CONFIDENCE FOR CHOICES WITH AN IMPLAUSIBLE ALTERNATIVE

A Thesis

Presented in Partial Fulfillment of the Requirements for
the Degree Master of Arts in the
Graduate School of The Ohio State University

By

Kenneth C. Olson, B.S.

* * * * *

The Ohio State University

2008

Master’s Examination Committee:  
Trisha Van Zandt, Adviser
Hal Arkes
Jay Myung

Approved by

---

Adviser
Graduate Program in Psychology
In a laboratory setting, people viewed simple visual stimuli and chose the categories to which they belonged. A computer tracked people’s choices, confidence ratings, and response times (RTs) when either all alternatives available for a choice were reasonable or one alternative was very unlikely. While the objective probability of an alternative in a set being correct decreases with the introduction of another alternative, I expected that inclusion of an implausible alternative would increase confidence that the alternative chosen was correct. I further suspected that RT would be constant despite changes in the types of alternatives available. Accuracy, confidence, and RT were predominantly unchanged by manipulating alternatives in the choice set. This contradicted two of my hypotheses and the hypotheses of other investigators in this area of research. A final experiment focusing just on confidence obtained more contradictory results, suggesting that people process all combinations of alternatives in a set similarly.
ACKNOWLEDGMENTS

Particular thanks go to my mentor, Trisha Van Zandt, for her advice and patience.

I also extend thanks to my wife, Nicole, for her continuing support.
VITA

August 20, 1981 .............................. Born - Merriam, Kansas

2004 ........................................... B.S. Psychology,
                                  Baker University, Kansas

2005-2006 ................................. University Fellow,
                                  The Ohio State University

2006-present ............................... Graduate Teaching Associate,
                                  The Ohio State University

FIELDS OF STUDY

Major Field: Psychology
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>Vita</td>
<td>iv</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Dud-Alternative Effect</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Confidence Calibration</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Models of Confidence</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1 The Ecological Model</td>
<td>6</td>
</tr>
<tr>
<td>1.3.2 Sequential Sampling Models</td>
<td>8</td>
</tr>
<tr>
<td>1.4 A New Theory of the Dud-Alternative Effect</td>
<td>10</td>
</tr>
<tr>
<td>1.5 Hypotheses</td>
<td>11</td>
</tr>
<tr>
<td>1.6 Organization of This Thesis</td>
<td>13</td>
</tr>
<tr>
<td>2. Experiments</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Methods for Experiments 1 and 2</td>
<td>15</td>
</tr>
<tr>
<td>2.1.1 Subjects</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2 Apparatus</td>
<td>16</td>
</tr>
<tr>
<td>2.1.3 Stimuli</td>
<td>17</td>
</tr>
<tr>
<td>2.1.4 Procedure</td>
<td>17</td>
</tr>
</tbody>
</table>
2.2 Results for Experiments 1 and 2 ........................................ 21
  2.2.1 Accuracy and Confidence .......................................... 22
  2.2.2 Response Time ..................................................... 28
  2.2.3 Discussion ......................................................... 31
2.3 Methods for Experiments 3 and 4 ..................................... 33
  2.3.1 Subjects ........................................................... 34
  2.3.2 Apparatus ........................................................ 34
  2.3.3 Stimuli ............................................................ 34
  2.3.4 Procedure ......................................................... 35
2.4 Results for Experiments 3 and 4 ..................................... 39
  2.4.1 Accuracy and Confidence .......................................... 39
  2.4.2 Detecting the Dud-Alternative Effect ......................... 40
  2.4.3 Response Time ..................................................... 48
  2.4.4 Discussion ......................................................... 53
2.5 Methods for Experiment 5 ............................................. 56
  2.5.1 Subjects ........................................................... 57
  2.5.2 Stimuli ............................................................ 57
  2.5.3 Procedure ......................................................... 57
2.6 Results for Experiment 5 ............................................. 58
  2.6.1 Discussion ......................................................... 60
3. General Discussion ..................................................... 62
  3.1 Motivation ............................................................ 62
  3.2 Experiments 1 and 2 ................................................ 63
  3.3 Experiments 3 and 4 ................................................ 65
    3.3.1 Choice .......................................................... 65
    3.3.2 Singular Duds .................................................. 66
    3.3.3 Implausibility .................................................. 67
    3.3.4 Learning ........................................................ 67
    3.3.5 Deliberation ..................................................... 68
  3.4 Experiment 5 ........................................................ 70
  3.5 Future Directions ................................................... 71

Bibliography .............................................................. 73


## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Raffle Types and Sets</td>
<td>59</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Experiment 1 Stimuli</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Experiment 1 Sequence</td>
<td>20</td>
</tr>
<tr>
<td>2.3 Confidence Histograms</td>
<td>24</td>
</tr>
<tr>
<td>2.4 Overconfidence by Category</td>
<td>25</td>
</tr>
<tr>
<td>2.5 Confidence across Stimuli</td>
<td>26</td>
</tr>
<tr>
<td>2.6 Difference between Confidence and Accuracy</td>
<td>27</td>
</tr>
<tr>
<td>2.7 Relationship between RT and Confidence</td>
<td>29</td>
</tr>
<tr>
<td>2.8 RT across Stimuli</td>
<td>30</td>
</tr>
<tr>
<td>2.9 Experiment 3 Stimuli</td>
<td>35</td>
</tr>
<tr>
<td>2.10 Experiment 3 Sequence</td>
<td>36</td>
</tr>
<tr>
<td>2.11 Experiment 3 Accuracy</td>
<td>41</td>
</tr>
<tr>
<td>2.12 Experiment 4 Accuracy</td>
<td>42</td>
</tr>
<tr>
<td>2.13 Confidence Versus Accuracy</td>
<td>43</td>
</tr>
<tr>
<td>2.14 Calibration</td>
<td>44</td>
</tr>
<tr>
<td>2.15 Experiment 4 Confidence</td>
<td>45</td>
</tr>
</tbody>
</table>
2.16 Experiment 3 Confidence between Conditions .................. 47
2.17 Experiment 3 Confidence across Stimuli Characteristics .......... 48
2.18 Experiment 4 Confidence between Conditions .................. 49
2.19 Experiment 4 Confidence across Stimuli Characteristics .......... 50
2.20 Experiment 3 RT between Conditions ............................ 51
2.21 Experiment 4 RT between Conditions ............................ 52
2.22 Experiment 4 RT and Confidence Relationship .................. 53
2.23 RT across Stimuli Characteristics ............................... 54
CHAPTER 1

INTRODUCTION

People make probability judgments about events every day, such as how likely it is that there will be space for their cars in the company parking garage or how likely it is that the weather will be sunny. As another example, a person may doubt that a pair of shoes is the most durable out of the available alternatives at a store. His assessments of attributes such as water-resistance and quality of construction contribute to his feeling that one pair is probably better than the others. Because probabilities are relevant to nearly all human activity, psychologists question how people think about them. Specifically, researchers have been concerned with how people assess their confidence that their choices are accurate or their expectations are correct (e.g. Dawes, 1980, Einhorn and Hogarth, 1978, Keren, 1991).

Research on confidence generally makes use of simple choice tasks in a laboratory setting. Participants react to a stimulus with a limited set of responses. After each response, they make confidence ratings on an arbitrary scale. Confidence is generally defined as the strength of the belief that a choice is correct. They could be asked to pick which of two doors conceals a prize and indicate their confidence with integers 50 to 100 before the truth is revealed, or they may answer a general knowledge question such as, “Is Chicago or Baltimore farther north?” and then point to verbal labels of
confidence such as “unsure” or “certain”. These tasks may seem artificial, but they mimic many confidence judgments outside a laboratory setting.

While statements like, “I’m not sure these shoes will keep my feet dry,” come naturally on occasion, very important judgments almost always have an explicit confidence estimate attached to them, such as an eyewitness’s confidence in identification of a criminal (Brewer et al., 2002, Wells et al., 1981) and a physician’s certainty that a medical diagnosis is correct (Arkes et al., 1995, Goldberg, 1968).

In medical matters, a job of the general practitioner is to decide what condition afflicts a patient and offer treatment based on symptoms. Some popular television shows have been based around cases of people with unusual health problems who are suffering and must seek medical attention. Though the doctor’s role is dramatized, it is not wrong to emphasize the uncertainty associated with deciding what disease is present and predicting treatment outcomes. Diagnostic and prognostic judgments in clinical psychology and medicine have come under scrutiny many times for their reliability and validity (e.g. Dawson et al., 1993, Goldberg, 1968, Goldman et al., 1983). “Real-world” studies of physicians’ decisions, from heart catheterization to ulcer diagnosis confirm earlier psychology experiments showing that decision-makers’ beliefs about their accuracy do not correspond to real accuracy.

How a decision process gives rise to feelings of confidence is of great significance to criminal justice. Consider as an example a police lineup of suspects. Witnesses to a crime are asked to choose whom they saw commit the act and tell how sure they are of what they remember. If they testify in a later trial, judges and juries will place a great deal of trust in the witnesses’ statements of personal confidence, even when that confidence does not match accuracy of judgment (Wells et al., 1981). Surveys
of the legal system show that police officers, attorneys, and jurors hold a common belief that eyewitness confidence is indicative of accuracy in memory (Deffenbacher and Loftus, 1982, Potter and Brewer, 1999). Police lineups often contain a single primary suspect with the remaining members serving as distractors. How lineups are filled with distractors can greatly alter eyewitness choices and confidence. Wells et al. (1993) used several types of lineups for studying how available options in this choice task affect accuracy and confidence. Subjects tried to pick a criminal out of varying sets of people, combinations of possible suspects chosen for each type of condition. The main suspect could be present or absent from the lineup, and the distractors could be chosen to look more or less like the criminal described. The lineups pertinent to the present research were ones in which the main suspect was matched to the description of the criminal while distractors were intentionally mismatched to the description.

This experiment by Wells et al. (1993) led to a troubling discovery about the effect distractors can have on people. They found that for the experimental condition using distractors from a pool which did not strongly resemble the criminal, subjects made far more false identifications yet showed a bit more overall confidence in their choices. Even a slight increase in confidence misrepresents the probability that the witness identified the real criminal. The theorized reason for these results was the presence of a single person who appeared like the description of the culprit surrounded by distractors who clearly did not.
1.1 Dud-Alternative Effect

A similar effect has been found by Windschitl and Chambers (2004) using general knowledge questions. Unconcerned with accuracy of judgment, they asked subjects to rate the likelihood that a given option was correct. This preselected focal choice was generally the best and most popular option but not necessarily the right answer. Despite reduced normative probability that the selected option was correct, likelihood ratings were greater when the alternative choices were obviously far from the truth. They labeled this the “dud-alternative effect,” a boost to confidence occurring when at least one alternative is a “dud.” One question they used stated, “One, and only one of the following is the most popular choice of color for a new automobile,” and listed the options, “Lime, Blue, Orange, Red.” When the less likely options of lime and orange were not provided, red and blue looked like fairly good choices. With red as the focal option, people placed imaginary bets on the likelihood that it was the correct answer. They were likely to bet less money when lime and orange were not listed and more money when these dud alternatives were listed. If the bets had been real, they stood to lose more money when more alternatives were listed. Indeed, a normative account of subjective probability would have people judge the chance for the focal outcome to be right as lower when weak alternatives are added to the list.

A normative theory of subjective probability states how people should judge probability and bet their money. For example, imagine a hat contains four colors of cards: two red, two blue, one yellow, and one green. If a person draws a card, the chance of drawing a red card is two out of six. If the yellow card is removed and the number of red cards remains the same, when he draws another card, the chance of drawing a red one again is two out of five. If someone bets on the chance of drawing a red
card, he should bet more when only three colors of cards are available, because the
probability of winning is greater. He should feel more confident about drawing a red
card when the yellow card is not in the hat. The dud-alternative effect suggests that
sometimes people feel just the opposite way about their chances of success.

The dud-alternative effect is not a response reversal. Because confidence ratings
apply to just the focal choice, subjects are not changing their minds. Changes in
certainty rather than choice were the core result of the Windschitl and Chambers
(2004) experiments. When people choose the same option twice, the normative ac-
count of subjective probability cannot explain why subjects show significantly higher
certainty for the first choice than they do for the second which has a greater chance
of being right.

The dud-alternative effect is best seen when people have two options from which
to choose and subsequently have third and fourth options which detract from the
probability of the others being correct. For example, if a subject is informed that he
holds 20 tickets for a raffle to win a prize, and another individual entered holds 14
tickets, he might say he is moderately confident of winning. However, if told that
two other people have tickets, 2 and 1 respectively, the subject might say he is very
confident of winning. He feels more secure in the latter situation although his chance
of winning is slightly lower, resulting in overconfidence.

1.2 Confidence Calibration

The dud-alternative effect is a type of overconfidence arising when duds are in-
cluded in a choice set. Overconfidence refers to instances when confidence is higher
than the probability of success. Underconfidence refers to instances when confidence
is lower than the probability of success. Often, the two are compared by first averaging each for a collection of choices. How closely confidence and probability of success match is called calibration.

Subjects' confidence calibration will vary somewhat based on the task. People are usually overconfident on general knowledge questions such as which of two cities has a larger population (Lichtenstein and Fischoff, 1977) but underconfident in judgments about sensory stimuli such as which of two lines is longer (Dawes, 1980, Juslin et al., 1997). The possibility of systematic over- or underconfidence based only on the type of stimulus suggests that poor confidence calibration is not motivational. The mode by which information regarding a choice comes to a person should not provide a motivation to inflate or deflate feelings of confidence. Rather than emotion and motivation, some psychologist have used cognitive processing to explain changes in confidence. Below, I present major theories for how mental computations result in observed confidence.

1.3 Models of Confidence

Theories on confidence which may have the precision to make predictions about confidence under dud-alternative conditions include the models of Gigerenzer et al. (1991), Merkle and Van Zandt (2006), and Tversky and Koehler (1981). I describe how these and other theories shaped my hypotheses about the dud-alternative effect.

1.3.1 The Ecological Model

The earliest description of the relationship between accuracy and confidence may be the verbal model of Peirce and Jastrow (1884) in which they proposed that the two variables could be equated in the long run, meaning that over a large set of
events, average confidence and average accuracy for predictions should be nearly equal. One adaptation of this idea appears in the probabilistic mental model or ecological model by Gigerenzer et al. (1991). The basis of the model is environmental cue validity: people extract useful “ecological” cues from stimuli to make decisions. Confidence in the form of a probability judgment emerges from the predictive power of the cues. Both choice and confidence are determined by the probability that a decision is correct, conditional upon the information at hand. Cue validity is defined by the empirical relative frequency of a correct prediction from a cue. Gigerenzer et al. took the parsimonious view that a person uses the cue with the highest validity to make a decision and expresses confidence equal to that validity.

Instances of overconfidence arise in the ecological model when commonly effective cues fail (Gigerenzer et al., 1991). The cues’ predictive validities are high in most – but not all – circumstances. Trick questions are those which quickly bring to mind misleading information. The terms of the question may feel familiar and elicit high confidence but ultimately not help a person to answer correctly. Griffin and Tversky (1992) found evidence against the ecological model when they observed that questions not specially selected to have trick cues could still create overconfidence in subjects. People were consistently overconfident on ordinary, trivia-type questions. Although the basic assumptions of the ecological model are not disputed here, thus far it does little to explain ubiquitous overconfidence, and even less to explain how higher confidence comes about in the dud-alternative effect.

The ecological cues on which people base decisions are unlikely to be different when comparing alternatives that include a dud; hence, the same cue validities inform confidence when a dud alternative is present as when a dud alternative is absent. If
there were a way people used different cues to make the choice, it might explain the dud-alternative effect. I incorporate a similar idea about people using different reasons for choosing an alternative into the next series of models to develop two major hypotheses.

1.3.2 Sequential Sampling Models

Another class of models can account for confidence and accuracy along with response latency. Sequential sampling models have been the most successful at accommodating reaction times (RTs) and accuracies in decisions between two alternative responses. They are well established and popular process accounts of decision-making in many tasks, such as vigilance, perceptual judgment, and categorization (Busemeyer and Townsend, 1993, Diederich and Busemeyer, 2006, Nosofsky, 1997, Ratcliff, 1978, Smith, 2000, Usher and McClelland, 2001) and have been vital to evaluation of theory regarding basic cognition. A useful feature of sequential sampling models is that they can incorporate the static description of a decision through signal detection theory with a dynamic account of reaction time and choice probabilities (Laming, 1968, Pike, 1973). They are founded on the idea that the representation of stimuli is noisy and assume the stimuli are mapped onto a theoretical activation dimension, which has commonly been called stimulation or information. Subjects repeatedly sample from the noisy activation to obtain evidence for each alternative. An instantaneous evidence level exists for the choice, which fluctuates over time.

Sequential sampling models generally are able to predict the relationships between RTs and probabilities of correct and error responses along with the shapes of the RT distributions. Individual models may be evaluated by how well their unique
predictions match experimental data. Several models also explain and predict confidence (e.g. Baranski and Petrusic, 1998, Merkle and Van Zandt, 2006, Van Zandt and Maldonado-Molina, 2004). These models postulate that the evidence levels for competing responses at or shortly after the time a choice is made enter into some mental assessment of confidence.

Under the sequential sampling paradigm, Merkle and Van Zandt (2006) proposed a modification of the Vickers (1979) balance-of-evidence hypothesis to successfully model confidence. They used a ratio of evidence for the chosen alternative to the sum of evidence for all alternatives.

\[
Confidence = \frac{X_{\text{win}}}{X_{\text{win}} + X_{\text{lose}}} \quad (1.1)
\]

Evidence \((X)\) originates from an underlying parameter for rate of evidence accumulation. Individual people have individual parameters, and evidence changes from moment to moment by a partially random process. Because the final level of evidence for each alternative in each choice is inconsistent, the ratio is not always reflective of the probability that the choices are correct. No one has applied this model of confidence to the dud-alternative effect, but the approach shows promise for explaining occasions on which people exhibit non-normative confidence in their choices.

Though the work presented here does not model timing, accuracy, and confidence in choices, I have theorized about what results would be found for these three measures under the dud-alternative effect. The foregoing process account for the effect is heavily influenced by these dynamic models of choice.
1.4 A New Theory of the Dud-Alternative Effect

Merkle and Van Zandt (2006) used a ratio for confidence which alone cannot account for the dud-alternative effect. No matter how small one makes the evidence for alternatives not chosen, it will increase the denominator in the equation and decrease the confidence for the alternative chosen. This is true under the assumption that evidence for an alternative is always favorable. I propose that evidence can count against an alternative just as it can count toward an alternative. Unfavorable evidence for an option not chosen would be negative evidence and would decrease the value of the denominator and increase confidence. Only dud alternatives are more likely to garner negative evidence than positive evidence.

Before elaborating on how this process works, I must state that Windschitl and Chambers (2004) dismissed the use of a similar ratio from support theory (Tversky and Koehler, 1981) to explain a subjective judgment of probability, \( P \), of a focal option, \( A \).

\[
P(A, \overline{A}) = \frac{s(A)}{s(A) + s(\overline{A})}
\]  

(1.2)

According to Tversky and Koehler (1981), how a person feels about the likelihood of outcome \( A \) develops from the support for that outcome, \( s(A) \), divided by the support for all outcomes, including support for the non-focal options, \( s(\overline{A}) \), also called the “residual.” Originally, support theory did not identify how support for the non-focal options in a choice is created or combined as it enters the equation. Unlike Windschitl and Chambers (2004), I maintain that such a simple ratio of evidence
for the best alternative versus total evidence for all alternatives may explain the dud-alternative effect if evidence can disaffirm as well as affirm options.

I imagine the confidence for a multi-alternative choice comes about by each alternative gathering evidence which can be both positive and negative. The assumption that people can think of both supporting and contradicting reasons for an alternative is the foundation of work by Koriat et al. (1980). They demonstrated significant changes in confidence when task instructions had people focus on thoughts that disaffirm one option in a choice. Subjects were better calibrated when considering evidence to contradict the best alternative. Is it possible for subjects to have confidence further from objective probability when considering evidence to contradict the worst alternative? People naturally have an easier time thinking of reasons not to choose dud alternatives than reasons to choose them. If reasons are the basis for confidence, duds are clearly worse than the other alternatives because of mostly disaffirming evidence.

1.5 Hypotheses

For the experiments performed, I expected the typical relationship between confidence and accuracy to hold for most simple choice tasks. As accuracy increases, confidence often does as well. This may occur for individual choices, but I hypothesized the opposite would be apparent between conditions with and without a dud alternative. Many experiments show that accuracy and confidence do not have a straightforward relationship (e.g. Busey et al., 2000, Griffin and Tversky, 1992, Henmon, 1911), which could be especially true for the dud-alternative effect. I hypothesized that while confidence would increase for a choice including a dud alternative,
accuracy would decrease, something implied by Windschitl and Chambers (2004) with the raw expected probabilities of outcomes.

The relationship between confidence and time to make a decision is task-dependent, so while there are exceptions (e.g. Van Zandt, 2000, Vickers, 1979), confidence decreases as RT increases (Henmon, 1911, Petrusic and Baranski, 2003). Often, this relationship appears to be uniformly linear and monotonic, so much so that Audley (1960) considered a measure of RT to be a good estimator of confidence. Confidence and RT may be so closely connected as to have a common underlying cause. In sequential sampling models, some have suggested that both variables are tied to the intensity of the activation toward the alternative which is chosen (e.g. Link, 1992). In the dud-alternative effect, I thought the strengths of activation and evidence for the focal alternative would not change, so no change in average RT would occur (The shapes of the distributions could be different.). From one choice to another, RT and confidence would maintain their typical inverse relationship, but for a whole class of choices defined by whether or not the dud was present, I expected there would be no strong pattern. I hypothesized that while confidence would increase for a choice including a dud alternative, RT would stay the same.

“... We assumed that the processes driving the dud-alternative effect ... are general processes that apply to almost any form of likelihood judgment involving multiple alternative hypotheses” (Windschitl and Chambers, 2004, p.205). To see how widespread the effect is, my experiments presented subjects with dud alternatives in sensory identification tasks somewhat different from the original experiments. I provided people very strong alternatives, strong alternatives, weak alternatives, and dud
alternatives. I even had alternatives so implausible that subjects could readily disregard them. I intended these experiments to provide new information on how people’s evaluations of varying evidence for multiple alternatives influence their accuracy and speed along with confirming the way they influence confidence.

1.6 Organization of This Thesis

The remainder of this thesis is organized as follows: Chapter 2 will describe four experiments exploring the pervasiveness of the dud-alternative effect in choice tasks, including methods, results, and discussion as well as a fifth experiment to replicate one from Windschitl and Chambers (2004), and Chapter 3 will explain future directions for this research in light of results from Chapter 2.
CHAPTER 2

EXPERIMENTS

I designed these experiments to elicit simple choices and confidence estimates from decision makers in a controlled setting. The methods are established procedures for perceptually based choices. Subjects performed discrimination or identification tasks with an added confidence component. I used novel stimuli so that participants could not draw on previous experience or memory associations to perform the assigned tasks. In all cases, the subjects chose one of three or four possible responses and then rated their feelings of confidence about that choice.

The task for the first two experiments required subjects to categorize a random pattern of asterisks based on the number of characters in the pattern and pre-established categories (Merkle and Van Zandt, 2006). The number of asterisks in each pattern was randomly drawn from one of four normal distributions which overlapped on the single attribute of numerosity. The variance in the distributions ensured that no response scheme would result in all correct choices. As in a signal detection task, subjects were expected to set criteria for different responses to discriminate among types of stimuli.

I chose this task because I expected less variability in performance between and within subjects than in other tasks that have been used to study confidence. With
a general knowledge task, all questions must undergo screening to ensure that their answers are about equally apparent to everyone, but with this perceptual task, the question is always the same, and subjects attained similar, near-peak performance after little training. More importantly, this task lends itself well to any future modeling of the response selection process. Conditions are quantifiable, and the distributions underlying the categories can determine the activation for a choice in a sequential sampling paradigm (Ratcliff et al., 1999).

The task for the third and fourth experiments required subjects to categorize stimuli based on more than one attribute. Again, the exact stimuli were created based on underlying normal distributions for greater control of what subjects perceived and the way it might be modeled in the future.

The fifth experiment took a very different approach from the other four by adopting methods already used for studying the dud-alternative effect. It was designed as a simple follow-up to the other experiments that might answer lingering questions.

2.1 Methods for Experiments 1 and 2

These experiments had subjects always make choices from among four category alternatives. Not changing the number of alternatives made them different from the experiments of Windschitl and Chambers (2004). When a stimulus strongly affirmed one of two of the alternatives, it strongly contradicted the other of the two. If one category seemed very similar to a stimulus, another category would seem very dissimilar. The similarity of an image to category exemplars (Nosofsky, 1997) possibly determined the support for choosing a category in the same way that similarity of a
criminal’s appearance to that of people in a lineup probably determined the support for naming one of those people a suspect in the work of Wells et al. (1993).

I hypothesized that a dud-alternative effect for these experiments would appear as a large increase in confidence for alternatives on the ends of the numerosity spectrum, categories 1 and 4, accompanied by very little change in accuracy and RT. Although a decrease in accuracy would be more fitting, the way categories related to each other may have made recognizing stimuli from categories 1 and 4 easier than recognizing stimuli from categories 2 and 3. Subjects would be most overconfident for the outer categories.

2.1.1 Subjects

Fifteen subjects from the OSU undergraduate subject pool served in each experiment in exchange for course credit. All subjects reported normal or corrected-to-normal vision as well as basic motor skills. Naive to the purpose of the experiments but informed of their role, they were tested individually after receiving both oral and written instructions. Experiment 1 had an approximate duration of 45 minutes per session, and Experiment 2 lasted about 55 minutes per session.

2.1.2 Apparatus

The display of stimuli occurred on a Viewsonic 6 video monitor with timing of presentation maintained by an IBM PC/SX desktop computer. The computer served also to record data from subjects, who made responses on a standard NMB keyboard with all nonessential keys removed. Because the pads to detect finger strokes are inset into the plastic construction of the keyboard, this avoided many accidental key presses.
2.1.3 Stimuli

Stimuli were arrays presented in the middle of the computer screen. Each array consisted of 200 characters in a $10 \times 20$ grid where every other column contained only spaces. This produced a roughly square display of asterisks and spaces. The number of asterisks for each trial was between 1 and 100 and was drawn from one of the four normal distributions shown in Figure 2.1. The locations of the asterisks were selected at random from the positions on the grid not otherwise occupied by spaces. The four normal distributions all had standard deviations of 7 with means of 35, 45, 55, and 65. For Experiment 2, a slight change was made to the lowest and highest distributions by forcing a few extra numbers toward extreme values to ensure sufficient data points for analysis of the tails.

2.1.4 Procedure

The experiments had three phases. The first training phase was a block of trials in which subjects became familiar with the choice task, and the second training phase added the confidence judgments. Data were collected during the third testing phase, in which subjects completed 10 blocks of 60 trials each. In the training phase, subjects were told that each pattern was produced by one of four generators, and that their job was to quickly and accurately decide which generator created the pattern of asterisks on the screen. They had four generators from which to choose, numbered one through four in a vertical arrangement in the number pad section of the keyboard. They were shown a typical example from each category to familiarize them with the appearance prior to training. Each training trial was to identify which of the four options listed was the category to which the stimulus belonged.
Figure 2.1: Experiment 1 stimuli. Four underlying distributions for categories of stimuli are shown.
Subjects performed 30 primary training trials in Experiment 1 and 60 primary training trials in Experiment 2. After a stimulus appeared, it remained visible until a key was pressed for a choice. During all training trials, participants were provided continuous feedback so that after a response a single word appeared where a stimulus had previously been. “Correct” was displayed for a correct choice, and “Error!” was displayed for an incorrect choice. Feedback was visible for 500 milliseconds (ms), and a new stimulus appeared 500 ms after feedback terminated.

In the secondary training phase, participants received instructions regarding how to make confidence judgments. For Experiment 1, subjects provided confidence ratings using the numbers 1 through 9 and 0 in a standard horizontal arrangement, left to right, at the top of the keyboard. The zero key stood for a rating of 10, which meant subjects had complete confidence that their choice was correct. Pressing the 1 key meant that the subjects were nearly certain they made the wrong choice, pressing the 2 key meant they probably made a mistake, and pressing the 3 key meant they were just guessing. Increasing from 3 to 10 meant increasing confidence from nearly no confidence to nearly complete confidence, and subjects were advised that the 6 key meant medium or “normal” confidence for this task. Because subjects in Experiment 1 stated that concentrating on accuracy early in the trials had made them forget the meaning of points along the confidence scale, for Experiment 2, this description was repeated after the first training block.

After confidence ratings feedback was given. The sequence of events for each trial of this practice phase is shown in Figure 2.2. Experiment 1 had 30 secondary trials with confidence ratings, and Experiment 2 had 60 secondary trials with confidence ratings.
Figure 2.2: Order of events for Experiment 1. Subjects viewed the stimulus until they made a key press for their choice, then the computer waited until they made a confidence rating before feedback was displayed for 500 ms. Another 500 ms pause came between trials.
Following the two training phases, subjects were given an opportunity to rest, then they completed 10 blocks of 60 trials each. For Experiment 1, the computer gave trial-by-trial feedback for each of 600 recorded choices. For Experiment 2, feedback came at the end of each 60-trial block, stating correct choices as a fraction of total choices for that block.

2.2 Results for Experiments 1 and 2

Some subjects found this task difficult while others were unsympathetic to the research. Many performed quite badly because they failed to follow instructions. I excluded data for any subject who had poor overall accuracy as measured by the mean number of correct choices, the $d-prime$ between categories 1 and 2, or the $d-prime$ between categories 3 and 4. Other data were not used because the subject gave the same confidence rating on more than half of the trials or because the subject’s accuracy and confidence were uncorrelated ($r < 0.05$). Many of the subjects removed from analysis did not meet two or more of the standards set. I excluded data for 8 subjects from Experiment 1 and data for 6 subjects from Experiment 2.

Trials with RTs for choice or confidence-in-choice greater than 10,000 ms or less than 200 ms were excluded. Furthermore, response times more than 3 standard deviations from a subject’s mean were trimmed. Confidence ratings were transformed from whole numbers into values on a scale from 0 to 1 by dividing by the largest possible rating. For example, a confidence rating of 2 became 0.2, and a confidence rating of 8 became 0.8.
2.2.1 Accuracy and Confidence

Accuracy between subjects in Experiment 1 varied from an average of 0.46 to 0.55. The averages across categories of stimuli also showed little variation. A within-subjects ANOVA revealed no significant differences in accuracy for the four categories ($F(3,18) = 0.21; p > 0.05$).

Mean overall accuracy in Experiment 1 was 0.50, and mean confidence was 0.71. The distribution of confidence ratings is shown in Figure 2.3, a histogram of confidence from all analyzed trials. Ratings below the average accuracy are very uncommon in Experiment 1. For confidence, there are differences in means among the categories to which the subjects responded ($F(3,18) = 21.17; p < 0.01$). A Tukey test showed that ratings on trials using stimuli from categories 2 and 3 were not significantly different, though all other differences were, including the mean difference between the outer categories and the inner categories. As expected, subjects were more confident in responses to categories averaging very high or very low numbers of asterisks. This would not be surprising if subjects made more correct responses for these stimuli, but they did not. One way to compare the two variables of interest here is by subtracting accuracy from confidence on each trial. As can be seen in Figure 2.4, subjects rated their confidence more highly for categories 1 and 4 despite limited change in the proportion of correct choices. Although overconfidence is common in this task, it is strongest for the outer categories, but differences are not significant ($F(3,18) = 1.98; p > 0.05$), which is partially explained by high variance from subject to subject for this measure. Subjects’ did generally have in common how often they chose categories 2 and 3. Subjects chose the central categories about twice as frequently as the outer
categories, meaning that many times they selected wrong responses when stimuli from the outer categories were presented, yet showed high confidence.

Another way to examine the data is to analyze the variables on the defining feature of the stimuli, number of asterisks. An example of this analysis for confidence for one subject is provided in Figure 2.5. The majority of subjects confounded this fashion of analysis by introducing a ceiling effect. Generally, their ratings were so high as to leave no room for their highest confidence to be indicated on the scale provided. On the far edges of the distributions, where one might expect overconfidence to be most apparent, it was attenuated by the limits of the measuring device. The range of the confidence ratings forces kinks into what might otherwise be a simple curve.

This problem was overcome in the Experiment 2 when no obvious ceiling effect was present. Data from more subjects were used in analyses, and the confidence matched better with accuracy. Mean accuracy was 0.54, while mean confidence was 0.68. A discrete distribution of confidence is in Figure 2.3, showing reduced relative use of the highest ratings. Still there was no significant effect of category on accuracy \((F(3,18) = 0.70; p > 0.05)\), but there was an effect on confidence \((F(3,18) = 26.72; p < 0.01)\). Number of responses to outer categories increased so that choices were more balanced.

Even without a ceiling on confidence, the variation in ratings cannot match that of accuracy, making a comparison of the two variables difficult without first normalizing their values. By aligning their spread and central tendency with a z-score conversion, general overconfidence is removed from consideration to reveal other trends because overconfidence may only be mistranslation of the subjects’ feelings onto the scale provided. Figure 2.6 is a plot of mean of confidence minus accuracy for each stimulus.
Figure 2.3: Histogram of confidence ratings. Part A shows the distribution of ratings for all subjects in Experiment 1. Part B shows the distribution for all subjects in Experiment 2.
Figure 2.4: Overconfidence by category. Part A shows confidence and accuracy grouped by categories of stimuli. Means with standard error bars are shown for subjects’ performance in Experiment 1. In Part B, confidence minus accuracy was computed per trial before grouping by categories of stimuli.
Figure 2.5: Confidence across stimuli. Circles indicate mean confidence per number of asterisks for a single subject in Experiment 1. Against the upper limit of the scale, the pattern flattens at ceiling.
Figure 2.6: Difference between confidence and accuracy across stimuli. Points are Experiment 2 subjects’ normalized confidence minus normalized accuracy averaged for each number of asterisks displayed then averaged across subjects.

Subjects saw. A U-shape emerges among the points as a result of the pattern of confidence rather than accuracy. As the number of asterisks approached extreme values, correct choices and confidence both increased, but confidence increased far more in proportion to its overall variability. After artificially improving calibration through the transformation of accuracy and confidence, the graph indicates that confidence is notably greater than accuracy for very high and very low numbers of asterisks.
2.2.2 Response Time

The third measure obtained from subjects was RT. Subjects could take as long as they needed for the decision of how to categorize stimuli, so average RT was 1385 ms in Experiment 1, and 1631 ms in Experiment 2. RT distributions for subjects had a common skew toward higher numbers. A log transformation made the RTs roughly normal to fulfill assumptions for some statistical tests. The categories did not have a significant effect on RT in Experiment 1 \((F(3,18) = 2.32; \ p > 0.05)\) but did in Experiment 2 \((F(3,18) = 7.55; \ p < 0.01)\). The trend for both was in the opposite direction of that for confidence and accuracy. RTs were lower for categories 1 and 4 as subjects made faster decisions.

As accuracy and confidence increased, RT decreased. This relationship did not hold for individual trials which was clear from weak correlations. Even for single subjects, correlations between any two of the three dependent variables did not exceed 0.40. The strongest relationship can be seen in Figure 2.7. For Experiment 2, higher confidence was paired with greater speed of responding and decreased variability in speed of responding.

Again there is a trend across individual stimuli. RTs were larger for middling numbers of asterisks and decreased as the stimulus had either more or fewer asterisks. In Figure 2.8, mean RTs are plotted for each number of asterisks, and least-squares trend lines are added. The dashed line represents the best fitting curve for points from all trials.
Figure 2.7: Relationship between RT and confidence. For Experiment 2, quantiles of RT are plotted against each level of confidence. Open circles represent the 0.1 quantiles; upward pointing triangles represent the 0.3 quantiles; solid circles represent the medians or 0.5 quantiles; inverted triangles represent the 0.7 quantiles; open diamonds represent the 0.9 quantiles.
Figure 2.8: RT across stimuli. The open circles represent mean RTs for each number of asterisks for just one subject. The solid line is the best-fitting line through those points of the form $rt = a \times (\text{number of asterisks} - b)^2 + c$. The $a$, $b$, and $c$ parameters were set by minimizing the sum of squared error. The dashed line is the best-fitting line for data from all subjects in Experiment 2.
2.2.3 Discussion

Results for Experiments 1 and 2 were partly as expected. I had hypothesized that stimuli from one end of the numerosity spectrum would make the category at the other end of the spectrum akin to a dud alternative so that confidence for these stimuli would be higher than for others while accuracy and RT would be mostly unchanged. Confidence was higher, and accuracy was not, but RT was lower.

Subjects did not perform better at choosing the two outer categories. Although most of the asterisk patterns of high or low numerosity might avoid confusion of category membership, subjects showed little improvement at identification. Instead, confidence had a large relative increase, and RT had a large decrease for extreme stimuli. Even without considering the specifics of stimuli, confidence and RT had a notable negative relationship. It would be simple to think subjects made some classifications so easily as to choose quickly and feel more certain that their responses were correct, but because their responses were not significantly more likely to be correct, the speed and determination with which they made their choices must have been driven by more than just how hard or easy each trial was. Confidence did not equate to accuracy in this task, especially for the trials when evidence strongly indicated one option and nearly eliminated another.

It can be assumed that for subjects following instructions, perceived numerosity of asterisks provided the primary rationale for choices, hence I attempted examination of the three dependent measures by this feature of the stimuli. Stimuli could have numbers of asterisks similar or dissimilar to the prototypical or mode number of asterisks for a category. A very large number of asterisks in a stimulus made it noticeably different from the category with very small numbers of asterisks. Conversely,
a very small number of asterisks in a stimulus made it noticeably different from the
category with very large numbers of asterisks. These two extremes of stimuli brought
the clearest changes in subjects’ behavior.

Is this evidence to support the dud-alternative effect? Ultimately, I did not find a
dud-alternative effect. The basic way to obtain that effect is simply to provide a dud-
alternative some times but not others. Here, when the stimulus was at one extreme,
judging it to be at the other extreme was unlikely. Perhaps subjects saw much evi-
dence in favor of category 1 as reducing the possibility of category 4, and vice versa.
Very many asterisks could make category 1 a dud-alternative, and very few asterisks
could make category 4 a dud-alternative. For numbers in between, any alternative
could seem reasonable so no dud-alternative would be apparent. Providing an implau-
sible alternative in exchange for a decent alternative may have increased confidence
more than accuracy and decreased RT, but the dud-alternative effect was previously
evoked by removal of dud alternatives without replacement. In Experiments 1 and 2,
subjects always had four alternatives from which to choose, so one cannot compare
confidence in this task to cases when subjects had only three alternatives from which
to choose.

Another limitation of these experiments was the way I solicited confidence ratings.
Some subjects said that the method by which they provided confidence ratings was
still confusing after changes in Experiment 2. Mostly, confidence on a zero to one
scale was greater than probability of correct responding, but this could be due to
misuse of the original rating scale. If subjects saw the middle of their 0-10 scale as
indicating a feeling of being more right than wrong, then it would not correspond
to probability. With four options from which to choose, a slight feeling that one is
making the correct decision should be just above random chance, 0.25. To use the measurement span properly, such a feeling would cause a subject to rate confidence near the lowest fourth of the scale. The rating 3 would mean the subjects had almost no confidence in choosing the alternatives they did, and ratings around 5 would be closer to their average performance. Experiments 3 and 4 attempted to resolve this issue with a shorter scale.

2.3 Methods for Experiments 3 and 4

I made several major changes to earlier experiments with the intention of obtaining more conclusive results in Experiment 3. First, categories needed to overlap nearly equally with all other categories, so that when subjects viewed a stimulus, one option for identification was best, but all options were reasonable (except under particular circumstances). This was accomplished by converting the one-dimensional stimuli from earlier experiments into two-dimensional stimuli, meaning that subjects judged more than just numerosity in order to make a choice. Instead of a single character filling the array, two characters in varying proportions appeared on screen. Then the categorization decision was made by how many characters were present and what fraction were of a type. Rather than just asterisks, subjects viewed patterns of letters. The dimensions of numerosity and proportion could each be divided into two parts so that four categories were established, each characterized by either the high or low end of dimensions. Figure 2.9 illustrates how the stimuli space is split into four quadrants.

Second, to address concerns about direct testing for the phenomenon of primary importance, any dud-alternative had to be obvious. To create the case for a dud-alternative, a dichotomous variable was presented with the pattern, which would
indicate that one of the four alternatives was extremely unlikely. Third, I made the
dud-alternative clearly present or absent from subjects’ choice sets. Half of trials
listed it as a potential choice, and half did not. Fourth, the confidence scale was
abbreviated for easier interpretation. Fifth, I provided a backstory for the task to
increase subjects’ interest and motivation.

2.3.1 Subjects

59 subjects from the OSU undergraduate subject pool participated in Experiment
3 in exchange for course credit, and 67 subjects participated in Experiment 4. All
subjects reported normal or corrected-to-normal vision as well as basic motor skills.
They were individually tested in 30-minute sessions after receiving both oral and
written instructions.

2.3.2 Apparatus

Desktop computers with monitors displayed stimuli and recorded data for subjects.
Responses were on a modified keyboard as in Experiments 1 and 2 but with different
lettering on keys.

2.3.3 Stimuli

Subjects saw patterns of “o”s and “.”s or “o”s and “v”s, which were generated
by distributions differing according to category. Categories were constructed to have
either large or small numbers of characters and have either large or small proportions
of “o” characters out of their totals. The total numbers of characters came from two
normal distributions with means of 19 and 31 and standard deviations of 4. The
numbers of “o”s were made by proportions taken from normal distributions with
Figure 2.9: Stimuli and categories. Part A illustrates the basic divisions between categories. Part B shows the whole distribution from which stimuli could be drawn.

means of .25 and .75 and standard deviations of 0.16. On average, category 1 had 14.25 “o”s and 4.75 of the other; category 2 had 4.75 “o”s and 14.25 of the other; category 3 had 23.25 “o”s and 7.75 of the other; and category 4 had 7.75 “o”s and 23.25 of the other.

The patterns of characters were presented in arrays in the middle of the computer screen. Arrays again consisted of 200 characters in a 10 × 20 grid where every other column contained only spaces. The locations of the characters were selected at random from the positions on the grid not otherwise occupied by spaces.

2.3.4 Procedure

Subjects were told that each pattern was a simulation of a tissue sample a doctor might see with only certain cells showing after a dye was applied. The sample therefore
Figure 2.10: A single trial. The sequence of events in a trial for Experiments 3 and 4 is shown.
indicated which type of cancer was causing the cell production. The “o” characters were referred to as “large cells” in Experiment 3 or “round cells” in Experiment 4 while the “.” characters were referred to as “small cells” in Experiment 3 and the “v” characters were referred to as “pointy cells” in Experiment 4. It was the task of the subjects to decide quickly and accurately which cancer generated which samples, hence identifying the category to which stimuli belonged. Categories 1, 2, 3, and 4 corresponded to ovarian, lung, stomach, and testicular cancers, with the keys O, L, S, and T in a vertical arrangement in the number pad section of the keyboard. They were shown a typical example of each cancer to familiarize them with the appearances prior to training and warned that men and women do not always develop the same cancers.

The experiments had four phases. The first phase was training, a block of trials in which subjects learned the basic choice task. A second training phase added feedback about speed, and a third training phase required confidence judgments. Primary data were collected during the fourth phase, testing in which subjects completed three more blocks. In all, subjects saw 6 blocks of 48 trials each and could take breaks between blocks.

On any trial, a stimulus pattern appeared with an indication of gender and a set of options which randomly did or did not contain the dud-alternative. Individual trials first showed on the right of the screen for 500 ms the set of alternatives from which to choose. Next, subjects saw “M” for male or “F” for female appear just to the left of the choice set. After 500 ms more, the stimulus appeared centrally and disappeared along with everything else as soon as a choice key was pressed.
During all training trials, participants received continuous feedback. “Correct” was displayed for a correct choice, and “Error!” was displayed for an incorrect choice. For the first phase, “Impossible” was displayed for a choice that did not fit the gender, and “Not Listed” was displayed for a choice that was not one of the available alternatives. Feedback was visible for 500 ms, and a new stimulus appeared 500 ms after feedback terminated. In the second phase, prior to information about the accuracy of a choice, “Too Fast” displayed for 1000 ms after RTs less than 500 ms, and “Too Slow” displayed for 500 ms after RTs greater than 2000 ms, and “Good Speed” displayed for 500 ms after other RTs. In the third phase, a confidence rating had to be made between timing feedback and accuracy feedback. For the long final phase, 250 ms of information regarding the RT of the selection was the only trial-by-trial feedback, and the computer immediately requested a confidence rating. Again, a 500 ms pause came between trials, but subjects waited until the end of the block to read how many of their choices were correct out of 48. Figure 2.10 shows the sequence of events.

Based on subjects’ comments in Experiments 1 and 2, these later experiments had a revised confidence rating system. 10 keys had been confusing and underutilized by many people, so 8 keys were used in Experiments 3 and 4. Subjects were told that pressing the 1 key meant they thought they had made a mistake, while pressing the 2 key meant they were completely uncertain about their choice. Keys 3 through 8 showed increasing confidence so that 5 meant medium confidence about the choice and 8 meant very high confidence about the choice.
In Experiment 4, as a manipulation check at the end of the experiment, I asked two questions of the subjects: “After training, which one of the cancers did a male (female) rarely or never have?”

**2.4 Results for Experiments 3 and 4**

**2.4.1 Accuracy and Confidence**

Data from subjects who used the same confidence rating on more than half of all trials were excluded from analysis. Data from subjects whose proportions of correct responses were less than 0.45 or whose accuracy and confidence showed a negative correlation were also not used. In Experiment 4, specific subjects’ data were excluded if the subjects did not answer the final questions regarding gender correctly. Generally, subjects removed from analysis failed to meet more than one of the criteria, indicating failure to follow instructions. For Experiment 3, 20 subjects remained, and for Experiment 4, 27 subjects remained.

Trials with RTs for choice or confidence-in-choice greater than 10,000 ms or less than 200 ms were excluded. Furthermore, RTs more than 3 standard deviations from a subject’s mean were trimmed. Confidence ratings were transformed from whole numbers into values on a scale from 0 to 1 by dividing by the largest possible rating.

Mean accuracy in Experiment 3 was 0.60. Accuracy varied for which category was chosen \( F(3,54) = 10.11; \ p < 0.01 \) but not for which set of alternatives was available \( F(3,16) = 0.01; \ p > 0.05 \). Accuracy was quite high \( M = 0.66 \) for choice 3, stomach cancer. There was a significant interaction \( F(3,57) = 8.90; \ p < 0.01 \), showing that for choice 1, ovarian cancer, accuracy was much lower in the dud-present condition \( M = 0.48 \) than in the dud-absent condition \( M = 0.69 \), but, conversely, for choice
4, testicular cancer, accuracy was higher in the dud-present condition ($M = 0.57$) than in the dud-absent condition ($M = 0.50$). Figure 2.11 graphs the differences between conditions, broken down by gender in the stimulus and choice made by the subjects. A small part of why Experiment 4 was necessary was to check whether this was random variation or due to particulars of the stimuli.

Mean accuracy for Experiment 4 was 0.62. Category choice made a significant difference ($F(3,74) = 19.42; p < 0.01$), with categories 2 and 3, gender-neutral cancers, more often correctly chosen than categories 1 and 4, gender-specific cancers. The main effect for inclusion or exclusion of the dud-alternative did not reach significance ($F(3,23) = 0.29; p >0.05$). The interaction in this experiment was significant ($F(3,78) = 8.30; p <0.01$), but I offer no explanation of this effect because people’s accuracy alone was not the main concern of these experiments.

For Experiment 3, confidence and accuracy had an overall positive relationship ($r = 0.20$) with the exception of choice 4, testicular cancer. Choosing category 4 was associated with high confidence but low accuracy. Compare graphs in Figure 2.13. Subjects very often felt good about this choice when it was incorrect. This irregularity was reduced in Experiment 4, when the two variables still had a weak correlation ($r = 0.23$) but the linear relationship was made clear in the calibration graph of Figure 2.14.

### 2.4.2 Detecting the Dud-Alternative Effect

Confidence for both experiments followed a similar distribution, which can be seen in the histogram in Figure 2.15. Roughly half the trials resulted in one of the top two confidence ratings, while the low ratings were very rarely used. Trials with the lowest rating were removed because participants only provided it when they had
Figure 2.11: Accuracy between conditions. Points plotted are for the accuracy difference between the dud-present condition and the dud-absent condition of Experiment 3. Averaged across subjects, data was split by gender of stimuli and category choice so that the number of trials for each point was approximately the same.
Figure 2.12: Accuracy between conditions. Points plotted are for the difference between the dud-present condition and the dud-absent conditions, averaged across subjects of Experiment 4. Compare to Figure ref3acdif.
Figure 2.13: Confidence and accuracy. Part A shows confidence broken down by gender of the stimuli and choice then averaged across subjects in Experiment 3. Part B does the same for accuracy.
Figure 2.14: Calibration. Accuracy was assessed over all trials for each combination of confidence level and condition. Shown are means for both conditions and standard error bars for the dud-present condition.
made obvious errors such as accidentally pressing the wrong key for their choices. Mean confidence for Experiment 3 was 0.74 though this varied greatly based on which choice was made \((F(3,54) = 17.30; \ p < 0.01)\). Again, confidence for choice 4, testicular cancer, was higher \((M = 0.81)\) than for other choices. No main effect for the presence or absence of a dud alternative was found \((F(3,16) = 0.56, \ p > 0.05)\), and no significant interaction emerged \((F(3,57) = 1.56, \ p > 0.05)\).

Figure 2.15: Confidence distribution. The histogram of confidence for Experiment 4 is shown.
Confidence for Experiment 3 was 0.740 when three options were available to choose and 0.734 when four options were available. This insignificant result is in the opposite direction that the dud-alternative effect would predict, but when broken down by subject, choice, gender of the stimulus, and number of alternatives available, then averaged across subjects, the pattern seemed reversed. Figure 2.16 plots the differences between times when the dud was present and absent from the choice set. Most of the points fall to the positive side of the reference line, indicating that confidence was generally higher when the dud-alternative was present, but this difference was small and unreliable.

Because overconfidence had been greatest in Experiment 2 when the stimulus strongly favored one option, I show in Figure 2.17 a breakdown of confidence for Experiment 3 based on the specific dimensions of the stimuli. Part A shows how confidence is higher around the edges of the stimuli space, but rather than the lowest confidence being in the center, where the categories overlap most, it is lowest in an area near category 1, ovarian cancer. It is highest in the area near category 4, testicular cancer. Part B shows the difference in confidence between the dud-present and dud-absent conditions. The region which had shown lowest confidence in Part A showed no dud-alternative effect in Part B, but the opposite may also be true. Some stimuli on the fringes or in the region of category 4 qualitatively showed an increase in confidence when the dud alternative was present.

Mean confidence for Experiment 4 was 0.77. As in Experiment 3, it was strongly affected by which choice the subjects made (\(F(3,74) = 21.02; p < 0.01\)), but not significantly different between the dud-present and dud-absent conditions (\(F(3,23) = 1.06; p > 0.05\)); nor did an interaction occur (\(F = 1.77, p > 0.05\)). Confidence
Figure 2.16: Confidence between conditions. Points plotted are for the confidence difference between the dud-present condition and the dud-absent condition of Experiment 3. Averaged across subjects, data was split by gender of stimuli and choice.
was 0.773 when only three options were available for the choice and 0.769 when four options were available. Figure 2.18 displays the confidence differences for Experiment 4. Changes in confidence between the dud-present and dud-absent conditions are very close to zero. A look at how confidence typically manifests over dimensions of the stimuli demonstrated a cleaner pattern than in Experiment 3. In Part A of Figure 2.19, a singular dip in confidence lies near the center of the graph where stimuli are the most difficult to categorize. In Part B, differences between times when the dud was present and absent are almost universally near zero.

2.4.3 Response Time

For most subjects, the choices typically took between 1 and 2 seconds. The mean of RTs for Experiment 3 was 1512 ms. As with other dependent variables,
Figure 2.18: Confidence between conditions. Points represent mean confidence differences between the dud-present condition and the dud-absent condition of Experiment 4. Data were split by gender of stimuli and choice then averaged across subjects.
Figure 2.19: Confidence over stimuli. As with Figure 2.16, in Part A, the intersections of lines represent confidence for a specific total number of cells and a 0.03 range of proportions of cell type. In Part B, intersections represent the difference in confidence between when the dud was present and absent.

Variance related to choice overwhelmed variance between the dud-present and dud-absent conditions. When subjects chose categories 1 and 4, RTs were smaller than when subjects chose categories 2 and 3. The effect of choice was significant ($F(3,53) = 33.38, p < 0.01$), but the alternatives offered made little difference ($F(3,16) = 0.79, p > 0.05$) with no significant interaction of the two factors ($F(3,16) = 0.62, p > 0.05$).

The mean of RTs for Experiment 4 was 1650 ms, with significant differences among category choices ($F(3,74) = 13.68; p < 0.01$). The choice of category 2 had the highest mean RT at 1757 ms while the choice of category 4 had the lowest mean RT at 1574 ms. In the presence of the dud alternative, RTs averaged 1641, and in the absence of the dud, RTs averaged 1659. While there was no effect for whether the dud was present or not ($F(3,23) = 0.04; p > 0.05$), there was an interaction ($F(3,23) = 4.40; p$
Figure 2.20: RT between conditions. Points represent mean RT differences between the dud-present condition and the dud-absent condition of Experiment 3. Data were split by gender of stimuli and choice then averaged across subjects.

Choosing category 1 when the dud alternative was offered resulted in a mean of 1675 ms but when the dud alternative was not offered a mean of 1534 ms.

Differences between dud-present and dud-absent conditions are shown in Figure 2.20 for Experiment 3 and Figure 2.21 for Experiment 4. Differences were around zero, especially for Experiment 4.

RT had an inverse relationship with confidence. RT quantiles are plotted for each level of confidence in Figure 2.22 to provide a sense of how subjects’ speed coincided with their feelings of being correct. As confidence ratings decreased, RTs
Figure 2.21: RT between conditions. This graph plots mean RT differences between conditions as in Figure 2.20, but the scale is smaller.
Figure 2.22: Relationship between RT and confidence. For Experiment 4, quantiles of RT are plotted against each level of confidence. Open circles represent the 0.1 quantiles; upward pointing triangles represent the 0.3 quantiles; solid circles represent the medians or 0.5 quantiles; inverted triangles represent the 0.7 quantiles; open diamonds represent the 0.9 quantiles.

usually increased, but the distribution of RTs also spread out slightly. Likewise, plotting response latency over the range of stimuli produced a pattern opposite that for confidence. Compare Figure 2.23 to Part A of Figure 2.19.

### 2.4.4 Discussion

I had expected that confidence and accuracy would have a weak positive relationship across trials, and confidence and RT would have a weak negative relationship across trials, and these data supported such ideas. However, my main hypotheses
Figure 2.23: RT over stimuli. The intersection of lines represent RT for a specific total number of cells and a 0.03 range of proportions of cell type. With smoothing, these points represent moving averages of RT over a wider range of stimuli.
pertained to situations evidencing the dud-alternative effect. I predicted that when a dud-alternative was available in the choice set, confidence would increase, accuracy would decrease, and RT would remain the same. My theory for how a dud-alternative causes these outcomes stated that a dud is removed from consideration as a choice because there is a clear reason to ignore it. How confident the person feels about the choice involves reflection on the original possibilities, so alternatives which have been discounted give a boost to confidence in the chosen alternative. If any of this proposed process had occurred, it would likely have shown up in the plots of differences between dud/no-dud conditions.

In Experiments 3 and 4, differences in confidence in choice correctness and differences in proportions of correct responses were negligible between conditions. Most points in Figure 2.18 are not above the reference line and most points in Figure 2.12 are not below the reference line. These results were inconsistent with my hypotheses.

As time is a more continuous variable than either accuracy or confidence, its distributions can be studied more precisely. Whether the dud was present or absent, the centers of the RT distributions were very close, and the overall shapes of the distributions were very similar. This singular result for RT may be in line with my third hypothesis, but without evidence of the dud-alternative effect on confidence, I cannot say what stable RTs mean.

Differences in RT distributions have long been taken to suggest differences in cognitive processing (Diederich and Busemeyer, 2006, LaBerge, 1962, Laming, 1968, Link and Heath, 1975, Ratcliff, 1978, Smith and Vickers, 1988, Usher and McClelland, 2001), but my expectation was that changes in the way a set of alternatives are processed under the dud-alternative effect would not be noticeable. Because there
are no clear RT differences between conditions in these experiments, subjects most likely only considered combinations of alternatives in one fashion, doing less to confirm my theory than disconfirm the dud-alternative effect in general. Taken as a whole, these experiments contradict my proposed processing account for the dud-alternative effect.

2.5 Methods for Experiment 5

Although the first four experiments were certainly in the spirit of testing the dud-alternative effect, they were not identical to the original research. Many differences between them and the work of Windschitl and Chambers (2004) could have made an impact on the results. Before theorizing on why any particular difference would prevent the dud-alternative effect from resulting, I felt it would be prudent to test for the effect in a situation in which it was previously obtained. Consistent failure to find the effect in Experiments 1 through 4 indicated that the effect is not robust to minor changes in the tasks subjects perform. Experiments were modified to reduce the differences between my methods and those of the earlier research, but my results became less supportive of theirs, so I became more determined to test for what I had first assumed would be present in the data – the dud-alternative effect. The next step was to replicate an experiment of Windschitl and Chambers (2004), so I chose Experiment 5, in which subjects indicated how good they felt about their chances of winning a series of raffles. Using the description of the booklet of raffles, I carried out my own experiment. The expectation was for a mild but statistically significant effect to emerge.
2.5.1 Subjects

43 participants were taken from The Ohio State University undergraduate subject pool, students who were enrolled in introductory psychology classes. They received course credit as part of their research experience requirement.

2.5.2 Stimuli

Table 2.1 shows the 24 critical raffles copied from Windschitl and Chambers (2004). The baseline/dud-absent raffles had only two people holding similar numbers of tickets. The dud-present raffles had four to five people. Two of the players were the same as those in the baseline raffles of the same sets, but two or three new players each held very few tickets. The mediocre-present raffles had the two original players, but the two or three new players each held slightly more tickets than the new players in the dud-present raffles. The raffles included caricatures of the players, with the leftmost player labeled “You.”

2.5.3 Procedure

Each participant received a booklet with 24 critical raffles and 9 filler raffles, in total taking about ten minutes to complete. The 24 raffles were divided into 8 sets, each set containing one baseline/dud-absent raffle, one dud-present raffle, and one mediocre-present raffle. The order of the 24 raffles was randomized with a constraint that raffles of a particular set were separated by several pages in the booklet. Each raffle was led by the question: “At a gut level, how would you feel about your chances of winning this raffle?” Participants respond by circling one of 9 asterisks on a scale with “not good at all” on the left and “very good” on the right. Written instructions on the first page of the booklet explained how to use the scale.
and stressed to participants that “We are interested in your initial impressions and your gut-level responses. We are not interested in your careful analysis of exactly how optimistic you should feel or in your precise assessments of the objective likelihood of winning.”

2.6 Results for Experiment 5

One subject’s data was excluded from analysis because nearly all responses were one of two ratings, but all other data were used.

Table 2.1 also shows the mean ratings for the 24 main raffles. Responses were analyzed by a within-subject ANOVA as in Experiments 1 through 4. Confidence ratings differed significantly across raffle sets ($F(7,287) = 74.16; p <0.01$). The set × raffle type interaction was significant ($F(14,574) = 9.37; p<0.01$), and the effect for raffle type was significant ($F(7,287) = 21.70; p <0.01$).

The mean confidence for baseline raffles without any weak alternatives was 6.17, while the mean confidence for raffles with dud alternatives was 6.13, and the mean confidence for raffles with mediocre alternatives was 5.29. These results are in the same direction as the objective probabilities of winning those raffles. To see where the major differences on the raffle type factor lay, two planned comparisons were performed. Subjects were significantly less confident about winning the mediocre-present raffles than they were about winning the baseline raffles ($F(1,82) = 34.10; p <0.01$). Subjects were neither significantly more or less confident about winning the dud-present raffles than they were about winning the baseline raffles ($F(1,82) = 0.08; p >0.05$).
<table>
<thead>
<tr>
<th>Raffle type</th>
<th>“You”</th>
<th>Other players</th>
<th>Objective probability</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Set 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>14</td>
<td>16</td>
<td>0.467</td>
<td>4.83</td>
</tr>
<tr>
<td>Dud present</td>
<td>14</td>
<td>2 16 1</td>
<td>0.424</td>
<td>5.64</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>14</td>
<td>6 16 5</td>
<td>0.341</td>
<td>4.60</td>
</tr>
<tr>
<td><strong>Set 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>23</td>
<td>15</td>
<td>0.605</td>
<td>7.21</td>
</tr>
<tr>
<td>Dud present</td>
<td>23</td>
<td>15 3 2</td>
<td>0.535</td>
<td>6.64</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>23</td>
<td>15 6 6</td>
<td>0.460</td>
<td>6.55</td>
</tr>
<tr>
<td><strong>Set 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>12</td>
<td>14</td>
<td>0.462</td>
<td>4.29</td>
</tr>
<tr>
<td>Dud present</td>
<td>12</td>
<td>1 2 14</td>
<td>0.414</td>
<td>5.14</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>12</td>
<td>3 4 14</td>
<td>0.364</td>
<td>4.71</td>
</tr>
<tr>
<td><strong>Set 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>39</td>
<td>31</td>
<td>0.557</td>
<td>7.26</td>
</tr>
<tr>
<td>Dud present</td>
<td>39</td>
<td>31 2 3</td>
<td>0.520</td>
<td>6.71</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>39</td>
<td>31 4 8</td>
<td>0.476</td>
<td>5.98</td>
</tr>
<tr>
<td><strong>Set 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>17</td>
<td>7</td>
<td>0.708</td>
<td>7.71</td>
</tr>
<tr>
<td>Dud present</td>
<td>17</td>
<td>1 7 1 2</td>
<td>0.607</td>
<td>7.26</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>17</td>
<td>3 7 3 4</td>
<td>0.500</td>
<td>6.86</td>
</tr>
<tr>
<td><strong>Set 6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>20</td>
<td>14</td>
<td>0.588</td>
<td>6.79</td>
</tr>
<tr>
<td>Dud present</td>
<td>20</td>
<td>2 14 1</td>
<td>0.541</td>
<td>6.60</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>20</td>
<td>6 14 6</td>
<td>0.435</td>
<td>5.83</td>
</tr>
<tr>
<td><strong>Set 7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>45</td>
<td>53</td>
<td>0.459</td>
<td>4.62</td>
</tr>
<tr>
<td>Dud present</td>
<td>45</td>
<td>3 4 53 3</td>
<td>0.417</td>
<td>5.00</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>45</td>
<td>23 28 53 20</td>
<td>0.266</td>
<td>3.55</td>
</tr>
<tr>
<td><strong>Set 8</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>41</td>
<td>36</td>
<td>0.532</td>
<td>6.62</td>
</tr>
<tr>
<td>Dud present</td>
<td>41</td>
<td>36 2 2 1</td>
<td>0.500</td>
<td>6.00</td>
</tr>
<tr>
<td>Mediocre present</td>
<td>41</td>
<td>36 29 22 22</td>
<td>0.273</td>
<td>4.29</td>
</tr>
</tbody>
</table>

Table 2.1: Ticket distributions, probabilities of winning, and confidence for three types of raffles are shown for 8 sets.
2.6.1 Discussion

These results are inconsistent with my hypothesis regarding confidence and inconsistent with the dud-alternative effect. Subjects’ feelings about the chances of winning the raffles roughly corresponded to the true chances of winning the raffles. The presence of dud-alternatives did not significantly increase confidence.

Windschitl and Chambers (2004) found that inclusion of mediocre alternatives did not change confidence, but inclusion of dud-alternatives increased confidence. The replication in Experiment 5 provided completely different results. Confidence decreased with mediocre alternatives and was unchanged with dud-alternatives. Previously, I had been concerned that the dud-alternative effect was not robust to changes in the tasks people performed, but now it seems especially fragile. This experiment refutes the generality and ubiquity of the dud-alternative effect.

Still, I cannot draw firm conclusions about the existence or non-existence of the effect. Although confidence was lower in the dud-present raffles, the difference was negligible. If people were basing their confidence ratings on the statistical probabilities of winning, a significant and consistent decrease in confidence should have emerged for raffles with duds; instead, on three out of eight sets, people felt better about dud-present raffles. Sets 1, 3, and 7 in Table 2.1 show what could be the dud-alternative effect.

Notice that there is a major difference between these three sets and the other five. Average confidence ratings for the baseline raffles are extremely low. Such low numbers necessarily tend to be lower than the confidence ratings for the dud-present raffles within the same sets. These results may relate directly to a property of the sets. On those three sets with a positive confidence difference between dud-present
raffles and baseline raffles, the greatest number of tickets belongs to one of the “other players.” The number of tickets subjects (“you”) held were lower than the number of tickets another player held. This is not true for any other sets.

If there is any sign of the dud-alternative effect in these data, it is for situations when the focal alternative has a lower chance of success than another alternative. Windschitl and Wells (1998) have discovered another effect in these judgments of the likelihood of winning raffles, which they termed the “alternative-outcomes effect.” They claim that with objective probability held constant, the subjective likelihood of a focal alternative decreases when the likelihood of the most similar alternative increases. In three sets of Experiment 5 the most similar number of tickets surpassed the focal number of tickets, and in each case, inclusion of dud-like numbers of tickets increased confidence. Assuming that both effects are individually weak in this experiment, their combined strength may be enough to show a change in confidence. Alone, the dud-alternative effect was not detectable.

When the three sets in which people had more confidence with a dud-alternative were combined with the five other sets in which people had less confidence with a dud-alternative, the average showed no effect. This lack of difference between the baseline and dud-present conditions meant subjects’ confidence ratings did not conform to normative theory for likelihood judgments, but they were far closer than anticipated. An increase to confidence due to duds was not apparent.
CHAPTER 3

GENERAL DISCUSSION

After repeated attempts to adapt the work of Windschitl and Chambers (2004) to a perceptual categorization task, few conclusions can be reached. I set out to expand the known realm in which the dud-alternative effect operates, but each new experiment instead revealed that the effect is hard to obtain – even in the limited domain where it previously appeared.

3.1 Motivation

The purpose of this series of experiments was to extend the dud-alternative effect to other judgment and choice tasks while measuring more than just feelings about the likelihood that an alternative was best. First, by tracking confidence, I expected to find the effect for choice sets with more variation in evidence favoring or opposing the worst option. With a great variety of more specific, objective amounts of evidence to support one option over another, I might have seen when the dud-alternative effect was strongest and weakest.

Second, by measuring accuracy and RT, I hoped to learn what effect, if any, dud-alternatives had on them. One of my main hypotheses was that the historic negative relationship between confidence and response latency would hold in most
circumstances but not for the dud-alternative effect. A dud in a set of options would not change the time needed for responding because a person need only devote a small amount of mental resources to considering that option. I further hypothesized that confidence and accuracy would have a positive relationship overall, but the proportion of correct responses when the dud alternative was present would be smaller than when it was absent. The choice would seem easier when evidence strongly disaffirmed one or more options, including a dud alternative listed, but the probability of choosing the best alternative would decrease. This second hypothesis is most in keeping with the original work on the dud-alternative effect, and yet it has found the lesser support of the two.

3.2 Experiments 1 and 2

Experiments 1 and 2 did not directly address the dud-alternative effect but did reveal how confidence, accuracy, and RT change in relation to evidence available for a choice among multiple alternatives. The main merit of these experiments was in establishing the utility of my basic procedure. The method of soliciting confidence ratings and the apparatus for measuring all the variables of interest were refined.

These studies confirmed the typical relationships among the three dependent variables. Although I was unable to test my hypotheses because they hinged upon finding a clear dud-alternative effect, my basic expectations were shown to be correct for how people would feel and behave when choosing among four options with varying evidence. As subjects’ observations favored the likelihood of one alternative but not another alternative, accuracy, confidence, and speed increased. The ease with which
the choice was made revealed itself in all three measures. Easy choices had increased accuracy, increased confidence, and decreased RTs.

More importantly, choices for which people quickly recognized one option as best and one option as worst gave rise to strong feelings of confidence even though the probability of responding correctly did not increase dramatically. One could report this as support for the dud-alternative effect by noting that when evidence for one alternative was very low, people experienced unwarranted confidence in their choices, but two characteristics of the situation belie this interpretation.

First, counter to my expectations, these cases had dramatically decreased RTs along with the increased confidence ratings. In all other cases, I saw this as a result of both the strength of the alternative chosen and the weakness of the alternatives not chosen. I cannot safely cite the near-impossibility of the worst alternative as the sole reason for the changes in dependent variables.

Second, what inspired these experiments was the work of Wells et al. (1993) with subjects choosing a suspect out of a lineup. They found that when an option was very dissimilar to a description in a person’s memory, confidence increased while accuracy became questionable. For Experiments 1 and 2, confidence and accuracy maintained a positive relationship even when one option was especially bad. For extreme stimuli in which one alternative became very unlikely, if the choices seemed easier, it’s because they were. Accuracy was higher on those trials of the experiments. Confidence ratings are subjective, so how people interpreted the confidence scale cannot be known, but so as long as both accuracy and confidence moved in the same direction on most trials, this was not truly a dud-alternative effect.
I discussed before how confidence may have increased unduly when one alternative was quickly dismissed, but this should not be surprising in light of normal calibration error. The larger proportional increase to confidence than to accuracy is typical of people's systematic mistakes in self-assessment. Overconfidence is greatest at higher accuracy because people experience large increases in confidence with moderate increases in success (Lichtenstein et al., 1982). Confidence may reach the top of its scale when accuracy is as low as 0.60.

Because of these alternative explanations of the data from Experiments 1 and 2 and the limitations of the study designs, the veracity of my hypotheses remained unknown. Stronger tests were needed.

3.3 Experiments 3 and 4

Experiments 3 and 4 more closely mimicked the work of Windschitl and Chambers (2004). Certain alternatives were sometimes not available, and subjects had to use what they knew about the categories to assess how they felt about one being correct. Still, in an effort to expand on the earlier work, some aspects of the procedure were different. Any of a number of these differences could have prevented the dud-alternative effect from fully manifesting, but most are insufficient to explain the lack of hypothesized results.

3.3.1 Choice

First, the best option was not always obvious and never was specified by the task. Instead of just judging the strength or likelihood of a focal alternative, subjects had to decide for themselves which option had the highest chance of being correct. This was the most important addition to all four of the early experiments. Although
a probability can be mathematically estimated for a focal alternative being the best choice, an empirical probability is more desirable in many cases. To find that empirical probability, I asked people to initially choose the best alternative, and then rate their confidence for that choice. It is possible that having people make a choice before expressing confidence results in different confidence than when the choice is made for them.

Upon the belief that people know how they should judge the likelihood of outcomes but are less likely to do so when rushed, time pressure was used to enhanced the dud-alternative effect (Windschitl and Chambers, 2004). If taking a normative approach to confidence requires more careful mental processing of stimuli, subjects in my experiments had ample time during the confidence judgment and added time during the choice to do so. For each trial, this experimental design may have encouraged deliberation which would increase awareness of the rules of probability.

However, increasing careful adherence to rules of probability just by introducing choice seems unlikely since subjects were under time pressure in an already difficult task. With hundreds of trials to complete, subjects probably focused on choosing correctly rather than whether their confidence ratings were internally consistent.

3.3.2 Singular Duds

A second difference on Experiments 3 and 4 lay in the number of dud alternatives presented. Windschitl and Chambers usually had two duds in a set while I had only one. Though the stimulus in some trials may have made a second dud of one category, I did not intend for this to happen in any predictable fashion. Instead, there was consistently one dud either present or not. This alternative should have
been seen as distinctly a dud by subjects while other alternatives remained reasonable to choose.

It may be the case that one dud-alternative is not enough to increase confidence, but the severity of the dud-alternative in each trial was intended to overcome that obstacle. Dud-alternatives and the reasons they were dud-alternatives were obvious to the majority of subjects.

3.3.3 Implausibility

Third, the dud-alternatives I presented were more implausible than before. For someone with a basic knowledge of human anatomy, the dud alternative was not just an extremely unlikely diagnosis, it was impossible. Knowing that one diagnosis would always be false for one gender may have allowed people to ignore that option completely after seeing the gender indicator on a trial.

However, the dud would have been hard to ignore because subjects rapidly responded to trials in which the category of the dud switched, the listing of the dud as an option switched, and the type of stimulus changed. The fact that subjects occasionally chose the dud alternative, indicates that they did considered it on some trials.

3.3.4 Learning

A simple comprehension of connections between gender and cancer in the real world was the only pre-existing knowledge subjects brought to bear on the experiment tasks, but deciding how to categorize stimuli required more understanding. Learning about the nature of the categories and stimuli as well as how to perform the task
during the experiment made an interesting fourth difference between Experiments 3 and 4 and experiments by Windschitl and Chambers (2004).

In large part, subjects learned how to carry out the task on training trials or otherwise did not perform well. For training to be effective, subjects needed trial-by-trial feedback, which may have altered the relationship between accuracy and confidence. Confidence scales can be arbitrary in their design, so knowing how to use one takes practice. Because subjects were learning to rate confidence shortly after learning how to choose a category and while feedback was still frequent, they may have associated the two more closely than anticipated.

Whether intentional or unintentional, matching of confidence to accuracy when a dud-alternative was present seems unlikely when confidence and accuracy were not very well calibrated overall. Subjects were overconfident about their choices, especially for more accurate choices, and the correlation between the two measures was low. From these results, I believe that people did not fully use feedback about accuracy to modify ratings of confidence.

3.3.5 Deliberation

If requiring subjects to make choices had promoted more in-depth consideration about confidence, then the opposite could be said of the way information was presented to subjects. The fifth major difference from the original research is that all information relevant to the confidence ratings disappeared after the choices and before the confidence ratings. This occurred in only one previous experiments by Windschitl and Chambers (2004) when subjects took too much time to respond.
If confidence is assumed to be a metacognitive task, then people have to think about their own reasoning for why they chose one alternative rather than another. The mental process for one’s confidence works on the output of the mental process for one’s decision. My theory for how confidence comes about is that instantly after a choice is made, a person reflects on how much evidence supports each alternative. Although I don’t imagine this is a drawn-out process, if the stimulus is removed, people may have difficulty retaining their own rationale for a choice. Properties of the stimulus and the dud-alternative which was rejected may be less salient to the subjects when the screen goes blank.

More difficulty retaining reasons for why to choose and why not to choose alternatives after the choice is done would make comparison of the alternatives difficult. My processing account states that evidence against the dud-alternative is to blame for increases in confidence, but if people can’t recall their reasons for rejecting the dud-alternative, no confidence effect would be found. In line with my own theory, this explanation of how erasing the stimulus and choice set after a decision is reached could reduce the dud-alternative effect would be more persuasive if Windschitl and Chambers (2004) had not employed a similar method in their final experiment.

None of these points provide cogent arguments for why the dud-alternative effect did not arise in the results of Experiments 3 and 4. In light of the results from Experiment 5, concluding that differences in stated methods are the sole culprit is not reasonable.
3.4 Experiment 5

Experiment 5’s methods were nearly identical to those in an experiment by Windischtl and Chambers (2004). Subjects were drawn from a similar population, stimuli were copied and ordered alike, and instructions did not noticeably differ. Great care was taken to replicate the procedure, yet the results were not the same. My results showed no confidence effect for the dud-alternative condition. They suggest that the dud-alternative effect is not as robust as expected.

How might these results be interpreted differently? First, there was a significant interaction of the raffle type and raffle set, meaning that some sets influenced the raffle type manipulation. For the dud-alternative effect, this is noticeable with 3 sets. Subjects had increased confidence on dud raffles for those raffle sets with the lowest overall confidence. Therefore, the process leading to the effect may operate most reliably in a small range of situations but less reliably in others.

Similarly, a second interpretation of the results is that they may have occurred by random chance. When testing any hypotheses, there is the possibility that an effect will not be statistically significant though it is in fact real. One may argue that my data are unique without any known cause, and that another replication would yield better results, but this repeated-measures design with over 40 subjects provides great power for detecting differences in means. Assuming there truly is a widespread increase in confidence when dud-alternatives are present in choice sets, the probability that my data would randomly not show the effect is small. A better interpretation is that the effect appears inconsistently under these particular methods, so the chance of not finding it is greater than anticipated.
My theory for how dud-alternatives affect confidence was based on the idea that people use reasons to make choices and evaluate how good those choices are. I cite the research of Koriat et al. (1980), showing that increasing awareness of reasons contradicting a choice strongly impacts confidence. I believe that duds typically elicit these kinds of negative reasons in people’s thinking and alter the balance of evidence for and against the options people choose. Making a connection between the dud-alternative effect and the reasons for confidence effect is very appropriate since both effects may not be robust. Just as I have experienced trouble tapping into the dud-alternative effect, Fischhoff and MacGregor (1982) struggled to replicate the effect of the reasons manipulation for confidence. They set out to relate confidence in general knowledge questions to confidence in event forecasting, and discovered that while the two are quite similar, asking subjects to state reasons for and/or against a choice was not a sure way to obtain a difference in confidence. Most of their results were inconclusive until they used a stronger manipulation. Likewise, for the dud-alternative effect, several changes to experiment procedures could provide greater support for my theory.

3.5 Future Directions

Taken together, these five experiments cast doubt on the dud-alternative effect, but to conclude when/if it works to change people’s confidence, more experiments are needed.

Another replication of any experiment by Windschitl and Chambers (2004) would be a quick way to determine whether my results from Experiment 5 were incidental.
This would also eliminate the possibility that some detail which seemed insignificant in the original research caused the effect. Null results would suggest that dud-alternatives have no special influence on confidence, while other results would support the conclusion of my study, which is that dud-alternatives may have weak effects on confidence.

Under the latter case there would be reason to believe the dud-alternative effect is real, and future research could reproduce experiments in which the effect was substantial while introducing one new manipulation at a time. To see if a choice component attenuates confidence effects, an experiment of Windschitl and Chambers (2004) should be used mostly in its previous form with an adaptation to include picking the best alternative before rating confidence. In this same fashion one could also test whether removing one dud-alternative and leaving one dud-alternative reduces the effect or test what happens when an impossible alternative is the dud.
BIBLIOGRAPHY


