown examples of the experimental stimuli. Participants were told that following each stimulus presentation they would be asked to make a decision corresponding to an instruction appearing on the screen. If the instruction read “MOST”, the participant’s
task was to decide which colour was represented most often in the array of colours. Alternatively, if the instruction read “LEAST”, participants were told to select the colour that was represented least often in the display. Instruction presentation preceded stimulus presentation by 500 ms, and instructions remained visible until a decision was made (see Figure 5).

Following each decision participants were asked to express their level of confidence that they had just made a correct choice. Participants were told to press the button on the response panel that best corresponded to this evaluation. The “GUESS” button was to be used when participants were guessing, the “CERTAIN” button was to be used when participants were positive they had not made a mistake, and the buttons in-between these extremes were to be used if participant confidence lay somewhere between a guess and certainty. The numbers written beneath the confidence response buttons on the response panel (see Figure 3) corresponded to probabilities of having made a correct choice. For example, when guessing in a 2AFC task there is a 50% probability that a correct choice will be made. Similarly, when guessing in a 3AFC task there is a 33% probability that a correct choice will be made, and when guessing in a 4AFC task there is a 25% probability that a correct choice will be made. Hence the numbers 50, 33, and 25 were written beneath the button “GUESS”. While these probability values were explained to participants, they were also told to use the mnemonic “low numbers mean less certainty, high numbers mean more certainty, press ‘GUESS’ if you’re guessing, press ‘CERTAIN’ if you’re positive you got it right”. Participants were also told to use
the button marked “X” if they were certain they had made a mistake. Note that, for ease of comparison, in the confidence related graphs and analyses in the sections that follow all confidence values have been converted to their 2AFC equivalents. For example, 3AFC guesses were given a value "50" before being analyzed. These 2AFC values, and their 3AFC and 4AFC equivalents, are noted in Figure 3.

Participants

Participants were drawn from a pool of students enrolled in first-year psychology classes at Carleton University. 17 female and 7 male students completed Experiment 1, 14 female and 10 male students completed Experiment 2, and 16 female and 8 male students completed Experiment 3. The median age of the participant pool was 20.23 years. Three participants failed to complete Experiment 3 and were replaced.

Analytical Method and Results

In all three experiments, trials where RT was less than 200 ms and trials where RT was more than 3 standard deviations above each participant's mean within each level of decisional complexity and decisional difficulty were removed from analysis. Removed trials accounted for 1.82% of the observations made in Experiment 1, 2.33% of the observations made in Experiment 2, and 2.09% of the observations made in Experiment 3. Block order, the order in which levels of complexity were presented to participants, was included in all analyses as a between subject factor but, as this factor...
never achieved significance, the results of any between subjects analyses are not reported. This was also true of the set order variable in Experiment 3. In all analyses of variance (ANOVAs) presented herein, in order to protect against violations of the assumption of sphericity, Huynh-Feldt degrees of freedom (df) were used to assess statistical significance though, for the sake of clarity, the df values actually reported are defined by the experiments themselves. All p(correct) values used in all ANOVAs were adjusted using Berkson’s correction (1953), which was used to attenuate extreme values of 0.0 and 1.0, and then subjected to an arcsine transformation in order to eliminate potential ceiling effects. The values plotted in the figures and tables, however, are the actual values observed. 95% confidence intervals were calculated, where applicable, using the Loftus and Masson (1994) method for within-participant designs. Post-hoc analyses of significant main effects consisted of a comprehensive series of all possible pairwise comparisons within each effect, with the appropriate Bonferroni adjustment to the alpha-per-comparison made within each family of comparisons as required to maintain a family-wise significance level of .05. Post-hoc analyses of significant interactions consisted of the derivation of Tukey’s HSD values and the subsequent evaluation of all possible pairwise comparisons of all cells involved in the interaction. While results will be discussed more thoroughly in the sections following, for the sake of convenience, the results for all analyses are compiled in Table 2.
Table 2. **Statistical results from analyses of Experiments 1-3.**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Response Time</th>
<th>Accuracy</th>
<th>Confidence</th>
<th>Time to Decide Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>F</td>
<td>(\eta^2)</td>
<td>F</td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 36</td>
<td>21.63***</td>
<td>.55</td>
<td>42.52***</td>
</tr>
<tr>
<td>Complexity x Difficulty</td>
<td>4, 72</td>
<td>1.87</td>
<td>.09</td>
<td>8.88***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 36</td>
<td>13.12***</td>
<td>.42</td>
<td>237.38***</td>
</tr>
<tr>
<td>Difficulty</td>
<td>2, 36</td>
<td>14.30***</td>
<td>.44</td>
<td>110.10***</td>
</tr>
<tr>
<td>Complexity x Difficulty</td>
<td>4, 72</td>
<td>1.87</td>
<td>.09</td>
<td>8.88***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 24</td>
<td>25.04***</td>
<td>.68</td>
<td>371.49***</td>
</tr>
<tr>
<td>Difficulty</td>
<td>2, 24</td>
<td>12.68***</td>
<td>.51</td>
<td>182.19***</td>
</tr>
<tr>
<td>Confidence Requirement</td>
<td>1, 12</td>
<td>9.13**</td>
<td>.43</td>
<td>3.88</td>
</tr>
<tr>
<td>Complexity x Difficulty</td>
<td>4, 48</td>
<td>2.99*</td>
<td>.20</td>
<td>7.93***</td>
</tr>
</tbody>
</table>

**Note.** Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see text); While non-significant main effects are reported, only significant interactions are reported. *** \(p < .001\); ** \(p < .01\); * \(p < .05\).
Complex decisions

Chapter 3: Experiments 1 – 3

Experiment 1

Purpose. The stimuli described in Chapter 2 are novel and, as such, it was important to ascertain whether they would be suitable for use in a series of experiments designed to examine the affects of manipulating decisional complexity on decisional RT, decisional accuracy, confidence, and post-decisional times to render confidence. The primary purpose of Experiment 1 was to determine whether participants would be able to make MOST/LEAST judgements based on these stimuli while maintaining above chance levels of performance. If so, a second goal of this experiment was to determine whether the Hick-like pattern of decisional RTs would emerge as a result. If both goals were met, then the experimental paradigm described in Chapter 2 would be an acceptable tool for investigating the decisional complexity phenomenon first observed by Hick (1954).

Beyond these more practical goals, this experiment represents the first time all four dependent variables of interest to those who study models of sensory-based decision-making processes were recorded within the context of a decisional complexity experiment.

Method. Participants were asked to make three blocks of decisions regarding the stimuli described in Chapter 2, one block for each level of decisional complexity. Block order was counterbalanced between participants. 48 three colour stimuli were created: 4 colours $\text{Permute}$ 3 colours = 24 possible permutations, and $24 \times 2$ instructions (MOST/LEAST) = 48 stimuli. In order to insure that participants did not make a fewer number of 2AFC decisions than 3AFC decisions, a comparable number of two colour
stimuli were created: \(4P^2 = 12\), \(12 \times 2\) instructions = 24, and \(24 \times 2\) replications of each stimulus = 48 stimuli. 48 four colour stimuli were also created in a similar fashion. Note that, although all four colours were used in the generation of the four colour stimuli, and \(4P^4 = 24\), only \(4P^2 = 12\) stimulus variants are possible since only one colour can be “MOST” and only one colour can be “LEAST”, while the remaining colours make up a middle 'noise' category wherein neither middle colour is represented more often than its pair. A completely permuted set of four colour stimuli were used in Experiment 4 (see Chapter 6 following). In total, each participant was asked to make \(48 \times 3 = 144\) decisions within an experimental session that lasted approximately 40 minutes.

**Results.** Experimental results are illustrated in Figure 6 and listed in Table 3.

**Mean decisional response time.** A significant effect of decisional complexity on mean decisional RT was found. Importantly, a Hick-like effect of decisional complexity on mean decisional RT was observed in that RT increased in a nonlinear fashion along with increases in the number of available alternative responses (Figure 6, top-left panel). In other words, the difference in mean decisional RT between the 3AFC and 2AFC levels of complexity, \(12011.5\) ms – \(6967.6\) ms = \(5043.9\) ms, was less than the difference in mean decisional RT between the 4AFC and 3AFC levels of complexity, \(15013.7\) ms – \(12011.5\) ms = \(3002.2\) ms. Post-hoc tests revealed significant differences between all compared mean decisional RTs.

**Mean decisional accuracy.** A significant effect of decisional complexity on mean decisional accuracy was also found, and post-hoc tests revealed significant differences between all compared pairs. It is worth noting that decisional accuracy,
though it decreased as more alternative responses were made available to participants, nevertheless remained above chance performance in all cases: 62% vs. 50% for 2AFC, 53% vs. 33% for 3AFC, and 43% vs. 25% for 4AFC decisions (see Figure 6, top-right panel).

**Mean confidence rating.** Mean reported confidence decreased reliably along with increases in decisional complexity (Figure 6, bottom-left panel), though comprehensive post-hoc tests revealed that the difference between 2AFC and 3AFC confidence failed to reach significance, \( t(21) = 1.65, p > .1 \). It is interesting to note that, contrary to the predictions of Vickers (1979), Vickers and Lee (2000), and Usher and McClelland (2001), neither confidence nor mean decisional accuracy remained constant in the face of changes in decisional complexity. Both confidence and accuracy were decreasing as decisions became more complex, yet participants were not taking the extra time required to bolster performance when making these more challenging decisions. Nevertheless, this finding implies that observed confidence ratings were carefully considered expressions of a real sensation, and were not being randomly generated by decision makers.

**Mean time to render confidence ratings.** Post-decisional times to render confidence were fast, with less than a second of post-decisional processing time being required before confidence ratings were expressed, and were not reliably affected by manipulations of decisional complexity (Figure 6, bottom-right). This finding suggests that confidence was being processed in parallel to the primary decision, that confidence
Complex decisions

processing was completed prior to the expression of the overt decision, and that
confidence ratings accurately reflected task performance.

Discussion. Taken together, these findings suggest that the stimuli were
effective: Accuracy decreased as complexity increased, but never fell below chance, and
observed decisional RT patterns were comparable to those observed by Hick (1954). As
such, I felt justified in proceeding to Experiment 2 wherein I replicated and extended
Experiment 1 by including decisional difficulty as an additional independent variable.
Figure 6. Experimental results from Experiment 1. Upper-left: the effect of changes in decisional complexity on decisional RT. Upper-right: the effect of changes in decisional complexity on decisional accuracy. Lower-left: the effects of changes in decisional complexity on mean confidence. Lower-right: the effect of changes in decisional complexity on post-decisional times to express a confidence rating. 95% confidence intervals are shown.
Complex decisions

Table 3. *Statistical results from analyses of Experiment 1.*

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Response Time</th>
<th>Accuracy</th>
<th>Confidence</th>
<th>Time to Decide</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>F</td>
<td>$\eta_p^2$</td>
<td>F</td>
<td>$\eta_p^2$</td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 36</td>
<td>21.63***</td>
<td>.55</td>
<td>42.52***</td>
<td>.70</td>
</tr>
</tbody>
</table>

*Note.* Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see Chapter 2); *** $p < .001$; ** $p < .01$; * $p < .05$. 

44
Experiment 2

**Purpose.** Experiment 2 replicated and extended the findings of Experiment 1 by adding a manipulation of decisional difficulty. Note that the medium-level difficulty stimuli used in this experiment were the same stimuli designed for use in Experiment 1. As such, Experiment 2 contained an exact replication of Experiment 1.

**Method.** As was the case in Experiment 1 Participants were asked to make three blocks of decisions, one for each level of decisional complexity. Block order was counterbalanced between participants. 48 three colour stimuli were created for each of three levels of decisional difficulty, and a comparable number of two and four coloured stimuli were also generated. In total, each participant was asked to make $48 \times 3 \times 3 = 432$ decisions within an experimental session that lasted approximately 90 minutes. Manipulations of decisional difficulty were governed by the methods described in Chapter 2.

**Results.** The results of the experiment can be found in Figure 7 and Table 5.

**Mean decisional response time.** As will become apparent following a review of Figure 7 (top-left), both decisional difficulty and decisional complexity had a significant effect on mean decisional RT. Predictably, participants took longer to make hard-level decisions (mean RT = 6787.4 ms) than it took them to make either medium-level decisions (mean RT = 6683.7 ms) or easy-level decisions (mean RT = 6088.7 ms), though post-hoc tests revealed that the difference in decisional RT between medium and hard level decisions was not reliable, $t(21) = .97, p > .2$. The Hick effect was also observed here: At every level of decisional difficulty, the difference in mean decisional
Complex decisions

RT between the 3AFC and 2AFC levels of complexity was less than the difference in mean decisional RT between the 4AFC and 3AFC levels of complexity (see Table 4). It is worth noting that post-hoc tests revealed how the overall differences between 3AFC and 4AFC mean decisional RTs failed to reach significance, \( t(21) = 1.07, p > .2 \) and, as such, these distributions of mean decisional RTs are nonlinear and are of the type predicted by Hick. The interaction between complexity and difficulty on decisional RT was not reliable.

Table 4. The emergence of Hick-like RT patterns at each level of decisional difficulty.

<table>
<thead>
<tr>
<th>Easy Decisions</th>
<th>Medium Decisions</th>
<th>Hard Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT_3 - RT_2 = 2262.5 ms</td>
<td>RT_4 - RT_3 = 879.0 ms</td>
<td>RT_3 - RT_2 = 2785.8 ms</td>
</tr>
<tr>
<td>RT_4 - RT_3 = 672.4 ms</td>
<td>RT_4 - RT_3 = 2810.5 ms</td>
<td>RT_4 - RT_3 = 1248.7 ms</td>
</tr>
</tbody>
</table>

Interestingly, participants were much quicker making their decisions in Experiment 2 (grand mean RT: 6519.7 ms) than participants in Experiment 1 (grand mean RT: 11330.9 ms). This finding will be discussed in greater detail in the General Discussion section following the results section of Experiment 3.

Mean decisional accuracy. The manipulation of decisional difficulty had a predictable effect in that, at all levels of decisional complexity, participants were more accurate making easy decisions than medium-level or hard-level decisions. Post-hoc tests revealed significant differences in levels of accuracy between all levels of decisional difficulty and between all levels of decisional complexity. A significant interaction between complexity and difficulty on decisional accuracy was also observed, in that
participants seemed to be equally accurate when making 2AFC medium-level and hard-level decisions (note the overlapping confidence intervals in Figure 7, top-right), but performed differently making medium-level and hard-level 3AFC and 4AFC decisions (again, Figure 7, top-right). This supposition was borne out by a series of Tukey's HSD post-hoc comparisons, which revealed that the only difference between means which failed to reach statistical significance was the difference between 2AFC medium level accuracy (59.48%) and hard level accuracy (58.64%) decisions, HSD = 5.07%, q(.05, 9, 72) = 4.536. Importantly, participants exhibited greater than chance levels of performance at all levels of decisional difficulty and complexity, a finding which implies participants were attempting to make correct decisions.

**Mean confidence rating.** Both decisional difficulty and decisional complexity had a reliable effect on mean confidence ratings (Figure 7, bottom-left). Predictably, participants expressed reliably lower confidence for more difficult decisions as well as for more complex decisions. Post-hoc tests revealed significant differences in mean confidence ratings between all levels of decisional difficulty and between all levels of decisional complexity. As was the case in Experiment 1, no apparent attempt was made by participants to maintain either a constant level of confidence or a constant level of accuracy.

**Mean time to render a confidence rating.** Post-decisional times to render confidence were fast and not reliably affected by either manipulations of decisional complexity or decisional difficulty (Figure 7, bottom-right). Again, this suggests that confidence ratings were available to decision makers as soon as an overt decision had
Complex decisions have been expressed. And, since confidence ratings were reliably affected by manipulations of difficulty and complexity, these fast post-decisional response times were not the product of participants randomly selecting confidence ratings.

**Discussion.** It would appear that, overall, the inclusion of decisional difficulty had the intended effect. Generally speaking, performance slowed, became less accurate, and confidence decreased as decisions became more difficult: Phenomena which cannot be explained by Vickers (1979), Vickers and Lee (2000), and Usher and McClelland (2001). Since this manipulation of decisional difficulty proved successful, the dataset derived from this experiment can serve as a standard to which subsequently designed models can be compared: It contains a manipulation of both complexity and difficulty, and it provides a measure of decisional RT, accuracy, confidence, and times to express confidence. Models of this data would have to produce nonlinear plots of decisional RTs which are faster for easy decisions, yield accuracy and confidence rates which are affected by both decisional complexity and decisional difficulty, and generate post decisional times to express confidence which are fast and platykurtic.

Experiment 3 replicated and extended the findings of Experiment 2 by having participants make an additional set of decisions where complexity and difficulty were manipulated as in Experiment 2, but where no confidence judgements were required. As such, Experiment 3 was designed to determine whether the RT and accuracy effects noted in Experiment 2 were simply artefacts of the extraneous requirement of having to rendering confidence judgements following every decision.
Figure 7. Experimental results from Experiment 2. Upper-left: the effect of changes in decisional complexity on decisional RT. Upper-right: the effect of changes in decisional complexity on decisional accuracy. Lower-left: the effect of changes in decisional complexity on mean confidence. Lower-right: the effect of changes in decisional complexity on post-decisional times to express a confidence rating. 95% confidence intervals are shown.
Table 5. *Statistical results from analyses of Experiment 2.*

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Response Time</th>
<th>Accuracy</th>
<th>Confidence</th>
<th>Time to Decide Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>$F$</td>
<td>$\eta_p^2$</td>
<td>df</td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 36</td>
<td>13.12***</td>
<td>.42</td>
<td>237.38***</td>
</tr>
<tr>
<td>Difficulty</td>
<td>2, 36</td>
<td>14.30***</td>
<td>.44</td>
<td>110.10***</td>
</tr>
<tr>
<td>Complexity x Difficulty</td>
<td>4, 72</td>
<td>1.87</td>
<td>.09</td>
<td>8.88***</td>
</tr>
</tbody>
</table>

*Note.* Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see Chapter 2); *** $p < .001$; ** $p < .01$; * $p < .05$. 
Experiment 3

**Purpose.** With regards to models of decisional complexity, psychophysicists have primarily been concerned with modelling decisional RT and accuracy. While some of these models do make reference to confidence, it must again be stressed that this author has not been able to find any data within this body of literature specifically concerned with the effects of asking participants to render confidence judgements. As such, since the effect of adding a confidence requirement to a paradigm which systematically varies decisional complexity has not yet been established, it seemed prudent to determine whether there were any unintended confidence-related consequences to the paradigm employed in Experiments 1 and 2. To this end, Experiment 3 replicated Experiment 2 but included blocks of trials wherein participants were not required to render confidence.

**Method.** Stimuli were identical to those used in Experiment 2. Participants were required to make two sets of three blocks of decisions, with sets being administered on different days. Within one set of blocks, participants were always asked to render confidence judgements following each decision. Within the second set, participants never had to express confidence. Set and block orders were counterbalanced between participants. In total, each participant was asked to make \(48 \times 3 \times 3 \times 2 = 864\) decisions within two experimental sessions, each of which lasted approximately 90 minutes.

**Results.** The results of the experiment can be found in Figures 8 - 10 and Table 7.
Complex decisions

**Mean decisional response time.** The observed effects of decisional complexity and decisional difficulty on decisional RT were entirely comparable to those observed in Experiment 2 (Figure 8, top-left and Figure 9, left). Main effects of both complexity and difficulty on mean decisional RTs were observed, and post-hoc tests revealed significant differences between all levels of decisional difficulty and between all levels of decisional complexity. The one additional and significant interaction observed was a two-way interaction between decisional complexity and difficulty. Collapsed over sets, it would appear that the only within complexity level differences between means to achieve statistical significance were the difference between 3AFC easy (mean RT = 4215.5 ms) and hard decisions (mean RT = 4682.9 ms), and the difference between 4AFC easy (mean RT = 4988.1 ms) and hard decisions (mean RT = 5440.2 ms), HSD = 401.28, \( q(.05, 9, 48) = 4.598 \) (see Figure 10). This finding, coupled with a review of Table 6 below, suggests that the nonlinear effect of decisional complexity on decisional RT predicted by Hick (1952) increased in intensity along with increases in decisional difficulty. This tendency was also observed in Experiment 2 (Figure 7, top-left) though, as noted, the interaction between decisional complexity and difficulty did not achieve statistical significance. This is likely due to the relatively small effect size of the interaction (see Tables 5 and 7).

Note that, collapsed over sets, the only between complexity levels to fail to achieve significance was the difference between 3AFC hard decisions and 4AFC easy decisions, HSD = 401.28, \( q(.05, 9, 48) = 4.598 \) (see Figure 10).
Importantly, there was a significant effect of the requirement of rendering confidence judgements on mean decisional RT. Mean decisional response time decreased when participants knew they would not have to render confidence judgements (mean RT easy: 3849.5 ms; mean RT medium: 4063.7 ms; mean RT hard: 4147.8 ms) compared to when they knew they had to do so (mean RT easy: 4458.0 ms; mean RT medium: 4721.2 ms; mean RT hard: 4846.2 ms).

Table 6. The emergence of Hick-like RT patterns at each level of decisional difficulty for each set of blocks. Note how the size of the "difference between differences" grows at each level of decisional difficulty.

<table>
<thead>
<tr>
<th>Confidence Blocks</th>
<th>Easy Decisions</th>
<th>Medium Decisions</th>
<th>Hard Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\text{RT}_3 - \text{RT}_2$</td>
<td>$\text{RT}_4 - \text{RT}_3$</td>
<td>$\text{RT}_3 - \text{RT}_2$</td>
</tr>
<tr>
<td></td>
<td>970.4 ms</td>
<td>948.1 ms</td>
<td>1151.8 ms</td>
</tr>
<tr>
<td>Difference =</td>
<td>22.3 ms</td>
<td>40.5 ms</td>
<td>261.2 ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\text{RT}_3 - \text{RT}_2$</td>
<td>$\text{RT}_4 - \text{RT}_3$</td>
<td>$\text{RT}_3 - \text{RT}_2$</td>
</tr>
<tr>
<td></td>
<td>948.4 ms</td>
<td>604.6 ms</td>
<td>1169.4 ms</td>
</tr>
<tr>
<td>Difference =</td>
<td>343.8 ms</td>
<td>554.5 ms</td>
<td>853.8 ms</td>
</tr>
</tbody>
</table>

It is again interesting to note that participants made faster decisions in Experiment 3's confidence block sets (grand mean RT: 4675.1 ms) than did participants in Experiment 2 (grand mean RT: 6519.7 ms) or Experiment 1 (grand mean RT: 11330.9 ms). The finding that RTs decreased as the overall number of decisions made by
Complex decisions

participants increased will be discussed in greater detail in the General Discussion section following.

**Mean decisional accuracy.** As was the case in Experiment 2, participants were more accurate making easy decisions than medium-level or hard-level decisions at all levels of decisional complexity regardless of whether confidence judgements were required. Participants also became reliably less accurate as decisional difficulty increased. Post-hoc tests revealed significant differences in mean accuracy levels between all levels of decisional difficulty and between all levels of decisional complexity (Figure 8, top-right; Figure 9, right; Figure 10, right). The observed complexity by difficulty interaction resulted in a pattern of accuracy comparable to that observed in Experiment 2: Participants displayed similar levels of accuracy when making 2AFC medium-level and hard-level decisions, but performed differently making medium-level and hard-level 3AFC and 4AFC decisions. As was the case in Experiment 2, Tukey's HSD post-hoc comparisons revealed that the only difference between means which failed to reach statistical significance was the difference between 2AFC medium level accuracy (62.66%) and hard level accuracy (59.45%) decisions, $q(0.05, 9, 48) = 4.598$, HSD = 5.94%.

Critically, there was no reliable effect of requirement to render confidence judgements of decisional accuracy. Coupled with the findings that decisional RT increased when confidence judgements are required and that times to render confidence judgements were again fast and flat (Figure 8, bottom-right), this finding suggests that confidence is processed in parallel to the primary decision *but does not affect the outcome*
Complex decisions

of that decision. In other words, it seems reasonable to conclude that confidence was not automatically processed during the course of these decisions.

**Mean confidence rating.** The observed effects of both decisional complexity and decisional difficulty on mean confidence rating replicated the findings of Experiment 2 (see Figure 8, bottom-left). Participants exhibited less confidence whenever complexity or difficulty increased, and post-hoc tests revealed that the differences were reliable between all levels of difficulty and all levels of complexity. Again, this suggests confidence ratings are principled and predictably influenced by the stimulus itself.

**Mean time to render a confidence rating.** As Figure 8 shows (bottom-right), and as has been noted above, participants again required less than a second to render confidence judgements, and this time was not affected by either manipulations of decisional difficulty or complexity.

**Discussion.** Having observed that when participants were asked to render confidence judgements post-decisionally they produced slower decisional RTs compared to when they were not asked to judge confidence, one is forced to consider the possibility that confidence is not processed automatically. If it were, decisional RT would be comparable regardless of whether an explicit expression of confidence was required. This phenomenon has been observed before within the context of a 2AFC decision-making experiment (Petrusic & Baranski, 2000) and, as such, this experiment serves as replication and extension of this result. More importantly, this finding suggests that, if a decision can be made without a decision maker computing confidence, then any model which suggests confidence regulates the decision-making process must be re-evaluated.
Figure 8. Results of Experiment 3 blocks where confidence judgements were required. Upper-left: the effect of changes in decisional complexity on decisional RT. Upper-right: the effect of changes in decisional complexity on decisional accuracy. Lower-left: the effect of changes in decisional complexity on mean confidence. Lower-right: the effect of changes in decisional complexity on post-decisional times to express a confidence rating. 95% confidence intervals are shown.
Figure 9. Results of Experiment 3 blocks where confidence judgements were not required. Left: the effect of changes in decisional complexity on decisional RT. Right: the effect of changes in decisional complexity on decisional accuracy. 95% confidence intervals are shown.

Figure 10. Results of Experiment 3, both block types collapsed. Left: the effect of changes in decisional complexity on decisional RT. Right: the effect of changes in decisional complexity on decisional accuracy. 95% confidence intervals are shown.
Table 7. Statistical results from analyses of Experiment 3.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Response Time</th>
<th>Accuracy</th>
<th>Confidence</th>
<th>Time to Decide</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>F</td>
<td>$\eta_p^2$</td>
<td>F</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2, 24</td>
<td>25.04***</td>
<td>.68</td>
<td>371.49***</td>
<td>.97</td>
</tr>
<tr>
<td>Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2, 24</td>
<td>12.68***</td>
<td>.51</td>
<td>182.19***</td>
<td>.94</td>
</tr>
<tr>
<td>Confidence Requirement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 12</td>
<td>9.13**</td>
<td>.43</td>
<td>3.88</td>
<td>.24</td>
</tr>
<tr>
<td>Complexity $\times$ Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4, 48</td>
<td>2.99*</td>
<td>.20</td>
<td>7.93***</td>
<td>.40</td>
</tr>
</tbody>
</table>

Note. Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see Chapter 2); While non-significant main effects are reported, only significant interactions are reported.

*** p < .001; ** p < .01; * p < .05.
General Discussion of Experiments 1-3

To summarize the results of these experiments, I observed that across all three experiments decisional RT increased, while confidence and accuracy decreased, as a function of decisional complexity. The patterns of RT increases were generally nonlinear as predicted by Hick (1952), with the differences between 3AFC and 2AFC RTs being greater than the differences between 4AFC and 3AFC RTs. Experiments 2 and 3 demonstrated additional effects of decisional difficulty on these variables where, predictably, decisions were made more slowly, with less confidence, and with less accuracy whenever decisional difficulty increased. There was a reliable complexity by difficulty interaction with decisional accuracy for both Experiments 2 and 3, as well as a complexity by difficulty interaction on decisional RT within Experiment 3. Mean times

Figure 11. Comparison of decisional RTs, collapsed over levels of decisional difficulty where applicable, from Experiments 1-3.
Complex decisions to render confidence were flat and very fast throughout the experiments, with neither complexity nor difficulty exerting any effect on this variable.

Interestingly, over the span of the three described experiments, mean decisional RTs decreased as the number of required decisions increased in each successive experiment (Figure 11). This finding is particularly interesting when one considers that the medium difficulty decisions made in Experiment 2 (mean RT: 6680.0 ms) were an exact replication of Experiment 1 (mean RT: 11330.9), and that the confidence set in Experiment 3 (mean RT: 4675.1 ms) was an exact replication of Experiment 2 (mean RT: 6519.7 ms). In other words, the requirements of the experiment as a whole seemed to have been used by participants to judge the speed with which individual decisions should be made. While the extra time spent making decisions seemed to have resulted in comparatively improved decisional accuracy in Experiment 1 (Table 8), no such improvement was derived from the longer response times in Experiment 2 relative to Experiment 3. In fact, relative to Experiment 2, accuracy was improved for the faster decisions made in Experiment 3 (Table 8).

Since decisional accuracy was always above chance performance, and accuracy did not always improve as a function of increased RTs, then one may draw three tentative conclusions from these findings. First, that the relationship between RT and accuracy is likely not the result of participants getting better with practice, since accuracy did not seem to be logically correlated with changes in RT across experiments. Second, it is possible that participants were spending less time "double-checking" their evidence when they knew they had a larger overall workload to perform. They were, perhaps, becoming
less unnecessarily vigilant in the face of a longer experimental session. Third, some of the time participants spent processing stimuli in Experiment 2 was unnecessary, since performance improved and RTs decreased in Experiment 3. While these between experiment differences would be difficult to model, the idea that participants were calculating an overall anticipated level of general effort – or, perhaps more accurately, anticipated discomfort – and were using this calculation to regulate overall time to make each decision, informed the EPIC model to be described in the section following.

Table 8. Proportion of correct responses as a function of decisional difficulty and complexity.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Complexity</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3 (Confidence Blocks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>2</td>
<td></td>
<td>.69</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.59</td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>.59</td>
<td>.59</td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>.67</td>
<td>.64</td>
<td>.65</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.56</td>
<td>.49</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.45</td>
<td>.42</td>
<td>.44</td>
</tr>
<tr>
<td>Hard</td>
<td>2</td>
<td></td>
<td>.53</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>.39</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td>.31</td>
<td>.31</td>
</tr>
</tbody>
</table>

The results of these three experiments have set the stage for the modelling work to be presented in the following section. Experiment 1 demonstrated that participants can, indeed, differentiate between the stimuli presented, and that they are able to make confidence judgements which accurately reflect changes in decisional complexity. Experiment 3, importantly, demonstrated that decisional RT increases when confidence
complex decisions

judgements are required but that confidence generation has no effect on decisional accuracy, nor does it interact with decisional complexity or decisional difficulty. These findings suggest that confidence is processed independently of, and in parallel to, the primary decision-making process: The confidence processing system seems to divert resources away from the primary decision-making process, thereby slowing it down, without actually interfering with the efficacy of the process itself. As such, a logical approach to constructing a complete model of complex decision-making would be to first model primary decision-making processes, generating RT and accuracy, and then to construct a confidence 'module' which could be activated as required and whose only effect on the primary model would be to slow down decisional RT.
Complex decisions

Chapter 4: Evidence Accrual Based on the Power-Law and Information Theory

by Carroll

The EPIC model was designed to provide a principled means of describing the results of Experiments 1-3. The basic version of the model (Figure 14) was designed to reproduce the effects of decisional complexity on decisional RT and accuracy that were generated within Experiment 1. A more advanced version of the model was developed to reproduce the additional main effects of decisional difficulty on RT and accuracy, as well as the pervasive interaction between decisional difficulty and decisional complexity on these variables as observed in both Experiments 2 and 3. Finally, a confidence 'module' was designed to be added, when required, to the EPIC model (see Chapter 6). This module was designed to reproduce the effects of decisional difficulty and decisional complexity on mean confidence ratings, as well as the main effect of increasing mean decisional RT whenever confidence ratings are required. As will be noted, the module is also capable of producing flat and fast post-decisional times to render confidence which are unaffected by either decisional difficulty or confidence level.

The following sections describe the theoretical underpinnings of the EPIC model, a description of the basic and advanced versions of the model, and a comparison of derived model data to the results described in Chapter 3.

Genesis of the EPIC Model

The persistent effect of decisional complexity on RT and accuracy suggested that an evidence accrual, or "race", model would provide a better fit to the data than would a
Complex decisions

diffusion, or "random walk", model. The essential difference between these two families of models is that race models postulate the existence of a series of evidence accrual counters (e.g., Vickers, 1979; see Chapter 1) whereas random walk models assume the existence of only a single counter which can register either positive or negative information (e.g., Nosofsky & Palmeri, 1997). More specifically, a 2AFC random walk model wherein \(|C|\) is the criterion amount of evidence required before an overt decision can be made, would have a decision maker choose alternative A if \(+C\) evidence has been accrued, and alternative B if \(-C\) evidence has been accrued. This implies that when choice A is made, no evidence has been retained in support of choice B and vice versa. This paradigm differs from 2AFC race models where, as noted in Chapter 1, evidence supporting both alternatives is retained throughout the decision-making process.

The ultimate decision to make the EPIC model a race model instead of a random walk model was motivated in part by Van Zandt, Colonius, and Proctor (2000). These researchers compared the descriptive capabilities of both race and diffusion models, and found neither class of model outperformed the other. As a result, these researchers have advocated the use of the computationally more parsimonious race models.

Further, it is remains unclear whether a diffusion model would be able to simulate confidence judgements. Because a random walk model has but a single evidence counter, confidence would have to be based on the number of 'steps taken' before a decision was made, such that fewer steps would be indicative of greater confidence. This method would be identical to "RT scaling", which is a theory of confidence generation that suggests people express higher mean confidence ratings whenever RT is relatively
Complex decisions

short. Petrusic and Baransi (2009) demonstrated that RT scaling is not likely involved in confidence generation (for more information see Chapter 6).

Other researchers have also observed that diffusion models were not designed to account for complex meta-cognitive processes such as confidence. For example, Ratcliff and Tuerlinckx (2002) note that diffusion models "best apply to situations in which subjects make two-choice decisions that are based on a single, 'one-shot' cognitive process, decisions for which reaction times do not average much over 1 sec" (pp. 439). Clearly, such decisional RTs were not observed in Experiments 1-3 and, as such, a random walk model would be an inappropriate candidate for a complete model of complex decision-making.

All things considered, race models were deemed to be more parsimonious and, as such, served as the logical foundation for the EPIC model.

**On decisional criterion settings in race models.** As noted by Vickers (1979), accuracy and decisional RT will vary as a function of the criterion amount of evidence a decision maker requires to support her/his choice prior to the expression of an overt decision (see Figure 1). In the 2AFC case, a larger decisional criterion will result in longer RTs since each evidence accrual event takes time to process and, on average, more evidence will have to be accrued before a decision can be made. A larger criterion will also result in greater decisional accuracy. Consider a 2AFC example where the correct response alternative is choice A instead of choice B, and the probability that any particular evidence accrual event will support choice A is .52. If only three supporting
Complex decisions

pieces of evidence are required before an overt decision is made, the probability that alternative A will be chosen is:

\[
p(\text{overt decision } A|\text{stimulus is of class } A) = p(A|A) =
\]

\[
p(\text{counter } A \text{ reaches criterion in precisely 3 steps}|A) + p(\text{counter } A \text{ reaches criterion in precisely 4 steps}|A) + p(\text{counter } A \text{ reaches criterion in precisely 5 steps}|A) =
\]

\[
\left( \binom{3}{3} \cdot 0.52^3 \cdot 0.48^0 \right) + \cdot 0.52 \left( \binom{3}{2} \cdot 0.52^2 \cdot 0.48^1 \right) + \cdot 0.52 \left( \binom{4}{2} \cdot 0.52^3 \cdot 0.48^2 \right) = 0.1406 + 0.2025 + 0.1944 = 0.5375
\]

Requiring a single additional piece of evidence to support the correct decision will increase the probability that a correct response is made:

\[
p(A|A) = \left( \binom{4}{4} \cdot 0.52^4 \cdot 0.48^0 \right) + \cdot 0.52 \left( \binom{4}{3} \cdot 0.52^3 \cdot 0.48^1 \right) + \cdot 0.52 \left( \binom{5}{3} \cdot 0.52^4 \cdot 0.48^2 \right) + \cdot 0.52 \left( \binom{6}{3} \cdot 0.52^4 \cdot 0.48^3 \right) =
\]

\[
0.0731 + 0.1404 + 0.1685 + 0.1617 = 0.5437
\]

Having established the relationship between criterion level and decisional accuracy in the 2AFC case, of concern becomes the question "Is it better to have a constant criterion across all levels of decisional complexity, or is it better to have varying criteria?" It will shortly become clear that a model attempting to reproduce the results of Experiments 1-3 will have to allow decisional criteria to vary along with decisional complexity in order to generate observed decisional RT and accuracy patterns. Before reaching this conclusion, however, one must consider the nature of the stimuli used in these experiments.

Participants in Experiments 1-3 were asked to assess 20×20 arrays of squares, each square being 20×20 pixels in size. Given the size of these stimuli, it seemed
reasonable to conclude that participants would be forced to consider smaller subsections of each colour array and to combine the results of this series of sub-decisions into a final decision regarding the stimulus as a whole. This “reasonable” assumption, however, forces a consideration of whether the colour ratios within any given sample of coloured squares would be comparable to the colour ratios found in the population of coloured squares.

Consider, for example, a four colour stimulus of the type used in Experiment 1. The probability that any one square will be a member of the most likely colour class is .27 (108 dominant colour squares / 400 squares total). If an observer samples a 10×10 = 100 square subsection of the entire stimulus, the multinomial probability that the sample is exactly representative of the colour population is:

\[ p(.27,.25,.25,.23) = \frac{100!}{27!25!25!23!}(.27^{27})(.25^{25})(.25^{25})(.23^{23}) = .0010 \]

In other words, the probability that exactly 27 squares of 100 will share the most likely colouring, exactly 25 will share the second-most likely colour, exactly 25 the third, and exactly 23 fourth is only .001. This is not to say that the probability that a majority of the observed squares will contain the dominant colour is .001. Instead, it is meant to illustrate the very unlikely probability that a given sample of size 100 will perfectly represent the population.
Complex decisions

Table 9

The derivation of \( p(\text{accrual event } A|A) \) for a sample of 5 coloured squares.

<table>
<thead>
<tr>
<th></th>
<th>( p(\text{colour 1}) = .27 )</th>
<th>( p(\text{colour 2}) = .25 )</th>
<th>( p(\text{colour 3}) = .25 )</th>
<th>( p(\text{colour 4}) = .23 )</th>
<th>( p(n_1, n_2, n_3, n_4) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>( n_1 )</td>
<td>( n_2 )</td>
<td>( n_3 )</td>
<td>( n_4 )</td>
<td>( p )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0629</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.0226</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0.0226</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0246</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.0123</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0.0123</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.0104</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0066</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0014</td>
</tr>
</tbody>
</table>

\[ \Sigma = p(\text{colour 1 is most represented in a sample of size 5}) = .1886 \]

The derivation of the multinomial probability that the most likely colour would dominate any given sample regardless of whether the proportions of colours observed are identical to the overall stimulus proportions, \( p(\text{accrual event } A|A) \), is more easily demonstrated via a consideration of a smaller sample of only five coloured squares. As demonstrated in Table 9, \( p(\text{accrual event } A|A) \) for a sample of 5 squares would be .1886. This value is considerably less than the population's .27!

\( p(\text{accrual event } A|A) \) varies as a function of sample size (see Table 10), and the expected number of evidence accrual events required before an overt decision is made, \( E(n_{ae}) \), varies as a function of these multinomial probabilities. Table 10 describes how, given the \( p(\text{accrual event } A|A) \) derived from the stimuli used in Experiment 1, one would not expect to consistently see a 'Hick-like' increasing but nonlinear pattern of \( E(n_{ae}) \) if the decisional criterion were held constant over all levels of decisional complexity. What is more, since expected decisional response time, \( E(\text{RT}) \), varies as a function of \( E(n_{ae}) \),
Complex decisions

that one could suggest $E(RT) = E(n_{ae}) \times T_{ae}$, where $T_{ae}$ refers to a constant time it takes to accrue each unit of evidence, $E(RT)$ should not consistently reflect Hick's observations if decisional criteria were held constant over all levels of decision complexity. As will be noted in the sections following, the EPIC model was best able to replicate observed RT patterns when it varied decisional criterion along with decisional complexity.
Table 10

Results of a Monte Carlo simulation in which the number of evidence accrual events required before an overt decision was made (Obtained(n<sub>ae</sub>)) and the proportion of correct responses observed (Obtained p(correct<sub>ae</sub>)) were recorded. The criterion level of evidence required was held constant at 4. Sample sizes for each evidence accrual event were systematically varied from 1 to 25. 50000 simulated 'MOST' decisions were made at each level of complexity for each sample size. p(accrual event A|A) is based on the sample probabilities used in Experiment 1, and are included for reference purposes.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
<th>Obtained(n&lt;sub&gt;ae&lt;/sub&gt;)</th>
<th>Obtained p(correct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.52</td>
<td>0.36</td>
<td>0.27</td>
<td>5.802</td>
<td>7.331</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>0.13</td>
<td>0.07</td>
<td>5.795</td>
<td>7.273</td>
</tr>
<tr>
<td>3</td>
<td>0.53</td>
<td>0.3</td>
<td>0.18</td>
<td>5.801</td>
<td>7.310</td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>0.3</td>
<td>0.22</td>
<td>5.776</td>
<td>7.278</td>
</tr>
<tr>
<td>5</td>
<td>0.54</td>
<td>0.25</td>
<td>0.19</td>
<td>5.798</td>
<td>7.257</td>
</tr>
<tr>
<td>6</td>
<td>0.38</td>
<td>0.31</td>
<td>0.19</td>
<td>5.778</td>
<td>7.261</td>
</tr>
<tr>
<td>7</td>
<td>0.54</td>
<td>0.32</td>
<td>0.22</td>
<td>5.789</td>
<td>7.264</td>
</tr>
<tr>
<td>8</td>
<td>0.41</td>
<td>0.29</td>
<td>0.22</td>
<td>5.776</td>
<td>7.219</td>
</tr>
<tr>
<td>9</td>
<td>0.55</td>
<td>0.33</td>
<td>0.22</td>
<td>5.788</td>
<td>7.245</td>
</tr>
<tr>
<td>10</td>
<td>0.43</td>
<td>0.34</td>
<td>0.22</td>
<td>5.774</td>
<td>7.234</td>
</tr>
<tr>
<td>11</td>
<td>0.55</td>
<td>0.32</td>
<td>0.24</td>
<td>5.772</td>
<td>7.187</td>
</tr>
<tr>
<td>12</td>
<td>0.44</td>
<td>0.35</td>
<td>0.24</td>
<td>5.765</td>
<td>7.212</td>
</tr>
<tr>
<td>13</td>
<td>0.56</td>
<td>0.35</td>
<td>0.24</td>
<td>5.773</td>
<td>7.191</td>
</tr>
<tr>
<td>14</td>
<td>0.45</td>
<td>0.33</td>
<td>0.24</td>
<td>5.756</td>
<td>7.172</td>
</tr>
<tr>
<td>15</td>
<td>0.56</td>
<td>0.36</td>
<td>0.25</td>
<td>5.768</td>
<td>7.181</td>
</tr>
<tr>
<td>16</td>
<td>0.47</td>
<td>0.36</td>
<td>0.25</td>
<td>5.766</td>
<td>7.164</td>
</tr>
<tr>
<td>17</td>
<td>0.57</td>
<td>0.35</td>
<td>0.25</td>
<td>5.764</td>
<td>7.130</td>
</tr>
<tr>
<td>18</td>
<td>0.48</td>
<td>0.37</td>
<td>0.25</td>
<td>5.744</td>
<td>7.144</td>
</tr>
<tr>
<td>19</td>
<td>0.57</td>
<td>0.37</td>
<td>0.26</td>
<td>5.752</td>
<td>7.136</td>
</tr>
<tr>
<td>20</td>
<td>0.48</td>
<td>0.36</td>
<td>0.26</td>
<td>5.724</td>
<td>7.116</td>
</tr>
<tr>
<td>21</td>
<td>0.57</td>
<td>0.38</td>
<td>0.26</td>
<td>5.754</td>
<td>7.111</td>
</tr>
<tr>
<td>22</td>
<td>0.49</td>
<td>0.38</td>
<td>0.26</td>
<td>5.721</td>
<td>7.110</td>
</tr>
<tr>
<td>23</td>
<td>0.58</td>
<td>0.37</td>
<td>0.27</td>
<td>5.741</td>
<td>7.086</td>
</tr>
<tr>
<td>24</td>
<td>0.5</td>
<td>0.39</td>
<td>0.27</td>
<td>5.715</td>
<td>7.092</td>
</tr>
<tr>
<td>25</td>
<td>0.58</td>
<td>0.39</td>
<td>0.27</td>
<td>5.732</td>
<td>7.081</td>
</tr>
</tbody>
</table>

*Where a Hick's-type pattern is said to have occurred when Obtained(n<sub>ae</sub>) 3AFC - 2AFC > 4AFC - 3AFC

**The derivation of this value is illustrated in Table 9."
At this juncture, it bears mentioning that the values listed in Table 10 were derived via a Monte Carlo simulation using the base probabilities of coloured square occurrence employed in Experiment 1. As such, this table would most accurately reflect the situation when Experiment 1 participants were asked to decide which of \( n \) colours was represented most often in the display. When the instruction was "MOST", every instance of a particular colour represented a piece of evidence supporting the conclusion that the observed colour was dominant. As such, the evidentiary contribution of any particular square in a sample is simply \( 1/m \), where \( m \) is the number of coloured squares in the sample.

When the instruction was "LEAST", however, the presence of a given colour is actually evidence that some other colour is least dominant within the sample. During the development of the EPIC model (see sections following), two means of generating "LEAST" instruction probability distributions were considered. In the first, when a particularly coloured square was observed, a partial amount of evidence was accrued for all other valid response options, with the increase for each alternative being equal to \( (1/m) / (nAFC - 1) \). In the second, when a particularly coloured square was observed, a full portion of evidence was accrued for only one other randomly chosen valid response. As will be noted later in this chapter, the latter option provided the better fit to all observed data.

In any case, Table 10 illustrates how larger samples of coloured squares, which would provide a more accurate reflection of the distribution of colours in the population.
Complex decisions than would smaller samples, do not tend to yield nonlinearly increasing $E(n_{ae})$ with changes in decisional complexity.

The Search for a Principled Means of Altering Decisional Criteria

Both Stevens’ Power Law (1957) and Shannon’s Information Theory (1948) informed the rules governing criteria adjustment within the EPIC model.

Stevens’ Power Law and the EPIC model. Coombs and Avrunin (1988) note that, in psychophysics, “good things satiate... bad things escalate”. This quote describes the long understood relationship between psychological sensation and different types of physical stimulation: It typically takes greater and greater magnitudes of pleasurable stimuli to create perceivable increases in the sensation of pleasure, while it takes smaller and smaller increases in the magnitude of painful stimuli to create noticeable differences in perceived pain. Stevens’ (1957) Power Law is capable of describing both of these psychophysical relationships:

$$\Psi = k\phi^a$$  \hspace{1cm} (8)

where $\Psi$ is the perception of stimulus intensity, $k$ varies as a function of the unit of measurement employed, and $\phi$ is the physical magnitude of the stimulus observed. The exponent “$a$” varies as a function of the stimulus set being investigated. When $a < 1.0$ the described relationship between perceived magnitude and physical magnitude is nonlinear, decreasing (i.e., the pleasure pattern) and when $a > 1.0$ the relationship is nonlinear, increasing (i.e., the pain pattern).
Following informal interviews with those who had participated in Experiment 1, it became clear that participants shared a consistent dislike for the more complex decisions they were asked to make. In particular, they seemed to find the 4AFC decisions quite frustrating, yet seemed less upset by the 2AFC and 3AFC blocks of trials. In other words, participant discomfort seemed to increase nonlinearly with linear increases in decisional complexity. As such, a version of Stevens' Power Law was used to describe the manner in which participants were altering their decisional criteria in response to shifts in relative psychological discomfort. To do so required the generation of an operational definition of 'discomfort'.

Imagine 'total discomfort experienced throughout the experiment' could be operationally defined as:

\[ \text{total discomfort} = \sum (n^a) \]  \hspace{1cm} (9)

where 'n' refers to the levels \( n_{AFC} \) through the entire range of observed complexity (here, \( n = \{2,3,4\} \)), and 'a' > 1.0 in accordance with the Power Law. As such, the proportion of relative discomfort experienced within a given level of decisional complexity can be described this way:

\[ \text{proportional discomfort} = \frac{n^a}{\text{total discomfort}} \]  \hspace{1cm} (10)

As frustration with the decision-making task increases, a perhaps unconscious though principled adjustment to a base decisional criterion \( C_{\text{base}} \) could be made as follows:

\[ C_{n_{AFC}} = C_{\text{base}} - (C_{\text{base}} \times C_{\text{weight}} \times \text{proportional discomfort}) \]  \hspace{1cm} (11)
Complex decisions

where $C_{\text{weight}}$ can be thought of as the proportion of total attention given to the decision-making task that is susceptible to the effects of discomfort.

The effects of altering exponent 'a' on the number of evidence accrual events preceding an overt decision is illustrated in Figure 12. As 'a' increases so too does the degree of nonlinearity of the plots. Exponent 'a' can be thought of as the extent to which relative discomfort increases along with increases in decisional complexity. When $a = 1$, for example, discomfort is spread more evenly amongst all levels of decisional complexity, simulating a decision maker who feels equally uncomfortable making 2AFC decisions and 4AFC decisions. When 'a' is raised as high as 5, however, the 4AFC decisions have become so uncomfortable compared to either 2 or 3AFC decisions that a decision maker begins to sacrifice accuracy in favour of speed (see Figure 12 where $a = 5$). This possible explanation for exponent 'a' was tested by Experiment 4 (see Chapter 7).
Figure 12. Mean number of evidence accrual events for correct decisions when the instruction is 'MOST' for stimuli used in Experiment 1, where obtained values were governed by Equation 11. Values were derived from a Monte Carlo simulation which made 10000 decisions per level of complexity for each of 5 exponents 'a'.

The relationship between 'a' and $C_{base}$ can be thought of as "principled" because it allows that a decision maker might actually increase base decisional criterion $C_{base}$ when decisions get more complex, in effect making an honest effort to maintain a constant level of performance, while the net criterion level employed, $C_{nafe}$, actually decreases despite this effort (see Table 11).
Complex decisions

Table 11

A demonstration of the relationship between proportional discomfort, $C_{nAFC}$, and $C_{base}$.

<table>
<thead>
<tr>
<th>Complexity</th>
<th>$C_{base}$</th>
<th>Discomfort</th>
<th>Exponent 'a'</th>
<th>$C_{weight}$</th>
<th>$C_{nAFC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
<td>.5</td>
<td>3.7241</td>
</tr>
<tr>
<td>3</td>
<td>4.25</td>
<td>2</td>
<td></td>
<td>.5</td>
<td>3.5905</td>
</tr>
<tr>
<td>4</td>
<td>4.5</td>
<td>2</td>
<td></td>
<td>.5</td>
<td>3.2586</td>
</tr>
</tbody>
</table>

I developed a race model employing Equation 11 above, but it was only marginally capable of reproducing the RT patterns described in Chapter 3. The reasons for this shortfall are hinted at in Figure 12. The difference between the mean number of obtained evidence accrual events for 2AFC and 4AFC decisions is very slight when the plots become appreciably nonlinear, while observed 4AFC decisional RTs were more than double the observed 2AFC decisional RTs in Experiment 1. I was, however, able to vastly improve model performance when I allowed Shannon's Information Theory (1948) to govern the rate of information processing within the model.

**Shannon's Information Theory and the EPIC model.** A small part of Shannon's work involved the calculation of "bit rates". A bit rate can be defined as the maximum amount of information which can be transmitted by a single unit member of a system of units, when each unit in the system can only assume one of a finite number of discrete states. For example, binary units are units which can assume only one of two states. A binary unit can only supply a single "bit" of information. Typically, such units are described as being either "on" or "off", or are assigned a value of either "1" or "0".
Complex decisions

Groups of binary units considered in totem can assume up to $2^n$ discrete states. As such, they are capable of representing increasingly complex patterns of information. For example, 2 units can transmit up to 4 distinct messages:

<table>
<thead>
<tr>
<th>Unit 1 State</th>
<th>Unit 2 State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The bit rate of information being transmitted by a single unit is defined by the following formula:

$$H = - \sum p_i \log_2 p_i$$

(12)

where $p_i$ is the expectation that event $i$ will occur, and $H$, as noted in the Chapter 1, is a measure of system entropy.

Consider a binary unit where $p(1) = p(0) = .5$. In this case, $H = 1.0$ and entropy is maximized because the state of the unit at any given time is governed by some completely random process. On the other hand, consider a binary unit where $p(1) = .52$ and $p(0) = .48$. Here $H = .9988$, indicating a greater ability to predict the state of the system at any given time and, therefore, a comparatively less chaotic system. Of course, when $p(1) = 1.0$ and $p(0) = 0$, or vice versa, $H = 0$ indicating an absence of entropy within the system, since the state of the system at any given time can be predicted with absolute certainty.

With regards to the present Experiments 1-3, the information provided by the observation of a single coloured square in the multi-coloured array is
Complex decisions

\[ H = -1 \sum [p(\text{colour occurring}) \log_2(p(\text{colour occurring})] \]  

(13)

and the colour proportions of Experiment 1-3 suggest the following levels of entropy:

Table 12

Entropy levels for Experiment 1-3. Note that only the "Medium Decisions" were presented in Experiment 1.

<table>
<thead>
<tr>
<th>Easy Decisions</th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Decisions</td>
<td>.9959</td>
<td>1.5741</td>
<td>1.9919</td>
</tr>
<tr>
<td>Hard Decisions</td>
<td>.9989</td>
<td>1.5820</td>
<td>1.9977</td>
</tr>
</tbody>
</table>

While the direct implication of Equation 13 is that 4AFC stimulus colour patterns are almost twice as unpredictable as 2AFC patterns, these nonlinear increases in entropy as a function of decisional complexity suggest a possible explanation for the RT patterns first noted by Hick (1952). Indeed, Hick himself suggested that if "H" were used as a measure of the amount of information that needed to be processed, and if the rate of information processing were held constant, then his observed RT patterns would be a logical consequence of this phenomenon. And while observed bit rates are often less than the maximum possible bit rates detailed within Equation 12, they have been observed to be comparable when \( nAFC < 5 \) (Stewart, Brown, & Chater, 2005).

Figure 13 illustrates how Information Theory can be used improve the performance of a race model. Recall how expected response time was conceptualized as \( E(\text{RT}) = E(n_{ae}) \times T_{ae} \). Information Theory was used to designate a value to a coefficient which would be comparable to coefficient \( T_{ae} \). Specifically, rather than allowing
evidence accrual counters to increase by 1 following each evidence accrual event, these values increased by $1 \times .9989$ for 2AFC decisions, by $1 \times 1.5820$ for 3AFC decisions, and by $1 \times 1.9977$ for 4AFC decisions. As Figure 13 illustrates, the result of allowing evidence accrual counters to increase disproportionately as a function of complexity is a greater difference between amounts of evidence accrued for 2AFC and 4AFC decisions relative to those illustrated in Figure 12. Given the described relationship between $n_{ae}$ and $E(RT)$, this difference represents a far better description of Experiment 1 observed RT data than was derived using only Equation 11.

Note that these accrued evidence values are not true $E(RT)$ values per se in that they are not scaled to fall within the range of true human RTs. As such, these will be described as "simulated decisional RTs" in the modelling work which follows, with the understanding that "simulated RT" = ($n_{ae} \times$ information bit rate) and that values $E(RT) = \text{simulated RT} \times T_{ae}$ could be derived if to scaled simulated RT values were required.
Complex decisions

Mean (Evidence Accrual Events x Bit Rate) for Experiment 1
for various exponents 'a', $C_{\text{base}} = 4$, $C_{\text{weight}} = .5$

Figure 13. Mean number of evidence accrual events weighted by the derived bit rate for correct decisions when the instruction is 'MOST' for stimuli used in Experiment 1, where obtained values were governed by Equations 11 and 13. Values were derived from a Monte Carlo simulation which made 10000 decisions per level of complexity for each of 5 exponents 'a'.

The Basic EPIC Model: Decisional RT and Accuracy

Following is an overview of the design and operation of the basic EPIC model, which was created to produce simulated RT and accuracy data for a series of decisions comparable to those made by human participants in Experiment 1. The basic operation of this model is illustrated in Figure 14.

From stimulus presentation to overt decision: The flow of the EPIC model. EPIC decisions involve four separate stages: Sample selection, sample feature evaluation, sample categorization, and summed evidence evaluation.
Complex decisions

Sample selection. A sample of 25 squares is generated for each evidence accrual event. In the case of the example presented below, the probabilities of colour square occurrence matched those of Experiment 1.

Sample feature evaluation. The observation of each individual square within the sample prompts a fractional increase in accrued evidence. When the instruction is “MOST”, evidence for the colour observed increases by $1/25$. When the instruction is “LEAST”, evidence for a random other valid colour response increases by $1/25$.

Consideration of a single square causes RT to increase by an amount equal to the stimulus bit rate divided by the size of the sample, \( RT = RT + H/25 \) (see Table 12 above).

Sample categorization. Once all 25 squares are evaluated, a decision is made regarding which colour is represented most often in the sample, and evidence is accrued accordingly. When the instruction is “MOST”, evidence for the majority colour observed increases by 1. When the instruction is “LEAST”, evidence for a random other colour increases by 1. Categorization of the whole sample causes RT to increase by the bit rate such that \( RT = RT + H \).

Summed evidence evaluation. The criterion amount of evidence required before an overt decision could be made is governed by Equation 11. Once a criterion level of evidence is accrued for one of the response alternatives, an overt decision is made.
Complex decisions

Figure 14. The flow of the basic EPIC model, when the instruction is “MOST”.

\[ C_{4AFC} = C_{base} - \left( C_{base} \times \text{Weight} \times 4^a/(2^a+3^a+4^a) \right) \]
Model parameter selection. An iterative process was used to find a parameter set which best described the RT and accuracy data generated in Experiment 1. Specifically, the parameter ‘C_{weight}’ from Equation 11 was held constant at .5 while the parameters ‘C_{base}’ from Equation 11 and ‘a’ from Equations 9 and 10 were varied and tested systematically using a program designed to run in MATLAB. In total, 511 parameter combinations were tested. Each test forced the model to make 10000 decisions at each of three levels of decisional complexity. In total, the 15,330,000 decisions required approximately 18 hours of computer run-time.

Best parameter selection required the calculation of the Bayesian Information Criterion (BIC), which served as a measure of each parameter set’s descriptive power (as per Leite & Ratcliff, 2010). These BIC values reflected the extent to which the proportion of correct and incorrect responses generated by each model deviated from the proportion of observed correct and incorrect responses observed in Experiment 1. BIC was calculated as follows:

\[
BIC = -2 \sum_i Np_i \ln (\pi_i) + M\ln(N) \tag{14}
\]

“where \( p_i \) and \( \pi_i \) are the proportions of observed and predicted data in the \( i^{th} \) bin, \( N \) is the number of observations in the condition, and \( M \) is the number of free parameters in the model” (Leite & Ratcliff, 2010, pp. 259 – 260). The term “bin” refers to the summed number of observations falling within a quantile of decisional RTs within each relevant condition. This somewhat cryptic explanation can be simplified with an example:
Table 13

The number of observations within each of the conditions in Experiment 1.

<table>
<thead>
<tr>
<th>Instruction:</th>
<th>Least</th>
<th>Most</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response:</td>
<td>incorrect</td>
<td>correct</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nAFC:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>222</td>
<td>344</td>
</tr>
<tr>
<td>3</td>
<td>267</td>
<td>295</td>
</tr>
<tr>
<td>4</td>
<td>329</td>
<td>238</td>
</tr>
</tbody>
</table>

To derive a BIC value, each of the summed observations presented in Table 13 were separated into six bins representing different RT categories (slowest to fastest) within each condition. For example, the 300 correct responses observed for 3AFC decisions where the instruction was “MOST” would be divided into 6 bins: the 50 slowest responses in bin 1, the 50 next slowest responses in bin 2, and so forth ending with the 50 fastest responses in bin 6. These values were then converted into proportions: For example, of all correct 3AFC decisions made for the instruction “MOST”, 60 / 300 = .2 were in the "slowest" bin. This proportion was then compared to the proportion of slowest/MOST/3AFC responses generated by the model being tested by using Equation 14. Note that the closer the expected proportions were to the observed proportions, the lower the BIC value becomes, though BIC can never reach zero because the value M in (14) penalises models for each additional free parameter added.
Complex decisions

Observed

Decisional Response Time

Decisional Accuracy

Modelled

Decisional Response Time

Decisional Accuracy

Figure 15. A comparison of obtained Experiment 1 decisional RT and accuracy plots to those generated by the basic EPIC model.
**Model results.** Figure 15 provides a comparison of the data observed in Experiment 1 and those produced by the basic EPIC model. The best of the 511 derived BIC values was 16921.61, which occurred when \( C_{\text{base}} = 5 \) and exponent \( a = 1.7 \). These parameters are associated with a \( C_{2\text{AFC}} \) of 4.5994, a \( C_{3\text{AFC}} \) of 4.2020, and a \( C_{4\text{AFC}} \) of 3.6986.

The best-fitting parameters do seem to provide an adequate description of the observed data. The distribution of modelled RT values seems comparable to those observed and, as indicated in Table 14, and shifts in model accuracy paralleled participant accuracy, though even the best-fitting parameters forced the EPIC model to become less accurate than expected as complexity increased. Nevertheless, the basic EPIC model seems to provide a reasonable approximation of human performance in Experiment 1.

<table>
<thead>
<tr>
<th></th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed p(correct)</td>
<td>.62</td>
<td>.53</td>
<td>.43</td>
</tr>
<tr>
<td>Modelled p(correct)</td>
<td>.65</td>
<td>.50</td>
<td>.37</td>
</tr>
</tbody>
</table>

**The Advanced EPIC Model: Decisional RT and Accuracy**

Unfortunately, the basic EPIC model required modification before it was able to reproduce the results of Experiment 2. Consider Figure 16, which illustrates the basic
model's best-fitting attempt to replicate the joint effects of decisional complexity and difficulty:

Figure 16. A comparison of obtained Experiment 2 decisional RT and accuracy plots to those generated by the basic EPIC model.
It is clear that the basic EPIC model was unable to capture the observed interactions between decisional difficulty and decisional complexity. As such, a more advanced version of the basic EPIC model was designed. This advanced model allowed both $C_{\text{base}}$ and $'a'$ from Equation 11 to vary as a function of decisional difficulty. Though less parsimonious than the basic EPIC model, this alteration is theoretically justifiable in that it is reasonable to assume that a decision maker would require more evidence to make a comparatively difficult decision. It is also possible that discomfort would be more evenly distributed over all levels of complexity for hard choices, in that even 2AFC hard choices might be considered unpleasant. If true, one would expect to see less nonlinearity for hard choices (see Figure 16, top-left panel).

The iterative program used to test various parameter values for the advanced EPIC model seemed to have reached a similar conclusion, in that the best fitting description of Experiment 2 data employed increases in $C_{\text{base}}$ and decreases in $'a'$ with each increase in decisional difficulty. This program was designed to run in MATLAB. With each pass, the program systematically varied the 3 different $C_{\text{base}}$ and 3 different $'a'$ parameter values, and then ran a simulated 10,000 decisions for each level of decisional complexity and each level of decisional difficulty. Decision simulations relied on the p(colour) values used in Experiment 2, and the decision-making process was identical to that illustrated in Figure 14. The total 90,000 decisions made for each one of 729 different parameter sets resulted in a total required 65,610,000 decisions and took approximately 6 days of computing time to run. The plots derived from the best-fitting parameter set are illustrated in Figure 17.
Complex decisions

Figure 17. A comparison of obtained Experiment 2 decisional RT and accuracy plots to those generated by the advanced EPIC model.

The iterative program settled on the following parameter settings: $C_{\text{base} \text{ easy}} = 3$, $C_{\text{base} \text{ medium}} = 5$, $C_{\text{base} \text{ hard}} = 5.5$, 'a' easy = 6.5, 'a' medium = 4, and 'a' difficult = 3, which yielded a BIC value of 50441.1842. The reason the BIC value is so high compared to the BIC from the Experiment 1 model test (BIC = 16921.61) is that the BIC formula
Complex decisions

favours parsimony and penalizes models with more parameters. As such, the basic EPIC model, with only 3 parameters ($C_{\text{weight}}, C_{\text{base}}, 'a'$), is at a considerable advantage compared to the advanced EPIC model, which has 7 parameters.

As was the case when Experiment 1 was modelled, the EPIC model becomes less accurate than expected as complexity increases (Table 15). Nevertheless, the model seems to do an adequate job reproducing the 'flavour' of the interactions noted in the Chapter 3 discussion of Experiment 2: RTs for easy decisions were faster than those of medium and hard decisions, which tended to 'fan' as complexity increased, and accuracy decreased as a function of both decisional complexity and difficulty.

Table 15

*Observed vs. modelled accuracy for Experiment 2 data. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th></th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Med</td>
<td>Hard</td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p(\text{correct})$</td>
<td>.69</td>
<td>.59</td>
<td>.59</td>
</tr>
<tr>
<td>Modelled</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p(\text{correct})$</td>
<td>.71</td>
<td><strong>.64</strong></td>
<td>.57</td>
</tr>
</tbody>
</table>
Complex decisions

Table 16

*Observed vs. modelled accuracy for Experiment 3. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th></th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Med</td>
<td>Hard</td>
</tr>
<tr>
<td>Observed p(correct)</td>
<td>.71</td>
<td>.63</td>
<td>.60</td>
</tr>
<tr>
<td>Modelled p(correct)</td>
<td>.72</td>
<td>.61</td>
<td>.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Med</td>
<td>Hard</td>
</tr>
<tr>
<td>Observed p(correct)</td>
<td>.72</td>
<td>.62</td>
<td>.59</td>
</tr>
<tr>
<td>Modelled p(correct)</td>
<td>.71</td>
<td>.62</td>
<td>.58</td>
</tr>
</tbody>
</table>

Finally, Figures 18 and 19 demonstrate the extent to which the advanced EPIC model was able to replicate the results of Experiment 3. The iterative process used was identical to that used for Experiment 2 data, though the model itself required one slight alteration. The confidence module, to be described in Chapter 6, was 'disconnected' from the EPIC model when it was making decisions that did not require a confidence rating.
Complex decisions

Note that, in order that the time required to compute the best fitting parameter set not require more than a week's worth of computing time, only 5000 simulated trials were run in each complexity × difficulty × block-type condition for each tested parameter set. Following the test, the best-fitting parameter sets were then re-run with 10000 simulated trials to derive a more stable estimate of BIC.
Figure 18. A comparison of obtained Experiment 3 decisional RT and accuracy plots, for trials where confidence judgements were not required, to those generated by the advanced EPIC model.
Complex decisions

CONFIDENCE BLOCKS

Figure 19. A comparison of obtained Experiment 3 decisional RT and accuracy plots, for trials where confidence judgements were required, to those generated by the advanced EPIC model.
The best BIC derived for the non-confidence set was $50935.7429$, with $C_{base} \text{ easy} = 2.5$, $C_{base} \text{ medium} = 3$, $C_{base} \text{ hard} = 4$, 'a' easy = 3.2, 'a' medium = 3, and 'a' difficult = 2.6. The best BIC for the confidence set was $50724.5271$, with $C_{base} \text{ easy} = 3$, $C_{base} \text{ medium} = 3.5$, $C_{base} \text{ hard} = 4$, 'a' easy = 1, 'a' medium = 1, and 'a' difficult = 1. Compared accuracy values are listed in Table 16. While the model again proved to be less accurate than desired when making more complex decisions (see Table 16), it is interesting to note that the iterative process used to select the parameter values which generated the best fit for observed accuracy nevertheless chose exponent parameters which reproduced the 'flavour' of the observed RT plots. Specifically, the exponent 'a' was held constant at 1 for all levels of decisional difficulty within the confidence blocks, which resulted in linear increases in RT with increases in decisional complexity.

Importantly, disconnecting the confidence module (see Chapter 6) when it was not required allowed the EPIC model parallel the observation from Experiment 3 that decisional RTs decreased for non-confidence trials while decisional accuracy remained uncompromised (Figures 18-19, Table 17). This result provides a "proof-of-concept" that confidence processes slow down the primary decision-making processes, but are not required by them. This will be discussed further in Chapter 6.

**Review of the Advanced EPIC Model**

The model was able to reproduce the observed nonlinear decisional RT patterns, as well as the interaction between decisional complexity and difficulty on RT. EPIC
model accuracy decreased along with both decisional complexity and difficulty, though it was repeatedly less accurate than desired when making 4AFC decisions. As such, it was unable to reproduce the interaction between decisional complexity and difficulty on accuracy. Nevertheless, model performance was reasonable in many regards.

A review of model performance and the subsequent comparison to observed data suggests that the answer I have given to the question "What does exponent 'a' represent?" is supported by circular logic. Recall that a small 'a' was said to imply general discomfort spread evenly over all levels of decisional complexity while a large 'a' implies disproportionately greater discomfort associated with more complex decisions. Following this explanation, Figure 19 suggests a low 'a' was employed for confidence decisions, while Figure 18 suggests a larger 'a' was employed for non-confidence decisions. Yet why should this be so? One could argue that non-confidence decisions were computationally less taxing than confidence decisions and so the 4AFC choices seemed comparatively more challenging within non-confidence blocks of trials. However, this explanation is being proffered *a posteriori*: "The higher exponent was employed because of disproportionately distributed discomfort, and I suspect there was disproportionately distributed discomfort because a higher exponent was employed". Experiment 4 will provide more concrete support for the idea that the exponent 'a' represents the distribution of discomfort over all levels of decisional complexity. In the Experiment 4 paradigm, half of the 4AFC decisions were designed to be more taxing than their associated less complex decisions. The EPIC model would predict far less linearity
in RT plots when discomfort is unevenly distributed, and the results of this hypothesis test are presented in Chapter 7.

Of course, the explanation offered for the assumed necessary shifts in decisional criteria might be completely wrong and the EPIC model invalid. The following chapter provides an account of my attempt to reproduce the RT and accuracy patterns observed in Experiment 2 using different sorts of race models via the model generation and testing method suggested by Leite and Ratcliff (2010). Included in these tests are evaluations of the descriptive abilities of the models suggested by Vickers (1979), Usher and McClelland (2001), and Leite and Ratcliff (2010).
Chapter 5: Other Models

Leite and Ratcliff's (2010) Modelling Technique

Leite and Ratcliff (2010) demonstrate a method by which the optimal parameter settings for any given model of nAFC decision-making can be computed, and through which the descriptive power of a number of these optimized models can be compared so that the model which best reflects a given set of observed accuracy data can eventually be selected. In essence, the Leite and Ratcliff method involves the arbitrary selection of various sets of parameter settings for a given candidate model, obtaining a measure of each set's descriptive power, comparing these values, altering the values of the set which fares worst, and then repeating the process until a desired level of descriptive power for at least one of the sets has been obtained. This process is repeated for each candidate model to be compared, and the model which generates the highest level of descriptive power is considered the "winner". This method was employed in an attempt to find a model which accurately reflected the data derived from Experiment 2.

Selection of a candidate model. The model first proposed by Usher and McClelland (2001) was selected as the basic model from which other models could be derived and compared. Specifically, the tests made use of the following equation, which describes how evidence is accrued within the context of a race model:

\[
d x_i = \left[ \rho_i - \kappa x_i - \beta \sum_{j \neq i} x_j \right] \frac{dt}{\tau} + \xi_i \sqrt{\frac{dt}{\tau}}
\]  

(15)

where \( \rho_i \) represents stimulus input strength to the \( i^{th} \) accumulator (i.e., the proportion of colours i relative to all observed colours in a given sample), \( \kappa \) represents the rate of
Complex decisions

‘memory’ decay within the $i^{th}$ accumulator, $\beta$ is the coefficient of lateral inhibition through which evidence accumulators compete with one another by inhibiting the activation of other accumulators to a degree proportionate to evidence accrued by the system, $\xi$ is the standard deviation of a distribution of noise within the system, and $dt/\tau$ is a time scalar set, as was the case in User and McClelland (2001) and in Leite and Ratcliff (2010), to 1/100 so that evidence was accrued by the model in something comparable to 10-msec steps.

Equation 15 was chosen because, with the proper parameter settings (i.e., $\kappa=0$, $\beta=0$, and $\xi=0$), Equation 15 reduces to the simplest of all possible race models:

$$dx_i = [p_i] \frac{dt}{\tau}$$

As such, a thorough investigation of Equation 15, via the methodology employed by Leite and Ratcliff, would allow for a comparison of a wide range of race models of varying degrees of theoretical complexity, from the basic race model described by Vickers (1979) to more advanced models employing such neurologically inspired mechanisms as lateral inhibition and memory decay.

The SIMPLEX parameter adjustment method. Leite and Ratcliff demonstrated how the SIMPLEX fitting method, first introduced by Nelder and Mead (1965), might be used to derive a best-fitting set of parameter values for any particular model.

This method is an iterative process. $N+1$ sets of parameter values, where $N$ is the number of free parameters in the models being compared, are tested and the BIC values for each are established. These values are subsequently ordered from lowest (i.e., the best fit) to highest (i.e., the worst fit), so that the worst combination of parameter values
Complex decisions may be replaced with a weighted mean set of values derived from the remaining $N$
“better than worst” sets of parameters. This entire process is repeated until a desired
level of descriptive accuracy is obtained, or until the parameter sets converge. Within the
present study, each run of the SIMPLEX routine was terminated following 40 successive
passes. The number 40 was chosen for two reasons: 40 passes produced relatively stable
(i.e., replicable) results and it took 24 hours for the computer to process the associated
3,240,000 decisions. With 32 separate sets of SIMPLEX tests planned (see below) this
meant that the model generation and results analysis phase of the project could be
conducted with, at most, slightly more than a single month of computer processing time.
This was deemed to be a reasonable compromise between test expediency and result
accuracy. This decision to terminate the routine after 24 hours of run-time was
influenced by Bogacz and Cohen (2001), who discuss how long a SIMPLEX, or in their
case the mathematically comparable SUBPLEX, routine can take to converge on a final
solution and how they chose to terminate their own routine after 24 hours of processing
time.

**The models tested.** A thorough evaluation of the ability of Equation 15 to
replicate the observations made during Experiment 2 required the testing of a series of
models, each with eight free parameters.

**Parameter 1: Inhibition.** Within each model, the inhibition parameter $\beta$ could
either be set to 0, mimicking the behaviour of a non-competing evidence accrual model,
or to some positive non-zero value $\beta \leq 1.0$ which would mimic the behaviour of a
competing accumulator model.
**Parameter 2: Decay.** Within each model, a decay value $\kappa$ of 0 would mimic the behaviour of a non-leaky evidence accumulation model, while a positive value $\kappa < 1.0$ would mimic memory leakage within the model.

**Parameters 3-5: Evidence accrual criteria.** Leite and Ratcliff (2010) suggest that the amount of evidence required before the overt expression a decision might vary along with changes in decisional complexity. Indeed, the EPIC model itself was designed to find a principled and systematic means of adjusting decisional criteria as a function of complexity (Chapter 4). As such, the tested models contained three separate evidence criterion parameters, one for each level decisional complexity. Within each test, these parameters were set to either the same value or to different values. When set to the same value, the resulting three criterion parameter model was identical to a model containing only a single criterion parameter.

**Parameters 6-8:** Noise not related to the decision-making process. This refers to $\xi_i$ in Equation 15. As with Parameters 3-5, there were three noise parameters included in each test which, if set homogenously, acted as a single, invariant noise parameter.
Complex decisions

Table 17

Model families tested based on Equation 15 $dx_t = [\rho_t - \kappa x_t - \beta \sum_{j=1}^{n} x_j] \frac{dt}{\tau} + \xi_t \sqrt{\frac{dt}{\tau}}$

<table>
<thead>
<tr>
<th>Test</th>
<th>Noise Parameters ($\xi_i$)</th>
<th>Number of Criterion</th>
<th>Decay ($\kappa$)?</th>
<th>Inhibition ($\beta$)?</th>
<th>“Least” Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>no</td>
<td>all others</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>yes</td>
<td>all others</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>no</td>
<td>random</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>random</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>yes</td>
<td>random</td>
</tr>
</tbody>
</table>
The modelled method of stimulus sampling was identical to that used for the EPIC model as described in Chapter 4 (Figure 14). Recall how the multinomial probability of generating a correct response differs depending on whether the instruction is "MOST" or "LEAST". In the case of the latter instruction, I noted how the EPIC model produced better results when the observation of one colour was allowed to provide evidence that some random other colour was, in fact, the "LEAST" colour. I was unsure whether this would be true when using Equation 15, and so I tested the entire family of models twice: Once where "LEAST" prompted evidence being accrued by a random other colour, and once where "LEAST" prompted partial evidence accrual to all other possible colour options. With the above discussion in mind, the families of models tested are presented in Table 17.

Model test results.

*Concerning inter-accumulator inhibition.* The SIMPLEX convergence pattern was comparable whenever a family of parameter sets were compared wherein any of the inhibition parameters held a value greater than zero: Regardless of the values of the other parameters in the sets, the largest BIC values were always associated with the highest levels of inhibition. As such, the SIMPLEX routine consistently adjusted these inhibition parameters so that, eventually, the inhibition parameters of all nine sets were reduced to zero.

As previously noted, each SIMPLEX convergence test took approximately 24 hours to complete and it was estimated that the entire set of tests would require over a
Complex decisions

month of processing time. Since it became clear following three complete inhibition tests, and 8 partial 3 hour tests, that the largest inhibition parameter values would consistently produce the largest BIC values, I decided that in the interests of expediency inhibition would be removed from further consideration. It is worth noting that Leite and Ratcliff (2010) also found the inhibition parameter ineffective, and likewise decided that this parameter had limited descriptive utility.

**Obtained BIC values.** The best BIC values generated for each of the remaining 16 tests of interest are presented in Table 18. The model selected by the SIMPLEX routine as providing the best fit to the accuracy data had a single noise component $\xi_i$ and a single criterion for all levels of complexity. Interestingly, though this was the model best able to reproduce observed accuracy levels, it produced very strange patterns of decisional RT (Figure 20). Conversely, those models which produced RT patterns that were Hick-like faired poorly in the test since they were incapable of accurately describing participant accuracy. The highest placed model capable of producing such a RT pattern came in 9th, and had three different noise components $\xi_i$ and varied decisional criteria along with decisional complexity (Figure 21). No model proved capable of simultaneously reproducing both observed accuracy and RT patterns.
Table 18

*Best BIC values produced by the SIMPLEX tests.*

<table>
<thead>
<tr>
<th>Descriptive Utility Rank</th>
<th>&quot;Least&quot; Method</th>
<th>Noise Parameters (ξₖ)</th>
<th>Number of Criterion</th>
<th>Decay (κ)?</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>all others</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>50469</td>
</tr>
<tr>
<td>2nd</td>
<td>all others</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>50486</td>
</tr>
<tr>
<td>3rd</td>
<td>random</td>
<td>1</td>
<td>1</td>
<td>no</td>
<td>50517</td>
</tr>
<tr>
<td>4th</td>
<td>random</td>
<td>1</td>
<td>1</td>
<td>yes</td>
<td>50548</td>
</tr>
<tr>
<td>5th</td>
<td>random</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>51215</td>
</tr>
<tr>
<td>6th</td>
<td>random</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>51318</td>
</tr>
<tr>
<td>7th</td>
<td>random</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>51452</td>
</tr>
<tr>
<td>8th</td>
<td>random</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>51596</td>
</tr>
<tr>
<td>9th</td>
<td>all others</td>
<td>3</td>
<td>3</td>
<td>yes</td>
<td>51628</td>
</tr>
<tr>
<td>10th</td>
<td>all others</td>
<td>3</td>
<td>3</td>
<td>no</td>
<td>51752</td>
</tr>
<tr>
<td>11th</td>
<td>random</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>51862</td>
</tr>
<tr>
<td>12th</td>
<td>random</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>52178</td>
</tr>
<tr>
<td>13th</td>
<td>all others</td>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>52486</td>
</tr>
<tr>
<td>14th</td>
<td>all others</td>
<td>1</td>
<td>3</td>
<td>no</td>
<td>52607</td>
</tr>
<tr>
<td>15th</td>
<td>all others</td>
<td>3</td>
<td>1</td>
<td>yes</td>
<td>52954</td>
</tr>
<tr>
<td>16th</td>
<td>all others</td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>52990</td>
</tr>
</tbody>
</table>

*The horizontal lines group families of tests which differ only according to decay level*
Figure 20. A comparison of Experiment 2 decisional RT and accuracy patterns with those produced by the 1st place model as selected by the method described by Leite and Ratcliff (2010).
Complex decisions

Figure 21. A comparison of Experiment 2 decisional RT and accuracy patterns with those produced by the highest placed model (9th place) which produced a Hick-like pattern of decisional RT as selected by the method described by Leite and Ratcliff (2010).
Also of interest was the finding that the inclusion of a decay parameter had a relatively inconsequential effect on model descriptive utility. Compare, for instance, the family pairs in Table 18. It is worth noting that Leite and Ratcliff reported a comparable finding (2010).

Finally, I remind readers that the BIC associated with the best-fitting EPIC model for Experiment 2 data was 50426. The lowest BIC value obtained in this series of tests was 50469. This suggests that the EPIC model provided at least, if not better, fit to the observed accuracy data than did the other models tested.

Discussion

Leite and Ratcliff's (2010) model testing routine produced unsatisfactory results as far as the present study was concerned: Generated models with good fits to observed accuracy patterns yielded poor fits to decisional RT patterns and vice versa. This was not entirely unexpected. The above tests were based on the LCAM, a model which, as discussed in Chapter 1, assumes a high and constant level of decisional accuracy whenever a decision maker is asked to emphasize accuracy over speed. Experiments 1-3, however, produced no observable proof that participants were attempting to maintain a constant and high level of decisional accuracy. Recall how confidence ratings decreased as decisional complexity increased (see Figures 6-8). This pattern suggests that participants were aware that their performance was decreasing as decisional complexity was increasing. Yet the observed RT data revealed that participants did not take the exponentially longer time that LCAM dictates would have been required to maintain constant levels of confidence (see Figures 6-8).
Also of some concern is the purely descriptive nature of the models presented above. The SIMPLEX routine chose as the best-fitting model one with a single non-decisional component with a value of .8122 and a single decisional criterion with a value of .3127. It did not assign these values for any principled or theoretically justifiable reason. This is not an uncommon practice among modellers. Brown and Heathcote (2008), for example, present a "linear ballistic accumulator model", whose parameters are set so as to describe their observed data without any consideration for the philosophies of decision-making or psychophysics, claiming that a theory articulating the manner in which parameters should be set was "outside the scope" of their article (p. 169).

Given the generally unsatisfactory results of the SIMPLEX model generation routine, and the computationally intensive nature of the months-long parameter setting process it requires, I decided to forego using this method to attempt a replication of the findings of Experiments 1 and 3. It would seem that neither LCAM nor APEP, as represented within the model testing method suggested by Leite and Ratcliff (2010), are sufficiently powerful to explain the data generated by Experiment 2. There is no reason to believe they would fare better in other tests of virtually identical data sets.
Chapter 6: The Confidence Module

Genesis of the Confidence Module

As noted in the Chapter 3 discussion of Experiment 3, observed decisional RT increased when confidence judgements were required but confidence generation had no effect on decisional accuracy, and did not interact with decisional complexity or decisional difficulty. As such, a confidence module was designed to act independently of the primary EPIC model so that, when the EPIC base program was instructed to call the confidence subroutine during a decision, the effect was to slow down decisional processing without affecting any other aspect of the primary decision itself.

Recall from the discussion in Chapter 1 that confidence within the context of the present discussion has been defined as a subjective evaluation of the integrity of the decision-making process itself. Recall as well that there has been some debate in the literature regarding which aspects of the decision-making process influence the generation of confidence ratings. Two possibilities have been discussed in the psychophysical literature. First, confidence may be a post hoc scaling of decisional RT (e.g., Volkmann, 1934; Reed, 1951) such that slow decisions lead to low confidence and quick decisions lead to high confidence. Second, confidence may be a function of some physical aspect of the judged stimuli themselves (e.g., Garrett, 1922; Vickers, 1979) such that comparisons of two very different stimuli would lead to high confidence while judgements of similar stimuli would lead to low confidence. Each of these possibilities was used to generate a confidence subroutine for the EPIC model, and the results of these simulations are discussed in the sections following.
Complex decisions

**Confidence based on scaled decisional RT.** The RT scaling confidence module assumed the following relationship between confidence and decisional RT: If the system took a relatively long time to make a decision, confidence was low; if the system made a relatively quick decision, confidence was high. The term "relative" here refers to comparisons made between a given decisional RT and other decisional RTs generated within a particular level of decisional complexity. Between complexity level comparisons were not permitted because, for Experiments 1-3, complexity level presentation order was a between participant IV, and was varied systematically. As such, while one participant may have been able to compare 4AFC RTs to previously made 2AFC RTs, other participants would not have had that luxury.

The flow of the RT-Scaling Confidence Module subroutine is illustrated in Figure 21. The system would begin the confidence generation process by recording decisional RT each time a decision maker expressed an overt decision. These RTs were saved in a vector of decisional RTs, and this vector served as the basis for future RT comparisons within the current level of decisional complexity. Following a given decision, the resultant RT was compared to both the shortest and longest recorded RTs to date. The difference between these extreme values was divided by 5, thereby creating the boundaries for six confidence values:

\[
\text{RT boundary width} = \frac{\text{MAXrt} - \text{MINrt}}{5}
\]

Boundary 1 = MINrt + RT boundary, Boundary 2 = MINrt + 2(RT boundary), ..., Boundary 5 = MINrt + 5(RT boundary), and Boundary 6 = MAXrt.
Complex decisions

Figure 22. Overview of the method by which the RT-Scaling Confidence Module generates a confidence rating.
Complex decisions

If a given RT value falls between MINrt and Boundary 1, then confidence = 100.

Similarly, if a given RT value falls between Boundary 1 and Boundary 2, then confidence = 90. In this way, any given RT value can be associated with a confidence category.

As has been noted, the increased decisional RT data generated by Experiment 3 and the fast, flat times to render confidence generated by all three experiments suggest that confidence was being generated prior to the expression of an overt decision. Since the confidence module needed to make five distinct operations/calculations before it could generate a confidence rating, it was assumed that the module required at least five simulated moments of RT to complete this task. If decisional RT was larger than 5, then 5 was added decisional RT and post-decisional time to render confidence was recorded as 1. If decisional RT was less than 5, then RT was doubled and post-decisional time to render confidence was set to 5 – RT.

The astute reader will wonder how a decision maker could possibly make an RT comparison prior to the expression of an overt decision; since it is only at that point that the decider would be aware of the length of time that was required to make the decision. This is, of course, a valid question, and not the only theoretical reason to question the validity of using RT scaling to produce a confidence rating. Consider an overly simplistic situation wherein a decision maker requires two pieces of evidence in order to express an overt decision. If each accrual event requires a single unit of RT then when deciding whether a stimulus is mostly 'red' or mostly 'green', where p(red) = .52, the following must be true:
Table 19.

The probabilities of various decisional RTs where nAFC = 2, p(red) = .52, p(green) = .48, and instruction = "MOST".

<table>
<thead>
<tr>
<th>RT Value</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(RT = 1) = 0</td>
<td>Impossible: At least 2 accrual events are required.</td>
</tr>
<tr>
<td>p(RT = 2) = .5008</td>
<td>([p(\text{red at time 1}) \times p(\text{red at time 2}) = .2704] + [p(\text{green at time 1}) \times p(\text{green at time 2}) = .2304])</td>
</tr>
<tr>
<td>p(RT = 3) = .4992</td>
<td>The sum of the probabilities of possible RT = 3 event sequences: red green red (.1298), green red red (.1298), green red green (.1198), and red green green (.1198).</td>
</tr>
<tr>
<td>p(RT = 4) = 0</td>
<td>Impossible: Three evidence accrual events will always produce a like pair.</td>
</tr>
</tbody>
</table>

As such, the only possible simulated RT values are 2 and 3, with p(RT = 2) being slightly higher than p(RT = 3). This would mean that, since only two confidence values are possible, the expected mean confidence for this decision maker would, over the long run, be 100(.5008) + 50(.4992) = 75.04.

Now consider the same situation at the 3AFC level of decisional complexity (Table 20). In this case, there are three possible confidence categories using RT-scaling: Confidence is 100 when RT = 2, confidence is 80 when RT is between 2.8 and 3.2, and confidence is 50 when RT = 4. As such, mean confidence in the long run is 100(.3347) + 80(.4444) + 50(.2209) = 80.07. In the long run, one would expect 3AFC mean decisional confidence to be greater than 2AFC mean decisional confidence! A review of Chapter 3 will demonstrate how this trend was never observed.
Complex decisions

Table 20.

The probabilities of various decisional RTs where nAFC = 3, p(red) = .36, p(green) = .3325, p(blue) = .3075 and instruction = "MOST".

<table>
<thead>
<tr>
<th>p(RT = 2) = .3347</th>
<th>red red (.1296) + green green (.1106) + blue blue (.0946)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(RT = 3) = .4444</td>
<td>There are 12 different ways for RT = 3: red green red, red green green, red blue red, etc... The sum of the probabilities of each possible event = .4444.</td>
</tr>
<tr>
<td>p(RT = 4) = .2209</td>
<td>There are 18 different ways for RT = 4: red green blue red, red blue green red, green red blue red, etc... The sum of the probabilities of each possible event = .2209.</td>
</tr>
</tbody>
</table>

The situation grows more extreme for 4AFC decisions, where an even higher mean confidence value would be expected. Imagine p(red) = .27, p(green) = .25, p(blue) = .25, and p(black) = .23. If so, then p(RT = 2) = .2508, p(RT = 3) = .3754, p(RT = 4) = .2807, and p(RT = 5) = .086 for the reasons explained in the paragraphs preceding. The associated confidence levels would then be RT 2 = 100, RT 3 = 90, RT 4 = 70, and RT 5 = 50. As such, when 4AFC decisions are made the, mean expected confidence rating would be 82.81.

Given these two objections, it is perhaps not surprising that the confidence module based on RT-scaling fared as poorly as it did.

**Experiment Simulations.** The best fitting parameter values found for the EPIC model simulations described in Chapter 4 were used to generate the confidence values displayed in Figures 23 to 26. The module was called by the EPIC program following
each decision made within each simulation. 10,000 decisions were made at each level of decisional complexity and, where required, at each level of decisional difficulty.

Since modelled decisional response times were always greater than 5, mean post-decisional times to render confidence were always 1. These flat, fast times to decide confidence match empirical observation (see Chapter 3). Mean modelled confidence ratings, however, bear little resemblance to those observed in the experiments. As predicted by the discussion above, there was a general – though not universal – tendency for mean modelled confidence ratings to rise along with increases in decisional complexity.

![Experiment 1: Observed Confidence vs. 'RT-Scaling' Modelled Confidence](image)

*Figure 23.* Observed mean confidence ratings and RT-scaling based modeled confidence ratings for Experiment 1.
Complex decisions

Experiment 2

Figure 24. Observed mean confidence ratings and RT-scaling based modeled confidence ratings for Experiment 2.

Experiment 3

Figure 25. Observed mean confidence ratings and RT-scaling based modeled confidence ratings for Experiment 3.
Complex decisions

Confidence based on the relative proportion of mode accrued evidence. The flow of the 'mode based' confidence module is presented in Figure 26. The means by which mean confidence and post-decisional time to render confidence judgements were generated depended on the following assumptions:

Memory of sampling events. Each time the simulated decision maker categorized a sample of size 25 squares, the result of the decision was noted by the confidence module. Events continued to be stored in memory until either "K" samples had been categorized or an overt decision was made, in which case no additional evidence would be available for processing. Whenever this occurred, samples were drawn at random from memory and treated as new samples until K total events had been evaluated. Note that the "confidence criterion" K need NOT be the same as the decisional criterion C. In the present simulations, K was set to 5 so as to allow for comparisons between the mode-based confidence module and the RT-scaling based confidence module.

Confidence as a proportion of mode evidence. The proportions of mode accrued evidence relative to total accrued evidence (p(mode)) were used to generate confidence ratings. Proportion boundaries corresponded to the numerical confidence categories written on the response panel that was used by participants in Experiments 1-3 (see Figure 3 in Chapter 2). For example, consider the numerical labels beneath the "GUESS" button: 50, 33, and 25. The labels beneath the button immediately adjacent to the "GUESS" button are: 60, 46, and 40. In the 2AFC case proportions ranging between .5 and .6 were assigned a confidence rating of "50". For the 3AFC case, proportions
Complex decisions ranging from .3333 and .46 were assigned a confidence rating of "50". For the 4AFC case, proportions ranging from .25 and .40 were assigned a confidence rating of "50".

It is interesting to note that, under this method of confidence evaluation, confidence would be expected to decrease as confidence criterion ‘K’ increased.

Consider the following table:

Table 21.

Expected mean confidence levels as a function of the proportion of mode evidence accrued relative to total confidence evidence accrued.

<table>
<thead>
<tr>
<th>Units of Mode Evidence</th>
<th>Proportion of Mode Evidence to Total Evidence:</th>
<th>Total Confidence Evidence K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.86</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.89</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[
E(\text{confidence}) = \text{mean of column} = 0.83 \quad 0.80 \quad 0.79 \quad 0.78
\]

This finding could explain the persistence of the idea that confidence is based on scaled decisional RT. If ‘K’ increased as a function of decisional difficulty or complexity, then confidence would shift in concert with RT since both of these DVs would be affected by complexity, but for completely different reasons. The fact that they tend to mimic one another is simply coincidence.
Complex decisions

**Post-decisional times to render confidence.** If confidence processing was completed prior to the expression of an overt decision, then post-decisional time to render confidence was set to 1, and decisional RT was increased by K. Otherwise, RT was doubled, and time to render confidence was set to K – RT.

**Experiment Simulations.** As was the case for the RT-scaling module simulations, the best fitting parameter values found for the EPIC model simulations described in Chapter 4 were used to generate the confidence values displayed in Figures 27 to 29. Again, 10,000 decisions were made at each level of decisional complexity and, where required, at each level of decisional difficulty.

Not surprisingly, mean post-decisional times to render confidence were always 1, once again simulating the observations of Experiments 1-3. The form of the module mean confidence plots does seem to match those based on observed data (see Figures 27-29), with the possible exception of hard decisions in Experiment 3 (Figure 29). In general, the module replicated the negative relationship between decisional complexity and mean confidence rating. While the module tended towards under-confidence for Experiment 1 (Figure 27 & Table 22) and over-confidence for Experiment 2 (Figure 28 & Table 22), it seems to have provided a much better fit to the data than did the confidence module based on scaled decisional RTs, with only 4 of the 21 modelled confidence values differing from the observed values by more than 5% (Table 22).
Figure 26. Overview of the method by which the Mode-Based Confidence Module generates a confidence rating. Note that the "K" amount of confidence evidence required before an overt decision is made need NOT be the same amount of evidence "C" required to make the actual decision.
Figure 27. Observed mean confidence ratings and mode-based modeled confidence ratings for Experiment 1.

Figure 28. Observed mean confidence ratings and mode-based modeled confidence ratings for Experiment 2.
Complex decisions

Experiment 3

![Graphs showing Observed Confidence and 'Mode-Based' Modelled Confidence](Figure 29).

*Figure 29.* Observed mean confidence ratings and mode-based modeled confidence ratings for Experiment 3.
Complex decisions

Table 22.

Observed vs. Mode-Based Modelled Confidence for Experiments 1-3. Differences greater than 5% are bolded.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Difficulty</th>
<th>Complexity</th>
<th>Observed Confidence %</th>
<th>Modelled Confidence %</th>
<th>Difference %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Medium</td>
<td>2</td>
<td>79.9</td>
<td>78.27</td>
<td>1.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>77.48</td>
<td>74.66</td>
<td>2.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>75.16</td>
<td>73.01</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td>Easy</td>
<td>2</td>
<td>76.18</td>
<td>79.14</td>
<td>-2.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>71.19</td>
<td>75.79</td>
<td>-4.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>67.53</td>
<td>72.65</td>
<td>-5.12</td>
</tr>
<tr>
<td>2</td>
<td>Medium</td>
<td>2</td>
<td>75.87</td>
<td>78.23</td>
<td>-2.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>69.18</td>
<td>74.97</td>
<td>-5.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>67.08</td>
<td>72.90</td>
<td>-5.82</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>2</td>
<td>74.88</td>
<td>77.78</td>
<td>-2.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>68.81</td>
<td>74.49</td>
<td>-5.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>65.78</td>
<td>72.90</td>
<td>-7.12</td>
</tr>
<tr>
<td></td>
<td>Easy</td>
<td>2</td>
<td>79.83</td>
<td>81.72</td>
<td>-1.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>76.45</td>
<td>76.20</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>75.04</td>
<td>73.14</td>
<td>1.9</td>
</tr>
<tr>
<td>3</td>
<td>Medium</td>
<td>2</td>
<td>79.07</td>
<td>80.53</td>
<td>-1.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>75.09</td>
<td>74.45</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>74.09</td>
<td>72.85</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>2</td>
<td>78.76</td>
<td>77.12</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>74.05</td>
<td>74.68</td>
<td>-.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>72.84</td>
<td>72.68</td>
<td>.16</td>
</tr>
</tbody>
</table>

General Discussion of the Confidence Modules

Contrary to the predictions of Volkmann (1934) and Reed (1951), the confidence module based on RT-scaling provided poor fits to observed mean confidence data. This finding would seem to support Petrusic and Baranski (2009), who observed how calibration and resolution measures derived from mean expressed confidence ratings were not entirely comparable to calibration and resolution measures derived from mean expressed RT ratings. What is more, the idea that confidence is based on a post hoc
scaling of decisional RTs runs counter to the observations, made in all three presented experiments, that post-decisional times to render confidence ratings were fast and unaffected by either decisional complexity or decisional difficulty. Participants in these experiments did not seem to require any post-decisional time to compute a confidence rating. This implies a decisional, not post-decisional, locus of confidence processing.

The logical conclusion, then, is that confidence judgements must be based on some other aspect of the decision-making process. The module which used the proportion of mode evidence to total evidence as the basis for confidence processing did an adequate job replicating the form and range of the observed confidence plots. As such, it is this module that was 'plugged into' the advanced EPIC model whose basic assumptions regarding participant discomfort were tested in a fourth experiment.
Chapter 7: Experiment 4

Purpose

The EPIC model is founded on three basic assumptions. First, as per Hick (1952), that the rate at which a decision maker is able to process visual information is constant. Second, as per Shannon (1948), that the amount of visual information offered by a stimulus increases nonlinearly as a function of decisional complexity. And third, as inferred from Stevens (1957) and Coombs and Avrunin (1988), that a decision maker’s patience for a task decreases nonlinearly as a task becomes more complicated.

It is the third assumption which stands as the EPIC model’s novel contribution to the field of race model design. The exponent ‘a’ provides an index of how much more psychologically uncomfortable a 4AFC decision is relative to similar, less complex decisions. In so doing, ‘a’ allows that the observed total quantity of information processed prior to the expression of an overt decision can be less than a decision maker’s selected ideal quantity of processed information. By employing this variable, EPIC can be made to replicate observed decisional RT and decisional accuracy. Furthermore, since this variable will affect the quantity of information processed at each level of decisional complexity, ‘a’ can also indirectly affect mean expressed confidence ratings and post-decisional times to make a confidence judgement, though only if the criterion amount of confidence evidence required, 'K', exceeds the criterion rate regulated by 'a'. When \( K < C_{nafc} \), DVs related to confidence are more directly affected by the level of K and by the probability of stimulus occurrence.
Complex decisions

Experiment 4 was designed to test the assumption that 'a' is an index of discomfort by observing participant performance in a 4AFC task where 4AFC decisions were designed to require more effort, and hence be less comfortable, than were the 3AFC or 2AFC decisions participants were asked to make. If the assumptions made by the EPIC model are correct, then participant decisional RT plots for the uncomfortable task should have been considerably less linear than would have been the plots observed in a basic replication of Experiment 2.

The paradigm employed in Experiment 4 was inspired by the work of Murdock (1960). As noted in the literature review presented in this dissertation's outset, Murdock hypothesized that decisional accuracy in a complex decision-making task involving stimuli arranged along a one-dimensional continuum, such as colour density, is a function of whether those stimuli are positioned near the ends of that continuum. One obvious test of this hypothesis would be whether, in a 2AFC task such as the ones presented in Experiments 1-3, decision makers are comparably accurate when asked to decide which colour is represented "MOST" as when asked which colour is represented "2\textsuperscript{nd} LEAST". Semantically, these instructions are identical and both refer to the same stimulus feature. As such, if Murdock is correct, accuracy under both instructions should be identical. It was a consideration of how the instructions "2\textsuperscript{nd} LEAST" and "2\textsuperscript{nd} MOST" would be interpreted as decisional complexity increased which inspired the Experiment 4 paradigm.

Consider the stimuli employed in Experiments 1-3. In a 2AFC task, the instruction "2\textsuperscript{nd} LEAST" refers to the "MOST" colour, and "2\textsuperscript{nd} MOST" refers to the
Complex decisions

“LEAST” colour. In a 3AFC task, both "2nd LEAST" and "2nd MOST" would refer to the ‘middle’ colour. In each of the 2AFC and 3AFC cases, an observant participant could employ a short-cut when interpreting the instructions: Either “LEAST means MOST” or “ignore the instruction and pick the middle colour”. No such cognitive short-cut exists when these instructions are used for 4AFC decisions. As such, these “irregular instructions” were employed as a means of forcing participants to apply a greater degree of concentration to the 4AFC task without simultaneously altering either decisional difficulty or decisional complexity.

The test of the EPIC model will be two-fold. If the discomfort inducing manipulation is successful, one would expect to observe less linear decisional RT plots for the irregular instructions than for the regular instructions. In other words, one would expect to observe a condition x complexity interaction on decisional RT. Second, having demonstrated the effect of disproportionately allocated discomfort on the linearity of these plots, a second test would be to see whether the model is able to replicate these effects via an appropriate manipulation of 'a'. Specifically, a larger 'a' should provide a better fit to the RT data for irregular instructions than would a small 'a'.

Participants

Twenty-four first-year psychology students from Carleton University, 13 female and 11 male drawn from a pool with a median age of 20.23, each participated in two 1.5 hour sessions in return for course credit.
Apparatus

The apparatus was the same as that used in Experiments 1-3.

Materials / Stimuli

The coloured square arrays were similar to those used in Experiments 1-3, though the 4AFC stimuli were modified. Rather than having two equally proportioned ‘middle colours’, colour proportions were arranged in each 4AFC stimulus so that there were colours represented MOST, 2nd MOST, 2nd LEAST, and LEAST in each display. Hard condition stimuli contained 106 squares of one colour, 102 squares of a second, 98 of a third, and 94 squares of the final colour (106 + 102 + 98 + 94 = 400; 106 / 102 = 1.0392, 102 / 98 = 1.0408, 98 / 94 = 1.0426); medium condition stimuli contained 111 squares of one colour, 104 squares of a second, 96 of a third, and 89 squares of the final colour (111 + 104 + 96 + 89 = 400; 111 / 104 = 1.0673, 104 / 96 = 1.0833, 96 / 89 = 1.0787); easy condition stimuli contained 123 squares of one colour, 106 squares of a second, 92 of a third, and 79 squares of the final colour (123 + 106 + 92 + 79 = 400; 123 / 106 = 1.1604, 106 / 92 = 1.1522, 92 / 79 = 1.1646).

Procedure

The two-session procedure employed was very similar to that used for Experiment 3. In one session, the Experiment 2 “MOST/LEAST” paradigm was replicated using the stimulus set designed for Experiment 4. In a second session, instead
of being presented with the instructions "MOST" and "LEAST", participants were asked
to evaluate stimuli according to the instructions "2nd MOST" and "2nd LEAST".

The regular instruction blocks of trials and the irregular instruction blocks of trials
were run on separate days, and within session block order was varied between
participants as defined by a Latin Square. Between session block order was varied
between participants.

With 24 stimuli × 3 levels of complexity × 2 instruction sets × 3 difficulty levels
× 2 sessions, each participant was asked to make 864 decisions total.

**Analytical Method**

As was the case for Experiments 1-3, within session trials where RT was less than
200 ms and trials where RT was more than 3 standard deviations above each participant's
mean within each level of decisional complexity and decisional difficulty were removed
from analysis. Removed trials accounted for 2.15% of the observations. Block order and
session order were included in all analyses as between-subjects factors but, as these
factors never achieved significance, the results of any between-subjects analyses are not
reported. Huynh-Feldt degrees of freedom (df) were again used to assess statistical
significance though, for the sake of clarity, the df values actually reported are defined by
the experiments themselves. As was the case for Experiments 1-3, all p(correct) values
used in all ANOVAs were adjusted using Berkson’s correction (1953), and the values
plotted in the figures and tables are the actual values observed. 95% confidence intervals
were calculated, where applicable, using the Loftus and Masson (1994) method for
within-participant designs. Post-hoc analyses of significant main effects consisted of a comprehensive series of all possible pairwise analyses within each effect, with the appropriate Bonferroni adjustment to the alpha-per-comparison made within each family of comparisons as required to maintain a family-wise significance level of .05. Post-hoc analyses of significant interactions consisted of the derivation of Tukey’s HSD values and the subsequent evaluation of all possible pairwise comparisons of all cells involved in the interaction.

Results

Results are presented in Tables 22-23 and in Figures 30-32.
Table 23. Statistical results from analyses of Experiment 4. The three-way interactions were not significant, and are not reported.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>df</th>
<th>F</th>
<th>η_p^2</th>
<th>df</th>
<th>F</th>
<th>η_p^2</th>
<th>df</th>
<th>F</th>
<th>η_p^2</th>
<th>df</th>
<th>F</th>
<th>η_p^2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 24</td>
<td>23.74***</td>
<td>.66</td>
<td></td>
<td>257.58***</td>
<td>.96</td>
<td></td>
<td>49.07***</td>
<td>.80</td>
<td></td>
<td>7.24**</td>
<td>.38</td>
</tr>
<tr>
<td>Difficulty</td>
<td>2, 24</td>
<td>10.37***</td>
<td>.46</td>
<td></td>
<td>83.89***</td>
<td>.88</td>
<td></td>
<td>41.50***</td>
<td>.78</td>
<td></td>
<td>.501 ns</td>
<td>.04</td>
</tr>
<tr>
<td>Instruction</td>
<td>1, 12</td>
<td>15.38**</td>
<td>.56</td>
<td></td>
<td>214.67***</td>
<td>.95</td>
<td></td>
<td>11.64**</td>
<td>.49</td>
<td></td>
<td>4.81*</td>
<td>.29</td>
</tr>
<tr>
<td>Complexity × Difficulty</td>
<td>4, 48</td>
<td>6.45**</td>
<td>.35</td>
<td></td>
<td>1.82 ns</td>
<td>.13</td>
<td></td>
<td>12.86**</td>
<td>.52</td>
<td></td>
<td>5.94**</td>
<td>.33</td>
</tr>
<tr>
<td>Complexity × Instruction</td>
<td>2, 24</td>
<td>7.18**</td>
<td>.37</td>
<td></td>
<td>8.66**</td>
<td>.42</td>
<td></td>
<td>9.75**</td>
<td>.45</td>
<td></td>
<td>1.163 ns</td>
<td>.09</td>
</tr>
<tr>
<td>Difficulty × Instruction</td>
<td>2, 24</td>
<td>3.60*</td>
<td>.23</td>
<td></td>
<td>15.89**</td>
<td>.57</td>
<td></td>
<td>9.00**</td>
<td>.43</td>
<td></td>
<td>.724 ns</td>
<td>.06</td>
</tr>
</tbody>
</table>

Note. Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see text); While non-significant main effects are reported, only significant interactions are reported. *** p < .001; ** p < .01; *p < .05.; ns p > .05
Table 24. Observed results from Experiment 4. RT = decisional RT (reported in ms), Accuracy = decisional accuracy (reported as a proportion), Confidence = mean reported confidence level (scale = 50, 60, 70, 80, 90, 100), and CT = post-decisional time to decide confidence (reported in ms).

<table>
<thead>
<tr>
<th>Instruction Type</th>
<th>Complexity</th>
<th>Difficulty</th>
<th>RT</th>
<th>Accuracy</th>
<th>Confidence</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>regular instructions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2AFC</td>
<td>easy</td>
<td>3149.1</td>
<td>0.70</td>
<td>85.48</td>
<td></td>
<td>486.04</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>3184.8</td>
<td>0.63</td>
<td>84.74</td>
<td></td>
<td>481.53</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>3153.6</td>
<td>0.58</td>
<td>85.14</td>
<td></td>
<td>452.67</td>
</tr>
<tr>
<td>3AFC</td>
<td>easy</td>
<td>3886.2</td>
<td>0.61</td>
<td>82.47</td>
<td></td>
<td>505.19</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>4112.5</td>
<td>0.50</td>
<td>80.38</td>
<td></td>
<td>497.56</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>4151.6</td>
<td>0.43</td>
<td>80.04</td>
<td></td>
<td>504.92</td>
</tr>
<tr>
<td>4AFC</td>
<td>easy</td>
<td>4496.9</td>
<td>0.58</td>
<td>81.45</td>
<td></td>
<td>502.85</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>5191.3</td>
<td>0.42</td>
<td>77.89</td>
<td></td>
<td>543.33</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>5385.3</td>
<td>0.39</td>
<td>77.82</td>
<td></td>
<td>582.42</td>
</tr>
<tr>
<td>collapsed 2AFC</td>
<td></td>
<td>3162.5</td>
<td>0.64</td>
<td>85.12</td>
<td></td>
<td>473.36</td>
</tr>
<tr>
<td>collapsed 3AFC</td>
<td></td>
<td>4049.9</td>
<td>0.51</td>
<td>80.97</td>
<td></td>
<td>502.56</td>
</tr>
<tr>
<td>collapsed 4AFC</td>
<td></td>
<td>5024.0</td>
<td>0.47</td>
<td>79.06</td>
<td></td>
<td>542.90</td>
</tr>
<tr>
<td>collapsed easy</td>
<td></td>
<td>3847.3</td>
<td>0.63</td>
<td>73.12</td>
<td></td>
<td>498.08</td>
</tr>
<tr>
<td>collapsed medium</td>
<td></td>
<td>4158.4</td>
<td>0.52</td>
<td>81.02</td>
<td></td>
<td>507.34</td>
</tr>
<tr>
<td>collapsed hard</td>
<td></td>
<td>4232.0</td>
<td>0.47</td>
<td>81.00</td>
<td></td>
<td>513.45</td>
</tr>
<tr>
<td>irregular instructions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2AFC</td>
<td>easy</td>
<td>3552.7</td>
<td>0.62</td>
<td>85.98</td>
<td></td>
<td>516.48</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>3588.0</td>
<td>0.57</td>
<td>84.10</td>
<td></td>
<td>498.37</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>3607.6</td>
<td>0.55</td>
<td>86.02</td>
<td></td>
<td>498.56</td>
</tr>
<tr>
<td>3AFC</td>
<td>easy</td>
<td>6007.4</td>
<td>0.44</td>
<td>79.04</td>
<td></td>
<td>588.41</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>6002.5</td>
<td>0.37</td>
<td>77.30</td>
<td></td>
<td>589.93</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>6033.4</td>
<td>0.35</td>
<td>77.49</td>
<td></td>
<td>573.05</td>
</tr>
<tr>
<td>4AFC</td>
<td>easy</td>
<td>6248.1</td>
<td>0.36</td>
<td>74.56</td>
<td></td>
<td>561.18</td>
</tr>
<tr>
<td></td>
<td>med</td>
<td>6449.2</td>
<td>0.28</td>
<td>73.23</td>
<td></td>
<td>573.55</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>6682.0</td>
<td>0.29</td>
<td>73.01</td>
<td></td>
<td>598.07</td>
</tr>
<tr>
<td>collapsed 2AFC</td>
<td></td>
<td>3582.8</td>
<td>0.58</td>
<td>86.03</td>
<td></td>
<td>504.49</td>
</tr>
<tr>
<td>collapsed 3AFC</td>
<td></td>
<td>6014.5</td>
<td>0.39</td>
<td>77.94</td>
<td></td>
<td>583.78</td>
</tr>
<tr>
<td>collapsed 4AFC</td>
<td></td>
<td>6460.0</td>
<td>0.31</td>
<td>73.60</td>
<td></td>
<td>577.64</td>
</tr>
<tr>
<td>collapsed easy</td>
<td></td>
<td>5263.9</td>
<td>0.47</td>
<td>79.88</td>
<td></td>
<td>555.26</td>
</tr>
<tr>
<td>collapsed medium</td>
<td></td>
<td>5340.5</td>
<td>0.41</td>
<td>78.91</td>
<td></td>
<td>553.87</td>
</tr>
<tr>
<td>collapsed hard</td>
<td></td>
<td>5439.8</td>
<td>0.39</td>
<td>78.85</td>
<td></td>
<td>556.51</td>
</tr>
</tbody>
</table>
Figure 30. Observed data for Experiment 4, regular instruction blocks. 95% confidence intervals are shown.
Figure 31. Observed data for Experiment 4, irregular instruction blocks. 95% confidence intervals are shown.
Figure 32. A detail of post-decisional times to render confidence ratings in Experiment 4. 95% confidence intervals are shown.

Decisional RTs. The instructional manipulation seems to have had the desired effect, in that the effect of instruction-type on decisional RTs was reliable, and the RT patterns for the irregular instruction set were considerably less linear than those observed for the regular instruction set (Figures 30-31, top-left panels). This finding accounts for the significant instruction-type × complexity interaction, and supports the EPIC model assumption that as discomfort is becomes less evenly distributed over all levels of decisional complexity, the greater becomes the observed patterns of nonlinearity.

Though the main effect of decisional complexity on decisional RTs was reliable, with post-hoc tests revealing significant differences in mean RT between all paired levels of complexity, a slightly ‘reversed Hick’ RT pattern was observed for the regular instruction set’s medium and hard level decisions (Figure 30). In each of these cases, the difference between 3AFC and 2AFC decisions was less than the difference between
Complex decisions

4AFC and 3AFC decisions (medium difficulty: 3AFC-2AFC = 927.68, 4AFC-3AFC = 1078.86; hard difficulty: 3AFC-2AFC = 998.03, 4AFC-3AFC = 1233.74). In the context of the EPIC model, this would require a difficulty exponent ‘a’ with a value of less than 1.0. An exponent ‘a’ < 1.0 would imply that 4AFC regular instruction decisions were more psychologically comfortable than were either their 3AFC or 2AFC counterparts. There is some reason to believe that this counter-intuitive suggestion may have actually been the case in Experiment 4 though not the case in Experiments 1-3.

When the 4AFC stimuli were designed for Experiment 4, the two middle colours in each display were altered so that, instead of being two equally proportioned ‘noise’ colours, a true “2\textsuperscript{nd} MOST” and true “2\textsuperscript{nd} LEAST” colour would be represented in each stimulus. The inadvertent side effect of this manipulation was that it created a greater difference between the number of 4AFC “MOST” and “LEAST” colours in each display. Consider the following table, which compares 4AFC stimuli from Experiments 2-4:

<table>
<thead>
<tr>
<th></th>
<th>Experiments 2 and 3</th>
<th>Experiment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MOST</td>
<td>LEAST</td>
</tr>
<tr>
<td>Easy</td>
<td>116</td>
<td>86</td>
</tr>
<tr>
<td>Medium</td>
<td>108</td>
<td>92</td>
</tr>
<tr>
<td>Hard</td>
<td>104</td>
<td>96</td>
</tr>
</tbody>
</table>

If Crossman’s (1952) assumption that decisional difficulty is a function of stimulus ratios is correct (see Chapter 1) then, within the regular instruction set, Experiment 4’s 4AFC decisions may have been comparably easier to make than
Complex decisions

Experiment 2 and 3’s 4AFC decisions. As such, Experiment 4’s 4AFC regular decision task may have been one of the easier conditions of the Experiment. A test of the EPIC model will be to see whether an exponent $a \leq 1.0$ is found which can provide a reasonable fit for the observed data set.

The main effect of decisional difficulty on decisional RTs was reliable, though post-hoc tests revealed the difference between medium and hard level difficulty was not significant, $t(21) = 1.61, p > .2$. The interaction between decisional difficulty and decisional complexity on RT is revealed in a noticeable "fanning effect" where 4AFC easy decisions are made more rapidly than 4AFC medium and hard decisions, though this difference is not reliable at the 2AFC and 3AFC levels (Figures 30 & 31). Although this effect is visually noticeable, it is worth mentioning that, statistically, Tukey's HSD post-hoc comparisons revealed that 2AFC mean RTs differ from 3AFC and 4AFC mean RTs at all levels of decisional difficulty, but that no other pairwise comparisons related to this interaction are significant, $HSD = 1132.2, q(.05, 9, 48) = 4.60$.

The interaction between decisional difficulty and instruction-type on RT can also be seen in a comparison of Figures 30 and 31. Tukey's HSD post-hoc comparisons revealed that there were no between-cell differences within either regular or irregular sets, but that irregular instruction RTs were reliably greater than regular instruction RTs between all levels of difficulty except irregular easy and regular hard, $HSD = 1075.59, q(.05, 6, 24) = 4.37$. Generally speaking, participants took longer making irregular choices than regular choices.
**Decisional accuracy.** Significant main effects of decisional complexity and decisional difficulty on decisional accuracy were observed, and post-hoc tests revealed reliable differences in accuracy between all levels of decisional complexity and between all levels of decisional difficulty. As was expected, participants were reliably less accurate whenever either complexity or difficulty increased.

Interestingly, a review of Figures 30-31 demonstrates how participants operating under regular instructions were significantly more accurate than when they were presented with irregular instructions. As Table 24 reveals, this was true at all levels of decisional complexity and at all levels of decisional difficulty.

Post-hoc tests revealed reliable differences in accuracy between all levels of complexity within and between both levels of instruction. Post-hoc tests also revealed significant differences between instruction levels at all difficulty levels except for both regular, hard decisions and regular, medium decisions which were not reliably more accurate than irregular, easy decisions.

**Mean confidence rating.** Significant main effects of all three IV’s on mean expressed confidence levels were observed. Participants were reliably less confident following irregular instructions compared to regular instructions. They also grew less confident whenever either decisional complexity or decisional difficulty increased. All tested pairwise comparisons were reliable except for the difference in mean expressed confidence following medium and hard difficulty decisions.

All two-way interactions involving confidence were found to be significant. While this lead to a rather convoluted series of post-hoc analyses, the story told by the
plots in Figures 30 and 31 can be summed up rather succinctly: Confidence plots were shallower and tended to fan more dramatically under regular instructions than under irregular instructions. This summation should be borne in mind while reading the following paragraphs.

Post-hoc tests of the complexity x difficulty interaction revealed that, while 2AFC decisions did not significantly vary from one another as a function of decisional difficulty, and while 3AFC medium and hard level decisions did not differ from each other or from 4AFC easy choices, all other pairwise comparisons related to this interaction were significant.

Analyses of the instruction type x complexity interaction revealed more variability within irregular instruction cells than within regular instruction cells: Both 4AFC and 3AFC irregular decisions were made with less confidence than 2AFC irregular decisions, while only 4AFC decisions differed from 2AFC decisions for the regular instruction cells. The result of these differences is the apparent fanning effect on confidence in the regular blocks of trials, and the absence of such an effect in the irregular blocks.

Between instruction types, 3AFC and 4AFC irregular instruction decisions were made with significantly less confidence than 2AFC regular decisions, and 4AFC irregular decisions were made with less confidence than 3AFC regular decisions. Only 2AFC irregular choices were made with more confidence than 4AFC regular decisions, HSD = 5.46, $q_{(.05, 6, 24)} = 4.37$. No other pairwise comparisons related to this interaction were found to be significant.
All but two pairwise comparisons between irregular and regular instruction × difficulty cells were significant, with only easy irregular decisions being made with comparable confidence to medium and hard regular decisions. Within the regular decision cells, easy decisions were made with reliably more confidence than were other harder choices. No other pairwise comparisons related to the instruction × difficulty interaction were found to be significant.

**Mean post-decisional time to decide confidence.** Unexpectedly, main effects of complexity and instruction on time to decide confidence were found, as was a significant interaction between decisional complexity and decisional difficulty (Figure 32). The effect of instruction was barely reliable ($p = .049$), with regular decisions resulting in confidence judgements made 45 ms faster compared to confidence judgements following irregular instructions. As well, 2AFC confidence judgements were made more quickly than were either 3AFC or 4AFC confidence judgements. The interaction between complexity and difficulty resulted in all 3AFC decisions being made more slowly than 2AFC hard decisions, all 2AFC decisions being made more quickly than all 4AFC decisions, and 3AFC medium and hard decisions being made more quickly than 4AFC medium difficulty decisions.

While statistically significant, it must be stressed than times to render confidence judgments were still very fast, with the range of observed mean times in Experiment 4 (452.7 - 598.1 ms) being faster than those observed in Experiment 2 (564.8 – 651.5 ms) and Experiment 3 (542.7-660.0 ms). Having already argued that Experiments 2 and 3 provided little evidence for post-decisional confidence processing, it seems incongruous.
Complex decisions

to suggest that these even faster response times allow it. In any case, it is unlikely that the EPIC model's confidence module, in the presented configuration, would be able to capture the granularity of these effects. Since the observed times are less than those of the previous experiments, and the module produced a constant post-decisional time to render confidence of 1 while trying to model observed behaviour in these early tests, then, given the observed range, the model would here have to produce times to render confidence ranging between .8046 and 1 to successfully replicate these time plots. As it stands, the EPIC model's module is incapable of producing a time to decide RT value of less than 1.

That said, it will shortly be noted that in order to replicate the mean confidence rating data, the EPIC confidence module was forced to employ varying confidence criterion levels. An interesting, if somewhat rough, approximation to the average observed times to render confidence plots can be made if one compares the changes in post-decisional times to render confidence to changes in confidence criteria K (Figure 33). If one allows that confidence processing does not begin at the outset of stimulus presentation, but instead starts at a later time and continues past the expression of an overt decision, then post decisional times to render confidence could be a function of a decision maker waiting for the appropriate amount of confidence evidence to accrue. Though the EPIC model's confidence module cannot capture this occurrence directly, it could nevertheless predict the general shape of post-decisional response times via the confidence criteria employed.
Complex decisions

Figure 33. A comparison of times to render confidence to confidence criteria employed.

The Model

Model parameter selection. The same type of iterative process used to select best-fitting parameter settings for the Experiment 1-3 models was again used here. The parameter ‘C\text{weight}' from Equation 11 was held constant at .5 while the parameters ‘C_{\text{base}}' from Equation 11 and ‘a' from Equations 9 and 10 were varied and tested systematically using a program designed to run in MATLAB. In total, 729 parameter combinations were tested for each instruction type. Each test forced the model to make 5000 decisions at each of three levels of decisional complexity and three levels of decisional difficulty. In total, the 32,805,000 decisions required approximately 9 days of computer run-time. As was the case for Experiment 3, following the series of tests, the selected best-fitting parameter sets were then re-run with 10000 simulated trials to derive a more stable estimate of BIC.
The EPIC model was able to follow irregular instructions by first ordering sampled evidence from “MOST” to “LEAST”, as it did when following regular instructions, and then incrementing the evidence counter associated with whichever colour occurred “2nd MOST” or “2nd LEAST” often as was appropriate.

Given the large number of significant confidence-related interactions observed in Experiment 4, a decision was made to allow the confidence criterion level ‘K’ to vary as a function of both decisional complexity and decisional difficulty. This decision was made following several failed attempts to generate the observed differences in confidence levels using a constant ‘K’. Once the optimal ‘$C_{\text{base}}$’ and ‘a’ parameters were derived, a systematic series of trial-and-error tests were performed to see which of a series of ‘K’ values would most closely fit the observed data. As it turned out, confidence was most effectively manipulated by allowing ‘K’ to alter as a function of decisional complexity, but not as a function of decisional difficulty. The results of these tests are presented below.

**Model results: Decisional RT and decisional accuracy.** The smallest BIC obtained for the regular instruction set was 50837.42. This was obtained with base criteria 3, 4, and 4, and discomfort exponents 2, .7, and .7 for easy, medium, and hard decisions respectively. Modelled decisional RTs are presented in Figure 34, and modelled decisional accuracy is presented in Figure 34 and Table 26. It is interesting to note that discomfort exponents of 2, 1, and 1 produced an almost identical BIC and very little difference in the linearity of the derived RT plots. In either case, the model was not able to reproduce the noted slight ‘reverse Hick’ effects observed in Experiment 4 for
Complex decisions

regular instructions and, as a result, both parameter sets left the model performing less accurately than expected when making 4AFC decisions. It is possible that a replication would have been obtained by using values of $a<0$, but negative values were not permitted by the EPIC model in its current configuration.

The smallest BIC obtained for the irregular instruction set was 50488.69. This was obtained with base criteria 5, 5, and 5, and discomfort exponents were 5, 4.5, and 4 for easy, medium, and hard decisions respectively. As can be seen in Figure 35, the model did a remarkable job replicating the observed 'fan' effect for irregular instruction RT plots, and also was able to replicate the observed instruction type $\times$ complexity interaction with RT (compare Figures 34 and 35). As noted in Table 26, the model was comparatively over-accurate making 2AFC easy and medium difficulty irregular decisions, and under-accurate making easy 3AFC and 4AFC decisions. Apart from these deviations, model performance was within the 95% CI ranges of all other categories of observed accuracy.
Complex decisions

Figure 34  A comparison of observed Experiment 4 decisional RT and accuracy plots, for trials where regular instructions were given, to those generated by the advanced EPIC model.
Figure 35. A comparison of observed Experiment 4 decisional RT and accuracy plots, for trials where irregular instructions were given, to those generated by the advanced EPIC model.
Table 26.

*Observed vs. modelled accuracy for Experiment 4. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th></th>
<th>Regular Instruction Trials</th>
<th></th>
<th>Irregular Instruction Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2AFC</td>
<td>3AFC</td>
<td>4AFC</td>
</tr>
<tr>
<td><strong>Observed p(correct)</strong></td>
<td>.70</td>
<td>.63</td>
<td>.58</td>
</tr>
<tr>
<td><strong>Modelled p(correct)</strong></td>
<td>.71</td>
<td>.64</td>
<td>.57</td>
</tr>
<tr>
<td><strong>Observed p(correct)</strong></td>
<td>.62</td>
<td>.57</td>
<td>.55</td>
</tr>
<tr>
<td><strong>Modelled p(correct)</strong></td>
<td>.76</td>
<td>.64</td>
<td>.58</td>
</tr>
</tbody>
</table>
Complex decisions

Model results: Mean confidence and post-decisional time to decide confidence. While various attempts were made to replicate the observed post-decisional time to render confidence effects, these tests were ultimately unsuccessful. In order for the module to generate a post-decisional time to render confidence greater than 1, it must have a criterion value ‘K’ which is larger than the total number of evidence accrual events, pro and con, generated at the time the overt decision is made. Since the module assumes that parallel processing of confidence doubles decisional RT for as long as confidence is being processed, the model was never able to generate a post-decisional effect without distorting the decisional RT plots. As such, modelled post-decisional response times were held to a fast and flat value of ‘1’ throughout the tests.

As noted in the introduction to this section, confidence criterion values ‘K’ were allowed to vary as a function of both decisional complexity and decisional difficulty. Systematic trial and error testing revealed that when K was allowed to vary as a function of decisional complexity, but not difficulty, the generated mean confidence plots approximated those observed (Table 27 and Figures 36 – 37).

For the regular instruction set, ‘K2AFC’ was set to 3 while ‘K3AFC’ and ‘K4AFC’ were set to 4. While these settings provided the closest fit to the observed data, the model tended towards under-confidence, most particularly when making hard decisions (Table 27, Figure 36). For the irregular instruction set, ‘K2AFC’ was set to 5 while ‘K3AFC’ and ‘K4AFC’ were set to 6. It is interesting to note that setting the levels ‘K’ higher for the irregular instruction set allowed the model to reproduce not only the general form of the irregular instruction plots, but also the observation that participants were less confident
Complex decisions following irregular instructions (mean confidence regular: 78.27, mean confidence irregular: 71.87). This is because confidence in the module is based on the relative proportion of mode evidence, and this value will tend to decrease as the quantity of evidence increases (see Chapter 6). It is also worth reiterating that these shifts in criterion tend to generally reflect changes in post-decisional times to render confidence.

The model was also able to mimic the main effect of complexity on confidence, in that it grew less confident as complexity increased (2AFC: 79.57, 3AFC: 74.37, 4AFC: 71.26), as well as the main effect of difficulty (easy: 75.81, medium: 74.83, hard: 74.57).

Table 27.

*Observed vs. Mode-Based Modelled Confidence for Experiment 4.*

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Complexity</th>
<th>Difficulty</th>
<th>Observed Confidence</th>
<th>Modelled Confidence</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular Instructions</td>
<td>2AFC</td>
<td>easy</td>
<td>85.48</td>
<td>81.97</td>
<td>3.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>84.74</td>
<td>79.54</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>85.14</td>
<td>76.25</td>
<td>8.89</td>
</tr>
<tr>
<td></td>
<td>3AFC</td>
<td>easy</td>
<td>82.47</td>
<td>80.35</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>80.38</td>
<td>78.23</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>80.04</td>
<td>75.25</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>4AFC</td>
<td>easy</td>
<td>81.45</td>
<td>80.12</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>77.89</td>
<td>77.95</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>77.82</td>
<td>74.76</td>
<td>3.06</td>
</tr>
<tr>
<td>Irregular Instructions</td>
<td>2AFC</td>
<td>easy</td>
<td>85.98</td>
<td>79.41</td>
<td>6.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>86.10</td>
<td>70.27</td>
<td>15.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>86.02</td>
<td>67.42</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>3AFC</td>
<td>easy</td>
<td>79.04</td>
<td>78.10</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>77.30</td>
<td>70.14</td>
<td>7.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>77.49</td>
<td>66.93</td>
<td>10.56</td>
</tr>
<tr>
<td></td>
<td>4AFC</td>
<td>easy</td>
<td>74.56</td>
<td>77.46</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>med</td>
<td>73.23</td>
<td>70.19</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hard</td>
<td>73.01</td>
<td>66.95</td>
<td>6.06</td>
</tr>
</tbody>
</table>
Figure 36. A comparison of observed Experiment 4 mean confidence and post-decisional time to render confidence plots, for trials where regular instructions were given, to those generated by the advanced EPIC model.
Figure 37. A comparison of observed Experiment 4 mean confidence and post-decisional time to render confidence plots, for trials where irregular instructions were given, to those generated by the advanced EPIC model.
Discussion of Experiment 4

Hick (1952) noted that mean decisional RTs increased in a nonlinear fashion along with increases in decisional complexity. The purpose of Experiment 4 was to determine whether the observed degree of nonlinearity could be affected by disproportionately allocating the levels of psychological discomfort, frustration, or sense of challenge that decision makers experience with increases in decisional complexity. The EPIC model predicts that, if the 4AFC task is disproportionately more challenging than the 3AFC of 2AFC tasks, then the degree of nonlinearity will be greater than when psychological discomfort is more evenly distributed throughout the levels of complexity.

The irregular instruction set was designed to give clever decision makers an opportunity to employ cognitive shortcuts when making 2AFC and 3AFC decisions. When asked whether such shortcuts were employed *all* participants said they used the 2AFC shortcut, and simply mentally reversed the displayed instruction: 2\(^{nd}\) MOST = LEAST, 2\(^{nd}\) LEAST = MOST. Twenty of the participants also realized that they could ignore the instruction in the 3AFC task, and simply pick the middle colour. Anecdotally, then, it would seem that the manipulation was successful. The decisional RT plots displayed in Figures 30 and 31 are of the type predicted by the assumptions underlying the EPIC model, and the model itself does an adequate job reproducing these plots (Figures 32 - 33).

Importantly, the experiment highlighted a short-coming in the EPIC mode-based confidence module. The effects of instruction and complexity on post-decisional times to render confidence were too fine to be captured by the module. The meaning of these
Complex decisions

effects, however, remains unclear. Certainly, decisional RTs for Experiment 4 were within the range of those observed in Experiment 2, so it would be hard to suggest that any more or less confidence processing was occurring in parallel to the primary decision-making processes within Experiment 4. The post decisional times to render confidence were typically faster than those observed in Experiments 2 and 3, so it would be equally hard to suggest that additional post-decisional confidence processing were occurring. And, yet, a significant effect of instruction and of complexity on time to render confidence was observed. These strange findings may prove fruitful ground for further study.

The utility of the EPIC model would be severely limited if it only proved capable of explaining the effect of inequitably proportioned discomfort over levels of decisional complexity. The final two experiments in this series were designed to test whether the EPIC model could be manipulated so as to account for such frequently observed psychophysical phenomena as the Semantic Congruity Effect (SCE).
Chapter 8: Experiments 5 - 6

The Semantic Congruity Effect

As early as 1911, Henmon noted the relative ease with which decision makers were able to choose the shorter of two relatively short lines compared to their performance when asked to choose the longer of two relatively short lines. This effect was tested empirically by Audley and Wallis in 1964 when they required their participants to make brighter/darker judgements of stimulus luminance for stimuli which varied along a scale of relative brightness. When participants were asked to choose the brighter of two relatively bright stimuli, responses were significantly faster than when these same participants were asked to choose the brighter of two relatively dark stimuli.

This interaction between stimulus magnitude and instruction has since become known as the Semantic Congruity Effect (SCE). SCEs occur when decision makers are asked to consider a dimension of a stimulus set which is typically associated with that set, as opposed to a dimension not typically considered associated with the set. For example, "wintry" is a representative feature of Ottawa weather in February, and it would be ‘natural’ to consider whether February 1\textsuperscript{st} was more representative of a winter day than was February 2\textsuperscript{nd}. In contrast, one would be slow to decide whether the local weather was "balmier" on February 1\textsuperscript{st} or February 2\textsuperscript{nd}, since one does not typically rate winter days in Ottawa along a scale of tropical "balminess".

It has been suggested that "(a)ny complete theory of the process of comparing either perceptual or remembered stimuli must provide an explanation for the semantic congruity effect" (Shaki, Leth-Steensen, & Petrusic, 2006; pp. 196). One might qualify
this statement by suggesting that, before any truly complete theory can be developed, it is
first necessary to demonstrate that SCEs are not simply artefacts of comparably simple
2AFC decisions, and whether SCEs vary as a function of decisional complexity. Once
these data have been obtained, a theory would have to explain these observed, or
unobserved, effects.

It is also be important to determine whether other, non-RT related SCE effects
extend to the nAFC case. For example, Petrusic and Baranski (1989) demonstrated how
mean confidence levels increased for judgements of semantically congruous stimuli
relative to judgements of semantically incongruous stimuli. If this effect is observed
within a 2AFC paradigm but not at higher levels of decision complexity, then a complete
type of decision-making would have to explain this phenomenon. Experiments 5 and
6 were designed to test whether SCE can occur in 3AFC or 4AFC decisions.

**General Method Experiments 5-6**

**Apparatus**

The apparatus used in these experiments was similar to that used in Experiments
1-4, the only difference being the layout of the response panel. In addition to the
confidence input keys, and instead of the colour response keys, the panel was labelled
with four directional “arrow keys” positioned over the keyboard number pad. The layout
of the response panel is illustrated in Figure 38.
Complex decisions

Figure 38. The layout of the response panel used in Experiments 5-6.

Materials / Stimuli

Examples of the stimuli used in this experiment are found in Figure 39. Each of the stimuli were comprised of four X’s arranged to form a cross around a central fixation point. No two X’s within a single stimulus were the same distance away from this point, so that one X was always nearest and one X was always farthest away. The reason four X’s were always presented, regardless of level of decisional complexity, was to control for changes in RT related to the visual scanning of the stimuli. Without this control, it could reasonably be assumed that it would take a decision maker comparatively longer to scan a stimulus display containing a greater number of X’s.

Two sets of stimuli were employed. The "CLOSER" set of X’s were randomly positioned 102, 104, 106, or 108 pixels away from the fixation point. The "FARTHER" set of X’s were randomly positioned 204, 208, 212, or 216 pixels away from the fixation point.

X’s were randomly coloured either black or red, with a minimum of 2 and a maximum of 4 X’s being coloured red on any given trial. Since participants were told to
Complex decisions

attend only to the red X's, the number of red X's was an indication of decisional complexity.

With 2 stimulus sets, 2 instructions, ("CLOSEST", "FARTHEST"; see Procedure section below), \(4C_2=6\) possible pairs of X's to colour red, and \(4C_2 = 6\) ways to arrange these pairs around the fixation point, there were \(2\times2\times6\times6 = 144\) possible 2AFC stimuli.

Sixty-four 3AFC stimuli were generated: 2 sets \(\times\) 2 instructions \(\times\) \((4C_3 = 4\) possible triplets) \(\times\) \((4C_3 = 4\) ways to arrange each triplet) = 64. These 64 stimuli were each presented twice, and a randomly selected 16 were presented a third time for a total 144 3AFC stimuli.

Since only \(4\times4\times1\times1 = 16\) 4AFC could be generated with the method used to generate the 2AFC and 3AFC stimuli, a slightly different method was used to generate the 4AFC stimuli. All possible permutations were used to arrange the X's about the fixation point, resulting in 2 sets \(\times\) 2 instructions \(\times\) 24 arrangements = 96 stimuli, a random 48 of which were presented twice for a total 144 4AFC stimuli.
Complex decisions

Figure 39. Stimulus examples. Panel A represents a 2AFC congruous decision, since the X's are relatively close to the centre of the display. Panel B represents a 3AFC incongruous decision, since the X's are relatively distant from the centre of the display.

Procedure

On each trial, participants were asked to decide which red × was either “CLOSEST” or “FARTHEST” from the central fixation point depending on the instruction printed below each stimulus. Following each decision, participants were asked to make confidence judgements. Stimuli were blocked according to level of decisional complexity, and complexity presentation order varied between participants. Trials were randomized within blocks.
Participants

Twenty-four first-year psychology students from Carleton University, 16 female and 8 male participated in Experiment 5, and were asked to take part in a single one-hour session in return for course credit. Twenty-four first-year psychology students from Carleton University, 15 female and 9 male participated in Experiment 6, and were asked to take part in two one-hour sessions in return for course credit. Participants were drawn from a pool with a median age of 20.23. Two participants failed to complete Experiment 6 and were replaced.

Analytical Method and Results

Again, trials where RT was less than 200 ms and trials where RT was more than 3 standard deviations above each participant's mean within each level of decisional complexity and decisional difficulty were removed from analysis. Removed trials accounted for 2.15% of the observations made in Experiment 5 and 1.58% of the observations made in Experiment 6. Block order was included in all analyses as a between-subject factor but, as this factor never achieved significance, the results of any between-subjects analyses are not reported. This was true of condition order in Experiment 6 as well. Huynh-Feldt dfs were again used to assess statistical significance, all p(correct) values used in all ANOVAs were adjusted using Berkson’s correction, and all 95% confidence intervals were calculated, where applicable, using the Loftus and Masson (1994) method for within-participant designs. Post-hoc analyses of significant main effects, consisted of a comprehensive series of all possible pairwise comparisons.
Complex decisions within each effect, with the appropriate Bonferroni adjustment to the alpha-per-comparison made within each family of comparisons as required to maintain a family-wise significance level of .05. Post-hoc analyses of significant interactions consisted either of the derivation of Tukey’s HSD values and the subsequent evaluation of all possible pairwise comparisons of all cells involved in the interaction, or of Scheffé comparisons of contrasting groups of interest where the latter set of tests provided a more meaningful interpretation of the results.

Experiment 5

Purpose

The point of Experiment 5 was to determine whether any significant interactions between stimulus set and instruction on either decisional RT or mean expressed confidence could be observed within an experiment that also manipulated decisional complexity. If both had been found, this experiment would have represented the first time such SCEs were observed in a complexity experiment.

While the relationship between decisional complexity and decisional RT has been observed in many studies, it is unclear whether this relationship would vary as a function of the SCE. If so, a reliable three-way interaction between decisional complexity, stimulus set, and instruction would emerge. In other words, the linearity of the complexity-related decisional RT plots would vary as a function of semantic congruity. Such a finding would lend support to the EPIC model by suggesting, as was done in
Complex decisions

Experiment 4, that decisional processes can be measurably affected by manipulating the relative ease with decisions of varying complexity are made.

Results

Results are listed in Table 28, and illustrated in Figures 40-44.
Table 28.

**Statistical results from analyses of Experiment 5. The bolded values are the effects related to semantic congruity.**

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Dependent variables</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Response Time</td>
<td>Accuracy</td>
<td>Confidence</td>
<td>Time to Decide Confidence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>F</td>
<td>$\eta_p^2$</td>
<td>F</td>
<td>$\eta_p^2$</td>
</tr>
<tr>
<td><strong>Experiment 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>2, 36</td>
<td>7.49**</td>
<td>.29</td>
<td>131.94***</td>
<td>.88</td>
</tr>
<tr>
<td>Set</td>
<td>1, 18</td>
<td>23.13***</td>
<td>.56</td>
<td>3.60 ns</td>
<td>.17</td>
</tr>
<tr>
<td>Instruction</td>
<td>1, 18</td>
<td>9.05**</td>
<td>.34</td>
<td>6.62*</td>
<td>.27</td>
</tr>
<tr>
<td>Complexity x Instruction</td>
<td>2, 36</td>
<td>2.19 ns</td>
<td>.11</td>
<td>6.87**</td>
<td>.28</td>
</tr>
<tr>
<td><strong>Set x Instruction</strong></td>
<td><strong>1, 18</strong></td>
<td><strong>3.06 ns</strong></td>
<td><strong>.15</strong></td>
<td><strong>1.82 ns</strong></td>
<td><strong>.19</strong></td>
</tr>
</tbody>
</table>

*Note.* Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see text); While non-significant main effects are reported, only significant interactions are reported. *** p < .001; ** p < .01; *p < .05; ns = not significant.
Experiment 5: Main Effects of Decisional Complexity

Figure 40. Observed main effects of decisional complexity for Experiment 5. 95% confidence intervals are shown.
Experiment 5: Main Effects of Stimulus Set

Figure 41. Observed main effects of stimulus set for Experiment 5. 95% confidence intervals are shown.

Complex decisions
Complex decisions

Experiment 5: Main Effects of Instruction

**Decisional Response Time**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Closest</th>
<th>Farthest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Confidence Rating (msec)</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>

**Decisional Accuracy**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Closest</th>
<th>Farthest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Decisional Accuracy</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**Mean Confidence Rating**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Closest</th>
<th>Farthest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Confidence Rating</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

**Time to Render Confidence**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Closest</th>
<th>Farthest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Decisional Response Time (msec)</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

*Figure 42.* Observed main effects of instruction for Experiment 5. 95% confidence intervals are shown.
Figure 43. Observed interactions between decisional complexity and instruction for Experiment 5. 95% confidence intervals are shown.
Interactions: Stimulus Set x Instruction

Figure 44. Observed interactions between decisional complexity and instruction for Experiment 5. 95% confidence intervals are shown.
**Decisional complexity.** The effects of decisional complexity are illustrated in Figure 40. Complexity had a reliable effect on decisional RT, and post-hoc tests showed that 2AFC decisions (4412.0 ms) were made more quickly than either 3AFC decisions (5526.9 ms) or 4AFC decisions (6136.2 ms). These later two groups did not significantly differ from one another, and the result was the expected pattern of nonlinear increases in decisional RTs. Complexity also had a reliable effect on decisional accuracy with post-hoc tests showing that all three groups differed significantly from one another, with accuracy decreasing as complexity increased. No other significant main effects of decisional complexity were found.

**Stimulus set.** The effects of stimulus set are illustrated in Figure 41. Set only seemed to reliably affect decisional RT, with decisions made for the "CLOSE" set being made more quickly than decisions for the "FAR" set. It is possible, though not certain, that this discrepancy was caused by the greater amount of visual scanning required to compare stimulus features situated farther from the central fixation point.

**Instruction.** The effects of instruction set are illustrated in Figure 42. Instruction had a reliable affect on all four DV's of interest. Participants were faster, more accurate, more confident, and took less time to express a confidence rating whenever the instruction was "FARTHEST". It is unclear why this should have been the case. Perhaps participants are naturally "hard-wired' to make judgements of distance rather than proximity, and that these latter types of judgements required more effort. If this is the case, and if the assumptions made by the EPIC model are correct, then one would expect to see an interaction between decisional complexity and instruction on RT, with
"FARTHEST" instruction RTs increasing in a less linear fashion with increases in complexity than "CLOSEST" instruction RTs. This would occur because "CLOSEST" decisions would be more uniformly uncomfortable than would "FARTHEST" decisions. Though this interaction was not reliable ($p = .1$), the pattern predicted by the EPIC model is nevertheless apparent in Figure 43.

**Decisional complexity × instruction.** The interactions between decisional complexity and instruction are illustrated in Figure 43. The only reliable interaction was on decisional accuracy: 2AFC decisions for both instructions were reliably more accurate than any other decisions, and 3AFC "FARTHEST" decisions were more accurate than 4AFC "CLOSEST" decisions, HSD = 4.23, $q(.05, 6, 36) = 4.23$.

**Stimulus set × instruction: the SCEs.** These interactions were the impetus of the present study. The point of Experiment 5 was to generate a decisional RT SCE and a mean confidence rating SCE, and then to see if these could be replicated with the EPIC model. Unfortunately only the later effect was observed. As can be seen in Figure 44, participants were more confident making congruent decisions than incongruent decisions, though the only reliable pairwise difference was between "CLOSER" set incongruent and "FARTHER" set congruent trials. In other words, participants were only reliably less confident responding to the instruction "CLOSEST" with "FARTHER" stimuli. No other pairwise differences related to the confidence SCE were reliable, however a Scheffé contrast between incongruent and congruent groups of means was significant, $t(18) = 4.56$, $p < .05$. 

170
The interaction on post-decisional time to decide confidence was reliable, though none of the pairwise comparisons were. A Scheffé contrast between incongruent and congruent groups of means, however, revealed that congruent decisions led to faster post-decisional times to render confidence judgements, $t(18) = 3.5735, p < .05$.

**Discussion of Experiment 5**

The anticipated three-way interaction between complexity, stimulus set, and instruction on RT did not emerge. This result might imply that the nonlinear relationship between decisional complexity and decisional RT does not vary as a function of semantic congruity. However, because the basic SCE on decisional RT was not observed, Experiment 5 proved to be an ineffective tool for allowing one to draw conclusions about this hypothesized interaction.

The SCE related to confidence did emerge, and participants were reliably less confident making decisions where the instruction was incongruous with the stimulus set. This finding, coupled with what appears to be a SCE related to post-decisional times to express confidence, implies that the two stimulus sets used in Experiment 5 were sufficiently different to generate congruent and incongruent stimulus classifications. Additionally, both the SCE on decisional RT and mean expressed confidence rating were observed in a sixth experiment, wherein the number of decisions required of participants was doubled. As such, it is possible that the otherwise ubiquitous SCE related to decisional RTs was not observed in Experiment 5 due to the lack of experimental power.
Complex decisions

associated with the test: $I-\beta = .38$. As such, the data obtained from Experiment 6 were used in the tests of the EPIC model to be described later in this chapter.

### Experiment 6

**Purpose**

The purpose of Henmon's aforementioned study of extent comparisons was to investigate the relation of the time of a judgment to its accuracy. He determined that there was an optimal response time at which point accuracy is maximized (1911). Garrett's (1922) "Study of the Relation of Accuracy to Speed" confirmed this finding, and demonstrated that accuracy in line-length comparisons suffered when participants took either too long or not enough time to make their decisions. Hick himself noted that participants could make quick decisions, but that they did so by sacrificing accuracy (1952).

Race models, such as the EPIC model, should be ideally suited to replicate speed-accuracy tradeoffs. Choosing an arbitrarily high decisional criterion setting in a race model will force the model to spend extra time gathering evidence, and will increase the probability that total evidence sampled is representative of the population from which it is drawn. On the other hand, setting a low criterion will result in a quickly drawn, small sample of evidence which is less likely to lead to an accurate decision. Experiment 6 was originally designed to generate speed-accuracy data within an experiment which also manipulated decisional accuracy, and then to see whether the EPIC race model was sufficiently powerful to replicate the observed data trends.
Complex decisions

As it turned out, the design of Experiment 6 also allowed for a replication of Experiment 5 but with a twice as many decisions made by each participant. These extra trials significantly increased the power of the experimental paradigm and, as will be seen, the SCEs which failed to emerge in Experiment 5 were reliable in Experiment 6.

Method

Procedure. One of the Experiment 6 sessions served as an exact replication of Experiment 5. Within the second session, participants were asked to make the same types of decision made in Experiment 5, but were instructed to emphasize decisional speed over decisional accuracy.

Speed condition blocks used the same stimuli and apparatus as the accuracy condition blocks. The deadlines for the speeded conditions were ½ of the mean decisional RT observed for each level of decisional complexity in Experiment 5 (i.e., 2206 ms, 2763 ms, and 3068 ms). Whenever participants made a decision which failed to meet these deadlines, a tone sounded and the message "TOO SLOW!" was flashed on the screen for two seconds. Participants were encouraged to "avoid the tone, even if you have to sacrifice decisional accuracy to do it". As was the case in Experiment 5, participants were asked to render confidence following each decision. Where necessary, these confidence judgements were made following the "TOO SLOW!" admonition.
Results

All IV's produced at least one significant effect, as did all possible two-way interactions between these variables. What is more, there were three significant three-way interactions. All of these results are listed in Table 29, and illustrated in Figures 45-56. Each significant effect will be discussed in turn, and this will be followed by a general review of the findings most relevant to the present thesis.
Table 29. Statistical results from analyses of Experiment 6. The bolded values are the effects related to semantic congruity.

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>df</th>
<th>F</th>
<th>η²p</th>
<th>F</th>
<th>η²p</th>
<th>F</th>
<th>η²p</th>
<th>F</th>
<th>η²p</th>
<th>F</th>
<th>η²p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity (nAFC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set</td>
<td>1, 12</td>
<td>18.80***</td>
<td>.61</td>
<td>2.45ns</td>
<td>.17</td>
<td>8.55*</td>
<td>.42</td>
<td>9.31**</td>
<td>.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instruction (I)</td>
<td>1, 12</td>
<td>5.58*</td>
<td>.32</td>
<td>3.51ns</td>
<td>.23</td>
<td>.00ns</td>
<td>.00</td>
<td>.01ns</td>
<td>.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition (s/a)</td>
<td>1, 12</td>
<td>115.89***</td>
<td>.91</td>
<td>99.21***</td>
<td>.89</td>
<td>14.17**</td>
<td>.54</td>
<td>2.64ns</td>
<td>.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nAFC × Set</td>
<td>2, 24</td>
<td>1.81ns</td>
<td>.13</td>
<td>4.07*</td>
<td>.25</td>
<td>.476ns</td>
<td>.04</td>
<td>1.82ns</td>
<td>.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nAFC × I</td>
<td>2, 24</td>
<td>3.58*</td>
<td>.23</td>
<td>2.00ns</td>
<td>.14</td>
<td>5.96*</td>
<td>.33</td>
<td>.74ns</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nAFC × s/a</td>
<td>2, 24</td>
<td>4.92*</td>
<td>.29</td>
<td>.08ns</td>
<td>.01</td>
<td>1.69ns</td>
<td>.12</td>
<td>1.30ns</td>
<td>.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set × I</td>
<td>1, 12</td>
<td>26.21***</td>
<td>.69</td>
<td>.86ns</td>
<td>.07</td>
<td>.17.66**</td>
<td>.60</td>
<td>.13ns</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set × s/a</td>
<td>1, 12</td>
<td>11.13**</td>
<td>.48</td>
<td>2.24ns</td>
<td>.16</td>
<td>.17ns</td>
<td>.01</td>
<td>1.70ns</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I × s/a</td>
<td>1, 12</td>
<td>4.98*</td>
<td>.29</td>
<td>.05ns</td>
<td>.00</td>
<td>4.85*</td>
<td>.29</td>
<td>1.43ns</td>
<td>.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nAFC × I × s/a</td>
<td>2, 24</td>
<td>2.13ns</td>
<td>.15</td>
<td>5.31*</td>
<td>.31</td>
<td>.69ns</td>
<td>.06</td>
<td>.42ns</td>
<td>.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set × I × s/a</td>
<td>1, 12</td>
<td>19.63***</td>
<td>.62</td>
<td>.45ns</td>
<td>.04</td>
<td>11.59**</td>
<td>.49</td>
<td>.69ns</td>
<td>.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Analysis of Response Times in ms; Confidence ratings were on 6-point scales (see text); While non-significant main effects are reported, only significant interactions are reported. *** p < .001; ** p < .01; *p < .05; ns = not significant.
Experiment 6: Main Effects of Decisional Complexity

Figure 45. Observed main effects of decisional complexity for Experiment 6. 95% confidence intervals are shown.
Experiment 6: Main Effects of Stimulus Set

![Graphs showing main effects of stimulus set for Experiment 6.](image)

Figure 46. Observed main effects of stimulus set for Experiment 6. 95% confidence intervals are shown.
Figure 47. Observed main effects of instruction for Experiment 6. 95% confidence intervals are shown.
Experiment 6: Main Effects of Speed/Accuracy Condition

**Decisional Response Time**

- Mean Decisional Response Time (msec)

**Decisional Accuracy**

- Mean Decisional Accuracy

**Mean Confidence Rating**

- Mean Confidence Rating

**Time to Render Confidence**

- Time to Render Confidence (msec)

Figure 48. Observed main effects of condition for Experiment 6. 95% confidence intervals are shown.
**Interactions: Decisional Complexity x Stimulus Set**

**Decisional Response Time**

![Graph showing Mean Decisional Response Time (msec) vs Decisional Complexity.](image)

**Decisional Accuracy**

![Graph showing Mean Decisional Accuracy vs Decisional Complexity.](image)

**Mean Confidence Rating**

![Graph showing Mean Confidence Rating vs Decisional Complexity.](image)

**Time to Render Confidence**

![Graph showing Mean Decisional Response Time (msec) vs Decisional Complexity.](image)

*Figure 49.* Observed interactions between decisional complexity and stimulus set for Experiment 6. 95% confidence intervals are shown.
Figure 50. Observed interactions between decisional complexity and instruction for Experiment 6. 95% confidence intervals are shown.
Interactions: Decisional Complexity x Condition

Figure 51. Observed interactions between decisional complexity and condition for Experiment 6. 95% confidence intervals are shown.
Complex decisions

Interactions: Stimulus Set x Instruction

![Graph showing interactions between stimulus set and instruction for Experiment 6.](image)

Figure 52. Observed interactions between stimulus set and instruction for Experiment 6. 95% confidence intervals are shown.
Complex decisions

**Interactions: Stimulus Set x Condition**

**Decisional Response Time**

- Mean Decisional Response Time (msec)
- Stimulus Set: Closer vs. Farther

**Decisional Accuracy**

- Mean Decisional Accuracy
- Stimulus Set: Closer vs. Farther

**Mean Confidence Rating**

- Mean Confidence Rating
- Stimulus Set: Closer vs. Farther

**Time to Render Confidence**

- Mean Decisional Response Time (msec)
- Stimulus Set: Closer vs. Farther

*Figure 53.* Observed interactions between stimulus set and condition for Experiment 6. 95% confidence intervals are shown.
**Interactions: Instruction x Condition**

![Graphs showing interactions between instruction and condition for Experiment 6. 95% confidence intervals are shown.](image)

*Figure 54.* Observed interactions between instruction and condition for Experiment 6. 95% confidence intervals are shown.
Figure 55. Observed interactions between decisional complexity, instruction, and condition for Experiment 6. 95% confidence intervals are shown.
Figure 56. Observed interactions between stimulus set, instruction, and condition for Experiment 6. 95% confidence intervals are shown.

Complex decisions: Stimulus Set x Instruction

Accuracy Condition: Stimulus Set x Instruction

Mean Decisional Response Time (msec)

Mean Confidence Rating

Mean Decisional Accuracy

Complex decisions
**Decisional complexity.** The effects of decisional complexity are illustrated in Figure 45. As was the case in Experiment 5, complexity had a reliable effect on decisional RT, with 2AFC decisions (2791.8 ms) made more quickly than either 3AFC decisions (3867.8 ms) or 4AFC decisions (3970.2 ms). These later two groups did not significantly differ from one another. Complexity also had a reliable effect on decisional accuracy and, again paralleling the related observations in Experiment 5, all three groups differed significantly from one another, with accuracy decreasing as complexity increased. Unlike Experiment 5, however, a reliable effect of decisional complexity on mean confidence rating was found with participants being reliably more confident making 2AFC decisions (mean confidence = 76.70) than either 3AFC (74.27) or 4AFC (73.18) decisions. The observed 3AFC mean confidence rating did not differ significantly from the 4AFC mean confidence rating.

**Stimulus Set.** As was the case in Experiment 5, participants were slower making decisions related to the "FAR" stimulus set (3738.7 ms) than when making decisions related to the "CLOSE" stimulus set (3323.5 ms). Additionally, participants in Experiment 6 were reliably less confident responding to the "FAR" stimuli (74.12 vs. 75.37) and took an average of 49 ms longer to make these confidence judgements (827.2 ms vs. 778.2 ms). These effects are illustrated in Figure 46.

The main effect of stimulus set on mean post-decisional time to express confidence was the only significant effect related to this DV observed in Experiment 6. Times to express confidence did not vary a function of complexity, instruction, or speed/accuracy condition. This would suggest that, with the exception of the 49 ms
effect of stimulus set, confidence processing by participants in Experiment 6 was largely
completed by the time an overt decision was made.

**Instruction.** Response to the instruction "FARTHEST" were made more quickly
than responses to the instruction "CLOSEST" (Figure 47; 3439.2 ms vs. 3647.2 ms).
This was the only observed main effect related to instruction. This would seem to
support the conclusion reached in Experiment 5, that judgements of distance are more
natural than judgements of proximity – at least as far as these stimulus sets are concerned.

**Speed/Accuracy condition.** This manipulation had the intended effects, in that
speeded decisions were made significantly more quickly (1670.1 ms vs. 5405.1 ms), less
accurately (.43 vs. .53), and with less confidence (72.16 vs. 77.35) than decisions where
accuracy was emphasized (Figure 48). It is interesting to note that participants are
perfectly capable of sacrificing accuracy in favour of speed, and that mean confidence
ratings are reliably affected when this occurs. A test of the EPIC model will be to see
whether it is capable of easily reflecting the effects of this manipulation.

**Decisional complexity × stimulus set.** Accuracy varied as a function of this
interaction, though the pattern of this interaction remains unclear. Figure 49 suggests that
3AFC close set decisions were made more accurately than 3AFC far set decisions, while
the reverse was true of 4AFC decisions. Post-hoc pairwise comparisons, however,
revealed no significant differences either within these levels of complexity, or between
the member means of these levels of complexity. Furthermore, as has already been
noted, no significant main effect of stimulus set on decisional accuracy was found. As
such, the difference must either be attributable to an unidentified contrast between groups
Complex decisions of means, or was a tested difference that failed to reach significance due to the conservative nature of post-hoc tests planned *a posteriori*.

**Decisional complexity × instruction.** Both decisional RT and confidence varied as a function of this interaction. Regarding decisional RT, while the only obvious difference between the two instruction plots presented in Figure 50 would seem to be the difference between 4AFC decisions ("CLOSEST" = 4145.7 ms, "FARTHEST" = 3754.7 ms), the contrast proved insignificant, $t(24) = 2.10, p > .05$. This is a function of the conservative nature of this test, since an unadjusted critical value of $t(2A) = 2.06$ would have yielded a significant result. As has been noted, the apparent interaction illustrated in Figure 50 can easily be replicated by the EPIC model by allowing discomfort exponent 'a' to assume a higher value when the model makes "FARTHEST" decisions, thereby implying discomfort is slightly more evenly distributed throughout all levels of decisional complexity when "CLOSEST" decisions are made.

Mean expressed confidence levels for 2AFC "CLOSEST" decisions were greater than for any 3AFC or 4AFC decisions, though post hoc tests revealed that only 2AFC "FARTHEST" decisions were significantly greater than 4AFC "CLOSEST" decisions, HSD = 2.62, $q(.05, 6, 24) = 4.37$. No other pairwise comparisons were found to be significant.

**Decisional complexity × speed accuracy condition.** Perhaps one of the most important findings in this experiment, at least so far as the present thesis is concerned, was the finding the speeded decisions produce a more linear decisional RT plot than do accurate decisions (Figure 51). This implies that the Hick-like pattern of RTs observed in
all investigations of decisional complexity are, in part, a function of the time a decision maker takes to make a correct decision and are NOT simply a product of information bit-rate as originally suggested by Hick (1952). This finding will be discussed more thoroughly in the sections dealing with the EPIC model replication of Experiment 6.

**Stimulus set × instruction: The SCEs.** Unlike experiment 5, SCEs were found for both decisional RTs and mean expressed confidence rating (Figure 52). While the general contrast between decisional RTs for congruous and incongruous decisions was not significant, \( t(12) = -2.09, p > .05 \), congruent "FARTHEST" decisions were made significantly faster than incongruent "FARTHEST" decisions, \( t(12) = -2.80, p < .05 \) (Figure 52). Importantly, the contrast between congruent and incongruent trial expressed levels of confidence was significant, \( t(12) = 4.43, p < .05 \). It should be noted that the likely reason that SCEs were found for decisional RTs in Experiment 6 and not Experiment 5 was that the power of Experiment 6 was so much larger: Decisional RTs \( 1 - \beta = .997 \), confidence \( 1 - \beta = .970 \). This increase in power was likely attributable to the fact that participants in Experiment 6 were asked to make twice as many decisions as were participants in Experiment 5.

**Stimulus set × speed/accuracy condition.** As can be seen in Figure 53, the significant interaction between stimulus set and condition on decisional RT is due to accurate decisions differing significantly as a function of stimulus set, \( t(12) = 2.19, p < .05 \), while speeded decisions did not. No other stimulus set × condition interaction was found to be significant.
Complex decisions

**Instruction \times speed/accuracy condition.** Both decisional RT and mean confidence rating varied as a function of this interaction. A review of Figure 54 shows a greater difference between "FARTHEST" and "CLOSEST" decisions made under accuracy stress than under speed stress but, unfortunately, post-hoc tests did not reveal any significant differences between or within these conditions. Comparably, though a funnel-shaped interaction related to mean confidence rating seems apparent in Figure 54, no post-hoc test revealed any significant differences between these four means. These failures to find significance are likely due to the conservative nature of post-hoc tests planned *a posteriori*.

**Three-way interaction between decisional complexity, instruction, and speed/accuracy condition on mean accuracy rating.** Again, no post-hoc tests planned *a posteriori* proved significant, but Figure 55 would suggest that speeded, 2AFC, "CLOSEST" decisions were faster than "FARTHEST" decisions, while the reverse was true for accurate, 2AFC decisions. It must be stressed, however, that no attempted Scheffé test of contrasting groups yielded a significant result.

**Three-way interaction between stimulus set, instruction, and speed/accuracy condition on decisional RTs and mean confidence rating: SCEs varying as a function of speed/accuracy condition.** These interactions are displayed in Figure 56. The most apparent cause of the three-way interaction related to decisional RT is the obvious SCE for accuracy blocks, $t(12) = 4.28, p < .05$, where none exists for the speeded blocks, $t(12) = .50, p > .05$. And while SCEs related to confidence were found for both speed, $t(12) = 4.99, p < .05$, and accuracy, $t(12) = 8.03, p < .05$, the difference between
means for the "FARTHER" stimulus set within the accuracy blocks was significant, $t(12) = 2.46$, $p < .05$, while this difference was not significant within the speeded set, $t(12) = .94$, $p > .05$.

**Discussion of Experiment 6**

The important findings, at least as far as the present thesis is concerned, are numerous. Not only were the observed main effects of decisional complexity on decisional RT, confidence, and accuracy similar to those observed in Experiments 1-3, but four additional sets of interactions between decisional complexity, stimulus set, instruction type, and speed/accuracy condition were also observed. Of these, perhaps the most pertinent finding is that the degree of nonlinearity in decisional RT plots decreased for speeded trials, while mean confidence plots reflected no comparable change.

Also of interest were the observed SCEs related to decisional RTs and to mean confidence ratings. While the observed two-way interactions between instruction and stimulus set were typical of SCEs (e.g., Shaki et al, 2006), the observed three-way interaction between speed/accuracy condition, stimulus set, and instruction were comparably novel. These showed that the SCE related to RT disappeared within the speed blocks, yet reliable confidence-based SCEs were found for both speed and accuracy blocks.

These two groups of findings provide yet more evidence that mean confidence ratings and decisional RTs are distinct phenomena. While main effect plots do not
Complex decisions

contradict the theory that confidence is a function of scaled RT, the observed interaction
plots make this theory seem a less likely explanation.

The final challenge for the EPIC model will be to see whether it is capable of
replicating the many observed significant main effects and interactions related to
decisional complexity and to the SCEs observed in Experiment 6.

The EPIC Model and Experiment 6

**Sample selection and probability values.** The experimental paradigm employed
in Experiments 5 and 6 necessitated a change in the way the EPIC model simulated
decisional difficulty. As demonstrated in Table 30, there were six extent length
comparisons possible in the 2AFC blocks and each of these were associated with
different probabilities that any sampled evidence would support a correct decision. On
the other hand, only four 3AFC comparisons and one 4AFC comparison was possible
given these stimulus sets. So while the model was altered to allow for six different
difficulty levels, it was also forced to generate an equal number of decisions at each level
of decisional complexity. As such, all 4AFC decisions were made using a single
difficulty level and two of the possible six 3AFC difficulty levels were repeated for two-
sixths of the trials.

"CLOSEST" and "FARTHEST" decisions each used the same probability levels.
For example, when the extents were 104 and 108 pixels long, it was assumed that there
was a 51% chance that the evidence would support the correct decision regardless of the
instruction.
Complex decisions

Table 30. Possible extent comparisons using the stimuli in Experiment 6, and the related probabilities associated with modelled sample selection for the instruction "FARTHER". Probabilities were calculated by dividing each extent length by the summed lengths of all extents being compared.

<table>
<thead>
<tr>
<th>nAFC</th>
<th>Possible extent comparisons: CLOSER stimuli</th>
<th>Possible extent comparisons: FARTHER stimuli</th>
<th>Probability sampled evidence supports extent as &quot;FARTHEST&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>102 – 104</td>
<td>204 – 208</td>
<td>0.4951 - 0.5049</td>
</tr>
<tr>
<td></td>
<td>102 – 106</td>
<td>204 – 212</td>
<td>0.4904 - 0.5096</td>
</tr>
<tr>
<td></td>
<td>102 – 108</td>
<td>204 – 216</td>
<td>0.4857 - 0.5143</td>
</tr>
<tr>
<td></td>
<td>104 – 106</td>
<td>208 – 212</td>
<td>0.4952 - 0.5048</td>
</tr>
<tr>
<td></td>
<td>104 – 108</td>
<td>208 – 216</td>
<td>0.4906 - 0.5094</td>
</tr>
<tr>
<td></td>
<td>106 – 108</td>
<td>212 – 216</td>
<td>0.4953 - 0.5047</td>
</tr>
<tr>
<td>3</td>
<td>102 – 104 – 106</td>
<td>204 – 208 – 212</td>
<td>0.3269 - 0.3333 - 0.3397</td>
</tr>
<tr>
<td></td>
<td>102 – 104 – 108</td>
<td>204 – 208 – 216</td>
<td>0.3248 - 0.3312 - 0.3439</td>
</tr>
<tr>
<td></td>
<td>102 – 106 – 108</td>
<td>204 – 212 – 216</td>
<td>0.3228 - 0.3354 - 0.3418</td>
</tr>
<tr>
<td></td>
<td>104 – 106 – 108</td>
<td>208 – 212 – 216</td>
<td>0.3270 - 0.3333 - 0.3396</td>
</tr>
<tr>
<td>4</td>
<td>102 – 104 – 106 – 108</td>
<td>204 – 208 – 212 – 216</td>
<td>0.2429 - 0.2476 - 0.2524 - 0.2571</td>
</tr>
</tbody>
</table>

**Number of parameters.** Since the number of available decisional difficulty levels varied as a function of decisional complexity, base decisional criteria were not allowed to vary as a function of decisional difficulty. Instead, base decisional criteria and the discomfort exponent 'a' were allowed to vary as a function of stimulus set, and separate parameter sets were derived for each instruction at each level of the speed/accuracy condition.

**Model parameter selection.** Four versions of the model were run, each attempting to find the best fit for the observed quantiles associated with a separate speed/accuracy condition × instruction cell. All told, the iterative process used to select
the best fitting parameter sets required the EPIC model had to make 972,000 separate decisions and took approximately four days to run. While parameter selection required only 1000 decisions per complexity × condition × instruction × stimulus set cell, the selected parameters were re-tested with 10,000 decisions per cell in order to confirm that the derived BIC values were stable. These were all deemed stable since, in all four cases, |BIC new − BIC old| < 100. The selected parameters are listed in Table 31.

Table 31.

*Best fitting base criterion* $C_{base}$ and discomfort exponent 'a' values for the EPIC model replication of Experiment 6.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Instruction</th>
<th>Stimulus Set</th>
<th>$C_{base}$</th>
<th>'a'</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Farthest</td>
<td>Farther</td>
<td>12</td>
<td>5</td>
<td>25599.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Closer</td>
<td>12</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Closest</td>
<td>Farther</td>
<td>14</td>
<td>4</td>
<td>25249.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Closer</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td>Farthest</td>
<td>Farther</td>
<td>3</td>
<td>4</td>
<td>25777.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Closer</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Closest</td>
<td>Farther</td>
<td>4</td>
<td>4</td>
<td>25795.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Closer</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

*Confidence criteria.* The sheer number of observed interactions involving mean expressed confidence level necessitated the implementation of varying confidence criteria within the confidence module. Interestingly, trial-and-error tests established that the same confidence adjustment made within all four versions of the EPIC model run was sufficient to reproduce the observed effects: Base confidence criterion 'K' was set to 5, was lowered by 1 for 2AFC decisions, and was lowered by 1 for semantically congruent decisions. As such, $K = 3$ for all 2AFC congruent decisions, $K = 4$ for all 2AFC
incongruent decisions or all 3AFC and 4AFC congruent decisions, and $K = 5$ for all 3AFC and 4AFC incongruent decisions. These values allowed modelled confidence to be within 5% of observed confidence in all of the simulations.

**Model results.** Model comparisons to observed data are plotted in Figures 57 – 64, and are described in the following sections.

**Decisional complexity.** Model replications of the main effects of decisional RT, decisional accuracy, and mean expressed confidence levels are plotted in Figure 57. The model did a reasonable job reproducing the observed decisional RT and confidence plots (see Table 33) but was quite under-accurate when making 4AFC decisions (Table 32). As will be seen in the rest of these analyses, the EPIC model struggled to reproduce any of the observed significant effects involving accuracy at the 4AFC level of complexity.

Table 32.

*Observed vs. modelled main effect of decisional complexity on accuracy for Experiment 6 data. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th></th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed p(correct)</td>
<td>.60</td>
<td>.44</td>
<td>.40</td>
</tr>
<tr>
<td>Modelled p(correct)</td>
<td>.58</td>
<td>.41</td>
<td>.32</td>
</tr>
</tbody>
</table>

**Decisional complexity × set.** Once again, as can be seen in Figure 58 and in Table 34, the EPIC model was unable to reproduce the observed accuracy levels for 4AFC decisions within the confines of this interaction. This is likely due to the close-to-
Complex decisions

chance-level difficulty values assigned 4AFC decisions within the model (see discussion above).

Table 33.

*Observed vs. modelled main effect of decisional complexity on mean confidence rating for Experiment 6.*

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Observed Confidence %</th>
<th>Modelled Confidence %</th>
<th>Difference %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>76.72</td>
<td>79.16</td>
<td>-2.44</td>
</tr>
<tr>
<td>3</td>
<td>74.31</td>
<td>75.95</td>
<td>-1.64</td>
</tr>
<tr>
<td>4</td>
<td>73.21</td>
<td>73.71</td>
<td>-.5</td>
</tr>
</tbody>
</table>

Table 34.

*Observed vs. modelled interaction between stimulus set and decisional complexity on accuracy for Experiment 6 data. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th>Stimulus set</th>
<th>p(correct) values</th>
<th>2AFC</th>
<th>3AFC</th>
<th>4AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FARTHER SET</td>
<td>Observed</td>
<td>.61</td>
<td>.46</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td>.58</td>
<td>.41</td>
<td>.32</td>
</tr>
<tr>
<td>CLOSER SET</td>
<td>Observed</td>
<td>.60</td>
<td>.42</td>
<td>.40</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td>.57</td>
<td>.40</td>
<td>.31</td>
</tr>
</tbody>
</table>

*Decisional complexity × instruction.* The model was able to reproduce the finding that responses to the instruction FARTHEST were faster and less linear across levels of decisional complexity than were responses to the instruction CLOSEST (Figure 59). The model was also able to approximate the effects of this interaction on mean
expressed confidence rating. The model tended toward overconfidence for FARTHEST 2AFC decisions, but was within 5% of the observed confidence ratings in all cases (Table 35).

Table 35.

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Complexity</th>
<th>Observed Confidence %</th>
<th>Modelled Confidence %</th>
<th>Difference %</th>
</tr>
</thead>
<tbody>
<tr>
<td>FARTHEST</td>
<td>2</td>
<td>76.20</td>
<td>79.50</td>
<td>-3.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>74.36</td>
<td>75.96</td>
<td>-1.6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>72.67</td>
<td>73.45</td>
<td>-0.78</td>
</tr>
<tr>
<td>CLOSEST</td>
<td>2</td>
<td>77.25</td>
<td>78.82</td>
<td>-1.57</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>74.26</td>
<td>75.93</td>
<td>-1.67</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>72.67</td>
<td>73.96</td>
<td>-1.29</td>
</tr>
</tbody>
</table>

**Decisional complexity × speed/accuracy condition.** Importantly, the EPIC model was able to reproduce the Hick-related interaction between decisional complexity and the speed/accuracy condition (Figure 60). As was the case with the observed data, the model decisional RT plots were less linear in the accuracy condition than in the speeded condition. This result is all the more fascinating when one considers that the derived discomfort exponent was, on average, not much higher in the accuracy condition (a = 4.5) than in the speeded condition (a = 4).

**Stimulus set × instruction: The SCEs.** The model was able to reproduce the observed SCE on decisional RT and on mean expressed confidence rating (Figure 61, Table 36). This is, of course, not surprising. As has been noted, race models are ideal
Complex decisions

candidates for replicating SCEs, since adjustments to the decisional criteria simultaneously affect both RT and accuracy. From a philosophical point-of-view, however, it should be noted that while these models can reproduce a SCE they do so without any actual understanding of the semantics of the instruction nor an appreciation of the difference between the stimulus sets. This point will be revisited in Chapter 9.

Table 36.

*Observed vs. modelled interaction between stimulus set and instruction on mean confidence rating for Experiment 6.*

<table>
<thead>
<tr>
<th>Stimulus set</th>
<th>Instruction</th>
<th>Observed Confidence</th>
<th>Modelled Confidence</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>FARTHER</td>
<td>FARTHEST</td>
<td>75.13</td>
<td>77.50</td>
<td>-2.37</td>
</tr>
<tr>
<td></td>
<td>CLOSEST</td>
<td>74.41</td>
<td>75.11</td>
<td>-0.70</td>
</tr>
<tr>
<td>CLOSER</td>
<td>FARTHEST</td>
<td>73.10</td>
<td>74.80</td>
<td>-1.70</td>
</tr>
<tr>
<td></td>
<td>CLOSEST</td>
<td>76.32</td>
<td>77.68</td>
<td>-1.36</td>
</tr>
</tbody>
</table>

*Decisional Complexity × stimulus set × instruction.* The results of the EPIC model's attempt to replicate the effect of this interaction on decisional accuracy are illustrated in Figure 62 and in Table 37. Again, the model fared poorly making 4AFC decisions, did slightly better making 2AFC and 3AFC accurate choices, and performed best of all making speeded 2AFC and 3AFC decisions. Though a within model comparison reveals that EPIC was slightly more accurate when in the accuracy condition relative to the speeded condition, it is worth noting that the mean base criterion level for the accuracy condition ($C_{base} = 12.5$) was much higher than for the speeded condition ($C_{base} = 3$). And yet, despite requiring four times the evidence while under accuracy
Complex decisions

constraints, the shift in actual accuracy was marginal. Again, this is likely a function of the low p(stimulus) values employed by the model.

Table 37.

*Observed vs. modelled interaction between condition, instruction, and decisional complexity on accuracy for Experiment 6 data. Bolded model values are outside the 95% CI calculated for the observed values.*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Instruction</th>
<th>p(correct) values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2AFC</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FARTHEST</td>
<td>Observed</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td>.60</td>
</tr>
<tr>
<td>CLOSEST</td>
<td>Observed</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td><strong>.60</strong></td>
</tr>
<tr>
<td>Speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FARTHEST</td>
<td>Observed</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td>.55</td>
</tr>
<tr>
<td>CLOSEST</td>
<td>Observed</td>
<td>.55</td>
</tr>
<tr>
<td></td>
<td>Modelled</td>
<td>.56</td>
</tr>
</tbody>
</table>

*Stimulus set × instruction × speed/accuracy condition: More SCEs.* Finally, Figures 63-64 reveal the extent to which the model was able to replicate the three-way SCE interaction of condition, stimulus set, and instruction on decisional RTs and mean expressed confidence. While it did a reasonable job with decisional RTs, with the SCE in the speeded condition being more subdued than in the accuracy condition, the EPIC model did not vary confidence as a function of condition. This is not surprising, as the
Complex decisions

confidence parameter 'K' was not programmed to shift as a function of condition. Trial-and-error tests where confidence was allowed to increase as a function of condition, while simultaneously altering as a function of complexity and congruity, yielded generally unsatisfactory results. Consequently, the model was not able to replicate this effect.
Figure 57. Observed and modelled main effects of decisional complexity for Experiment 6.
Figure 58. Observed and modelled interaction between decisional complexity and stimulus set for Experiment 6.
Complex decisions

**Decisional Complexity x Instruction**

**Observed**

![Observed graph showing interaction between decisional complexity and instruction](image)

**Model**

![Model graph showing interaction between decisional complexity and instruction](image)

*Figure 59.* Observed and modelled interaction between decisional complexity and instruction for Experiment 6.
Complex decisions

Figure 60. Observed and modelled interaction between decisional complexity and speed/accuracy condition for Experiment 6.
Complex decisions

Stimulus Set x Instruction

Observed

![Graph showing decisional response time and mean confidence rating for different stimulus sets and instructions.]

Model

![Graph showing decisional response time and mean confidence rating for different stimulus sets and instructions, modeled data.]

Figure 61. Observed and modelled interaction between stimulus set and instruction for Experiment 6. These are the SCEs.
Figure 62. Observed and modelled interaction between decisional complexity, instruction, and speed/accuracy condition for Experiment 6.
Complex decisions

**Figure 63.** Observed and modelled interaction between stimulus set, instruction, and speed/accuracy condition for Experiment 6. Accuracy condition.
Complex decisions

Figure 64. Observed and modelled interaction between stimulus set, instruction, and speed/accuracy condition for Experiment 6. Speed condition.
General discussion of the EPIC model's ability to replicate Experiment 6.

This experiment served to highlight both the strengths and the weaknesses of the EPIC model. As would have been the case with any simple race model, a parameter set was found which allowed the model to replicate the basic SCE on decisional RT. Unlike a simple race model, however, the EPIC model was also able to reproduce the nonlinear RT plots typical of experiments wherein decisional complexity is manipulated. More importantly, the model was able to generate a comparable interaction between decisional complexity and speed/accuracy condition on decisional RT, as well as the interaction between complexity and instruction on decisional RT. This is, perhaps, not surprising since the model's forte is the reproduction of nonlinear decisional RT plots.

The model was least able to reproduce observed decisional accuracy. The likely causes of this deficit are the probability values used to regulate random stimulus sampling. Recall how these values were a strict function of the extents used in the Experiments 5 and 6. It is possible that these strict values overestimated the difficulty of these decisions. Perhaps human participants simplified the task by using some sort of visual shortcut: Maybe X’s presented along a horizontal axis were compared more easily than X’s presented along a vertical axis, or perhaps X’s presented ‘kitty corner’ to one another were easier to compare than X’s at opposite ends of a dimension. In any case, this study highlights the fact that the EPIC model does not actually ‘see’ the stimuli being compared, and so is deprived of the rich universe of perceptual cues which can be employed by human decision makers.
Complex decisions

The final chapter of this thesis will revisit these experimental findings, and will discuss the limits and strengths of the EPIC model in light of all of the tests conducted throughout the six presented experiments.
Chapter 9: Conclusions and Implications

Summary of Experimental Findings

Plato is reputed to have once stated “a good decision is based on knowledge, not on numbers”. While those of us with an interest in using computer programs to replicate human decision-making behaviour might disagree with this statement, it nevertheless speaks to the raison d’être of this thesis. Several psychophysicists (e.g., Hicks, 1952; Vickers, 1979; Vickers & Lee, 1998; Usher & McClelland, 2001) have developed mathematical models designed to explain how changes in decisional complexity will affect decision-making behaviour. None of these have published a complete set of empirically derived data which could be used to validate these models. The purpose of this thesis was to generate a set of data which could be used to inform the development of a complete mathematical model of nAFC decision-making behaviour. Specifically, the six experiments presented herein have illustrated how decisional complexity affects decisional RT, mean decisional accuracy, mean expressed confidence ratings, and post-decisional times required to express a confidence rating. The following represents a brief summary of these findings.

Decisional RT. Replicating observations that have been made since the time of Hick (1952), Experiments 1-6 demonstrated how mean decisional RTs increased non-linearly as a function of decisional complexity. While Hick believed this RT pattern was a function of entropy as defined by Shannon (1948), it would appear that the situation is more complicated than Hick believed. Experiment 2 revealed that the degree of non-linearity can shift as a function of decisional difficulty. In Experiment 3, linearity...
Complex decisions

changed as a function of whether confidence judgements were required of decision makers. In Experiment 4, linearity was deliberately affected by altering the computational requirements of the decision such that 4AFC choices required more deliberate effort than did less complex decisions. Experiments 5 and 6 showed that linearity can shift as a function of instruction. Experiment 6 showed that linearity can be affected by SCEs and by the speed/accuracy stress placed upon a decision maker. Taken together, these findings suggest that decisional complexity cannot be considered in isolation and, conversely, that the well studied effects of such variables as decisional difficulty or instructional semantic congruity on decisional RT need to reviewed outside of the traditional 2AFC paradigm.

**Decisional accuracy.** Decisional RT and accuracy tend to be related (Leite & Ratcliff, 2010), so it is not surprising that observed decisional accuracy was as variable as decisional RT across these experiments. Accuracy decreased as a function of decisional complexity in Experiments 1-6, and as a function of decisional difficulty in Experiments 2-4. It was also shown to be affected by instruction in Experiments 4-5 and, of course, by the manipulation of speed and accuracy emphasis in Experiment 6. Unlike decisional RT, however, accuracy was not affected when participants were variably asked to render confidence judgements in Experiment 4. While this phenomenon had been observed under a 2AFC paradigm (Baranski & Petrusic, 1998), this finding has now been extended to the nAFC case. As such, Experiment 4 provides justification for modellers who wish to replicate only decisional RT and accuracy in situations where decision makers are not asked to express confidence. However, it must be stressed that these data also suggest
that confidence is not being processed automatically. If participants were calculating confidence ratings regardless of whether or not they are asked to do so, then no difference in decisional RT would have been observed in Experiment 4. As such, Experiment 4 data suggest that confidence cannot be used to regulate the primary decision-making process.

**Mean expressed confidence ratings.** Throughout the six experiments confidence was found to decrease as a function of decisional complexity. Confidence was also found to decrease nonlinearly as a function of decisional difficulty in Experiments 2-4, and as a function of both the speed/accuracy trade-off and of the SCE in Experiment 6. While main effect and two-way interaction plots illustrated throughout this thesis suggested that confidence ratings might be a function of inversely scaled decisional RTs, where slow responses equalled low confidence, the three-way interaction between condition, instruction, and stimulus set type in Experiment 6 cast this theory into doubt.

**Post-decisional times to express confidence ratings.** The six experiments provided evidence that confidence processing began prior to the expression of an overt decision, since times to express confidence were very fast throughout this study. Nevertheless, they also suggested that in cases where the time required to process confidence related evidence exceeded the time it took a decision maker to express an overt decision, time to express confidence can vary as a function of decisional complexity (Experiments 4-5), instruction (Experiments 4-5), and stimulus set (Experiment 6).
Modelling the Results of Experiments 1-6

Of the two broad classes of models typically used to replicate 2AFC experimental data, race models and random walk models, race models would seem better suited to model nAFC decision-making behaviours. Not only are race models computationally more parsimonious than random walk models (Van Zandt et al. 2000), but it is not immediately clear how a random walk model’s single evidence counter could be used to capture the intricacy of the data generated by Experiments 1-6. Race models, on the other hand, allow for as many evidence counters as there are alternatives to choose from. As such they seem to provide a logical structure within which a complete model of complex decision-making can be built.

Race models have the additional advantage of simultaneously generating both decisional RT and accuracy data: When the criterion levels of evidence required by the model are raised, both decisional accuracy and decisional RT increase. Importantly, this modelled relationship seems to reflect an oft observed correlation between these two DVs (see Experiments 1-6). This feature of race models has recently been shown to have another advantage: Since Leite and Ratcliff (2010) have developed a method by which two different race models can be compared based on how closely each mimics observed decisional accuracy, they have consequently also developed a means to test which model most accurately reproduces observed decisional RTs.

The problem faced by those who employ race models is not whether they are capable of replicating RT and accuracy data. This can usually be achieved via a trial-and-error manipulation of a race model’s decisional criteria parameters. Instead, the
Complex decisions

problem faced by these modellers is whether parameter manipulation is based on theory or whether it is simply a post-hoc selection of those parameters which provide the best fit to the data. In many respects, the EPIC model presented herein is guilty of the latter crime. Base criteria and discomfort exponent parameter selections were based on a series of competitions between parameter sets, and the final values chosen were those that most completely replicated observed accuracy levels. Confidence criterion values for the replications of Experiments 4 and 6 were selected via similar processes. I would argue, however, that in the case of the present study these post-hoc parameter selections were partially justifiable.

To begin with, I noted in Chapter 1 how I have not been able to find any examples of a sensory-based, decision-making experiment wherein decisional complexity was manipulated and both mean expressed confidence and mean time to render confidence judgement data were recorded. As such, it was impossible to know \textit{a priori} what the effects of IV manipulation would be on the four DVs of interest in these experiments. Only after the data from Experiments 1-3 had been obtained and analysed, was it possible to design and test a complete model of complex decision making behaviours. In this case, then, there was no chance to set the model parameters \textit{a priori} based on theories derived from empirical observation, because no one had yet observed the phenomenon to be described.

Once the EPIC model had been designed and tested, an attempt was made to generate experimental hypotheses based on the principles governing the EPIC model. Specifically, the model was designed to replicate the effects of changes in decisional
complexity on decisional RT and decisional accuracy. Within the context of a race model, it is clear that the nonlinear increases in decisional RT are a product of shifting decisional criteria. The problem with a simple race model is that, given the effects of increasing decisional complexity on the probability that any given sample of evidence will support the correct decision, the only way to replicate Hick’s observed RT patterns would be to lower decisional criteria with each increase in decisional complexity (see Chapter 4). Philosophically, the only way to explain this is by suggesting that decision makers do not try as hard when decisions become more complex.

The EPIC model provides an alternative explanation. Within EPIC, a base decisional criterion is selected. This can be thought of as the general amount of effort a decision maker is willing to put into making a set of decisions. EPIC also assigns a discomfort exponent 'a' to this decision maker. This exponent serves as an index of how evenly 'general effort' is to be spread over all levels of decisional complexity. As this exponent ‘a’ grows larger than 1, EPIC begins to reduce the criterion level of evidence required to make 4AFC decisions relative to 2 or 3AFC decisions. As such, even though the base decisional criteria is held constant over all levels of decisional complexity, perhaps indicating a decision maker's desire to sustain a constant level of effort throughout the experiment, when a > 1 all 4AFC choices are deemed to be too psychologically taxing to allow this sustained effort. Consequently, the relative criterion actually employed for these most complex decisions drops.

This assumption was subsequently tested. Recall that in Experiment 4, 4AFC decisions in one condition forced a greater degree of effort than was required of decision
Complex decisions

makers in a second condition. The EPIC model predicted that, in the high effort condition, total session discomfort would be disproportionately allocated to these 4AFC trials. As a result of this manipulation, mean RT plots should have appeared less linear for high effort blocks than low effort blocks. This prediction proved accurate (Chapter 7).

While the post-hoc assignation of best-fitting parameters has questionable theoretical value in some cases, this practice can be usefully employed when the goal of a researcher is to determine which of a series of models is most capable of replicating observed human behaviour. Chapter 5 details how such a determination can be made, as this chapter describes my attempt to determine which one of 32 different types of race model was best able to replicate Experiment 2 data. The base model employed in these tests was developed by Usher and McClelland (2001; see Equation 15). This model was selected because, with a proper adjustment of its parameter settings, it can be made to represent a broad range of race models ranging from a simple evidence accrual model to models which allow for memory decay and inter-accumulator inhibition. The SIMPLEX parameter selection method used to conduct these tests allows for both intra- and inter-model comparisons of parameter goodness-of-fit, so that a ‘best’ model can be selected from each class of model, and then all best fitting models can be compared to one another.

None of the models tested via this method proved capable of replicating the totality of Experiment 2 data. Neither intra-accumulator decay nor inter-accumulator inhibition improved the descriptive capabilities of Usher and McClelland’s LCAM
model. Without leakage or competition, LCAM (i.e., Equation 15) reduces to the race model described by Vickers (e.g., APEP, 1979), and this simple model also proved incapable of simultaneously reproducing both RT and accuracy data from Experiment 2. In short, the EPIC model provided a demonstrably better fit to the data than any of the other models tested.

The reasons EPIC outperformed both LCAM and APEP are perhaps grounded in the theoretical underpinnings of these models. LCAM and APEP assume that a decision maker attempts to maintain either a constant level of accuracy or a constant level of confidence throughout the course of an experiment. But neither accuracy nor confidence remained stable as decisional complexity shifted in Experiments 1-6. A race model can only replicate these observations by allowing criterion levels to drop as complexity increases. And while LCAM and APEP do not allow this, EPIC does: \( C_{\text{base}} \) can be held constant while the actual \( C_{\text{nAFC}} \) employed can decrease as a function of discomfort exponent ‘a’.

The confidence module was designed as a programmatic add-on to the EPIC model. Having decided that confidence cannot be based on scaled decisional RT (Chapter 6), I designed the module to base confidence judgements on the proportion of mode evidence gathered to total evidence gathered throughout the course of the decision-making process. This tactic seemed to provide a reasonable description of observed confidence levels for Experiments 1-3, but failed as more complex interactions involving confidence emerged in Experiments 4 and 6. The module was only able to reproduce these findings when the criterion level of evidence was allowed to shift as either a
function of instruction (Experiment 4) or of decisional complexity and semantic incongruity (Experiment 6). The finding that a race model was unable to reproduce the observed data without allowing for shifts in required levels of confidence related evidence is inconsistent with Vickers' (1979) suggestion that confidence criteria should be held constant within a race model (see Chapter 1).

Interestingly, experimental manipulation of the confidence criterion ‘K’ within the EPIC model demonstrated that mean confidence would decrease as ‘K’ increased. This is counter-intuitive finding defies any simple explanation from the point-of-view of a race model that links decisional criterion C to confidence criterion K (e.g., Vickers, 1979). These models homogenously associate higher accuracy with higher decisional criteria. Perhaps adjustments to ‘K’ reflect a general ambiguity concerning the environment within which a series of decisions are being made, such that globally confusing situations automatically prompt increases in the parameter. Note that modelled manipulations of ‘K’ were global, rather than specific: ‘K’ was adjusted for an entire block of trials or universally whenever a particular instruction appeared paired with a particular stimulus set. This is certainly consistent with the idea that ‘K’ is a function of global uncertainty.

These shifts in ‘K’ were offered as a potential explanation for the few observed significant changes in post-decisional times to express a confidence rating. Generally, post-decisional confidence selection times were flat and fast, suggesting that confidence had been computed prior to the expression of an overt decision. In the few instances where these times varied significantly, the shape of the derived distributions very roughly
Complex decisions

approximated corresponding shifts in ‘K’. That said, it should be noted that the EPIC model could not replicate any of the effects related to post-decisional confidence times, as the confidence module was designed to begin confidence processing at the outset of a decision. The only way to force a replication of these effects would be to set ‘K’ so high that confidence processing took longer than the primary decision. Not only would such a setting generate unrealistically low confidence ratings, but it would also skew decisional RT. This latter effect would occur because the confidence module would double decisional RT at each level of decisional complexity, and would consequently weaken the “Hick effect”. The implication is that confidence processing begins prior to the expression of an overt decision, but not necessarily at the moment of stimulus presentation.

Other considerations regarding the EPIC model. To have any kind of general utility, the EPIC model would have to be able to explain error-free decision-making behaviour. It does have this capacity. If the discomfort exponent ‘a’ in Equation 13 were set so that \( a \leq -1 \), then the EPIC model would allow criteria to increase with increases in decisional complexity, thereby permitting constant levels of decisional accuracy. It is interesting to note that, philosophically, a negative discomfort exponent implies a decision maker who is not only completely comfortable making decisions, but who increases her efforts in the face of rising decisional complexity, perhaps even enjoying the added challenge offered by more complex decisions. This author has yet to meet a group of experimental subjects collectively professing this level of commitment to a
Complex decisions

psychophysical experiment though, admittedly, this may be a function of the sorts of stimuli and motivational tools he has thus far employed.

This tongue-in-cheek reference to the emotional states of undergraduate participants highlights a potential flaw within the EPIC model: Without a means of measuring the emotional states of participants, the definition of the term "discomfort" remains somewhat nebulous. It is unlikely, for example, that the discomfort experienced by participants in these experiments was anything akin to physical pain. After all, Yerkes and Dodson (1908) have famously demonstrated how a rat's time to solve a discrimination task actually improves following the application of moderate electric shocks. Indeed, they showed how, when the discrimination task was simple, rat decisional RT became completely and positively correlated with the size of shock administered. This suggests that some sorts of physical discomfort might actually improve decision-making performance.

As noted in Chapter 4, the adoption of a "discomfort exponent" was inspired by informal conversations with the participants in these experiments. I assumed that discomfort was something akin to boredom or frustration, and that this sensation encouraged speeded, less accurate responses whenever participants were simply trying to "get it over with". So while there was some quasi-qualitative evidence to support the employment of Stevens' Power Law, a more formal method of measuring the emotional state of participating decision makers should be adopted as a part of future research. Perhaps this could be achieved through the monitoring of participant facial expressions as they performed their task, or through the collection of galvanic skin response data.
However it is measured, without a quantitative measurement of discomfort, the EPIC model is in danger of being based on a form of circular logic: "They lowered their criterion because they were uncomfortable, and we know they were uncomfortable because they lowered their criterion".

The EPIC model is also unable to explain SCE's even though it was able to replicate them. The parameter selection routine chose as best fitting parameters those that generated observed decisional RT SCEs. A trial-and-error process, governed by what was learned about the relationship between 'K' and confidence following the modelling of Experiment 4, was subsequently used to find parameters which reproduced the SCEs related to mean decisional confidence ratings. So while the figures printed in Chapter 8 suggest SCEs, they do not offer any explanation as to why or how decisional and confidence criteria come to be adjusted when a decision maker is faced with an instruction that is semantically incongruous with the stimuli being evaluated. Perhaps this is beyond the capacity of any basic race model. A semantic congruity effect likely requires three things to occur. First, an observer must – when presented with an instruction – be able to understand to which features of a given stimuli they must attend. Second, an observer must be able to categorize entire classes of stimuli. And third, it must be possible to 'prime' members of a given category when attention is drawn to one of the category’s defining features. All three of these things involve deeper levels of cognitive processing than EPIC can generate. The EPIC model does not ‘understand’ the deeper implications of any instruction more complicated than “if instruction = FARTHEST then do this, otherwise do that”. It did not classify the stimuli per se, nor
Complex decisions

were any remembered stimulus classes primed for faster response in specific circumstances.

So what, then, is the point of using a race model to replicate such a complicated effect? In this instance, EPIC can be thought of as a descriptive framework for a more complicated model of human cognitive processing. While it had no theoretical justification for setting a lower base criterion level for semantically congruent decisions, it did allow that doing so provided a better fit to the data than otherwise. As such, the utility of the model in this instance is to point out how some cognitive process $x$ must be evoked to effect change $Y$ such that $Y$ must shift in direction $Z$.

Beyond providing a roadmap for more ambitious future modellers, the EPIC model does have some theoretical merit in and of itself. For example, EPIC does not attribute observed error responses to random processes within the decision maker (e.g., Usher & McClelland, 2001; Busemeyer & Townsend, 1993; Van Zandt et al. 2000), nor to decision makers' levels of uncertainty regarding the integrity of their own nervous systems (e.g., Hick, 1952), nor to decision makers' randomly confusing one stimulus element with another (e.g., Crossman, 1955). Instead, the EPIC model is based on the assumption that an individual is perfectly capable of flawlessly telling the difference between, for example, a blue square and a red square, and that errors emerge when the samples selected for evaluation are unrepresentative of the overall distribution of colours within the population. By offering an explanation for how a flawless system can nevertheless generate a predictable distribution of errors, the EPIC model professes a degree of explicative utility which cannot be claimed by "random error" models.
Future research. The present study lays the groundwork for future research and for further development of the EPIC model. To begin with, recall how Crossman (1955) found no significant variation in decisional RT with increases in the absolute difference between stimulus intensities (e.g., \( |x - y| \) increases for all pairs \( x, y \)) so long as the stimulus ratio was held constant (e.g., \( x / y \) is held constant for all pairs \( x, y \)). It is worth noting how Petrusic and Jamieson (1979) did find an absolute difference by ratio interaction for decisional RT. Specifically, these researchers found an effect of decisional difficulty on decisional RT, and found the rate of change in decisional RT was a function of the differences in stimulus intensity. Although the stimuli generated for Experiments 1-4 rely on manipulations of the ratios between presented stimuli, it bears mention that the absolute differences between these stimuli are simultaneously altered whenever these ratios are changed (see Chapter 2). And while a comparable series of decisional complexity experiments with separate manipulations of ratio and absolute difference is conceivable, the paradigm would be quite complicated. Such a paradigm, however, might highlight a new constraint for the nonlinear evolution of RT plots.

This thesis has suggested that psychological discomfort is something which is either evenly distributed over all levels of decisional complexity, or is disproportionately allocated to 4AFC decisions. This seems logical, and it is not unreasonable to conclude that more complicated decisions would require more concentration and effort. Consider, however, an experimental paradigm wherein participants are offered rewards or punishments inversely correlated with level of decisional complexity. If, for example, participants were penalized $1.00 for every incorrect 2AFC decision, but only $.10 for
Complex decisions

every incorrect 4AFC decision, assuming decisional difficulty is held constant, it might be possible to generate greater psychological discomfort for less complicated tasks. If so, the resultant RT plots should provide inverted ‘Hick effect’ plots. Such an observation would provide support for the EPIC model’s assumptions related to the discomfort exponent ‘a’.

The decisional RT plots generated by the EPIC model owe their shape, in part, to Shannon’s Information Theory (1948). Recall how decisional RT increases with each evidence accrual event by an amount equal to the appropriate informational bit rate (see Chapter 4). Theoretically, it should be possible to alter the entropy of any visual stimulus, thus changing the bit rate of informational processing, by shifting the stimuli along some unattended dimension. If, for example, the colour stimuli presented in Experiments 1-4 were presented behind filters of varying levels of visual ‘noise’, it should take longer to process the extra information contained in each presentation. This manipulation of entropy should generate predictable (see Chapter 1, Equations 1 and 2) shifts in decisional RT unrelated to any shift in decisional difficulty.

General Conclusion

Regardless of whether the EPIC model ultimately proves to be viable, it is hoped that the data provided by Experiments 1-6 will prove useful in and of themselves. This dissertation offers multiple examples of the effects of manipulating decisional complexity on the four dependent variables typically referred to in the 2AFC literature: Decisional RT, decisional accuracy, mean expressed confidence ratings, and post-decisional times.
Complex decisions

required to express confidence. I believe that the observations made within this study
have set the boundaries for future nAFC model development, in that any complete model
of complex sensory-based decision-making will now have to explain the totality of the
data presented herein.
References


Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18, 459-482.
Appendix A

Consent Form and Debriefing Form

Consent Form:
Thank you for taking the time to participate in this study of decision-making processes. This study is being conducted by Steven Carroll as a part of a Ph.D. thesis in partial fulfillment of the Ph.D. degree at Carleton University. W. M. Petrusic, Ph.D., of the Department of Psychology at Carleton University, is supervising this project.

Participation in this study will require approximately 90 minutes of your time, and will earn you 2% towards your final grade.

The task requires you to make a series of decisions wherein you will compare certain features of a series of visual stimuli that will be presented to you. Following these decisions, you will be asked to express how confident you are that you have just made a correct decision. Some of these comparisons will be easy to make, most will be considerably more difficult. Please be aware that when you are confronted with a relatively difficult decision, it is quite natural to feel challenged. If at any time you feel any discomfort when performing these comparisons, it is your right to tell the experimenter that you need to take a break or that you wish to withdraw from the experiment completely. Participants who withdraw from the experiment will not be penalized in any way.

It is EXTREMELY important that you do your best to remain vigilant throughout this study. If at some point you no longer feel that you can properly complete the experiment (i.e. you start pushing buttons at random in order to “get it over with”) please inform the experimenter that you would like to withdraw from the study. You will not be penalized. Note that only participants who make a good faith effort to complete the entire experiment will be awarded full credit.

All data collected in this experiment will remain strictly confidential, and will be recorded in a fashion that protects the anonymity of the participants. All data collected will only be made available to researchers associated with this project.

If you have any questions about the study, please feel free to contact Steven Carroll at srcarrol@connect.carleton.ca, or Dr. Petrusic at 520-2600 ext. 2657. If you have any concerns about experimental ethics and/or participant treatment please call Dr. A. Parush, Chair, Carleton University Research Ethics Committee for Psychological Research, 520-2600 ext. 6026, or Dr. J. Mantler, Chair, Department of Psychology, 520-2600 ext. 4173.

Thank you for your participation.

Steven Carroll
Debriefing:

Thank you for participating in this study of decision-making processes.

In this study you were asked to make a series of perceptual judgements. Following these judgements you were asked to subjectively evaluate your performance by ‘rendering confidence’ (in other words, following your decisions you were asked to express how confident you were that you had just made a correct decision).

The goal of this experiment is to measure how primary decision response times, times to render confidence, decisional accuracy, and expressed confidence vary as a function of decisional complexity (i.e., I’m looking at the differences between 2 response decisions, 3 response decisions, and 4 response decisions). It is my hope that this data will allow for the development of a mathematical model of decision making that can accurately describe increasingly complex decision making processes. Such a model, if accurate, would have predictive utility as well as a possible implementation in research into artificial intelligence and/or computer models of cognitive processes.

All data collected in this experiment will remain strictly confidential, and will be recorded in a fashion that protects the anonymity of the participants. All data collected will only be made available to researchers associated with this project.

If you have any questions about the study, please feel free to contact Steven Carroll at srcarrol@connect.carleton.ca, or Dr. Petrusic at 520-2600 ext. 2657. If you have any concerns about experimental ethics and/or participant treatment please call Dr. A. Parush, Chair, Carleton University Research Ethics Committee for Psychological Research, 520-2600 ext. 6026, or Dr. J. Mantler, Chair, Department of Psychology, 520-2600 ext. 4173.

Thank you for your participation.

Steven Carroll
Appendix B

EPIC Model MATLAB Code (Experiment 2 Version)

clearvars;
rand('twister',sum(100*clock));

D1vector = [3.5];
E1vector = [6.5];
D2vector = [5];
E2vector = [4];
D3vector = [5.5];
E3vector = [3];

for(d1Loop = 1:1)
    for(d2Loop = 1:1)
        for(d3Loop = 1:1)
            for(e1Loop = 1:1)
                for(e2Loop = 1:1)
                    for(e3Loop = 1:1)

decisionMatrix = zeros(90000,9);
ex2quantiles;

bicMatrix = zeros(1,9);
bicCounter = 0;
modelNvector = zeros(1,36);
decisionCounter = 0;
freeParamsInModel = 7;
for(nAFC = 2:4)

    trialsPerBlock = 10000;
    for(diffCounter = 1:3)
        if diffCounter == 1
            baseDecCriterion = D1vector(1,d1Loop);
            discomfortExponent = E1vector(1,e1Loop);
        elseif diffCounter == 2
            baseDecCriterion = D2vector(1,d2Loop);
            discomfortExponent = E2vector(1,e2Loop);
        else
            baseDecCriterion = D3vector(1,d3Loop);
            discomfortExponent = E3vector(1,e3Loop);
        end

    trials = 0;

    sampleSize = 25;
    discomfortWeight = .5;

    %------------------
    % PROPORTIONAL DISCOMFORT
    %------------------
    sumDiscomfort = (2^discomfortExponent)+(3^discomfortExponent)+(4^discomfortExponent);
    currentDiscomfort = nAFC^discomfortExponent;
    proportionalDiscomfort = currentDiscomfort/sumDiscomfort;
    decCriterion = baseDecCriterion - (baseDecCriterion * discomfortWeight * proportionalDiscomfort);

    while(trials < trialsPerBlock)
trials = trials + 1;
decisionCounter = decisionCounter + 1;

decision = 0;
colourChosen = 0;
rtCounter = 0;
correctCount = 0;
colourCountVector = zeros(1,4);
colourVector = zeros(1,4);
sampCounter = 0;
instructionRand = rand(1,1);

if(instructionRand <= .5)
    instruction = 1; % instruction is "MOST"
else
    instruction = 2; % instruction is "LEAST"
end

if(instruction == 1)
    if(diffCounter == 1)
        colourPossibilities = [ 1 0 0 0; .5375 .4625 0 0; .385 .33 .285 0; .29 .2475 .2475 .215];
    elseif(diffCounter == 2)
        colourPossibilities = [ 1 0 0 0; .52 .48 0 0; .36 .3325 .3075 0; .27 .25 .25 .23];
    else
        colourPossibilities = [ 1 0 0 0; .51 .49 0 0; .3475 .3325 .32 0; .26 .25 .25 .24];
    end
else
    if(diffCounter == 1)
        colourPossibilities = [ 1 0 0 0; .4625 .5375 0 0; .285 .33 .385 0; .215 .2475 .2475 .29];
    elseif(diffCounter == 2)

Complex decisions

```
colourPossibilities = [ 1 0 0 0; .48 .52 0 0; .3075 .3325 .36 0; .23 .25 .25 .27];
else
    colourPossibilities = [ 1 0 0 0; .49 .51 0 0; .32 .3325 .3475 0; .24 .25 .25 .26];
end
end

thisBitRate = 0;
for(j=1:4)
    if((colourPossibilities(nAFC,j))>0)
        thisBitRate = thisBitRate + (colourPossibilities(nAFC,j) * log2(colourPossibilities(nAFC,j)));
    end
end
thisBitRate = -1 * thisBitRate;

pColour1 = colourPossibilities(nAFC,1);
pColour2 = colourPossibilities(nAFC,2);
pColour3 = colourPossibilities(nAFC,3);
pColour4 = colourPossibilities(nAFC,4);

while(decision<l)
    sampColourCounter = zeros(1,4);
decSampler = 0;
sampCounter = sampCounter + 1;

    while(decSampler < sampleSize)
        sampThisUnit = rand(1,1);
```

242
rtCounter = rtCounter + (thisBitRate/sampleSize);
decSampler = decSampler + 1;

if(instruction == 1)
  if(sampThisUnit <= pColour1)
    sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
    colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
  elseif((sampThisUnit <= (pColour2+pColour1))&&(sampThisUnit > pColour1))
    sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
    colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
  elseif((sampThisUnit <= (pColour3+pColour2+pColour1))&&(sampThisUnit > (pColour1+pColour2)))
    sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
    colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
  else
    sampColourCounter(1,4) = sampColourCounter(1,4) + 1;
    colourCountVector(1,4) = colourCountVector(1,4) + (1/sampleSize);
  end
else
  sampColourPicker = rand(1,1);

if(sampThisUnit <= pColour1)
  if(nAFC == 2)
    sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
    colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
  elseif(nAFC == 3)
    if(sampColourPicker <= .5)
      sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
      colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
    else
      sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
  end
Complex decisions

colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
end
else
    if(sampColourPicker <= .3333)
        sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
        colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
    elseif(sampColourPicker <= .6666)
        sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
        colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
    else
        sampColourCounter(1,4) = sampColourCounter(1,4) + 1;
        colourCountVector(1,4) = colourCountVector(1,4) + (1/sampleSize);
    end
end
elseif((sampThisUnit <= (pColour2+pColour1))&&((sampThisUnit > pColour1))
    if(nAFC == 2)
        sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
        colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
    elseif(nAFC == 3)
        if(sampColourPicker <= .5)
            sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
            colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
        else
            sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
            colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
        end
    else
        if(sampColourPicker <= .3333)
            sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
        end
    end
end
complex decisions

colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
elseif(sampColourPicker <= .6666)
sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
else
sampColourCounter(1,4) = sampColourCounter(1,4) + 1;
colourCountVector(1,4) = colourCountVector(1,4) + (1/sampleSize);
end
end

elseif((sampThisUnit <= (pColour3+pColour2+pColour1)) && (sampThisUnit > (pColour1+pColour2)))
if(nAFC == 3)
if(sampColourPicker <= .5)
sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
else
sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
end
else
if(sampColourPicker <= .3333)
sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
elseif(sampColourPicker <= .6666)
sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
else
sampColourCounter(1,4) = sampColourCounter(1,4) + 1;
colourCountVector(1,4) = colourCountVector(1,4) + (1/sampleSize);
end
end

245
else
    if(sampColourPicker <= .3333)
        sampColourCounter(1,1) = sampColourCounter(1,1) + 1;
        colourCountVector(1,1) = colourCountVector(1,1) + (1/sampleSize);
    elseif(sampColourPicker <= .6666)
        sampColourCounter(1,2) = sampColourCounter(1,2) + 1;
        colourCountVector(1,2) = colourCountVector(1,2) + (1/sampleSize);
    else
        sampColourCounter(1,3) = sampColourCounter(1,3) + 1;
        colourCountVector(1,3) = colourCountVector(1,3) + (1/sampleSize);
    end
end
end

if((sampColourCounter(1,1)==max(sampColourCounter))&&(sampColourCounter(1,1)>sampColourCounter(1,2))&&
(sampColourCounter(1,1)>sampColourCounter(1,3))&&(sampColourCounter(1,1)>sampColourCounter(1,4)))
    colourCountVector(1,1) = colourCountVector(1,1) + 1;
    confTracker(1,sampCounter) = 1;
elseif((sampColourCounter(1,2)==max(sampColourCounter))&&(sampColourCounter(1,2)>sampColourCounter(1,1))&&
(sampColourCounter(1,2)>sampColourCounter(1,3))&&(sampColourCounter(1,2)>sampColourCounter(1,4)))
    colourCountVector(1,2) = colourCountVector(1,2) + 1;
    confTracker(1,sampCounter) = 2;
elseif((sampColourCounter(1,3)==max(sampColourCounter))&&(sampColourCounter(1,3)>sampColourCounter(1,1))&&
(sampColourCounter(1,3)>sampColourCounter(1,2))&&(sampColourCounter(1,3)>sampColourCounter(1,4)))
    colourCountVector(1,3) = colourCountVector(1,3) + 1;
    confTracker(1,sampCounter) = 3;
elseif((sampColourCounter( 1,4)==max(sampColourCounter))&&(sampColourCounter( 1,4)>sampColourCounter( 1,1 ))&&(sampColourCounter(l ,4)>sampColourCounter(l ,2))&&(sampColourCounter(l ,4)>sampColourCounter(l ,3)))
    colourCountVector(l,4) = colourCountVector(l,4) + 1;
    confTracker(l,sampCounter) = 4;
end

rtCounter = rtCounter + thisBitRate;

if(max(colourCountVector) >= decCriterion)
    decision=1;
    if(colourCountVector(1,1)==max(colourCountVector))
        correct = 1;
        colourChosen = 1;
    else
        if(colourCountVector(1,2)==max(colourCountVector))
            colourChosen = 2;
        elseif(colourCountVector(1,3)==max(colourCountVector))
            colourChosen = 3;
        elseif(colourCountVector(1,4)==max(colourCountVector))
            colourChosen = 4;
        end
    correct = 0;
end

%-------------------------------------------------------------
% CONFIDENCE MODULE
%-------------------------------------------------------------
if(length(confTracker) < 5);
    confCriteria = length(confTracker);
else
    confCriteria = 5;
end

CT = 0;
modeCount = 0;
newMode = 0;
oldMode = newMode;

for(confer = 1:confCriteria)
    tempConfVector = confTracker(1,1:confer);
    if(tempConfVector(1,confer) == mode(tempConfVector))
        modeCount = modeCount + 1;
        newMode = mode(tempConfVector);
    end
    currentConf = modeCount/length(tempConfVector);
    oldMode = newMode;
end

tempCT = confCriteria;

cScaleMatrix = [.5 .6 .7 .8 .9 1; .3333 .4667 .5999 .7333 .8666 1; .25 .4 .55 .7 .85 1];
thisCscale = nAFC - 1;

if(currentConf < cScaleMatrix(thisCscale,2))
    scaledConf = 50;
elseif((currentConf >= cScaleMatrix(thisCscale,2))&&(currentConf < cScaleMatrix(thisCscale,3)))
    scaledConf = 60;
elseif((currentConf >= cScaleMatrix(thisCscale,3)) && (currentConf < cScaleMatrix(thisCscale,4)))
    scaledConf = 70;
elseif((currentConf >= cScaleMatrix(thisCscale,4)) && (currentConf < cScaleMatrix(thisCscale,5)))
    scaledConf = 80;
elseif((currentConf >= cScaleMatrix(thisCscale,5)) && (currentConf < cScaleMatrix(thisCscale,6)))
    scaledConf = 90;
else
    scaledConf = 100;
end

if(rtCounter>tempCT)
    rtCounter = rtCounter + CT;
    CT = 1;
else
    rtCounter = rtCounter + rtCounter;
    CT = tempCT - rtCounter; % bare minimum amount of processing required
end