WISDOM OF THE CROWD MECHANISMS

by

Jon DMC Walker

BA, Pomona College, 1972

MSIS, University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

School of Information Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Jon DMC Walker

It was defended on

March 4, 2016

and approved by

John Duffy, PhD, Professor, University of California, Irvine

Stephen C. Hirtle, PhD, Professor, Information Sciences and Technology

Michael B. Spring, PhD, Associate Professor, Information Sciences and Technology

Vladimir Zadorozhny, Associate Professor, Information Sciences and Technology

Dissertation Director: Stephen C. Hirtle, Professor, Information Sciences and Technology

WISDOM OF THE CROWD MECHANISMS

Jon DMC Walker, MSIS

University of Pittsburgh, 2016

As Web 2.0 facilitates the collection of a vast amount of interactions, a phenomena, known as the wisdom of the crowd, is increasingly enlisted to justify using those interactions as surrogates for expert opinions. This dissertation explores this phenomena through an analysis of the micro elements of two wisdom of the crowd simulations: (1) Hong and Page's (2004) Diversity trumps ability model and (2) Luan et al.'s (2012) Fast and Frugal simulation. The focus of this study is on the micro elements that contribute to those simulations' results. This focus leads to the identification of a search mechanism that favors exploitation as a first step followed by exploration as defined by March's (1991) Exploration/Exploitation simulation.

Three new methods for creating a group of experts were developed and were shown to be not only superior to the Top 10 agents but also superior to the more diverse random group of ten agents which consistently outperformed the Top 10 agents in the Hong-Page model. It was also shown that these expert groups were more efficient in incorporating the entire range of heuristics possessed by the universe of agents. The problem spaces were manipulated in various manners and the effect of such manipulations demonstrated. Additionally, group process losses were demonstrated through the simulation of a Hidden Profile scenario in which skills possessed by only one agent were ignored by the group. The effect of the dichotomization rate in the Fast and

Frugal paradigm was highlighted and the effect of an alternative dichotomization rate demonstrated along with increasing the number of cues and manipulating the degree of correlation among them. Additionally, a set of perfect cue weights was developed for the Fast and Frugal paradigm and a simulation showed how a single agent executing the paradigm to choose the correct alternative saw its ability deteriorate as the cue weights progressed from the perfect order to all cues being equally weighted while groups of agents experienced increasing accuracy over the same progression.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The World Wide Web allows individuals to push items into the (digital) public square and to observe and interact with items that others have pushed into that public square. As Web 2.0 developed, the public obtained an even greater ability to interact with the items in the digital public square. The investigations into these interactions has led some to dismiss the entire phenomena with a 'Garbage In – Garbage Out' attitude; others, however, believe that there is a wisdom of the crowd effect, which, similarly to evolutionary forces, is not forced or managed but arrives at a level of "correctness", efficiently and better than would be the case in a managed environment, and even more importantly in a manner that scales easily.

This dissertation will explore the mechanisms of wisdom of the crowd simulations. If the phenomena, commonly identified as the wisdom of the crowd, is based on a collection of individuals making their own decisions as isolated individuals, as members of a group, as observers of, or as influencers of other individual's decision making, the processes which influence that decision making need to be identified so that they can be incorporated in agents used to simulate the wisdom of the crowd process(es). This dissertation focuses on mechanisms that have been used in two wisdom of the crowd simulations and does not attempt to define the processes by which some group decisions are remarkably accurate while others are notoriously inaccurate.

This section discusses the processes which influence the quality of decisions made by individual decision makers beginning with the historical perspectives of Aristotle, Malthus, Adam Smith, and Condorcet. It continues with a review of Francis Galton's three articles (1907 a, b, c) in *Nature* at the turn of the 20th century and then focuses on several processes which have been identified that negatively influence group decision making, specifically with Caplan's (2001, 2007, 2009) observations that the crowd, systematically, makes errors, Strasser's (1985) work on how groups fail to use all the information available to them when they work towards reaching a group-wide decision, and positively by Servan-Schreiber's (2012 a, b) work on how the market mechanism, itself, works in the process of group decision making.

This is followed in section 2 by an exploration of the machine learning field of Ensemble Learning which has many properties similar to the wisdom of the crowd effect. The use of supervised learning is used to show ensemble learning in several examples. Larrick and Soll (2006) are presented with their reliance on the power of the mere mathematics of averaging. Larrick is the author whom Surowiecki (2005) cites as the expert who speaks against the fallacy of 'chasing the expert', which is the idea that problems are solved by finding an expert in the field and following his/her advice. The section ends with a quick presentation of some representative alternatives with entirely different approaches to explaining the wisdom of the crowd.

Section 3 presents a more detailed description of Hong and Page's (2004) simulation of how diversity trumps ability and a simulation from Gigerenzer's ABC group (1996) about how the fast and frugal paradigm can be simulated to demonstrate some conditions under which group diversity trumps individual ability and when it doesn't.

Section 4 presents extensions to the Hong Page (2004) Heuristic Simulations that introduces learning into the process of group formulation and alternative metrics. Specifically, two negative correlation learning type algorithms and one more of an ensemble learning method are developed which are used to create teams of experts which outperform the normal group of the top 10 agents and outperforms the group of random agents also. A 'Hidden Profile' effect is also demonstrated when the group of agents ignore any heuristic that only one agent possesses. Additionally, Luan's (2012) Fast and Frugal simulation is extended with alternative dichotomization rates, cue validities, and environmental settings. The median dichotomization rate is demonstrated to result in 8/15 of all possible decisions being made at each cue which, in turn, means that only very rarely are decisions still pending after five cues and this is independent of the amount of information in the cues. This median dichotomization rate is compared to a dichotomization that only identifies the object with the highest value for that cue (i.e., whereas the median dichotomization makes the maximum number of decisions possible being made at each cue, the alternative makes the minimum number of decisions at each cue level). This section also demonstrates using 25 instead of 5 cues and the results of having those cues correlated to different degrees in groups.

Section 5 introduces a different perspective on a simulation based on the Hong-Page model; this model highlights the exploration versus exploitation aspects of decision making and the effect of several group processes on group decisions.

Section 6 discusses the findings and the final section identifies major conclusions and some limitations and delimitations of the studies and areas for future investigation.

1.1 HISTORICAL REFERENCES

The standard classic Greek philosophy reference to effective group decision making comes from Aristotle's *Politics*

> the many, who not as individuals excellent men, nevertheless can, when they have come together, be better than the few best people, not individually but collectively, just as feasts to which many contribute are better than feasts provided at one person's expense (Landemore & Elster, 2012, p. 1)

This is, of course, to be contrasted with the general tone of the Platonic Dialogues in which the many not only flounder around making various mistakes but come together not to improve their decision making but to eliminate those who are trying to lead them to better decisions, for example in the *Republic's* cave and, of course, in the story of Socrates' death sentence by an Athenian jury (280 votes of guilty, 220 votes for not guilty, perhaps). Even Aristotle towards the end of his life leaves Athens before it has the opportunity to repeat this wisdom of crowds on him.

The manner in which groups of people reach decisions is a topic that has been an area of interest throughout just about the entire Western Canon. During the Enlightenment, Smith and Condorcet put forward ideas that groups (with the individual members acting independently) arrived at beneficial decisions whereas Malthus, famously, predicted otherwise.

The Reverend Thomas R. Malthus (1798) is perhaps only remembered as the author of *An Essay on the Principle of Population* (original editions from 1798 to 1826). To place him in intellectual historical time, *On the Origin of Species* was first published in 1859 (Darwin, 1859) and was directly influenced by Malthus. Malthus starts out his analysis with two postulates:

"First, That food is necessary to the existence of man. Secondly, That the passion between the sexes is necessary and will remain nearly in its present state." He then continues (Malthus, 1798, p. 6) with

> Assuming then my postulata as granted, I say, that the power of population is indefinitely greater than the power in the earth to produce subsistence for man.
>
> Population, when unchecked, increases in a geometrical ratio. Subsistence increases only in an arithmetical ratio.

Malthus thus using 'the passion between the sexes' as the driving force in the development of history paints a picture of a contemporary and future dystopia caused by man's (and woman's) failure to temper their 'passions'. Porter (1986) points out that Malthus was reacting against writers such as Süssmilch who were maintaining that the goal of each state should be to increase its population to as large extent as possible, that the size of a kingdom's population was the measure of its power. Süssmilch's main work was published from 1740 to 1798 in four editions and was derived from the divine imperative to 'be fruitful and multiply'. Porter distinguishes these two perspectives as being pre and post the French Revolution. Süssmilch was a German Protestant minister and predates the Revolution. Malthus, an early enthusiast of statistics and founding member of the Statistical Section of the British Academy of Sciences, spanned the French Revolution.

Adam Smith published his enormously influential *The Wealth of Nations* in 1776 and Condorcet published his *Essay on the Application of Analysis to the Probability of Majority Decisions* in 1785. Malthus comments at refuting the views of both of these two works. In fact the full title of his essay is "An Essay on the principle of population, as it affects the future improvement of society with remarks on the speculations of Mr. Godwin, M. Condorcet, and other writers." He also has chapters in his essay directed at Adam Smith's economic analysis.

Adam Smith's work is most famous for its conception of the 'invisible hand':

> "…he [the individual member of the state] intends only his own gain; and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for society that it was not part of it. By pursing his own interests he frequently promotes that of the society more effectually than when he really intends to promote it." (Smith, 1776, p. 184).

Specifically, he writes about markets which appear to satisfy the desires of the community of consumers without a central authority. Not only does the 'invisible hand' operate without a central authority, but additionally he notes that it functions more effectively than intentional actions, as a general rule. Note the phrase within commas, "as in many other cases", which is not generally further explored, but clearly implies that Adam Smith does not view the 'invisible hand' as a mere market economic force, but rather as a more general force that transforms individualistic micro-behaviors into far reaching macro outcomes[1].

Condorcet's jury theorem states that if each member of a voting group has a better than 50% chance of making the correct choice then the probability that a group of them will make the correct choice increases with the size of the voting group – wisdom of the crowd, in other words[2]. However, he also acknowledges the opposite: When the members of the group have less

---

[1] Wikipedia's article on Smith's "invisible hand" also cites this *The Theory of Moral Sentiments* where "… [Smith] describes a selfish landlord as being led by an invisible hand to distribute this harvest to those who work for him." (en.wikipedia.org/wiki/Invisible_Hand accessed 3/4/14)

[2] Wikipedia defines the Central Limit Theorem as "In probability theory, the central limit theorem (CLT) states that given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected valued and well-defined variance, will be approximately normally distributed. That is, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (commonly known as the "bell curve")." (en.wikipedia.org/wiki/Central_limit_theorem accessed 3/4/14)

than a 50% change of making the correct choice individually, the group will more and more likely make the wrong choice the larger the group gets. Malthus, when discussing Condorcet (Malthus, 1798), notes that Condorcet died at the hands of French revolutionists, in 1794, whom he was championing as travelling down the path of increased enlightenment. The wisdom of the crowd seems to have gone astray during the Reign of Terror.

In opposition to Malthus' "passion of the sexes", both Smith and Condorcet rely on crowd decisions to lead to correct decision making. In Smith's case, the 'invisible hand' is a precursor to Schelling's work on micro-motives and macro-behavior (Schelling, 1978). Condorcet's jury theorem assumes that the agents act independently, which is something they failed to do during the French Revolution and that they have better than a random chance of being correct, individually, something Americans tend to deny about the French. Aristotle, on the other hand, is generally assumed to referring to a collaborative attempt to achieve some goal. There is the assumption that the 'many' do not independently become better as much as when they collaborate 'the better' rises by mutual consensus, by its self-evident superiority. The 'feast' doesn't turn out like the pot-luck dinner with 15 different varieties of macaroni and cheese but is organized in some manner that results in an appropriate variety of dishes.

Additionally Wikipedia defines the law of large numbers as "In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trails are performed." (en.wikipedia.org/wiki/Law_of_large_numbers accessed 3/4/14)

1.2 EPISTEMOLOGICAL REFERENCES

An analysis of the wisdom of the crowd effect could also be motivated epistemologically. Starting with Plato's having Socrates claim that knowledge is "True belief with an account" in *Theaetetus*, through Aristotle's original work (*Organon*) in defining of the study of deductive logic, including Bacon's *Novum Organum*, 1620, in which Bacon presents his theory of inductive logic and the scientific method. This method of knowledge had to wait for the development of probability theory to reach a sufficient level before a full blown statistical underpinning could support it. (Porter, 1986)

McKenzie (2004) points out this statistical underpinning provided a basis for a normative theory of correct decisions and psychology evolved from believing that people generally made decisions in accordance with that normative theory to Kahneman and Tversky's series of highly influential articles focusing on biases and decisions made contrary to those norms. Following this 'heuristics and biases' program, psychology developed an environmental/evolutionary perspective

> "First, the emphasis in the heuristics-and-biases program on studying the cognitive processes underlying judgment and decision making behavior represents important progress. Second, comparing the two optimistic views, the 1960s perspective and the research post-1990 described earlier, there are clear and important differences. The latter stresses the importance of environment in determining what is normative and why people behave as they do. Content and context matter, both normatively and descriptively. The realization (by psychologists) that a given task might have multiple reasonable normative responses opens the door to better understanding of behavior." (McKenzie, 2004, p. 333)

This justification of a decision process makes the final proof of its appropriateness not its 'truth' but its reasonableness within a given context. A reliance on the wisdom of the crowd can be interpreted as an additional step towards merely making manifest a means of knowing that has already proven effective throughout human evolutionary history. When all your cavemen buddies come running for their lives around a boulder, you run with them without any need to find out for yourself what it is they are running from. In the post-modern context, the infallible 'True belief with an account' may not just be out-of-reach; it may well be either forever unreachable or merely just nonexistent, in which case the crowd defines the most usable alternative.

Arthur (1994) writes about how people use inductive reasoning when dealing with complex events. He asserts that when humans are confronted with a complex situation, they retrieve a mental model which approximates the current situation and then reason deductively about the situation. When the deductions/mental models do not effectively address the current requirements, they will switch to a different mental model while remembering that the prior mental model did not adequately function in that situation. This process allows for a weighting scheme to be developed for the different models. In complex adaptive systems, it may well be that there is no 'logically' valid response but that different people following different models with different weights will eventually arrive at an optimal solution without any individual having any idea of what the optimal solution is. The Santa Fe Institute example of this is the crowded bar problem where if the bar is too crowded, it isn't pleasant but if it is not too crowded then it is pleasant. So how do you decide to go to the bar or not, simulations of different agents with different rules end up with an optimal number of agents 'going' to the bar. The rules are neither deductive nor inductive but adaptive.

Klein (2008) presents a recognition primed decision model. The person recognizes a pattern. His exemplar agent is a fireground commander (the fireman who is in charge of the firefighting team when they are fighting a fire): this commander recognizes the fire as fitting a pattern and mentally runs down items that are necessary for the current situation to match the pattern, when they match that is okay, when they don't match, the question becomes can adjustments be made to deal with the lack of congruence and if not then he starts searching for a different pattern. This allows for quick pattern recognition (inductive reasoning) followed by a step by step analysis of how accurately does the current situation match the pattern (deductive reasoning).

Pearl (2009) has spent his career demonstrating that causality can be inferred from data and that causality is what is important rather than merely being able to assert probabilistic relationships. Of course, the mantra of 'correlation does not mean causation' is true and he is not denying it. He needs to introduce additional elements of information in order to make those causal statements.

> This distinction [between standard statistical analysis and causal analysis] implies that causal and associational concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change – say from observational to experimental setup – because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by causal assumptions which identify relationships that remain invariant when external conditions change….behind every causal conclusion there must lay some causal assumption that is not testable in observational studies. (Pearl, 2009, p. 99)

1.3 GALTON'S OX AND NUMBER OF JELLY BEANS IN A JAR

The current poster child for group decision making comes from Surowiecki's (2005) retelling of Francis Galton's story about a group's estimate of the dressed-out weight[3] of an ox. Galton has three publications in *Nature* relating to this event. *Nature* (28 Feb 1907) "One Vote, One Value" (Galton, 1907 a) *Nature* (7 March 1907) "Vox Populi" (Galton, 1907b) and in letters to the editor "The Ballot-Box" (28 March 1907) (Galton, 1907 c). The first article (Galton, 1907 a) "One Vote, One Value", presents his opinion that in awarding damages a jury of 12 people would do better by taking the mean between the 6^{th} and 7^{th} ranked values (i.e., the median of the recommended award amounts by the jury) instead of the mean of the entire 12 person jury since the extreme values have a larger effect on the mean than the more central values. The second article (Galton, 1907b), "Vox Populi", is the article dealing with the dressed out weight of the ox. The purpose for this publication is to support the prior "One Vote, One Value" article. The third reference (Galton, 1907 c) in the Letters to the Editor section, actually addresses two letters: one in which the writer informs Galton that people, at these fairs, take great pride in their skill at guessing dressed out weights and that it isn't something done on a lark, the second letter is Galton replying to R.H. Hooker's letter of 21 March 1907 in which he (Hooker) calculates the mean from the centiles provided in the "Vox Populi" article.

> A weight judging competition was carried on at the annual show of the West of England Fat Sock and Poultry Exhibition recently held at Plymouth. A

[3] Dressed-out weight is the animal's weight "after being killed, the hide, head, feet and gut are removed… Dressing percentages are highly variable…" www.thebeefsite/articles/759/dressing-percentage-of-slaughter-cattle (accessed Oct 21, 2013) Dressing percentage is in the 59% region. An ox is not a species of bovine or even a sub-species, but (usually) a bovine that can be hitched to a yoke to do some beast of burden task.

> fat ox having been selected, competitors bought stamped and numbered cards, for 6d. each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered and "…dressed." (Galton, 1907b, p. 450)

Six pence, 6d., in 1907 is worth £2.24 in 2012 currency using a retail price inflator (US $3.62, using Oct 2013 exchange rates) and £8.71 ($14.08) using average earnings.[4] Given the relatively high cost of entering an estimate, one imagines that would tend to limit those making estimates to those with some confidence in their estimates. In the first letter to the Editor in the Ballot-Box article (Galton, 1907 c) mentioned above, the author speaks to the fact that there is great pride in the area in their ability to predict 'dressed-out' weight. Rather than being a "guess how many jelly beans there are in the jar" kind of contest, it represents a valuable skill in the region. Page (2007) states "guessing the weight of a steer is not that difficult." From his experience, guessing the weight within a few hundred pounds is an easy skill to acquire; however, note from that the only estimates that were even 100 pounds away from the true value were the lowest 10% and the highest 5%, so that 85% of all the estimates were within 100 pounds, a much better demonstration of skill than is implied by Page's 'within a few hundred pounds' comment.

Galton produces the following table in the second article, "Vox Populi", (Galton, 1907b).

[4] Inflation conversion preformed at http://www.measuringworth.com/ppoweruk/ on Oct 19, 2013, exchange rate conversion between pounds and dollars as of the same date via Google's Chrome currency converter.

Table 1 From Galton (1907b) on distribution of weights

Degree of the length of Array 0-100	Estimates in lbs	Observed deviates from 1207 lbs	Normal p.e. = 37	Excess of Observed over Normal
5	1074	-133	-90	+43
10	1109	-98	-70	+28
15	1126	-81	-57	+24
20	1148	-59	-46	+13
25	1162	-45	-37	+8
30	1174	-33	-29	+4
35	1181	-26	-21	+5
40	1188	-19	-14	+5
45	1197	-10	-7	+3
50	1207	0	0	0
55	1214	+7	+7	0
60	1219	12	14	-2
65	1225	18	21	-3
70	1230	23	29	-6
75	1236	29	37	-8
80	1243	36	46	-10
85	1254	47	57	-10
90	1267	52	70	-18
95	1293	86	90	-4

"P.E." is defined in the article as probable error is half of 45 + 29: 45 is the error at the 25^{th} percentile and 29 is the error at the 75^{th} percentile. An estimate drawn at random from the set of valid estimates is equally likely to be within these bounds as being outside of these bounds, therefore the probable error is 37. This is Galton's calculation and not the currently popular 'Expected Error' which in this case would be the sum of the actual error value for each card times 1/787, representing the average error if a random card was drawn at random from the collection and the process repeated a large number of times.

If you take the average of the second column of this table "Estimates in lbs" you get 1197.211 and the actual dressed out weight was 1198 pounds. However, 1197.211 is not the average of the 787 estimates taken. The numbers in the second column represent the values of the 20^{th}-tiles. The same 'statistic' for quartiles (25, 50 and 75) is 1201.667; for quintiles (20, 40, 60, 80) is 1199.5; and the tens centiles (10, 20, 30, etc.) is 1198.333. What all these numbers principally leave out are the most extreme values; the tails which, by definition of the mean, hold the most power to move the average. However, all three of these numbers are quite close to the actual value and better than the mean at 1207.

Galton is explicitly claiming that average is the incorrect metric to use to determine a consensus measure due the leverage that the extremes have to move the average; however, he does note that the mean of the centiles was unexpectedly accurate in this instance. Galton also does not claim that the median as the consensus measure is the 'correct' measure, merely that it more clearly represents a 'one vote, one value' weighting. Galton does not believe in democracy to identify 'correct' answers, per se. In this case, it was more 'correct' than he would have expected; however, not quite as accurate as the 'correct to within half a pound' story implies. The second letter to which he is responding in "letters to the editor reply" (Galton, 1907 c) deals this number and its surprising accuracy.

The Galton story shows that people trying to guess the correct weight, when combined in a yet unspecified manner and estimating something with which they can be expected to be quite familiar, can do quite well. However, Krause (2010) demonstrates that crowds can be very wrong. Krause's experiment was to ask visitors at a biometrics exhibition in Berlin two questions: the first how many marbles are in a jar and the second was "how many times a coin needs to be tossed for the probability that the coin shows heads on all occasions to be roughly as

small as that of winning the German lotto" (Krause, Ruxton, & Krause, 2010, p. 29). The group, as expected, was close to the number of marbles (within 1.5% percent[5]); however, on the second question, remember this is a group at a biometrics exhibition, the group median was off the 400% and the mean was off by almost 2,000% (Correct response = 24, median = 100, mean = 498.30, n = 1,953)[6]. However, these statistics were obtained from data that was already censored: Krause's data collection mechanism had minimum and maximum allowed amounts and in 4.2% of the cases the estimate of the number of marbles exceeded those limits and the participant was told to reenter a number between 40 and 1,500 (correct value 562) and 6.3% of the time the participants entered a number that exceeded the limits for the coin toss estimate and they were told to reenter a number between 2 and 9,999 (correct value 24) (Krause, James, Faria, Ruxton, & Krause, 2011).[7] Including these extreme values would have moved the means and even the medians if we assume that were mostly too large, which is implied in the article but not explicitly stated.

---

[5] 1.5% would be within 18 pounds of ox in Galton's example which is a far larger error rate that was observed by Galton's measures of centrality – either the median, Galton's choice measure, or the averages of the different centiles. But whereas as Galton (1907c) points out there were experts in estimating the dressed out weight of an ox in Plymouth, there probably was not an equal abundance of experts in estimating the number of marbles in the jar at the German exhibition.

[6] The German Lotto is based on choosing 6 numbers out of 49 with order not mattering. There are 13,983,816 different combinations of possible winning tickets (49*48*47*46*45*44/6*5*4*3*2). 23 consecutive heads have a 1/8,388,608 probability and 24 consecutive heads have a 1/16,777,216 probability. (This, as does the Krause paper, omits the 'super' ball option which then adds another digit that must be matched that ranges from 0 to 9.)

[7] Krause's experiment also highlighted another problem with wisdom of the crowds: Participants entered their estimates on a computer and after they had entered their estimate they were shown a screen showing the distribution of prior estimates and were allowed to enter another estimate. If one was not entered within 120 seconds it was assumed that the participant did not reenter an estimate. If another participant starts within 30 seconds of the second estimate being entered, they assume that the same person is just repeating the experiment and they filtered out those entries. In their final data there were 6,568 responses to the number of marbles question and 4,511 of them were filtered out, a 69% rejection rate. For the coin toss estimate, there were 6,266 estimates and 4,313 of them filtered out, again a 69% rejection rate. This seems like a fairly high rejection rate for a process that is being assumed to work without a controlling authority. Galton, in comparison, eliminated very few estimates due to their illegibility.

Lorenz (2011) in a study asked 144 Swiss people questions about Switzerland. In two questions the arithmetic mean was off by ~1,300%, the next question ~300%, the next two ~150%, and the last question 60%. For example, "What is the population density of Switzerland in inhabitants per square kilometer" had true value of 184, but the group mean was 2,644 (+1,337%) and "How many assaults were officially registered in Switzerland in 2006" has a true value of 9,272 and the group mean was 135,501 (+1,156%). The raw data for the Lorenz study is available from PNAS and for the first question, one respondent started with an estimate of 250,000 for the population density, more than 100 times larger than the next value. In the assaults question, the highest initial response was 700,000 which was only slightly more than twice the value of next highest response. The actual value was ~10,000 so they were quite off. Both of these examples would probably have been censored in the Krause study, so the means reported in that study could well have approached those reported in the Lorenz study.

In the Lorenz study, the authors note that some entries were removed from the analysis due to their extreme values, which they attributed to misunderstanding of the question or deliberately trying to undermine the experiment. These examples are independent of the one in which the author's believed that the participant confused meters and kilometers (Lorenz, Rauhut, Schweitzer, & Helbing, 2011, p. 9024) in which his estimate was 400,000,000 for "What is the length of the border between Switzerland and Italy in kilometers?" and the actual true number was 734 for an error of 54,495%. The Lorenz study also uses the logarithm of the estimates in computing a group mean (the geometric mean) assuming that the individuals had a problem with what was the appropriate scale for the reply. One could imagine that the number of marbles in a jar would, in general, have a smaller relative variance than would number of grains of rice or

grains of sand in the same container attributable both to not having experience with that number of items and thereby not being sure of the appropriate scale.

Krause's finding is reminiscent of Kahneman and Tversky's "Belief in the Law of Small Numbers" (Tversky & Kahneman, 1971) where they experimented on a group attending conferences of the Mathematical Psychology Group and the American Psychological Association and found that those respondents' intuitions about statistical issues were quite incorrect. Additionally, Krause et al noted that since the participants had to pay € 4 to enter the exhibition, they "will be inclined to take the exhibits seriously." This is similar to Galton's comments about the fee required to enter an estimate of the ox's weight removing guesses on a lark.

Condorcet basically agrees with the common interpretation of the Galton story but adds the restriction that people judge correctly at least minimally above random level when grouped together will reach a correct judgment. Although this point is not explicitly addressed by Galton, he does repeatedly say that he would like to know the occupation of each person making an estimate of the dressed out weight; the assumption being that someone related to the industry would have something similar to Condorcet's minimally above random chance of guessing a reasonable weight. When Page was asserting that guessing the weight of steer was not that difficult, he was referring to his experience as a 'cattleman' (having owned 9 head of livestock) to justify his claim.

Smith's 'invisible hand' is quite different; in this case people trying individually to accomplish one task turn out to efficiently accomplish an entirely different and very important task. Malthus disagrees with both Condorcet and Smith in that he holds, that in regard at least to procreating, people are incapable of making the correct decision and will inevitably lead to a dystopian future. There is, of course, a lot of political and economic theorizing about how

Smith's 'Invisible hand' does not accomplish its task equally for all members of the society. Hardin (1968) famously painted the same dystopia as Malthus with the driving force being economic self-interest rather than 'passion of the sexes' and there has been a vast literature in opposition (Ostrom, 1990) to Hardin's 'Tragedy of the Commons' as, of course, there is against Malthus himself.

Malthus was very instrumental in the development of Darwin's concept of natural selection. Darwin rather than finding a dystopian future inevitable comes up with natural selection in which the ecological pressures lead to the more efficient extraction and utilization of available resources. Darwin's mechanism is much closer to the "Invisible Hand" in that individuals, finding a niche in which they can thrive, lead to the efficient exploitation of the available resources.

There is no assumption that the individuals are misinterpreting the environment when they act in what turns out to be a self-destructive manner in Malthus' and Hardin's scenarios. They are correctly interpreting what course of action leads to their immediate gains. Thus they would appear on the surface to satisfy Condorcet's correct minimally above 50% of the time requirement. Much like the classical single episode prisoner's dilemma in seeking to maximize their share, they create a situation in which everyone's share is minimized.

1.4 SYSTEMATIC ERRORS AND INFORMATION CASCADES

Caplan (2007) demonstrates that, in the United States, survey after survey has demonstrated that crowds have very biased opinions on various factual items when a correct answer is unequivocal and publically available. In a subsequent publication, Caplan (2009) in surveying the wisdom of

the crowd effect, which he labels the "Miracle of Aggregation", proposes testing the effect in three different manners. The first is to ask a question on a survey for which there is a factually correct answer, much like the probability of getting a winning combination in the German Lotto used by Krause above. Caplan is an economist and his questions tend to be economic: How large a portion of the US Federal budget does foreign aid represent? The true answer is around 1%; a common response (by 41% of the respondents) is that it is the largest item in a list being considered (foreign aid, welfare, interest on the federal debt, defense, Social Security, and health are the items on the list) whereas Social Security selected by approximately 22% of the respondents is the largest budget item in the group. So almost twice as many people misidentify the largest item as correctly identify it (22% vs. 41%). He notes that welfare was the second most frequently placed first item.

His second test bed for analyzing group decision making is to "compare the beliefs of the average citizen to the beliefs of the average well-informed citizen" (Caplan, 2009, p. 5) He attributes this method to Althaus (2003) who has labeled it: the "Enlightened Preference" approach. In this method, general questions are first asked with provide an estimate of how informed the respondent is about the general area of focus; subsequently the opinion questions are asked, the test is then "Are the responses to the opinion questions generally the same between the people with high levels of information on the subject matter and those people with low levels of information on those areas?" Althaus (2003) reports that in his survey addressing political questions there were consistent differences between the two groups. Assuming that the people with higher levels of information were unbiased, the uninformed portion must have been biased since their average response was quite different from the informed responses.

His third method is to compare public opinion with 'expert opinion' and in this arena again public opinion does not commonly coincide with his surrogate for truth. Caplan (2001) presents an alternative to Downs' Rational Ignorance (Downs, 1957) which is an economic analysis of the cost/benefit ratio of information which concludes that there are instances in which it is not efficient to expend scare resources to acquire useless information. However, as Caplan notes in these cases, the people who have chosen to remain without information should not, in general, exhibit similar biases nor have excessive certitude about their opinions which they have rationally chosen not to substantiate with expensive information. What Caplan wants to focus on is those cases in which someone is very certain and exhibits a non-random bias. These instances he labels as Rational Irrationality. He explains the persistence of these 'irrational' beliefs by the low private cost of maintaining those beliefs and other social forces for maintaining them. Were the costs for maintaining the belief to rise above the level of the benefit of maintaining the belief, the belief would be abandoned or modified to some form with a lower cost which may represent a permanent change in the belief or just a temporary change until the cost of holding that belief returns to its low cost. In this manner, groups of people can support grossly inefficient policies since the individual cost for belief supporting the policy is very small.

In the theory of information cascades, we find an explanation as to how some groups settle on incorrect decisions even when acting 'rationally rational' (to paraphrase Caplan) but still not wise.

> An informational cascade occurs when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behavior of those ahead of him, to follow the behavior of the preceding individual without regard to his own information. (Bikhchandani, Hirshleifer, & Welch, 1992, p. 992)

A key aspect of an information cascade is that individuals can see the effects of prior decisions or perhaps just the information that the preceding person had prior to making a decision. When a sufficiently large number of prior individuals have made a certain choice or received the same information, a person who believes that his sources of information has some non-negligible possibility of error will ignore his source of information and follow the prior decisions when his private information tells him to act differently. At that point, individuals who are following him will have no access to what his private information suggested and only see the effect of decisions made prior to his decision. His decision is uninformative. In the case where the signal the previous person receives is visible, the information does not get lost and subsequent people having access to all previous signals will be able to make decisions based on more complete information.

When only actions are available, once a person decides to ignore his private information and follow the crowd all following people using the same decision rule will make the same decision. When the signals are available, a person could rationally choose to ignore his personal signal and follow what the then dominate signal is, subsequent people will have that ignored signal available and will be able to self-correct if that ignored signal eventually becomes the dominate signal.

This process could be used to explain other research findings. During the Cold War era, academic research was not immune to the requirements to display appropriate anti-communist credentials. Asch (2003) referring to a 1952 experiment reports that when an individual subject was presented with the task of saying which line in target set matched a given line, the individual would frequently agree with others who had been enlisted to give the wrong answer. As Shiller (2003) explains the common interpretation of this fact was that the subject was giving into social

pressure and going along with the crowd which was wrong, rather than accepting that the probability that his or her information is correct when 6 others see it differently is vanishing small (assuming that the other 6 people aren't in cahoots with each other which turns out to be wrong in this case). Riesman's (1961) *The Lonely Crowd* is another example from the same general era of demonstrating a presumed superiority of individual decision making over using information available from the crowd. In popular literature, of course, there is a plethora of works such as Ayn Rand's 1957 *Atlas Shrugged* (Rand, 2005) in which an information cascade is seen as a personal failure rather than an efficient use of available information. An alternative explanation for this phenomenon could be that it was a reaction to the Nuremburg trials and shows too much deference to authority.

March (1991) presents a simulation that deals with the interrelationship between the mean competence of an organization and its variance when compared to a group of 2, 10, or 100 other organizations. The group members' competence scores are individually distributed IID ~N(0,1)[8]. For each trial, the maximum value of the group represents the group's competence. The question then becomes what level of competence does the individual organization have to achieve in order to compete equally with the group; that is if there are N members in the group what level of competence must the individual organization achieve to be the highest scoring organization $1/(N+1)^{th}$ of the times. If the individual organization's competence was distributed ~N(0,1), identical to all the members of the group, it would have an equal opportunity as any other organization to be the highest scoring organization for any trial. The issue then, for March, is, when the individual organization's competence's variance changes, how its mean competence

[8] Statistical distributions within this dissertation will be identified as ~U (Uniform) or ~N (Normal) followed by the bounds for a Uniform distribution and mean and standard deviation for Normal. ~N(0,1) is a normally distributed (pseudo-) random variable with a mean of 0 and a standard deviation of 1, a z-score. ~U(0,1) is a uniformly distributed (pseudo-) random number between, but not including, zero and one.

score must change in order to remain equally likely to be the top scoring group. In an organization, if the members' diversity changes, which is assumed to change the variance of the organization's competence, how does their (the members of the organization) average level of competence need to change in order for the organization to remain equally competitive?

A simulation of 2, 10, and 100 member groups was run with 5,000 trials each. The maximum score for each trial was sorted and the $1/(N+1)^{th}$ maximum score identified. This is the score that the individual organization had to achieve to be the highest scoring organization $1/(N+1)^{th}$ of the times. For example, if there were 10 organizations, the maximum scores were 1 through 10, then if an organization had a score greater than 1 it would expect to have the highest score at least $1/10^{th}$ of the times. The individual organization is not trying to be better than the highest values of the group scores $1/(N+1)^{th}$ of the times, it is merely trying to be higher than lowest $1/(N+1)^{th}$ of the values.

What this makes clear is that when taking the best score of a group of N organizations with a ~N(0,1) distribution a very large number of times, there will be a large number of not very good scores. When a simulation of 100,000 groups of 10 members with an effective score distributed ~N(0,1) is run, the average value of the maximum score is 1.54 which leaves approximately 6% in the right tail of a normal distribution. However a histogram of the maximum scores is shown below, quartiles are 25% = 1.13, 50% = 1.500, 75% = 1.91 with the minimum being -0.429 and the maximum being 4.52.

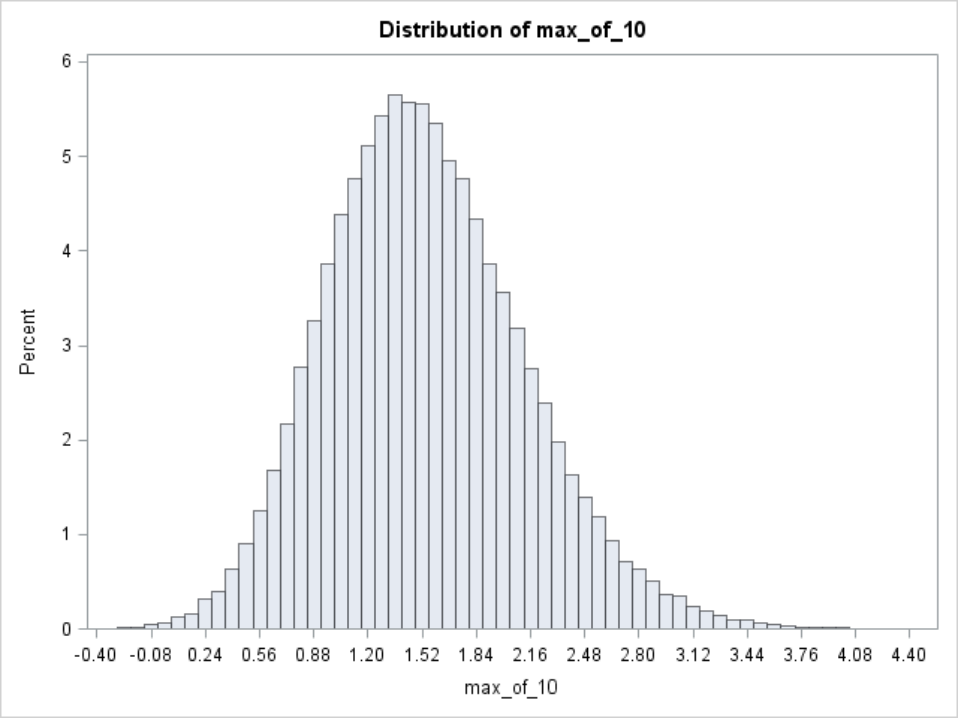

Figure 1 Histogram of Maximum Scores

So although on average the group is better than its individual members. The individual members' scores are distributed ~N(0,1) so their average score is 0, rather than 1.54 the group demonstrates. The group's maximum score is less than 1.335 more than 25% of the time. So to insure that our reference group has the best score $1/11^{th}$ of the times, it merely has to always do better than the bottom 11^{th} of the distribution. Which when it's variance is 1 means that its mean need only be .795 – 1.335, or -.54. So an individual organization with an average competence of more than ½ a standard deviation below the average will still score higher than the group in proportion to the population.

When the size of the group is increased to 100, the range of the maximum values goes from 1.09 to 5.28 with quartiles being 2.2 (25%), 2.5 (50%), and 2.8 (75%). The $1/11^{th}$ of the

population value associated with a group of 100 is 1.98[9] rather than the .795 which is the value associated with a group of 10. So with a group of 100, the individual wanting to be represented in equal proportion to the population with a variance of 1, must have a mean of 1.98 – 1.335 or .648 which is in the 74th percentile as opposed to the -.54 which was only in the 25th percentile.

It is not that 'Sometimes groups make bad decisions' as much as it is "On average groups of size 10 make decisions that would be better made by an individual with an expected competence level of ½ standard deviation lower 9.09% of the time." As the group size gets larger that percentage approaches zero one would expect from the Condorcet Jury Theorem, although the Jury Theorem is based on yes/no voting decision process rather than on a maximum value group decision process. Incidentally, this is not the point that March is making when he presents this analysis. He is interested in showing that increasing the variance of competence is a more effective way of dealing with competition against a large group than trying to increase the mean competence.

1.5 DECISIONS IN GROUPS: HIDDEN PROFILES

There is a substantial literature on group decision making (if not the entire discipline of Social Psychology) and one the subareas of most relevance to the wisdom of the crowd effect is called "hidden profile" (Strasser & Titus, 1985). In this paper, the Strasser and Titus discuss their study

[9] With 5,000 simulation of a group of 100, the average maximum value is 1.98 with a standard deviation of 0.008 and a maximum value of 2.01 and a minimum value of 1.95. March's Figure 6 points to something around 1.7 for this value. The numbers for the group of 10 were a mean of .795 (sd=0.12, max=.842, min=.750) the .795 was where March's Figure 6 was pointing to for a group of 10. The numbers for a group of two are the mean equals -.52 (sd=.019, max=-.45, and min=-.59). Again this value is does not match March's value of approximately .22 for a group of 2.

in which a group of individuals was provided with some information that was shared among all members of the group and some information that was uniquely provided to each individual member of the group in a simulated selection process. Based on only the shared information, one candidate should have been preferred; however, when all the information was available a different candidate should have been preferred. The point of the experiment being would all the available information be used in coming to a group decision or not.

The group discussion tended to focus only on the shared information and especially that information which supported their 'pre-meeting' choice of the first candidate. When the information was distributed among group members, the "correct" solution was only found 18% of time; whereas when all group members start out having sufficient information to detect the best solution, the "correct" solution is found 83% of the time. Note the almost identical 'random' noise components to these two statistics: the 18% that found the right answer is almost exactly the 17% that chose a non-"correct" option when provided with all the relevant information. The 18% that found the "correct" response is interpreted as just random chance. Strasser does not investigate the similarity in these two error rates in his paper.

The participants in this study were to choose between two political candidates in a 'caucus' type encounter. Strasser found that the groups tended to settle on the candidate that a plurality favored prior to the meeting. Discussions tended to be based on the information that supported their pre-meeting opinions rather than exploring the contributions that other group members' new sources of information could make. The discussion was used as a platform to verify a previous opinion rather than gather whatever information was available and make the best decision based on all the available information. This effect has also been studied as confirmation bias where people tend to seek information that supports their hypothesis or

overweight confirmatory evidence and underweight evidence that conflicts with their hypotheses (Lehner, Adelman, Cheikes, & Brown, 2008).

Wason (1960) demonstrated the confirmation bias effect with his experiment in which the numbers '2-4-6' were presented to subjects who then tried to guess the rule that was being followed by the experimenter to come up with the set of numbers. The participants propose any set of numbers and ask the experimenter if those numbers followed the rule or not. The results were that the subjects made up a hypothesis in their heads and only asked about sets of numbers that would confirm their hypothesis. The philosophical basis for this is, of course, Popper's falsification principle. Six of the 29 subjects in this study correctly identified the rule on their first guess; these subjects tended to propose more sets of numbers that did not match the rule.

The thrust of hidden profiles argument is that different group members with different sources of information when pooled will tell a partial story which when combined with other partial stories will be compelling enough to induce the group members to select the "correct" solution, rather than the some other mechanism to combine the members' preferences such as voting or averaging. Galton's theory was that the median was best way to combine group's opinions into a common opinion in which no individual member has a larger impact that any other member (i.e., the mean of a very highly skewed distribution is very sensitive to the absolute value of the most skewed value within the distribution.) Condorcet's theory was that given that the individual members are better than random judges, the average of a sufficiently large number of them will tend toward the 'correct' response since the individuals' error terms will tend to cancel each other out, since they are independent. All of these are different from the Miracle of Aggregation that Caplan refers to, which is that some high percentage of the groups are totally uninformed and have random opinions that are not biased and only a very small

percentage of ‘knowledgeable’ members will be able to insure that the group selects the ‘correct’ response. Of course, Caplan’s thesis is that the uninformed masses are not merely uninformed but actively misinformed so that their misinformed opinion is so biased that a small proportion of informed member won’t be able to keep the group from making an incorrect decision.

There is a substantial body of research reinforcing the finding that groups do not efficiently uncover hidden profiles. Lu et al (2012) conducted a meta-study of 65 studies ‘using the hidden profile paradigm’. The results of this meta-study showed

> (a) groups mentioned two standard deviations more pieces of common information than unique information; (b) hidden profile groups were eight times less likely to find the solution than were groups having full information; (c) two measures of information pooling, including the percentage of unique information mentioned out of total available information (the *information coverage* measure) and the percentage of unique information out of total discussion (the *discussion focus* measure), were positively related to decision quality, but the effect of information coverage was stronger than that of discussion focus; and communication medium did not affect (d) unique information pooling or (e) group decision quality. (Lu, Yuan, & Lauretta, 2012, p. 54)

They also tested to see if there was an effect due to the meetings being computer mediated or face-to-face and found no significant difference in either the amount of information pooled or in the rate of the group coming up with the correct solution (the communication medium mentioned in the above quoted text). Now these results do not deny that there are wisdom of the crowd effects but suggest that it not caused by pooling of information not available to all members of the crowd. In other words, the “hidden profile” is not uncovered just because it is available to some combination of members of the group.

Kerr (1996) in comparing judgment bias in groups and individuals identifies several systematic biases evident in group decision making: Commission – using information not relevant to the question; Omission – failing to use information diagnostic to the question; and

Imprecision – reaching a different conclusion than a normative model would propose (uses the right facts but comes to the wrong conclusion). The failure to use the uncovered information to find the 'correct' response is not unique to a 'hidden profile' scenario, as is evident in Strasser's 17% of the groups that chose a non-'correct' response even when everyone always had all the relevant information is an example of omission in Kerr's analysis of failures in group decision making processes.

Mojzisch et al (2010a) tried to separate out dysfunctional group processes from other issues in identifying the causes for the failure of groups to effectively use all the information available. In their study, each person was given some unique piece of information and all members shared other pieces of information. They asked the participants to write down the information they remembered after reading the scenarios and combined these into a single document and returned that document to each individual telling them that the document represented what they remembered and what all other group members remembered and then asked them to make a decision. Since the individuals never actually interacted with the other members of the group, any potential dysfunctional group interaction was (theoretically) eliminated. These groups performed better than groups which came to a decision via a group discussion. They still failed to use all the available information in a significant number of cases. Their conclusion was that the group members did not fail to pool their information as much as they failed to adjust their opinion from that which they held based on the information they originally received to a more informed opinion that could have been supported by the evidence presented during the group discussion or the presentation of the group information document.

Putnam (2007) clearly demonstrates that in the short run the principle of effect in racial diversity and, by extension, identify diversity is a decrease in social capital which he defines

simply as "social networks and the associated norms of reciprocity and trustworthiness" (Putnam, 2007, p. 137). He shows that not only is there less trust among the different groups (be that difference racial or otherwise), there is also less trust within one's own group. However, he still pushes an agenda that diversity is beneficial even in the face of temporarily decreased social capital. This decrease clearly leads to less information being shared and a general lowering of the quality of decision making based merely on the diversity of the group. Putnam, dealing mainly with immigration, shows how 1st and 2sd generation immigrants are vastly over-represented in many categories (such as Noble Prize winners) and therefore although their diversity may be lowering the quality of group decisions, their individual ability contributes to those decisions. Additionally, he believes that the decrease in social capital is a short term phenomena and that with more exposure that effect is ameliorated.

Wisdom of the crowds' skeptics are not limited to criticizing the mechanisms by which the decisions are reached. Gigone and Hastie point out in a meta-analysis of several group decision making articles that the nature of the task the group is focused is an important element in outcome. Note that just because a task is a "eureka" task doesn't mean that the "eureka" moment won't be incorrect. Incorrect results can be convincingly demonstrated when incorrect assumptions are accepted.

> Groups performing "eureka" tasks (…), tasks with demonstrable solutions, tend to outperform their average members and approach the performance of their best members. When one or two group members can demonstrate or effectively justify the correct answer to the rest of the members, the group will usually make a correct judgment. On the other hand, groups performing tasks that involve solutions that are not easily demonstrable tend to perform at the level of their average members. Thus, the accuracy of group judgment depends greatly on the nature of the judgment task. (Gigone & Hastie, 1997, p. 147)

The Lorenz study, mentioned above with respect to the size of errors with some group processes, focuses on the effect of social influence on wisdom of the crowd. The article's title is "How social influence can undermine the wisdom of crowd effect" (Lorenz, Rauhut, Schweitzer, & Helbing, 2011). They created 12 groups of 12 participants each with three group information sharing arrangements. In the first, there was no information shared among the group members, in the second, group members received the arithmetical mean of the other group's members, and in the third, the "subjects received a figure of the trajectories of all subjects' estimates from all previous rounds." (Lorenz, Rauhut, Schweitzer, & Helbing, 2011, p. 9021).

Each participant was first presented with the question and provided his/her estimate, then the participants received information about the other group members' estimates if they were in the second two conditions and asked to make a new estimate, this process was repeated until each participant had made 5 estimates. Participants were in a cubical working on a computer screen so they had no direct interaction with the other group members; however, since all participants were recruited from the same school (total of 6,000 students) they may have known some of the members of their group. The study design required that the estimates happen simultaneously in order to update the information after each estimate, so the entire group, probably, meet in a common area before being assigned to individual cubicles, giving the participants an opportunity of knowing that someone they knew was in their group. Additionally after the first and last estimates, participants were asked to estimate their confidence with their estimates. In only 21.3% of the cases was the arithmetic mean of the group's estimates after the fifth estimate better than the first estimate. When the arithmetic mean was replaced with the arithmetic mean of the logarithm of the ratio of the estimate to its true value (resultant being % of true value), the value jumps to 77.1%.

They demonstrate that the social influence effect is to lower the variance of the estimates: in fact the group members' estimates do converge with information about the estimates of other members of the group. However, this decrease in diversity of estimates is not accompanied by an increase in the accuracy of the group mean. As the range of the estimates decrease the true value lies at the less centrally located estimates. This leads to their third finding which is that there is greater confidence after the range has been decreased by the social interaction which is unwarranted by the actual value of the group's mean estimate. This experiment does not claim that there is no wisdom of the crowd but only that social interaction diminishes its accuracy.

1.6 DECISIONS IN GROUPS: MARKET MECHANISMS

Adam Smith's 'Invisible Hand' is not only a reification of market forces; it also idealizes them in a libertarian sense. There are many instances in which market forces fail and governmental regulation are necessary; public goods being the classic example. The line between where the "Invisible" hand of the market and the "big stick" of government regulation best function is not clear even in the most ideologically orthodox circles. "The government that governs least governs best" has not defined 'least' and does not claim that the government that "governs not governs best".

The 'invisible hand' operates in an environment of private property and 'private property', itself, is only what the authority of the state says it is. However, begging the question of the political implications of the "Invisible Hand", a similar force is enrolled in wisdom of the crowd problem solving strategies.

Emile Servan-Schreiber, a CMU PhD in Cognitive Psychology, was the founder and CEO of NewsFutures (2000-2010) a prediction market company. He believes

> "Rather than the particular trading mechanism used, the ultimate driver of accuracy [in prediction markets] seems to be the betting proposition itself: on the one hand, a wager attracts contrarians, which enhances the diversity of opinions that can be aggregated. On the other hand, the mere prospect of reward and loss promotes more objective, less passionate thinking, thereby enhancing the quality of the opinions that can be aggregated." (Servan-Schreiber, 2012b, p. 1)

The 'a wager attracts contrarians' refers to the fact that to maximize one's profits in a stock market type market place, it is not only important to buy or sell the correct stock, it is important to do it before others do and decrease the profit potential. But then others have to follow the contrarian in order for the contrarian to realize his/her expected profits. If other buyers follow the contrarian and buy more of the 'stock' the price will rise and the value of the 'stock' increases. If others sell the 'stock' the price will fall, and if the contrarian shorted the sale, s/he will be able to buy the stock back at a lower price that s/he received when it was previously sold thereby creating a net profit. A contrarian who buys a 'stock' that continues to fall into bankruptcy or 'shorts' a stock that continues to increase in value, is being diverse, but as Scott Page points out being diverse isn't sufficient: "Being diverse in a relevant way often proves hard. Being diverse and irrelevant is easy." (Page, 2007, p. 679).

The act of buying or selling an asset sends a signal to the market; however, exactly what the signal is isn't perfectly clear. Servan-Schreiber points out that when an agent buys a contract at 65 cents; all the market can reasonably be sure of is that the agent thinks it will be worth more (or at least not less than that much). The agent may actually believe that the "true" value of the contract is less than 65 cents but that other agents in the market place will pay more than 65 cents for it at a later time period. However, even in this case, the agent believes that s/he will

receive more than 65 cents for the contract at some later date. In a cash based market in which the currency is experiencing a high inflation rate, the actual 'value' being considered could be not nominal cash value since the cash standard is changing with the inflation rate, but some other 'inflation-adjusted' measure of value. So the person could buy the 65 cent contract thinking it will only be worth 60 cents later, but that all other alternative wealth holding options would even decrease more.

For the agent to make money in the market, s/he must believe that the current price for some asset is either too high or too low. Not only must s/he believe that, s/he must act on that belief and make a trade before too many other agents come to the same conclusion and drive the price of the asset to a more 'correct' level. Thus the agent must act quickly based on private opinion and must act truthfully. If s/he thought the price was going to fall, s/he needs to sell the contract (either from assets held or shorted). The act of selling the contract reflects her belief that the price is going to be lower in the future. This, of course, assumes that one agent does not have sufficient resources to corner the market.

Servan-Schreiber sums up predictions markets with

> Whichever design is chosen [the different pay out structures for prediction markets], the prediction market performs three tasks: provide incentives for research and knowledge discovery (the more informed you are the more you can profit), provide incentives for timely and truthful revelation (the sooner you act on your information, the more profit you can make), and provide an algorithm for aggregating opinions (into the trading price). (Servan-Schreiber, 2012a, p. 6)

Berg, from the Henry B. Tippie College of Business at the University of Iowa – home of the Iowa Electronic Markets, presents the theory of "marginal trader" to explain the efficiency of the electronic markets to predict future events (Berg, Forsythe, Nelson, & Rietz, 2008). In this review of political markets, she notes that although many traders clearly demonstrate biases

towards their favorite candidate, there is a group of traders who are more efficient traders and have a larger impact on the market. "Marginal" in this case does not mean irrelevant which could be sometimes be used as a synonym for the word. "Marginal" is used in the economic sense of the appropriate critical value, as in setting the price to the marginal cost of a unit: being how much did the last unit cost to produce which is not the average production price. The "marginal trader" is also called the "market maker" since their transactions are generally set the price where other traders just buy or sell at whatever is the current market price. "Marginal traders, not average traders, drive market prices and, therefore, predictions." (Berg, Forsythe, Nelson, & Rietz, 2008).

Surowiecki disagrees with this understanding of the mechanism at work in electronic prediction markets.

> The idea of the "marginal investor" is also invoked by many economists to explain why financial markets are relatively efficient. It is an intuitively appealing concept, because it allows us to retain our faith that a few smart people have the right answers while still allowing the market to work. But it's a myth. There is no marginal investor in the sense of a single investor (or small group of investors) who determines the prices that all investors buy and sell at. (Surowiecki, 2005, p. 4312)

Plott (2002), from Cal Tech where Page started his work on decision making in crowds, uses market based information processing as the basis for his "Information Aggregation Mechanism". The role of the Information Aggregation Mechanism is to aggregate information not to price some asset. This mechanism should function like a vacuum sweeper and collect all sources of information which are held by a variety of people and situations. He notes that there may be several organizational forces which prevent the efficient pooling of information from different sources. Similar to the 'hidden profiles' literature, just having the information available does not insure that it is effectively used to reach an optimal decision. In this instance, "The

mechanisms are supposed to aggregate information that is there and not create it from nothing. If the participants know nothing, the mechanism will produce nothing." (Plott & Chen, 2002, p. 4). This research effort was directed at implementing an "Information Aggregation Mechanism" (IAM) at Hewlett Packard; although similar to the Iowa Electronic Markets, the authors identified major differences. The Iowa Electronic Markets (IEM) focus on predicting public events, the results of elections, no one assumes that the participants in the IEM have insider, private information that leads them to buy and sells contracts as they do. Whereas in the Hewlett Packard IAM, they were selecting a few people to participate in the experiment because of the information they were expected to possess or have access to. Additionally they were selected from different parts of the business because the different parts of the business were expected to have access to different sources of information.

Sunstein (2007), who prefers prediction markets to deliberating groups since they are subject to less of negative influences that cause groups to make incorrect judgments, uses the selection of John Roberts to the Supreme Court by President Bush of an example of how prediction markets sometimes fail in predicting events: the primary reason in this case being that the information about who President Bush was going to select as a Supreme Court Justice was not available to those participating in the market. This echoes Plott's statement, from above, that the prediction markets aggregate information not create it (Plott & Chen, 2002). In *Infotopia*, Sunstein also writes about the market predicting a 65% chance of the Special Prosecutor, Patrick Fitzgerald, indicting Karl Rove over the disclosure of the CIA agent's name. "This is the most fundamental limitation of prediction markets: They cannot work well unless investors have dispersed information that can be aggregated." (Sunstein C. , 2006, p. 136)

2.0 MECHANISMS

2.1 ENSEMBLE LEARNING

Ensemble learning is a machine learning term describing the process of combining a collection of machine learning algorithms in order to obtain a better estimate of some unknown quantity. Specifically its main aim is to improve estimation by decreasing the variance of the algorithms, with the implicit assumption that the algorithms are unbiased and better estimators than a random function would be (Polikar, 2012).

If there was a source of perfect information (and, equally important, all interested parties interpreted the output of that source in a manner that agreed to its perfection), there would be no need to seek the wisdom of crowds or create an ensemble of estimators. The oracle at Delphi or perhaps Google could be asked any question and its response completely trusted to provide the final word on the issue. However, the problem with the oracle at Delphi was that its responses were ambiguous, to say the least; therefore it was the interpretations given by the oracle's priests that were considered not always to be correct. IBM's Watson still has a ways to go before it is accepted as the all-purpose question answerer. Even HAL from *2001: A Space Odyssey* made judgment errors and Dave had to consult other experts to see what course of action to take.

Polikar (2006) specifies several reasons for using ensemble based systems: decreased sensitivity to over-training of the estimators, too large of a data set to be handled by one estimator (his example is a process that produces 10 gigabytes of data for every 100 kilometers of pipeline in a system with over 2 million kilometers of pipeline), the opposite condition of too little data to train an estimator, too complicated of a classification problem for a single classifier to solve, data fusion cases where data with heterogeneous features need to be combined (for example, from the medical domain MRI, EEG, blood test, etc.).

The main goal in ensemble learning is to create a process whereby the responses obtained exhibit less variance from the correct response than a single response, assuming of course that the mechanism that creates the responses is unbiased and is more accurate than a random process. If the algorithm always provides the same response it is biased, much like the mantras of the different political parties, the variance in its response would be minimized.

Table 2 shows an example with the fixed response representing an extreme value (1) and the average value of the sample space (3), with the variance of the error term remaining the same as that of the sample space at 2.5. However, the variance in its response with respect to the correct response would just be equal to the variance associated with solution space to the original question. So neither the biased response of an extreme value nor the unbiased response of the expected value decreased the variance from that of the original response space.

Table 2 Invariability of Variances of Fixed Responses Example

	Value of interest	Error from Fixed Response	Error From Fixed Response average	Difference from Column Average
Fixed Value		1	2.5	
Case 1	1	0	-1.5	2
Case 2	2	1	-0.5	1
Case 3	3	2	+0.5	0
Case 4	4	3	1.5	1
Case 5	5	4	2.5	2
Column Average		2	0.5	
Variance	2.5	2.5	2.5	

By combining the responses in some yet unspecified manner, the variance of the responses with the correct answer can be decreased. Note, however, that, of course, this decreasing of the variance does not assert that the 'combined' estimate is better than the best estimate of the individuals. Thus if for the original question, we knew which individual estimator was the best, we would use its output and not intentionally dilute it with other estimators we knew to be inferior.

However, this 'best' estimator has to be at least as good as every other estimator for every trial; otherwise there would be cases in which it wouldn't be 'best' and we would be better off combining it with other estimates which may be better for one particular set of input values. However, the worst estimator is not just eliminated in a similar fashion, since in a binary estimation problem, the estimator that is wrong 100% of the time, provides us with a prefect estimator and by inverting its estimation we have the correct estimation. If it is a truly random

estimator that does not provide any useful information, it would be eliminated from the ensemble, even though, in the random binary example, it is, on average, correct 50% of the time.

Even in those cases in which our 'best' estimator is only the best one in some known proportion of the time, if we know beforehand that the question under consideration is one for which it is the best estimator, we can just use its response and know that we have the best response. In this case, we have merely partitioned our problem space into two subspaces: one is the problem space for which we have an estimator that provides the best response and one for which it doesn't.

For a regression problem in which we are estimating a 'continuous' value, the best estimator doesn't have to be perfect, it just needs to always predict a value closer than any other estimator. Depending on the problem, a regression based estimator also may need to always be one side of the correct response or the other. For example, it may need to predict the value up to, without being larger than, some target amount. For example, if we were using the estimator to predict the time that some event will happen and we need to take some action before it happens but as close to that time as possible, we need to have our estimator predict a time before the event, or predicting the value of a dosage that achieves it effect without being fatal to the patient. For a classification problem, where the estimator is assigning a membership in a class and the classes are not ordinal, the estimator needs to always assign the case into the correct class whenever any other estimator assigns the case into the correct class, in order to be the best estimator. The best estimator remains the best even when it makes errors, if all other estimators make the same errors plus some additional ones.

The mechanism that produces the average estimate being more correct than the estimate of the average estimator is that when the different estimators have errors that put them on

opposite sides of the true value; that is to say when their error terms are negatively correlated the average of their estimates is better than the estimate of their average estimator. For example, assume our ensemble includes two estimators and the true value they are trying to estimate is 10, Classifier A's estimate is 6 and Classifier B's estimate is 12. A has an absolute error of 4 and B has an absolute error of 2 for an average error between them of 3. The average of Classifier A's and Classifier B's estimates is 9 ((12 + 6) / 2) which has an absolute error of 1. Therefore the average estimate of the two classifiers, the average Classifier, produced an error of 1 and the average estimate produced an error three times as large. This effect is determined by two properties: 1) the estimates of the different classifiers bracket the true value – some are larger than the true value and some are smaller than the true value and 2) the measure of the error used is the absolute value of the difference between the estimate and the true value (although actually any convex function would produce the same result, the most common one being the square of the error term). (Larrick & Soll, 2006) If the estimates didn't bracket the true value then they (the average classifier and the average of the classifications) would be equal.

2.2 SUPERVISED ENSEMBLE LEARNING

In Supervised Learning, we start with a dataset in which the input values (independent variables, features, etc.) and the output values (dependent variable) are known with the assumption that this dataset is drawn from a universe in such a manner that the relationship between the input variables and the output variables in the sample database is the same as that relationship is with other exemplars in the universe under consideration but not included in our dataset. The task is to define an estimating process that uses the input values to generate an output value. There are

many ways this is done with ensemble learning, one method, classification selection, is to partition the input domain into categories and develop estimators that predict well within each category. When actually being used to predict an unknown event, the estimator from the category associated with that particular set of input values could be used to generate the expected value or it could just be more heavily weighted than an estimator which 'specialize' in further away regions of the input space. This is much like having specialized medical practitioners who are experts in a small subfield.

An alternative method, classification fusion, is to take many different samples out of the sample base and create an estimator on the basis of each of those samples and then use each of those estimators to generate estimates when presented with input values for which outputs are unknown. Several different modifications to this method have been developed: Bagging, Random Forrest, and Boosting. There is the additional possibility of using a mixture of algorithms to create the ensemble rather than just using different instantiations of the same algorithm based on different training sets or training the estimator on different subsets of the input variables.

Kuncheva (2004, p. 105) presents a four way break down of the process as: Combination Level – design of different combination mechanisms, Classifier Level – use of different base classifiers, Feature Level – use of different features, and Data Level – use of different data sets. As she notes, some of the decisions at the data level such as bagging and boosting are defined in relation with specific classifiers, so the levels are not completely independent in practice whereas they do deal with different aspects of what needs to be addressed when creating an ensemble.

The purpose of an ensemble in ensemble learning is to get a better estimate than one could get from a single algorithm. If each member of the ensemble always returned the same

value it would not be an improvement over only using one estimator, which is why some people do not choose to get a second medical opinion, believing that the second person is just going to give the same answer as the first. The most common way used to obtain different estimates from the individual estimators is to train the estimators with different data sources or for those estimators that rely on initial values to create different estimators from different initial values. (Polikar, 2006)

In her textbook on combining classifiers in ensemble learning systems, Kuncheva (2004, p. 101) quotes Ho (from Bell Labs)

> Instead of looking for the best set of features and the best classifier, now we look for the best set of classifiers and then the best combination method. One can imagine that very soon we will be looking for the best of combination methods and then the best way to use them all. If we do not take the chance to review the fundamental problems arising from this challenge, we are bound to be driven into such an infinite recurrence, dragging along more and more complicated combination schemes and theories and gradually losing sight of the original problem. (Ho, 2002)

This quote identifies one possible problem with the philosophy of ensemble learning and that is that there does not seem to be a sound theoretical grounding for it. It works, that is not the problem, but why it works and what parts work under what conditions are not addressed as much as some other 'trick' to make it work better. It addresses the engineering problem of making something that works better without addressing the scientific problem of what are the underlying causes and effects; although there is probably some question of the role of a 'mere' technician as opposed to a 'engineer' in the process of confronting a problem. But the entire question from Polikar's initial comments about a perfect source of information and Ho's comments about not going back to the original question and solve it are also reflected in Surowiecki's campaign against experts (c.f. his continual references to the failure of 'chasing the experts'.) Surowiecki

actively believes that there is no single 'best' source for a response, Polikar admits to the possibility of one and that it could/should be used when it exists, and Ho believes that it is worth looking for.

2.3 AVERAGING, BIAS, AND VARIANCE

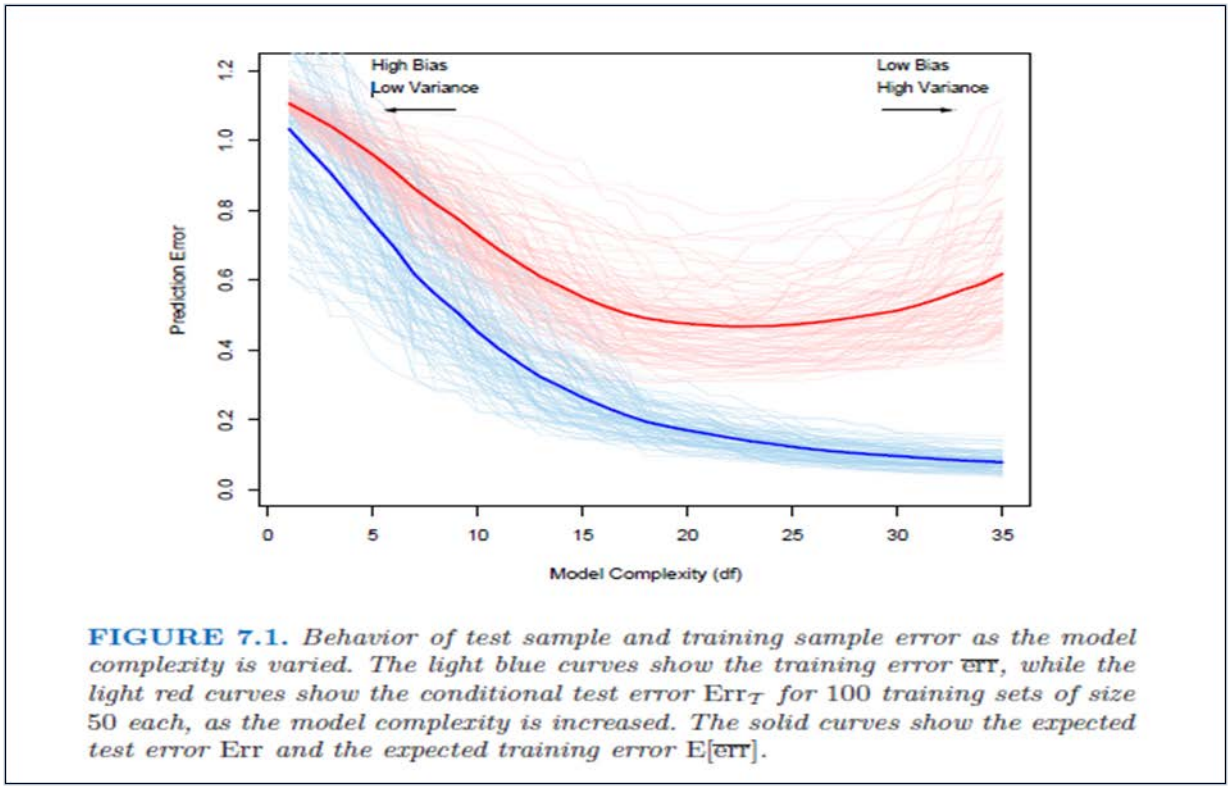

FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Figure 2 Bias Variance tradeoff (Hastie, Tibshirani, & Friedman, 2008)

Figure 2, taken from Hastie's *The Elements of Statistical Learning* (Hastie, Tibshirani, & Friedman, 2008) clearly shows the tradeoff between bias and variance within the training dataset and test dataset and then between those two datasets. Within the training dataset (the blue lines),

as the number of items in the model increases the model's error decreases, the distribution of the light blue lines narrows. The darker heavy blue line is the average of all the light blue lines (in copies without color, the blue lines are the ones on the bottom of the graph; red lines are on the top). The red lines represent the error rates when the model is used to predict data not seen during the training period. Note that whereas the variance decreases as the model becomes more complex with the training data (blue lines); it decreases with the test data when the model becomes simpler. The other important point about the chart is that the lowest error in the test dataset is not with the most complex model, not with the model with the lowest error in the training phase. The model was overtrained.

Larrick and Soll (2006) demonstrate that averaging different opinions is a generally superior method that trying to pick out the single best opinion. Their initial example has two weather forecasters: one predicts a high of 60° and the other 70°, now the actual temperature was 73°. The first was off by 13 and the second was off by 3 (for an average of 8 = (13+3)/2), the average of the two predictions was 65° and that average was off by 8. The first number is the average error and the second is the error associated with the average weather forecaster. The average error will always be equal to or less than the error associated with the average estimator. In the case where the estimates are all either larger than the true value or less than the true value, the average error will equal the error of the average estimator; however, whenever the true value is smaller than at least one of the estimates and larger than at least one of the estimates, the average error will be absolutely larger than the error of the average estimator. In the above case, assume that the actual temperature was 68°. The first estimator's error is 8 and the second estimator's error is 2 for an average error of 5; whereas the average estimator's estimate is 65°

and the average estimator's error is then 3 whereas the average of error of the two estimators is 5.

Mathematically the average estimator's error always be less than or equal to the is caused by using the absolute deviation as the error measurement; if the arithmetic error was used the errors for the two estimators would be 8 and -2 which would lead to an average error of 3 which is the same as the average estimator. Larrick and Soll (2006, p. 112) point out that the mean absolute deviation is not the only measure with this property any convex loss function would have the same property. The most commonly used other convex loss function is the squared deviation, which in this case would lead to (64+4=68/2) 34 for the average of error and 9 for the error of the average estimator: a 9:34 ratio rather than the 3:5 ratio with absolute deviation since the squared deviation obviously weighs the larger errors significantly more than the smaller errors.

Soll and Larrick (2009) present a model of how people should choose between two opinions (and then easily extendable to more than two opinions). In their model, they, which they call the PAR (probability, accuracy, redundancy) model, use the probability of the actor being able to distinguish the most accurate of two estimates, the average accuracy of the estimators, and the redundancy between the two estimators which is a measure of the correlation of the error terms of their estimates. They find that people in general, when faced with one judge with a different estimate from their own estimate, they most frequently stick with their own estimate, and sometimes either take the average of their estimate with the judges or completely accept the judge's estimate. The average weight of 70% own judgment 30% other judge does not reflect a general weighting of 30% other judge and 70% own judgment as much as 100% own judgment (70% of the time) and 100% other judge's estimate (30% of the time). The 30%

figure has often been quoted (Soll & Larrick, 2009, p. 781) as the average weighting that people apply when comparing their judgments to others. However, as they show on each individual item the weights are more frequently 100% one way or the other and then 50% weighting in the remaining cases, rather than generally applying a 70%/30% weighting.

Hong and Page (2012) presents a statistical model of the wisdom of the crowds. Their first theorem is "[Diversity Prediction Theorem] The squared error of the collective prediction equals the average squared error minus the predictive diversity." The collective prediction in this instance is the average of the individual components predictions.

1

$$(\boldsymbol{c}-\boldsymbol{\theta})^{\boldsymbol{2}}=\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum(\boldsymbol{s_i}-\boldsymbol{\theta})^{\boldsymbol{2}}-\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum(\boldsymbol{s_i}-\boldsymbol{c})^{\boldsymbol{2}}$$

$$Predicitive\ Diversity\ \equiv\frac{1}{N}\sum(s_i-c)^2$$

2

$$Average\ Squared\ Error\ \equiv\frac{1}{N}\sum(s_i-\theta)^2$$

3

Where θ is the true value, c is the collective prediction; s_i is i^{th} component estimation of θ over N components. This formula is reminiscent of the decomposition of the sum of squares

from the ANOVA formula where the difference between the individuals and the overall value is equal to sum of the differences from the individuals and their group plus the sum of the difference between the group effect and the overall mean times the number of groups, where the group effect is represented by c in this formula rather than by two or more groups for the items in the analysis.

$$\sum(s_i - \theta)^2 = \sum(s_i - c)^2 + N(c - \theta)^2 \qquad 4$$

The Predictive Diversity is always a positive value, therefore the larger this value the smaller the squared error of the collective and the better the estimate. However, there is no free lunch; the size of the Predictive Diversity influences the value of c which influences the size of the collective's squared error. Predictive diversity contributes to decreasing the total squared error of the prediction when the diversity component is larger than the error in c. Remember that the Predictive diversity can easily be maximized by choosing the most extreme values allowed in the range of the function. This would move c to the mean of the extreme values which may or may not be a reasonable estimate of θ. Their "Corollary 1 (Crowd Beats Averages Law)" (Hong & Page, 2012, p. 61) is what we have already seen in Larrick and Soll (2006) cited above.

Hong and Page's second theorem

> Theorem 2: (Bias-Variance-Covariance (BVC) Decomposition). Given n generated signals with average bias $\bar{b}$, average variance $\bar{V}$, and average covariance $\bar{C}$, the following identity holds:
>
> $$E((c - \theta)^2) = \bar{b}^2 + \frac{1}{N}\bar{V} + \frac{n-1}{N}\bar{C} \qquad 5$$
>
> (Hong & Page, 2012, p. 63)

> Corollary 2 (Large Population Accuracy). Assume that for each individual average bias $\bar{b}$ = 0, average variance $\bar{V}$ is bounded from above and that average covariance $\bar{C}$ is weakly less than zero. As the number of individuals goes to infinity, the expected collective squared error goes to zero.
>
> (Hong & Page, 2012, p. 64)

"Assume that for each individual average bias b (bar) = 0" means that the individual components on average correctly estimate value which is a heroic assumption. The normal explanation for the efficiency of negatively correlated estimators is that some will generally overestimate the true and others will generally underestimate the true value and the average of the two will cancel out each other's errors. If we assume one set of estimators to have an expected value 2 below the true value and the other set of estimators to have an expected value two above, the estimated bias squared would be 4 not zero, however the expected bias of the average of the estimators would be 0. Nevertheless what this equation clearly shows is that negatively correlated estimators decrease the expected error of the combined estimator, to the extent that they don't move the combined estimator away from the true value as mentioned in the previous paragraph.

Hong, Page and Riolo (2012) present an analysis of how incentives influence the components of an ensemble in the predictions they make. In this analysis, they demonstrate that when the rewards are based on market mechanisms or pari-mutuel betting schemes, the aggregate (average) ensemble estimates are more accurate even though the individual estimates are less so. This is the same effect that Servan-Schreiber was referring to when he stated that it was the actual market mechanism that leads to the wisdom of the crowds (Servan-Schreiber, 2012b). These mechanisms reward agents for having the correct information and acting on that information before others, in the case of the market mechanism. The amount of the reward that

the agent receives is a function of how different the truth is from what the common consensus of the truth is. In a market, what the current price of an asset is reflects the common consensus of its value, if its true value is very close to that consensus, it will be hard to make money on it; however, if the value is very different, then an opportunity to make a profit exists. In horse racing, betting on the favorite does not lead to a large payoff, but betting on a long shot does – if it wins.

The horse racing scenario and the stock market play out differently. In the horse racing scenario, if other people come along and bet on the same house after our agent did, the agent's pay off would decrease since the payoff is determined at post-time, at least in horse betting in the United States. But this is because there is an event that establishes the true value of the asset, the race ends and a winner is declared. In the stock market example, the agent who buys a grossly undervalued stock needs other agents to follow him in buying that stock and driving up its price in order to be rewarded for his action. The assumption is that 'the true value of the asset being higher than its current value' means that at some point in the future, its current value will increase towards its 'true' value. When that happens, the agent will be able to capitalize on the difference between the price at which he acquired the asset and the price that some other agent is willing to pay for it.

Krause (2011) performed an experiment in which a group of people were asked to estimate the number of beans in 10 differently shaped jars (n=50). Everyone made their estimates in the same order. Now with 50 estimates of the number of beans in 9 jars, it is possible to obtain weights for each of the 50 people in order to minimize the absolute error between the group's average estimate of the number of beans in each of the nine jars and the actual numbers, a simple linear regression. So with the 10^{th} jar, it is then possible to address the

issue of what do you want to weight the most: the performance of the individuals that is give a higher weight to those people who guessed closest to the true value or group impact of the individuals that is give a higher weight to those people who moved the average closer to the true value. In order words, do you weight the diversity of the individual or the ability? Note that diversity isn't being weighted in a randomly, as Page noted above being diverse and irrelevant is not useful, you may want to give more weight to diversity that is relevant. The results were that the weights based on contribution to the group mean lead to a better group score than weights only based on performance. They go on to state in their paper "However, it is important to keep in mind that while diversity seems to be a necessary condition for swarm intelligence, it is clearly not a sufficient one." (Krause, James, Faria, Ruxton, & Krause, 2011, p. 947) This is also more than mere negative correlation, since an estimate that moves the mean away from the true value and negatively correlated with the group's mean with that value being included would not be highly weighted. Brown (2004) studies the effectiveness of Negative Correlation Learning on ensemble learning. The Negative Correlation Learning used by Brown is technique developed by Liu in his PhD thesis (Liu, 1998).

2.4 OTHER MECHANISMS

Other theories have been forwarded as to the factors in the wisdom of the crowds other than the mathematics of averaging. For example, Woolley, currently at CMU's Tepper School of Business, when working at MIT's Center for Collective Intelligence with Thomas Malone put forward the notion of a "Collective Intelligence" factor in groups (Woolley A. W., 2010). This factor springs from the "general intelligence" factor that has been documented in individuals.

Individual people appear to have a factor labeled "c" which is positively related with diverse cognitive tasks and accounts for up to 50% of the variance observed in the various tasks. She looks for, and finds, a factor she labels "g" that is positively related at approximately the same level as "c" to group performance. It is not related (in a statistically significantly manner) to either the average IQ of the group or to the highest IQ in the group. It is mostly related to the average social sensitivity of the group members, measured by the "Reading the Mind in the Eyes" test[10], and the degree of uniformity in the number of speaking turns taken by each of the group members during the problem solving session (i.e., if one person dominated the discussion the group achieved a lower score on its task that if all members of the group participated more equally). (Engel, 2014)

In a prior study Woolley et al. (2008) presents a study in which she demonstrates that for an analytic team when controlling for both the introduction of task specific experts and instruction in effective group decision process both were important in raising the level of the groups' performance. They found with only the introduction of task specific experts actual group performance decreased from the baseline whereas with only instruction in group decision processes group performance increased. When both elements were added to the baseline group, however, performance increased the most. A principle understanding here is that analysis relies on the application of analytic tools to data and both the analytic tools and access to data are (hopefully) distributed throughout different members of an analytic team. The more the team draws on the experience of the different members of the team, the more diverse will be the

[10] "Reading the Mind in the Eyes" is an advanced test in the theory of mind. It shows a series of photographs, cropped so that only the person's eyes are seen, and the subject then chooses what emotion is being displayed in the photograph. It has high test-retest reliability. (Ferandez-Abascal, 2013) This is related to the Theory of Mind which deals with a person's ability to think about what a different person mental state is.

possible solution set. If the team doesn't have a wide range of experience, that will, of course, limit the benefit of the crowd.

Mojzisch (2010a) comes to a different conclusion in "Process Gains in group decision making". He identifies the typical reasons for the failure of groups to use hidden information to come to a correct decision: shared information is discussed more since more people have access to it and group members enter a group discussion with the goal of negotiating for their solution which is based on partial information. To simulate a group decision process without individual interactions, he had group members write down what they thought were salient points to be considered in the decision to be made. The experimenter gathered these documents and created a summary document which was then passed out to the individuals. This was compared with the normal process of having the group sit around a table and come to a consensus about the decision to be made. Additionally the individual group members' initial decisions were recorded. The individuals had the lowest score followed by the traditional group at the table followed by group that only saw summaries of pooled information. The three differences were statistically significant at the .007 level or below. His experiment also measured the amount of information used in reaching these decisions and found that the groups mediated by the written information notes used more information to reach their decisions than the groups around a table. Shared information was discussed more than hidden information in the groups around the table; however, it wasn't clear if this was just because more people knew the shared information and could speak to it as opposed to some intentional mechanism to reach a consensus.

The main observed difference was that the groups interacting via the written document presented and used more information than the groups around the table. Thus the conclusion of the study was that groups around the table did not pool sufficient information to

make the correct decision. He refers to (Mojzisch & Schulz-Hardt, 2010b) to recommend that group members not make their preferences known before the discussion in order to improve the quality of the discussion and the amount of information presented: to minimize the well-known anchoring effect.

Wittenbaum (1999) believes that sharing mutual information reinforces the social fabric of the group whereas sharing uniquely held information strains those social relations. Thus it is an individual's psychological need to be validated as knowledgeable and competent that determines what information is discussed, rather than what information would lead to a better decision. In a subsequent study (Wittenbaum G. M., 2000), she found that this effect was limited to the least experienced group members. Whereas Woolley found a 'g' factor that accounted for reaching 'good' decisions (the ‘g’ apparently stands group in her study rather than ‘good’), Mojzisch's study removes the interpersonal elements, which incidentally he labels dysfunctional group processes, and finds that the groups make better decisions with they have no direct human to human interaction.

InnoCentive is a company that posts problems and payments that will be made to the individual/group that solves the problem. However, as Page (2007) notes InnoCentive is not doing any type of data aggregation to obtain its wisdom. In Page’s terms, it is using the crowd to find the needle in the hay stack. He refers to Lakhani study of InnoCentive which found that “postings that are solved successfully tend to attract a diverse and differentiated pool of solvers. If a problem attracted a physical chemist, a molecular biologist, and a biophysicist, it was far more likely to be solved that if it attracted only chemists.” (Page, 2007, p. 335) Note, however, that the group in which the diversity exists was self-formed prior to solving the problem. InnoCentive is a market-maker which brings together problems and solvers. It does not organize

individuals into teams to become solvers. In fact, Lakhani's (2008) Harvard Business School case study, InnoCentive.com (A), deals with the issue having the company get into the business of facilitating group processes, which as the case study identifies is quite different from the matching problems with solvers which is their primary business. This type of wisdom of the crowd comes from the crowd recognizing a preferred solution when one is presented; the power comes from having enough solutions presented that one is accepted as the preferred solution by consensus. However, as InnoCentive demonstrates having a large number of groups working on the same problem does not guarantee that one of the groups will find the answer.

Krause includes Wikipedia in this group of pseudo-wisdom of crowd's effects. "…the very purpose of Wikipedia is largely incompatible with the SI [Swarm Intelligence] concept because Wikipedia is meant to represent what is already known about a particular topic, rather than generate new insights." (Krause, Ruxton, & Krause, 2010, p. 32) Although he admits that the writing and editing of the Wikipedia's articles are joint processes, his definition of SI is "two or more individuals independently, or at least partially independently, acquire information and these different packages of information are combined and processed through social interaction, which provides a solution to a cognitive problem in a way that cannot be implemented by isolated individuals." (Krause, Ruxton, & Krause, 2010, p. 29) The Wikipedia model, however, seems more likely to be influenced by Woolley's factors. Anonymity is the exception rather than the norm in Wikipedia articles.

Another pseudo-wisdom of the crowds mechanism can be found in Webb et al. (2000) in a text designed for social science researchers. They want those researchers to expand their primary data sources away from just interviews and questionnaires and therefore present a variety of different data sources and the questions they were used to address. The most famous

of these is the maintenance records at Chicago's Museum of Science and Industry. From the records of the frequency of floor tile replacement, researchers were able to demonstrate what the traffic patterns at the museum were. Their collection of 'unobtrusive measures' include physical traces, archives, and observations. The physical traces (wear and tear on floor tiles) are the most reminiscent of the wisdom of crowds with the paths that foraging insects use to find food sources. These paths are not intentional (for example, the way a bee's wiggle dance could be said to be intentional) but are nevertheless an effective means for information about a rich food source to be transmitted throughout the foraging party. Using Krause's SI metaphor, what is lacking from the unobtrusive measures is the reaching a decision on the basis of the multiple sources of information. The same museum records show that people usually turn right to go into an exhibition hall (at least people in Chicago do so). This might be a carryover effect from driving on the right hand side of the road or it might not, but merely because lots of people do it does not mean that it is a wisdom of crowds' mechanism at work. In the Invisible Hand metaphor, the people make decision to accomplish their own private ends and those micro-motives combine to produce an observable macro-behavior. But there is no similar observable evident in the context of 'unobtrusive measures'.

3.0 BASE SIMULATIONS

3.1 HONG AND PAGE'S SIMULATION

Hong and Page present arguments supporting the power of a group to reach better decisions than individuals. In one of their papers, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers" (Hong & Page, 2004), the main point they address is that when faced with the option of hiring 20 randomly selected individuals from a pool of qualified individuals or the "most qualified" 20 individuals from that pool, there are (fairly robust) conditions under which the randomly selected 20 individuals outperform the 20 highest scoring individuals. Hence the main demonstration of their paper is "diversity trumps ability", under certain conditions.

This is, of course, partially a demonstration of the logical fallacy of composition which infers that a group will exhibit a given trait if some subset of the group's constituent parts exhibits that trait. "All Star" teams are almost never as good as the best teams (composition) and the best teams don't necessarily have the individually best players (decomposition). However, what they demonstrate, in addition to this obvious point, is that the group made up of randomly selected (qualified) agents outperforms the group of the most qualified agents. Using the team analogy, a team of randomly selected players will beat a team of 'All-Stars'. The 'All-Star'

analogy is not optimal in this instance since All-Star teams are created to span the diversity of positions required for a team, if rather than an 'All-Star' team, the group of the highest paid professional players or the players receiving the most votes in some selection process are used, it would be obvious that a limited skill set would be over represented and other necessary skills for a team to function would be entirely lacking. Very rarely are left guards among the highest paid or most popular athletes in an American football league. Additionally, the "under certain conditions" restricts, among other things, that every agent under consideration is capable of performing the task to an acceptable level.

This effect could also be dismissed as merely an example of over-fitting a model to one data set and it should not be surprising that an over-fitted model is outperformed by a more generally robust model. There is a substantial literature in the field of machine learning about over fitting; it is a well understood phenomena. In this case the group of experts all became 'experts' because of their ability to solve a given set of problems and their skill set had been honed to that given set. However, when faced with problems not in that given set, as a group they may not have some critical skill that is needed to address the new problems.

They demonstrate 'diversity trumps ability' with a simulation which is reproduced in this study and then they also support their simulation with a mathematical demonstration. The agents in this simulation have 3 heuristics which they apply in searching for the highest return from a solution space. A real world situation could be that the agent is designing some item for which a test exists which rates the item on an open ended interval of 0 to 100. The agent has three ways to design the required item and applies those ways until s/he gets an item that scores better than the current best item on the test. S/he then continues applying her/his three ways of improving that new and improved design, repeating this process until applying the 3 design techniques does

not create an item that gets a higher score from the test. The score of the final product is a function of what the three methods the agent possesses and the position of the initial starting design.

The world of possible designs in the simulation is represented by a vector placed in a ring formation with the indices increasing in a clock-wise manner with the first element, element 1, being immediately to the right of the last element, element 2000, so that the next element after element 2,000 is the element 1. The elements in the vector are random numbers uniformly distributed from the open interval of 0 to 100. Heuristics, ways to design in the above example, are represented by integers from 1 to 12 without any repeating values; since "order matters", there are 12 * 11 * 10 possible sets of three different heuristics (1,320). The same heuristic cannot be assigned more than once in an agent's set.

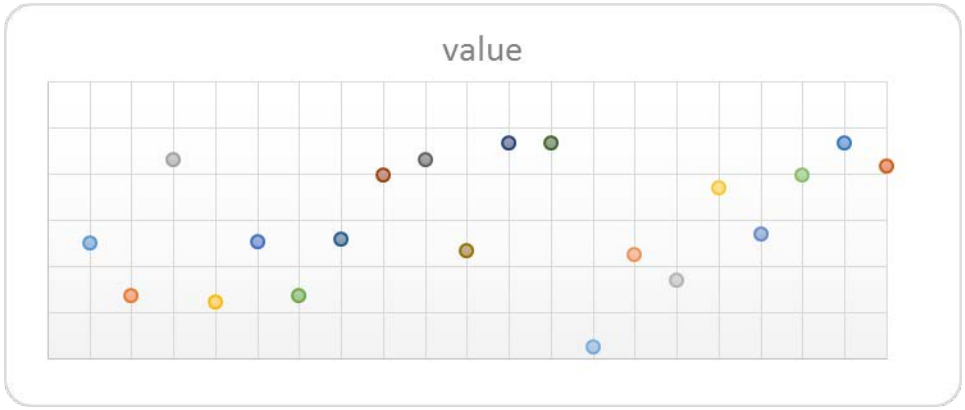

Figure 3 First 20 Values of Design Space

Figure 3 shows the values of the first 20 places in one sample execution of the simulation. Each possible set of heuristics (1,320 different sets) starts at each location in the

solution space (2,000 locations in the ring) and determines the value associated with at set of heuristics in that order when starting at that spot. The first agent of the 1,320 has the heuristic set (1, 2, 3). When it starts at position 1, which has a value of 50.1, it applies its first heuristic which is 1, by adding that value to its current location (1 + 1= 2). It then checks the value of the sample space at position 2, 27.4 in this case, and since it is not larger than its current high value of 50.1 from position 1, it keeps position 1 as the best solution. It then tries heuristic #2, which has a value of 2, 1 + 2 = 3, the value of the design space at position 3 is 86.2, which is higher than the current high value, so position 3 becomes the current best design and 86.2 its value. It then applies the third heuristic, 3, which is added to the current best design position of 3 (3 + 3 = 6), the value at 6 is 27.2. Not higher than 86.2, so position 3 remains the best design. The agent then returns to the 1st heuristic, 1, adds that to the current location, 3 + 1 = 4, whose value is 24.4, smaller than 86.2 so no change, next heuristic #2, 3 + 2 = 5, the value at 5 is 50.9, not larger so no change, and the third has already been applied from this position. So we know that no improvement is possible given the set of three heuristics. Therefore the value for heuristic (1, 2, 3) at position 1 is 86.2 from position 3 and 5 positions were checked to arrive at that value (2, 3, 6, 4, and 5). This process is repeated for each of the 2,000 positions on design space ring and each location on the ring has a value which represents the design with the highest value that was obtained when applying the agent's heuristics from that starting position. The average of these 2,000 values are then averaged together to create an expected value for that agent. This process is then repeated for all 1,320 possible agents. The agents can then be ranked on the basis of this average expected value over the entire design space.

The top ten agents from the simulation are combined into a group and labeled "Experts" and a random group of 10 agents is also selected and labeled "Generalists". There is a slight

chance that an agent could be a member of both groups, which happened in one of the initial executions of the algorithm but not in the one used for demonstration purposes.

There is a second statistic which measures the diversity of the two groups. It is defined in a general form in the Hong and Page's paper (2004), but in the specific example where there are three heuristics for each agent, the value of the diversity metric is 1 if two agents have no heuristics in common in the same position, .66 if they have one, .33 if they have two, and 0 if all three (which is not possible since the pool of agents consists of the 1,320 unique permutations of the integers from 1 to 12). For example, the agent with heuristics (1, 2, 3) when compared to agent (1, 2, 4) would have a diversity metric of .33 between themselves since heuristics 1 and 2 both occupy the 1st and 2sd positions respectively. Both of the above agents would have a diversity index of 1 compared with the agent with heuristic (3, 1, 2); since although they all share several of the same heuristics, none are in the same position. In a sample running from Hong and Page's paper, a group of "Experts" had an average diversity of .72 and the corresponding group of generalists had an average diversity of .92. In the sample used above, the average Expert diversity is .85 and the corresponding average for the Generalists is .94.

In the example used above, the "Experts" in descending order have the following heuristics: (11, 2, 3), (2, 11, 3), (3, 11, 2), (4, 5, 11), (2, 3, 11), (5, 4, 11), (11, 3, 2), (2, 11, 10), (4, 11, 5), and (11, 4, 12). Note that all six versions of 2, 3, 11 are included in the "Experts" group. The other four include three versions of the 4, 5, 11 group and the final one is 2, 11, 10. Clearly, this group of "Experts" shares a group of heuristics to a non-random degree. Table 1, below, shows the distribution of the heuristics for the group of Experts and Generalists. Note that number 1, 6, 7, 8, and 9 never appear in any of the heuristics for the Experts whereas the Generalists only have one value, 2, that never appears. Additionally note that every one of the 10

experts has the number 11 as one of their heuristics. For the generalists, the number 1 appears in 6 of the ten which is its most frequent heuristic.

Table 3 Distribution of Experts/Generalists heuristics

Value	1st Heuristic	2sd Heuristic	3rd Heuristic	Total
1	0 / 2	0 / 3	0 / 1	0 / 6
2	3 / 0	1 / 0	2 / 0	6 / 0
3	1 / 1	2 / 0	2 / 2	5 / 3
4	2 / 0	2 / 1	0 / 0	4 / 1
5	1 / 1	1 / 1	1 / 0	3 / 2
6	0 / 2	0 / 0	0 / 0	0 / 2
7	0 / 2	0 / 0	0 / 1	0 / 3
8	0 / 1	0 / 1	0 / 1	0 / 3
9	0 / 0	0 / 1	0 / 1	0 / 2
10	0 / 1	0 / 1	1 / 2	1 / 4
11	3 / 0	4 / 2	3 / 1	10/ 3
12	0 / 0	0 / 0	1 / 1	1 / 1
Total	10/10	10/10	10/10	30/30

A similar table would not be meaningful for a summary of 100 runs because the clustering of the heuristics to a few digits would be lost when agents from different runs were combined. Therefore Table 2 shows the distribution of the number of times a given number of agents shared a heuristic in a run of 100 trails. The table shows that 38 times, out of the 100 possible, a digit was shared by all ten agents and 25 times 9 agents shared the same digit. The columns labeled Expert Product and Generalists Product are the result of multiplying the number of agents by the frequency; there are 100 trials, with 10 agents in the expert group each of which has 3 heuristics, therefore there are 3,000 heuristics to be accounted for and the two product columns sum to 3,000.

Table 4 Distribution of digits in Experts and Generalists Heuristics in 100 trials

# of agents	Expert Frequency	Expert Product	Generalists Frequency	Generalists Product
1	151	151	226	226
2	119	238	347	694
3	87	261	316	948
4	82	328	161	644
5	83	415	71	355
6	62	372	21	126
7	58	406	1	7
8	28	224	0	0
9	25	225	0	0
10	38	380	0	0
Sum		3000		3000

Table 4 shows the lack of diversity in the distribution of the expert's heuristics. The chi square for the number of agents versus expert-generalist cross tabulation is, of course, quite significant (chi sq= 378, df=9, p<.0001). The values and frequencies for the 38 heuristics that are shared by all ten 'experts' in an expert group are (heuristic value/frequency that value is used) 1/1, 2/1, 3/1, 4/1, 5/0, 6/2, 7/5, 8/5, 9/8, 10/5, 11/3, 12/6. The number, 1, appeared once in each of the ten experts' sets of heuristics, the same for numbers 2, 3, and 4. Five never appeared, 6 appeared twice, etc. There seems to be a clear preference for the higher value heuristics with the heuristics with values 1 to 6 appearing in only 6 cases and the values 7 to 12 appearing in 32 cases. When the simulation is run for a series of 500 cycles rather than 100, the corresponding numbers are 1/11, 2/8, 3/11, 4/11, 5/22, 6/5, 7/18, 8/12, 9/23, 10/19, 11/17, and 12/18 which almost perfectly divides the group by frequency at 6-7 boundary (5 with a frequency of 22 is the exception other all values above 6 have frequencies higher than those below 7) and there were

175 groups of experts in which each of the ten members of the group shared one of the previous numbers, Table 5 Distribution of Heuristic Value when shared by all group members. There is no equivalent set of numbers for the generalists group since there are no generalist groups in which each member of the group shares a digit with all other members of the group. Thus it is clear that the clustering is not only a function of a set of heuristics but also of the numerical value of the heuristic with the higher value heuristics being more likely to be the shared heuristic.

Table 5 Distribution of Heuristic Value when shared by all group members

Values of heuristic shared by all members of the Experts Group	Simulation run of 100 cycles	Simulation run of 500 cycles
1	1	11
2	1	8
3	1	11
4	1	11
5	0	22
6	1	5
7	5	18
8	5	12
9	8	23
10	5	19
11	3	17
12	6	18

Each member of both of these groups has its own set of three heuristics. Hong and Page implement the group's activity as a sequential activity although they state that implementing it as a simultaneous activity did not produce a significantly different story. The simulation continues with each team starting at each point in the ring and computing its value for each starting position. The order in which the team members search appears to be fixed but the order isn't

specified in Hong and Page's paper (i.e. – "the first agent search until she attains a local optimum. The second agent …" (2004, p. 16386) without how the first and the second are defined.) The process is that each agent uses its heuristics to attempt to find a higher value than the one it started with and when it can no longer improve from the value at some location, the agent passes that location on to the next agent in line. When all ten agents in that group have tried and failed to achieve a higher score, that score is deemed to be value of starting at the starting point. As in the initial simulation, each of the 2,000 points in the ring is used as a starting point for each of the two groups, Experts and Generalists. The 2,000 values are then averaged for each of the two groups. This process is run 50 times and an average of the averages is computed. In Hong and Page's paper the average for the Experts is 92.56 (sd: 0.020) and for the Generalists 94.53 (sd: 0.007).

From Table 1, it should be clear that the Generalists with 11 distinct heuristics should do better than the experts who only have 7 distinct heuristics. It isn't however essential since the one heuristic that the Generalists are lacking, heuristic 2 in this case, is possessed by the experts and it could be key. For example, say that the value at position x was 90 and the value at x+2 was 99, if 90 is higher than the preceding 12 positions every agent would end up occupying it and if all positions following were lower except for position x+2 then it would require a heuristic of 2 to land on it.

3.2 JAR OF MARBLES VERSUS REASONING QUESTIONS

Krause (2011) used the data they gathered from the marbles in a jar and coin tosses to match the German Lotto experiment to investigate Hong and Page's findings. In guessing the marbles in a

jar experiment, exemplars from the top 25% of a random grouping were generally better than the rest of their group until the group size got up to 40 individuals from that point on the group mean was better than the 'expert' (the expert being a random person chosen form the best 25% of the people in the group). In the coin toss verse the German Lotto experiment, the exemplar from the top 25% of the group was always better than the group and the error rate for the group got larger the larger the group got.

They then compared an expert from the coin toss/German Lotto experiment with the other participants for the number of marbles experiment. The steps were determined by the size of the absolute error on the coin toss/German Lotto experiment. The correct response was 24 and 24 was the second most frequently 'guessed' value. So the 7% who knew the answer to the coin toss/German Lotto question, the 'experts', were compared to all those who didn't. The 'experts' (Lotto question experts) were on the order of 20 times further from the correct value (number of marbles) than the non-experts. As the absolute error increased from 0 to 10,000 (actually 9,999 – 24 which was the largest possible error due to censoring of data) the size of the 'expert' error decreases (Hong and Page attribute this to the fact that the expert group is getting larger than therefore more diverse) whereas the error rate for the non-expert group remains fairly constant throughout the entire range except for the very last values, the most extreme.

The experts, however, always underestimate the number of marbles in the jar; the non-experts start underestimating but switch to over estimating for most of the range. Krause points out that this is a systematic bias on the part of the experts, but that the experts' estimation of the number of marbles also always has a smaller variance than the non-experts. "Regarding the marble problem, averaging of opinions can result in a good approximation of the correct value

because the guesses are imprecise but not fundamentally biased as they are in the coin-flipping problem." (Krause, James, Faria, Ruxton, & Krause, 2011, p. 946)

The fact that the estimates for the coin toss/German Lotto were so extremely high for the entire group but generally lower for the coin toss/German Lotto 'experts' would immediately lead one to think that something like Kahneman and Tversky's anchoring effect (Tversky & Kahneman, 1974) where the experts had just computed/estimated a smallish number for the coin toss/German Lotto estimate and therefore underestimated the number of marbles in the jar, except that the order of the questions was reversed with the marble question being first. So the anchoring effect that Krause may have observed was that the non-experts were unduly influenced by their estimates of the number of marbles in the jar and thus overestimated the coin toss/German Lotto value. The 'experts' were more able to access different sources of information about the coin toss/German Lotto value and therefore were not as influenced by their number of marbles estimate.

An obvious difference between the Krause and the Hong - Page simulations is that in the Hong - Page simulation the solution is self-evident to the group. Every agent in the group uses the same metric to determine what agent has the highest value. Whereas in Krause simulation, the 'most' correct answer is not self-evident; therefore the group of agents' best estimate is determined by the average of each group member's estimate. However, this representation of the group's best estimate as the average is the same in both the number of marbles and coin toss/German Lotto experiments, one of which demonstrates a wisdom of the crowd effect and one that doesn't.

3.3 FAST AND FRUGAL SIMULATIONS

Gigerenzer and Goldstein (1996) present three models of how 'fast and frugal' decisions work quite well. As they point out, they are not the first theorists to define a model which is not based on maximizing some normative measure of correctness. Simon's (1955) theory of satisficing, rather than maximizing, is the most famous prior articulation of the theory. In Simon's theory, an agent has an aspiration level and when the agent has found a solution that exceeds that aspiration level, s/he stops searching. There is no attempt to obtain the maximum level of 'utility', merely a level above some threshold value. This is a function of the limited capabilities of individual agents. According to Gigerenzer, Simon posits that this behavior is also a function of the environment in which the agent finds him or herself (Simon, 1956). This depends on the belief that the environment is structured in such a manner that finding an acceptable level of utility requires less expenditure of resources than finding a maximum level. Simon's example is selling a house, if you list your house for sale for $15,000 (remember these are 1955 prices he is dealing with) and you get two offers say of $16,000 and $25,000 you'll take the $25,000[11] offer (*ceteris paribus*) (Simon, 1955, p. 104). You, of course, take the offer with the most utility; however, what you don't do is continue keeping your house on the market until you are sure that you have achieved the maximum utility from the transaction. This 'satisficing' heuristic can be also seen as being the most efficient in environments in which "it is difficult to return to previously seen options and tell what options lie ahead" (Todd & Gigerenzer, 2007, p. 170). For example, in the

[11] There is the condition that the $25,000 is not so high that it changes one's basic belief about the value of the house. An offer significantly above the 'satisficing' level would lead to questioning of that level.

above case you are more likely to accept an offer if you believe that you'll not receive a better one and you may be quicker to accept that offer if it is above your aspiration level.

Gigerenzer's and Goldstein's models are based on induction. An agent uses the information available to it to make decisions. In their example, pairs of cities in Germany (their home country) are presented to agents who must choose which city has the larger population. All the cities in Germany with a population larger than 100,000 are used in the experiment (N=83). Now the agents may not know the relative population of the cities, so they have to use other clues. "Clues" and "cues" are used interchangeably in this analysis. A clue provides a cue (or vice versa) as to the relative size of the city. Kahneman (2011) identifies this as the intuitive heuristic: when faced with a hard question, substitute it for an easier one: the participants don't know the populations of the cities but they may know that one city has a soccer team and the other doesn't and they use that information to make the decision (perhaps) without being conscious of having answered a different question than the one which was asked.

In this experiment, there are nine other clues: 1) Is it (one of the cities in the pair) the capital of the country? 2) Was the city once an exposition site? 3) Does the city have a team in a particular soccer league, etc.? Each agent either recognizes the city or does not recognize the city and knows 6 categories of the clues (0%, 10%, 20%, 50%, 75%, or 100%), i.e., they knew none of the clues, a random 10% of the clues, etc. A preliminary study showed that for University of Chicago students (not necessarily Germans), 80% of time if the person recognized only one city of the pair, it was the larger city that they recognized. Agent pools were then created with being able to recognize 0 to 83 of the cities and with the six different categories of clues. 500 agents were simulated from each category for each pair of cities: (83*82)/2=3,403 possible pairs of cities ((84 * 6) = 504 agents making 3,403 decisions each = 1,715,112 decisions * 500

simulations of each agent = 857,556,000 agents decisions). The 84 represents knowing 0 to 83 different cities; the 6 represents the different percentages of the number of cues the agent knows, two agents who both know 10% of clues will most likely know different clues since clues are randomly chosen.

Agents also needed to be able to rank the validity of their clues (Todd & Gigerenzer, 2007). In this example, Berlin with a population of over 3,000,000 is the largest city in Germany and the capital. Therefore the clue: "Capital of the country" will always produce the city with the highest population when answered yes. It is right 100% of the time: this measure is called the clue's validity[12], the number of times it is right divided by the sum of the number of times it is wrong and the number of times it is right. It has prefect validity; however, it also has low discriminatory power. It can only discriminate in those pairs in which Berlin was one of the cities being compared. With 83 cities there are (83 * 82 / 2) 3403 pairs and this clue can only be used in 82 of them. It discriminates in only 2.4% of the cases. The lowest validity clue was "Was it a city in East Germany" which was right 51% of time or barely above the value of 50% which would be obtained randomly. Note that if a clue was correct 49% of the time, it would be equal in information content to the East German City clue and the agent would just use the opposite encoding: if the clue said city A was larger than city B and the agent knew the clue was correct less than 50% of the time, the agent would know that city B having the larger population was more likely.

The agents, if it only recognized one city, choose it as having the larger population, if they didn't recognize either city they randomly chose one. When they recognized both cities,

[12] Clue validity is mathematically the same as the Goodman-Kruskal gamma coefficient. (Luan, Katsikopoulous, & Reimer, 2012)

they looked at the clues they had in rank order of the clue's validity, when one city had a positive answer and the other city's answer was negative, the city with the positive answer was chosen as the larger city. This model is called "Take the best" because the response is based on the response for the best clue which distinguishes the population of the two cities and solely on the response for that single clue. The other models involve choosing a random clue or choosing the clue that was used to distinguish two cities most recently and are described as "Minimalist".

Gigerenzer's ABC group (Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development in Berlin) published a simulation of his 'fast and frugal' heuristic in a group setting (Luan, Katsikopoulous, & Reimer, 2012). In this set of simulations they vary

- the heuristics which agents use
- the size of the groups
- how cue search order was learned
- the size of the errors in the information being searched

They found that by alternating several of their controls they could increase the error rate in the selections made by the individual agents and thereby increased the diversity of the group to which those agents belonged. At that point, they could then investigate under what conditions did "increased group diversity" make up for the decrease in individual accuracy in terms of overall group accuracy. They found that individual accuracy versus group diversity was mainly important when the diversity within the group of cues was large then individual accuracy was favored; when the all the cues were closer together in their validity, group diversity trumped individual accuracy. They, of course, used Gigerenzer's "Take the best" and the random cue selection minimalist heuristic as their test bed.

In their study, in their test bed on one axis they had four conditions. Where cue diversity ranged from no difference, through small and moderate to large difference. Validities are measured as the proportion of times the cue identifies the correct response. The betas are the weights that each individual cue has on creation the criterion variable along with an error term.

Table 6 (Luan, Katsikopoulous, & Reimer, 2012, p. 6) The Linear beta coefficients, validities, and explained variance for the four different task environments.

Environment	Beta's 1 to 5	Validities 1 to 5	% of variance accounted for by cues
Large difference	0.37, 0.23, 0.11, 0.07, 0.04	0.86, 0.71, 0.60, 0.57, 0.54	0.865
Medium difference	0.23, 0.20, 0.16, 0.13, 0.11	0.78, 0.71, 0.67, 0.64, 0.61	0.864
Small difference	0.19, 0.18, 0.17, 0.16, 0.15	0.71, 0.70, 0.69, 0.68, 0.67	0.866
No difference	0.17, 0.17, 0.17, 0.17, 0.17	0.69, 0.69, 0.69, 0.69, 0.69	0.865

The differences in the validity of the different cues was (.86-.54=.32) .32 for the large difference environment, .17 for the medium difference environment, .04 for the small difference environment and 0 for the no difference environment. Note that the intergroup differences vary significantly .04 followed by .17 followed by .32. Additionally these differences only reflect the extreme values. The differences between the different cues are:

Table 7 Differences between successive cue validities

Environment	Cue 1 to 2	Cue 2 to 3	Cue 3 to 4	Cue 4 to 5	Cue 1 to 5
Large difference	.15	.09	.03	.03	.32
Medium difference	.07	.04	.03	0.4	.17
Small difference	.01	.01	.01	.01	.04
No difference	0.00	0.00	0.00	0.00	0.00

Almost 50% of the difference between the cue validities comes between the first and second cues in the large difference environment (.15/.32) and 41% in the medium difference (.07/.17). This relatively overweighting of the first cue may be responsible for a significant part of the results they report in their paper. The observation that the minimalist's group never matches the 'take the best' group in the large difference environment may reflect that the degree to which the first and second cues dominate rather than having the .32 difference observed equally distributed among the four different inter-cue gaps.

Groups of 15 and more from the medium diversity group match or exceed the performance of the 'take the best' group (which is an individual agent since all agents would make the same choice) and the 5 member group passes it in the small difference environment. In the no difference environment all groups exceed the 'take the best' agent which is tied with the minimalists group of one agent since all cues have the same validity.

The distribution of information in the cues in Gigerenzer's original article (1996) is given below in Table 8Table 8Table 8Table 8Table 8. The complete data for the cues and city populations is available in an appendix of their paper. The following two tables are computed from that data (with the correction of Munich's population being 1,229,026 rather than the 1,229.026 as is entered in the appendix, the other errors in the value of a state capital cue and a soccer team cue are ignored as the authors state they did in their analysis.) (Gigerenzer & Goldstein, 1996, p. 33)

Table 8 German City Cues

	Soccer	State Capital	East German	Indus-trial Belt	License Plate	Inter City Train	Expo-sition	National Capital	Univer-sity
Soccer Team	15	5	1	3	5	3	8	0	11
State Capital	.19	15	4	0	6	15	7	1	10
East German	-.12	0.14	13	0	6	10	1	0	3
Industrial Belt	0.02	-0.22*	-0.20	15	2	10	2	0	5
License Plate	0.13	0.21	0.26*	-0.10	18	16	10	1	14
Inter City Train	0.13	0.27*	0.22	-0.09	0.17	62	13	1	41
Exposition	0.46*	0.37*	-0.11	-0.11	0.54*	0.19	14	1	13
National capital	-0.05	0.24*	-0.05	-0.05	0.21	0.06	0.25*	1	1
University	0.20	0.14	-0.25*	-0.25	0.27*	.49*	0.37*	0.11	43

This table shows the distribution of cues: the numbers below the diagonal are the correlation coefficients for the row and column cue, the numbers above the diagonal are the frequency counts of when both the column and row are true, the numbers on the diagonal are the number of times the cue is true. For example, the first item is on the diagonal and says that there are 15 soccer teams in the particular league being considered, the next number, 5, means that there are 5 state capitals with soccer teams, the number below the 15 (row 2, column 1) .19 means that the correlation between having a soccer team and being a state capital is 0.19, if there would have been an asterisk following the number, it would indicate that the probability of the correlation coefficient being zero was less than 0.05.

The mechanism used in the simulation correlates the cue values to the target values by the size of the betas used in the equations, but as this table shows there is also a significant amount of structure between the cues independent of the target variables. This relationship is missing from the simulation since the cues are ~N(0,1) distributed. This is not just an insignificant item. The development of the fast and frugal heuristic is based on a structure among the cues. In order for the cues to be positively related to the population of a city, they must be positively related to each other. Simon writes (1956, p. 129)

> "…a great deal can be learned about rational decision making by taking into account, at the outset, the limitations upon the capacities and complexity of the organism, and by taking into account the fact that the environments to which it must adapt possess properties that permit further simplification of its choice mechanisms."

Gigerenzer (1996, p. 3) quotes Simon: "Human rational behavior is shaped by a scissors whose two blades are the structure of the task environments and the computational capabilities of the actor." The structure of the task environment is a function of the validity of the cue, its discriminatory power, and the conditional information that cue has given previously used cues.

The value of the cues to identify the larger city varies also. The "National Capital" cue is always correct giving a validity score of 1 (validity is the number of times it correctly identifies the larger city out of all comparisons). The other metric of interest is "How often does it distinguish between two cities?" In the case of the 'National Capital' cue, it selects Berlin as being the larger city 82 times out of 3,403 possible comparisons for a selectivity score of 2.4%.

Table 9 German City Cue Metrics

	Soccer	State Capital	East German	Indus-trial Belt	Licen-se Plate	Inter City Train	Expo-sition	Nation-al Capital	Univer-sity
# of times makes a distinc-tion	1020	1020	910	1020	1170	1302	966	82	1720
# of times correct distinc-tion made	825	713	481	590	818	992	810	82	1188
Validity	80.9%	69.9%	52.9%	57.8%	69.9%	76.2%	83.9%	100%	69.1%
Selecti-vity	24%	21%	14%	17%	24%	29%	24%	2.4%	35%
p of F test for Type III SS	0.0092	0.6129	0.2465	0.1502	0.4713	0.3586	0.0006	<.0001	.1261

The cues, in the Fast and Frugal simulation (Luan, Katsikopoulous, & Reimer, 2012), are distributed N(0,1) and are independent from each other. This is very different from the data used in the Gigerenzer's (1996) model as show in the two tables above. In the German cities data, the cues have a true value in the following range of percentages: being the capital of the country is, of course, the lowest true value one time out of 83 cities (true value meaning having a value of 1 that is being true rather than a value of 0 which implies not true), the highest percentage is 75% associated with having an Inter-City Train Service (62 times), next is having a university 52% (43 times), all the other cues vary between 22% and 16% (18 to 13 times). However, the cues are not independent, Pearson Correlation Coefficients, in absolute value terms, range from a low of 0.02 (Inter-City Train Service and prior East German city) to a high of .54 (Hosted an

exposition and License Plate). The probability of the correlations being not zero is less than 0.05 in 12 cases out of a possible 36 (9 different cues, 9 * 2 / 2=36), which is to be expected as all the cues can be seen as being associated with economic activity. The Pearson Correlation Coefficients with the population for the 9 cues are below 0.05 six times with one being 0.058 and the other two .59 (prior East German City) and .91 (Industrial belt). So the cues are highly correlated with each other and with the target variable, population.

A general linear model of the population and the cues has an R squared of 0.84 (model F (9,73) =44, $p<0.0001$). The Type I SS has five cues at 0.05 significance and Type III SS has three cues significant at the 0.05 level. This isn't just a random by-product of this example, but a basic principle in the fast and frugal paradigm: the fast and frugal paradigm works because there is a structure to the environment which can be exploited. The differences in the betas in Table 6 (Luan, Katsikopoulous, & Reimer, 2012, p. 6) The Linear beta coefficients, validities, and explained variance for the four different task environments. above will create significant correlations between the cues and the target value, but not among the cues themselves, so the simulation of includes part of the environmental structure expected by the model.

Additionally they dichotomize the cues into two groups using the following process: Cues are randomly generated with from a normal distribution with a mean of zero and variance of one as is an error term. The value of the criterion value is the sum of those values times the appropriate betas (described in Table 6). The cue values are then recoded into 1 or 0 depending on the cue being above or below the median for that set of 15. This process makes the cue maximally informative. Since the cues are ~N(0,1), the expected value of the median is 0; so the recoded cues will be (approximately all positive cues recoded to 1 and all negative cues recoded to 0. For example, in one instance the mean of the first cue was -0.08 and the median was -0.15,

when recoded the average original value for the first cue that were recoded to 0 was -.88 and the average original value for the first cue that were record to 1 was .62. Thus when the first cues differ the value of the criterion value differs by (-.88 - .62 = 1.5 *.37 (1st beta for Large Differences) = .56). The criterion scores for this group range from -.93 to .58 (mean = -0.07, std dev = .46), so a difference of .56 is rather large. This size of this effect is partially the large beta (0.37 is the largest beta in the simulation) and partially the effect of dichotomizing the cue variables at the median. The effect of this dichotomizing is to artificially inflate the validity of the cue. There are 105 pairs from these 15 items and in 49 of them the value of the first cues match so that it isn't used to distinguish between the two items. In the remaining 56 pairs, the cue, with a value of 1, is associated with the higher criterion score 53 times, giving it a validity of 95%. Instead of using the median, if the 25th percentile is used for the dichotomizing point, the corresponding values for the 105 pairs are: 61 times the first cues match so it isn't used and in the remaining cases, 37 out of 44 times if its value is 1 it is associated with the higher criterion score for a validity of 84%.

They also test the effect of different learning scenarios but since the minimalist's class of agents just selects a random heuristic, there is not any meaningful comparison between the two classes of agents: the minimalist class never learns anything, it always acts randomly. The 'take the best' class monotonically increases its accuracy as it has a larger and larger learning sample to learn how to relatively rank the different clues in the large difference environment. In the small difference environment, the increase in accuracy was positive until the last step, but it was always very small. In the last step, when the 'take the best' individuals and groups both worked on the same ranking of cues their accuracy dropped as was seen in the group size experiment.

Their final experiment was to increase the error term and evaluate the effect of the increased error on the different decision processes. When the size the error term increases, the individual scores decrease monotonically; however, the group scores start to increase for both the 'take the best' and the minimalist's groups before falling. The same effect is observed in large difference environments and small difference environments. This was the experiment that demonstrates the 'miracle of aggregation' to use Caplan's term. "When there are random errors in cue values, different agents may make different decisions for the same pair of options, even when they all adopt the same decision heuristic and search cues in the same order." (Luan, Katsikopoulous, & Reimer, 2012, p. 4) implies that the error is in the cue itself as much as in the agent's perception of the cue. Since if the cue itself was incorrect, both agents would act on that same piece of misinformation in a similar manner, thus it has to be cue that is in error. For example, assume someone still believes that Bonn is the capital of Germany, when comparing Berlin with any other city; they would believe that both Berlin and the other city were not the capital of the country and go on to the next cue. In this case, since 'capital of the country' has a validity of 1, they have zero chance of making a better decision. However, in cases where the error in perceiving the cue value is for a cue with less than 1 as its validity, there is a chance that having made that error would lead to a better decision.

3.4 MARCH'S EXPLORATION/EXPLOITATION SIMULATION

March (1991) presents a simulation of organizational learning that has several similarities with group processes. The organization in March's examples are businesses with structure and purposes that all groups do not share but his simulation introduces an interplay between the

individual and the group whereby the individuals influence what the group 'knows' and the group influences what the individual's 'know'. The context for the March article is the tension between exploring new avenues to achieve the organization's goals versus increasing current skills in exploiting known avenues.

> "Adaptive systems that engage in exploration to the exclusion of exploitation are likely to find that they suffer the costs of experimentation with gaining many of its benefits. They exhibit too many underdeveloped new ideas and too little distinctive competence. Conversely, systems that engage in exploitation to the exclusion of exploration are likely to find themselves trapped in suboptimal stable equilibria. As a result, maintaining an appropriate balance between exploration and exploitation is a primary factor in system survival and prosperity." (March, 1991, p. 71)

The theoretical structure of March's simulation is that there exists a 'code' within an organization that represents its knowledge. New individuals entering the organization learn that code and come to believe it. The 'code', however, is not static nor was it created by fiat. It comes from the beliefs of the individuals who have made up the organization and who enter the organization: it is a mutual learning environment. The organization learns from those individuals who are more closely aligned with reality than it is.

In the simulation, there is a vector of values representing external reality; each value in the vector is valued either as -1 or +1 with equal probability. Individuals have a similar vector representing the external reality items; their vectors are valued -1 or +1 also, but they also have a neutral 0 value. The individuals modify their perception of reality as they become socialized into the organization. In each time period, any particular value in which the individual differs from the organization has a probability of changing to organization's value, p_1. This p_1 is the effectiveness of the individual's learning of the organization's value vector. If the individual's

value for any particular item is zero, the individual is indifferent about that item of reality and does not adjust his vector to reflect the organization's code.

At the same time, the organization recognizes that some people in the organization have a value structure which more closely matches external reality that the organization's code. The organization learns from those individuals who have a more externally validated value structure with a probability of p_2. The organization starts out with all zeroes in its value structure and learns its values from the individuals in the organization.

The simulation starts with 30 items in the reality vector and 50 individuals in the organization. The values of the reality vector have an equal probability of being -1 or +1. The values of the individuals' vectors are -1, 0, or +1 with equal probability. The organization's code starts out with all zeroes. An equilibrium is reached when the individuals in the organization and the organization share the same values in all 30 positions in their value vectors. Note that there is nothing requiring that the shared value vector of the organization and individuals match that of reality. The two learning rates are p_1, which is the rate at which the individuals become socialized into the organization and p_2 which is rate at which the organization learns from the individuals.

March found that higher learning rates lead to the two sets of value vectors matching faster than lower learning rates. However, an average correct percentage can be calculated at each time point. The percentage correct being the percentage of times that the individual's or the organization's vector match reality's vector. Faster learning rates, in general, lead to lower average correct percentage among the different vectors. In terms of exploitation and exploration, the faster learning rates exploits current knowledge whereas slower learning rates allows for differences to exist for a longer time and thereby have the opportunity of entering into the

organization's code. However, the fast learning rates are initially higher before falling below the slower learning rates.

March experimented with different learning rates for the individuals within the organization. When the learning rates differed, the final knowledge had a high percentage than when all members of the organization had the average learning rate of the two separate groups: for example, groups of the individual learning rate being .1 and .9 in equal proportions versus a group in which all members had a learning rate of .5. However, the distribution of the learning is not equal. The fast learning groups have better average percentages than the slow learners from whom they are learning how to improve their percentage correct. The slow learning are holding on to what they know (even when it is wrong) longer, whereas the fast learning gain the benefit from the organization's code being improved more quickly than the slow learners.

Two additional parameters were introduced into his model, p_3 and p_4. P_3 is the turnover rate. When new individuals enter an organization, it gives the organization a new opportunity to learn about the external world. High turnover, by itself, removes sustained learning opportunities; however, moderate turnover with fast organizational learning leads to a higher organizational knowledge. P_4 is environmental turbulence, the rate at which the values in the reality vector change values. Since the organization learns from individuals and individuals learn from the organization, after a period of no turnover they will reach an equilibrium. In which case there will be little pressure to adjust the organizational code to match a changing reality. It is in the presence of turnover that the change in external reality enters the organizational code. In an organization competing with other organizations, the correspondence between the organization's code and the environment would influence its ability to survive

which would present a pressure on its code independently from that caused by turnover of personnel; however, this source for organizational change is not modelled in this simulation.

4.0 EXTENSIONS TO BASE SIMULATIONS

4.1 HEURISTIC SIMULATIONS

Figure 4 shows the structure of the Heuristic simulations. The areas shaded in grey show the sections of the simulation that are iterated a varying number of times. For example, when the Training Size is 500 and the Testing Size is 1,000, in Step 1 the Training Phase is executed 500 times and then in Step 2 the Testing Phase is executed 1,000 times.

<table>
<tr><td colspan="4">Heuristic Simulations</td></tr>
<tr><td rowspan="3">Training Phase</td><td rowspan="3">Training Size</td><td>Create Training Problem Space</td><td>Uniform random
Sombrero (cycle, beta)</td></tr>
<tr><td>Calculate 1,320 agents' values in problem space</td><td>First improvement
Best improvement
Negative Correlation Learning Average
Negative Correlation Learning Cut</td></tr>
<tr><td>Calculate summary score</td><td>Average
Group Average</td></tr>
<tr><td colspan="3">Select groups based on training phase</td><td>Top 10
Top from groupings
Random Group</td></tr>
<tr><td rowspan="2">Testing Phase</td><td rowspan="2">Testing Size</td><td>Create Testing Problem Space</td><td>Uniform random
Sombrero (cycle, beta)
Reuse x% of training space</td></tr>
<tr><td>Calculate 'expert' group and random group values</td><td>Top10 & Random
Top from groupings
NCL Average
NCL Cut
First Improvement
Best Improvement</td></tr>
<tr><td rowspan="2">Collect Summary Statistics</td><td colspan="2">Calculate average score for each group over all tests</td><td></td></tr>
<tr><td colspan="2">Calculate diversity among groups</td><td>Diversity metric
Distribution of heuristics
Frequency of 0's</td></tr>
</table>

Figure 4 Heuristic Simulation Structure

The 'Training Phase' is similar to the 'Training Phase' in machine learning. It is the phase in which the parameters are calculated which will be used to select the varying 'experts' groups. The problem space it works with is randomly constructed as is the 'Testing Phase' problem space (except during the 'Reuse the training problem space examples'). The two spaces are labeled "Training" and "Testing" because they are the where the parameters are obtained and

then applied which is a common usage for these phases. The metrics developed for this dissertation will be discussed below in the appropriate section.

4.1.1 Heuristic Simulations Membership in 'Experts' group

The criterion for membership into an 'expert' group is explored with the creation of two additional grouping mechanisms. An 'Average' expert group is defined by combining executions of the simulation into groups and selecting the agent with the highest frequency of being in the top 10 of the individual groups. The group size that produces the 'best' group of expert agents is a group size of 1, raising questions about the value of averaging agents' scores over a set training samples. Additionally negatively correlated group mechanisms were created: Negative Correlation Learning – Cut Value was created in which expert groups are created by taking the top scoring agent for an execution of the simulation and then adding members to that group who contribute the largest number of new top values (top values are defined as have a score greater than 95) and Negative Correlation Learning – Average Value which, instead of using number of ring positions with a score above a cut value, uses the agents' scores on the remaining sites on the ring and incrementally adds the one agent at a time. These groupings are negatively correlated in the sense that the new members' contributions are negatively correlated with the contributions of the members already in the group. The results of this section are that the negatively correlated group out performs the random group with a (statistically) significantly lower level of heuristic diversity.

4.1.1.1 Group Average Definition of Experts

In the Hong-Page simulation, the expert group is defined as the ten agents with the highest average score in the 'training' dataset. The training period is exhaustive, in that during this phase each agent finds its return value for each possible starting location in the training problem space and each possible agent is simulated throughout the entire problem space: 'all' the agents, individually, explore 'all' the problem space. In the Group Average method, the following process was used to group the agents into several groups and use those group scores to determine which agents should be placed in the experts group.

Table 10 Group Average Creation Method

Step	Activity
1	Create multiple iterations of the problem space and compute the average score for each agent over the problem space
2	a – Get overall average for each agent over all training spaces b – Assign a Group ID number to each iteration of the training phase. Group ID = ceil (iteration/group size)
3	Get an average score for each agent for each group
4	Identify top 10 agents for each group by highest average score within the group.
5	Count number of times each agent is in any Top 10 group
6	Identify top 10 agents based on frequency of being in the Top 10 of a group, ties broken by overall average computed in Step 2a above
7	Compute diversity metric on group of agents identified
8	Iterate steps 1-6 100 times to get an average score for this process with each particular group size

An average score was computed for each agent, Step 1 above, in a varying number of different training problem spaces[13]: each training problem space contains 2,000 locations and each agent returns a value associated with each of those locations, the average of those 2,000 values is the agents' average score for that particular training set. In the group model, the training sets are grouped together and the agents' rank within those groups is used to select agents for inclusion into the 'experts' group (Steps 3 to 6). This mechanism attempts to selects agents which perform well (compared to other agents) in a variety of training spaces.

For example with 2,000 iterations of the training problem space, the 2,000 iterations could be grouped together into 200 groups of ten each: iterations 1-10 being group 1, 11-20 group 2, etc. The average score for each agent is computed over the ten iterations for each group and the agents with the highest scores are identified. The number of times that each agent is in the group of the top ten scoring agents is summed and the list is sorted by the number of times the agent placed first and the secondarily by the agent's average score over all 2,000 iteration. The top 10 agents from this list are then grouped together to form the Average Group. Regardless of the size of the groups, the end result is a group of ten agents that are grouped together to form the 'experts' group.

The following table shows the results of one test run of 20,000 iterations with 200 groups of 100 being used as the grouping factor compared with the original Hong-Page definition of the experts group, as the 10 agents with the highest score in the training phase (i.e., the first 100 iterations of the 20,000 total iterations created the first group, an agent's score for each iteration

[13] In the Hong-Page paper, the results are presented averaging over 50 trials. They varied their simulation with heuristics ranging from 12 different heuristics to 6-20 different heuristics, the number of heuristics varying from 3 to a range of 2 to 7, and the problem space varying from 2,000 to a range from 200 to 10,000. Although it is not explicitly stated, it appears that training is conducted on one instantiation of the problem space. (Hong & Page, 2004)

was its average over the 2,000 positions on the ring, each agent then had its iteration scores averaged to create an average score for that agent for that group, and the ten agents with the highest scores for that group of 100 iterations were identified as the top 10 agents for the first group of 100 – this process was completed 200 times for 200 groups of 100 or 20,000 iterations in total). The final column identifies those agents who are in the experts group using either grouping mechanism. In this example, the Hong-Page Top 10 method and the Group Average had an overlap of six agents which are identified in the last column of the table.

Table 11 Average Group Expert Example[14]

Agents	Individual Average over 2,000 iterations	# of times in top 10 agent group	Position in overall group (original experts)	In both groups of experts
10, 12, 11	85.146		1	
11, 10, 12	85.145	3	4	*****
12, 10, 11	85.145	2	5	*****
11, 10, 9	85.138	3		
1, 3, 11	85.146	2	2	*****
12, 11, 10	85.145	2	3	*****
1, 11, 3	85.143	2	6	*****
11, 12, 10	85.143		7	
10, 11, 9	85.143	2	8	*****
10, 11, 12	85.142		9	
11, 1, 3	85.140	2	10	*****
5, 6, 7	85.140	2		
8, 10, 12	85.136	2		

[14] The 13 agents in this table clearly demonstrate the bias towards high valued heuristics: 12 appears 7 times; 11 appears 11 times; 10 appears 9 times; 9 2 times; 8, 7, 6, and 5 once each; 3 and 1 three times each; and 2 and 4 not appearing at all. The random expected number of counts for each heuristic is 3.5 times each. Note that all six versions of 10, 11, 12 appear in the combined list.

The following table shows the effect of different group sizes on 2,000 iterations of the training phase on the average final scores the group of experts, defined by the Average Group method, achieves when it runs through the test phase 100 times. The results are from a general linear model of final score as a function of group size (as a class variable not as a numeric value). The model F statistic is 5,551 (df 7, 1.6E6; $p<.0001$)[15]. There were 1,600,000 observations in this regression: 8 different group sizes multiplied by 2,000 ring positions multiplied by 100 iterations in the testing phase. The r^2 is .02 showing that actually very little of the variance is explained by the model. Group size 1 creates a cohort of experts based on 2,000 groups; group size 2,000 creates the original definition of the experts group since there is one group of 2,000. In the model, the betas are being compared to Group size 2000 which is represented by the intercept term. All betas are significant at the $p<.0001$ level. The t-values for the betas range from 96 to 182. All the groups, other than the Group of 2000 which is the same as the Top 10 Expert Group, have a statistically significantly higher score than the Top 10 Expert group that was created by taking the ten best agents from the entire training sample.

[15] A second sample of 2,000 iterations was run with an F statistic of 1,951 ($p<.0001$), r^2 of .01, and the beta for group of 1 being 94.225 which was significantly larger than all other beta's except the group of 50, which was 94.815 and significantly larger than the group of 1 at the .0001 level.

Table 12 Group Regression Betas and Average Scores

Regression Term	Beta Estimate	Average Score
Group size 1	3.966	94.651
Group size 20	2.954	93.64
Group size 50	2.405	93.090
Group size 100	3.127	93.812
Group size 250	2.239	92.924
Group size 500	2.232	92.917
Group size 1,000	2.079	92.764
Group size 2,000/ (Intercept)	0 90.685	90.685

Group size 1 can be shown to be significantly larger than the other coefficients at the p=0.0001 level. A group size of 1 means that the effective measure being used is not the average score achieved over a group of trials, but the number of times that a particular agent was in the top ten in each trail. In this set of trials the following table shows the agents and the number of trials in which the agent was in the top ten agents for a trial (2,000 trials * 10 agents per trial = 20,000 top ten positions / 1,320 agents gives an expectation of each agent being in 15 top ten groups if each agent has an equal probability of being in each top ten group). There was an 11th agent (2, 6, 10) that was in the top ten 33 times which wasn't included in the average expert's group since its overall average was 85.1216 compared with 2, 7, 12's overall average of 85.1238.

Table 13 Frequency of agents in top ten group

Agents	Number of time in top ten group
2, 5, 9	36
3, 5, 9	36
1, 3, 11	35
4, 12, 3	35
2, 12, 7	35
12, 10 ,11	34
12, 11, 3	34
5, 8, 11	34
10, 12, 11	33
2, 7, 12	33

The top two agents in the above table share 2/3rds of their heuristics which leads to the question of how many of the times when each agent is in a top ten group is the other agent (of the top two) also in the same group. To investigate the possibility of a high level of redundancy the same process was run on a sample of 50,000 iterations and the cross frequency of being in the top ten groups counted. With 50,000 iterations there are 50,000 chances for an agent to be in a top ten of an iteration which produces an expected value of (50,000/1,320) * 10 = 379 (number of groups for each agent to be in the top 10 of). In this sample the agent most frequently in a top ten group is in 797 of them, almost a 2 fold increase over the expected value. The lowest ranking agent in this group of the top ten was in the top ten in 642 (of the 50,000 possible) groups. The following table shows how often the top ten agents were in groups with other top ten agents in the bottom triangle. The upper triangle shows the percentage of overlap between the row and column. For example, 2.6.10 and 1.3.5 shared 4 groups, thus 2.6.10 had 4 /797 % of

groups shared with 1.3.5. As can be seen from the size of the counts, the sample size had to be increased to 50,000 in order to have meaningful counts in the various cells.

Table 14 Co-occurrence of top ten agents in 50,000 samples

	2.6.10	1.3.5	2.5.8	1.4.7	2.7.12	1.6.11	1.3.8	4.7.10	1.5.9	3.7.11
2.6.10	797	.005	.009	.005	.006	.005	.010	.003	.019	.016
1.3.5	4	739	.008	.011	.004	.005	.060	.001	.034	.011
2.5.8	7	6	714	.02	.008	.006	.007	.027	.01	.003
1.4.7	4	8	14	677	.007	.015	.007	.043	.010	.004
2.7.12	5	3	6	5	667	.019	.015	.009	.010	.007
1.6.11	4	4	4	10	13	665	.017	.008	.005	.009
1.3.8	8	44	5	5	10	11	660	0	.006	.003
4.7.10	2	1	19	29	6	5	0	658	.010	.008
1.5.9	15	25	7	7	7	3	4	7	657	.002
3.7.11	13	8	2	3	5	6	2	5	14	642

The most frequent co-occurring agents are 1.3.5 and 1.3.8 (cells are shaded grey): 1.3.5 is in 739 top ten groups and 1.3.8 is in 660 top ten groups. They share 44 groups in common (6%), which is more than would be randomly expected, but these two agents also share the first two heuristics, so they have a Hong-Page diversity index of .33. The next highest is the 29 (4.2%) with agents 4.7.10 and 1.4.7: these agents since they don't have any heuristics in the same locations have a perfect diversity score of 1 even though the both share the 4 then 7 pair of heuristics. The next highest is the 25 (3.4%) with agents 1.5.9 and 1.3.5: here the agents both share heuristic 1 in position 1 and therefore have a diversity score of .66, however they also share another heuristic although that heuristic isn't sequential with the 5 being in the second

position for one agent and in the last position for the other[16]. From Table 17, I was expecting something on the order of 50% sharing of groups between agents with low diversity metrics rather than the 10% observed; for example I would have expected 1.3.5 and 1.3.8 to share approximately 300 groups in which they were both in the top 10 rather than just 44. The reason for this difference is that Table 17 shows how agents overlap within the same problem space whereas Table 14 shows how agents do not, in general, overlap across different problem spaces.

4.1.1.2 "Negative Correlation Learning" type definitions of experts

An additional mechanism for selecting the 'experts' was developed along the lines suggested by Negative Correlation Learning. The negative correlation learning method selects agents whose errors are negatively correlated with the other agents in the collection; the method developed here, selects agents that do well on different areas of the problem set from those areas that the agents that are already included in the expert group did well on. The first process, called NCL Cut Point (NCL C), is to count all the instances in which each agent returns a value of greater than a critical value in the problem space, the cut point. Individual agents generally return values between 80 and 85 for a problem space and groups return values 90 to 95 (c.f. Table 12). The problem space, itself, has an average score of 50. The cut point used in this dissertation is 95. This value was selected because it is slightly greater than the expected average score, and when the number was higher there were frequently simulations in which there were no candidates for the final positions, i.e., there were no ring positions left with any agent having a score over a value higher than 95.

[16] Note that all possible diversity scores, from the Hong-Page diversity metric, are present in these three examples: .33 when two heuristics match position, .66 when only one heuristic matches position, and 1.00 when none of the heuristics match. The only other possible value for their metric is 0.00 which is not a valid value in this scenario since all agents are unique and there no two agents could have all three heuristics match.

Table 15 Steps in Negative Correlation Learning - Cut Point

Step	Activity for each iteration of the training phase
1	Create Problem Space and compute each agent's sco in the problem space.
2	Load data array (1320,2000) with each agent's s position
3	Start selecting 'expert' agents
4	Cycle through problem space for each agent count times that agent returns > cut value for a ring positio
5	Add agent with the highest count to the expert's gro for the highest count, randomly choose one of the tie
6	For each ring position for which agent identified in S cut value (95) set all the value of that position to agents
7	Go back to step 4 computing each agent's average s ring, the ring positions which have been set to zero to the numerator but the denominator will re calculating the average
8	Stop when 10 agents have been identified

The agent with the highest count of "greater than 95" values is included in the experts group. The sites for which that agent returned a "greater than 95" value are then removed from consideration for all the remaining agents. The process is then repeated with obtaining the count of the number of times the agents return values "greater than 95" from the remaining sites and the next top agent is added to the expert set and the sites which lead that agent to be added to the expert set are then excluded from further consideration. This process continues until there are 10 agents in the 'expert' group (ties for the last positions are settled by which agent has the best overall average from all the training spaces used). This process, by removing sites for which other agents in the expert group return "greater than 95" values, leads to the agents in the expert group being negatively correlated with respect to the sites for which they values of 95 or higher. Although literally, what was being used as the criteria was having the highest marginal number of sites with a score above the cut point rather than a correlation of error terms.

This process is demonstrated with a simpler example. In this example, the problem space is five positions long (rather than 2,000). Each agent only has one heuristic. We have three agents with heuristics 1, 2, and 4. The problem space and the values associated with each ring location are shown below in Table 16. The problem space is outlined in the heavier lines to separate those values from the agent's heuristics and the counts associated with the process of choosing the agent with the highest number of sites above the cut value of 3 in this example.

- In step 1 (count the number of times agents return scores greater than 3), agent 2 returns a value of 4 or 5 for four of the ring positions, so it is selected first and positions 1-4 are grayed out in subsequent steps.
- In step 2, agent 3 returns a 4 or 5 once and there is selected as the second agent in the group.
- At this point the identified agents (2 and 3) have returned 4 or 5 for all positions on the ring, so agent 1 would never enter the group since it has nothing to add.

Table 16 Negative Correlation Group Example

	Position ->	1	2	3	4	5	# >3
Actual value at position		2	1	5	4	3	
Step 1	Heuristic						
Agent 1	1	2	5	5	4	3	3
Agent 2	2	**5**	**4**	**5**	**4**	3	**4**
Agent 3	4	3	2	5	5	4	3
Step 2							
Agent 1	1	~~2~~	~~5~~	~~5~~	~~4~~	3	0
Agent 2	2	~~5~~	~~4~~	~~5~~	~~4~~	~~3~~	0
Agent 3	4	~~3~~	~~2~~	~~5~~	~~5~~	4	1

Table 17 NCL Cut Example

Agents (Heuristics)	Number of positions with value of 95 or higher	Original number of positions with value of 95 or higher
1, 8, 10	642	642
6, 9, 10	296	595
7, 10, 11	196	524
4, 12, 1	142	542
10, 3, 5	103	538
9, 10, 11	78	586
11, 2, 12	50	469
9, 8, 6	37	573
8, 10, 11	29	606
5, 11, 12	25	454

This table shows that agent (1, 8, 10) returned values of 95 or higher 642 times from the possible 2,000 positions in the problem space. Once those 642 positions were removed from ring, the next agent (6, 9, 10) returned 296 positions with a value of 95 or higher. Agent (6, 9, 10) originally returned values of 95 or higher for 595 positions, but 299 were positions that agent (1, 8, 10) also returned a greater than 95 value for. For an average running of this simulation, each agent returns a value of 95 or higher for 540 positions and for 1,700 positions there will be an agent that returns a value of 95 or higher. In the above example, the ten agents returned 95 for higher for 1,598 positions.

The second Negative Correlation Learning type selection mechanism, called NCL Average (NCLa), is based on the agent's average score rather than the number of sites for which it is returning a value of 95 or higher. The process is to select the agent with the highest average score and then zero out those sites on the ring for which that agent returns a score higher than its average score, then proceed to the recomputed all the agents average scores with those zeroed

out ring positions not included and iteratively selecting the agent with the highest average score and then zeroing out the ring position for which it is returning a value higher than its average score.

Table 18 Negative Correlation Learning - Average Steps

Step	Activity: For each iteration of the training phase
1	Create Problem Space and compute each agent's score for each position in the problem space.
2	Load data array (1320,2000) with each agent's score in each ring position
3	Start selecting 'expert' agents
4	Cycle through problem space for each agent calculating that agent's average score
5	Add agent with the highest score to the expert's group
6	For each ring position for which agent identified in Step 5 has a score > its average score set all the value of that position to zero for all other agents
7	Go back to step 4 computing each agent's average score over the entire ring, the ring positions which have been set to zero will not contribute to the numerator but the denominator will remain 2,000 when calculating the average
8	Stop when 10 agents have been identified

An example from an actual execution of the simulation is shown below

Table 19 NCL Average Example

Agents (Heuristics)	Highest average score on remain positions	Original average score on all 2,000 positions
6, 11, 4	87.02	87.02
1, 2, 3	85.03	86.27
10, 12, 9	84.59	86.19
7, 8, 6	84.48	85.42
11, 5, 12	83.39	85.63
4, 12, 11	81.40	86.72
3, 6, 12	78.20	85.82
4, 9, 10	76.28	86.42
1, 3, 7	73.74	86.24
1, 2, 4	68.34	85.76

This table shows the same drop off in marginal contribution which was seen in Table 17. The average agent score for this running of the simulation was 85.72 with a standard deviation of 0.46. The range was 87.02 to 83.89.

A conceptual way of comparing these two NCL methods of creating a team is to think of the cut method as trying to increase the team's average score by increasing the number of positions with a score of 95 or higher. This increases the average by raising the top part of the distribution. An alternative method to raise an average score is to raise the bottom scores. In NCL Average, as the top scores are removed from the problem space, it is agents with higher scores for those positions in the ring for which the leading agents have low scores who contribute most to increase the group's average. While it does not explicitly only focus on raising low scores, the way NCL Cut focuses on including agents based on the number of positions in the ring they have with scores above 95, NCL Average by re-computing its average score by

iteratively removing high scoring ring positions effectively focuses on agents with higher scores in lower valued ring positions.

Since these two methods build 'teams' of experts rather than identify individual 'experts', the same method used to evaluate the Group Average selections will not work. In the Top 10 expert method, the agent's average score for the entire training space is the critical value evaluated for inclusion into the experts group; whereas, in the Negative Correlation Learning models the problem space is iteratively altered as agents are added to the experts group, it is computationally much more intensive.

To test this method, the training phase was run 1,000 times with 10 iterations in each execution, each execution built a team of experts and those 1,000 exemplars were used as input into a testing phase where the testing was run on a sample of 1,000 different problem spaces. This produced 1,000 average scores for each of the exemplars. This was done for both the Cut and the Average NCL methods along with a random control giving a total of 3,000 data points for the linear model.

The following table, Table 20, shows the results from a general linear model analysis of the team's score controlling for team creation mechanism along with a model controlling for the diversity within the team. The Team Score columns show that whereas NCL A and NCL C are not different from each other, both are, statistically, significantly better than a random team and since the random team is better than the Top 10 team, they are better than the Top 10 teams also. The effect size (θ=.088), however, is very small to non-existent (Cohen's small is 0.2 to 0.3). The R-squared show that a moderate amount of the total variance in team scores is explained by the mechanism used to create the team. The Team Diversity columns show that NCL C is more diverse than NCL A which is more diverse than the random team; however, the R-squared shows

that very little of the total variance in team diversity is explained by the team composition mechanism. However, note that NCL C is, statistically, significantly more diverse than NCL A however their team scores in the testing problem spaces are the same to the second decimal place. The Group size =1 beta from Table 12 is included in the following table to show how it compares with the NCLA and NCLC scores.

Table 20 Comparing NCL A and NCL C with Random Groups and Average Group Size=1

	Team Score		Team Diversity	
Model F statistic	509	<.0001	104	<.0001
R squared	.26		.07	
Intercept (Random)	94.5	<.0001	91.7	<.0001
NCLA	94.76	<.0001	92.25	<.0001
NCLC	94.76	<.0001	93.32	<.0001
From Table 12 Group Size 1	94.651			

4.1.1.3 Additional Team Composition Analysis and Problem Space Modifications

Figure 4 Heuristic Simulation Structure shows that a problem space is created during the training phase and during the testing phase. When the problem space is uniformly randomly distributed between 0 and 100, the testing problem space and the problem space which was used to select the different agents are not correlated and the fact that a group of agents performed well in one space would not lead one to believe that they would perform well in a completely unrelated space. It is possible to test how the similarity of the testing problem space to the training

problem space effects the results but varying the degree to which the testing problem space differs from the training problem space. In this dissertation, this is done by rather than recreating the problem space during the testing phase, reusing the problem space that was used in the training phase and merely randomly change a varying percentage of the values in the space. For example at 50%, for each position in the problem space a uniform random number between 0 and 1 would be generated and if its value was less than .5, a new random number would be generated and would replace the value at that location in the problem space. The baseline for this is running the simulation in the testing phase on exactly the same problem space as was used in the training phase. The testing phase runs a different algorithm than the training phase, so the outcome is not a foregone conclusion. On a sample run of the simulation, the following table shows the results.

Table 21 Individual and Group Scores

Top 10 Agents	Individual Score	NCL C Group	NCL score
1, 8, 5	86.194	9, 6, 1	642/642
1, 6, 9	86.165	12, 11, 8	332/599
9, 6, 1	86.154	2, 5, 12	203/620
9, 1, 6	86.132	11, 3, 7	126/552
6, 1, 9	86.099	5, 8, 10	85/526
1, 8, 11	86.085	4, 9, 12	51/558
6, 9, 1	86.078	7, 12, 10	39/528
1, 9, 6	86.078	8, 9, 11	31/545
1, 11, 8	86.078	11, 6, 12	22/590
1, 5, 8	86.060	3, 7, 9	19/531
Group	91.631	Group	94.955
Diversity	84.444	Diversity	94.815

A group of 'random' experts in this trial got a final score of 95.205 with a diversity score of 93.333, so in this instance it was higher than the NCL C defined expert group.

Note that the Top 10 Agents are missing heuristics 2, 3, 4, 7, 10 and 12[17] whereas the NCL C Group has all 12 heuristics represented. The 'random' group outperforms the 'experts' on the problem space which defined the experts which may be counterintuitive. This table demonstrates that the top ten agents do not outperform a random group of agents in the problem space in which they are the top 10. Note that as the NCL C team is built it is, generally, adding new capabilities whereas as agents as successively added to the top 10 group, it is repeating existing capabilities. The Top 10 and NCL C columns show the number of new heuristic introduced in that step followed by those heuristic in numeric order.

Table 22 Stepwise introduction of heuristics by Top 10 and NCL C

Step	Top 10	NCL C
1	3: 1, 5, 8	3: 1, 6, 9
2	2: 6, 9	3: 8, 11, 12
3	0:	2: 2, 5
4	0:	2: 3, 7
5	0:	1: 10
6	1: 11	1: 4
7	0:	----- all heuristic are present
8	0:	
9	0:	
10	0:	
	---- 2, 3, 4, 7, 10, 12 missing	

[17] Note that this sample also includes an example of when the top 10 includes all six possible agents with the same heuristics: 1, 6, 9 in this case.

Another approach to understanding the generally 'disappointing' performance of the Top 10 is to freeze the problem space used to define the group of experts and re-run that group of agents on the same problem space. Since it would be the same space on which they were the Top 10 agents, one could expect that they would perform better than the random group. Table 23, below, shows the incremental effect of each additional agent in the group of Top 10 agents and in a group of random agents when the problem space used in the training phase (1 iteration) is the same problem space as that used in the testing phase (1 iteration). The first three columns show the Expert group defined by the Top 10 method: column 1 is the actual set of heuristics for each agent, column 2 is the agent's individual score in the problem space, and column 3 is the group score when that agent is added to the group. For example, in the second row, 7.6.4 had the second highest individual score and when only 6.9.11 and 7.6.4 were used to define the expert group, their joint score was 91.07. The last three columns are the same as the first three except that they refer to a randomly created group of agents. By chance, the last agent in the random group was the highest scoring agent and is the first agent in the expert group, note, however, that this agent, the highest scoring agent in this problem space, makes no marginal contribution to the Generalists group score when added in the 10^{th} position. The 'equals' sign after the cumulative score means that the scores were equal to 9 decimal positions whereas when two successive values are the same only due to rounding off to the second decimal point, there is no equals sign after the number.

Obviously, the Expert group will be leading after the first row, since the first row for the experts group represents the agent with the highest score on that problem space. It is interesting to note that the Random Agent group has a higher group score by the third agent.

Table 23 Incremental Group Scores Change in Frozen Problem Space

Expert Heuristic	Expert Agent (Individually)	Cumulative Expert Score	Random Heuristic	Random Agent (Individually)	Cumulative Random Score
6.9.11	86.77	86.77	3.11.4	85.27	85.27
7.6.4	86.64	91.07	9.2.4	85.56	90.13
11.6.9	86.63	91.09	12.6.4	84.66	92.46
6.7.4	86.63	91.09 =	3.12.7	85.39	92.83
2.12.5	86.60	92.92	11.8.6	85.45	93.32
2.5.9	86.57	92.92 =	11.2.10	85.52	94.31
6.4.7	86.54	92.92 =	6.12.1	85.21	94.71
6.11.9	86.51	92.92 =	8.10.5	85.06	94.94
1.4.6	86.50	93.62	10.2.11	85.43	94.94 =
2.7.10	86.44	94.53	6.9.11	86.77	94.94 =

This observation of the Random group having a higher score than the Top 10 experts group on the problem space which defined the Top 10 experts would imply that ‘overtraining’ is not the reason for the Expert groups being out performed by the Random group. The redundancy of the heuristics in the Expert group is demonstrated by 4 of the 10 agents having no effect on the group’s score when added to the group whereas in the Random group only the last two agents are added with having an effect on the group’s score.

Table 24 Incremental effect of additional agents

Group Position	Top 10	Random	NCL A	NCL A Ind Avg	NCL C	NCL C #>95
1	86.83	85.96	86.83	86.83	86.52	707
2	86.83=	91.82	91.70	86.26	92.03	620
3	90.87	92.43	94.20	85.92	93.74	644
4	92.65	94.24	94.61	86.01	94.47	633
5	92.66	93.98 (decrease)	94.86	85.79	94.48	604
6	93.31	94.79	94.86=	85.81	94.48=	639
7	93.31	94.25	94.86=	86.01	94.56	597
8	94.33	95.08	94.86=	85.85	94.56=	598
9	94.33=	95.13	94.86=	86.04	94.89	599
10	94.33=	95.01 (decrease)	94.86=	85.87	94.89=	637

This table shows the effect of incrementally adding agents to the experts group for Top 10, Random, NCL A, and NCL C methods of selecting agents. One item that stands out in this table is that in the random group there were two instances in which adding a new member lead to a lower average score. This is caused by the new agent leading the group to a local maximum from which they cannot escape whereas prior to the addition of that agent, they were not able to reach that local maximum so they were not trapped there and were able to pass it up. Additionally, the fact that the NCL A (NCL A is the average negative correlation learning method which bases expert membership on what remaining agent has the highest score on those sites that no already selected agent has a score greater than that agent's mean score) reaches its maximum value at the 6^{th} agent even though the remaining 5 agents often have higher individual scores. In this sense, these agents preform worse as members of the experts group than agents they performed better than as individual agents. The NCL A process is ordered, the 5^{th} agent

who raised the team score from 94.61 to 94.86 had a higher score on the remaining positions in the ring (81.06) than the agent which entered the experts group at step 6 who had a score of 78.74.

These 'remaining positions in the ring' scores are not available from the table above, the NCL A Ind Avg column shows what the individual agent scored over the entire problem space whereas the 'remaining positions in the ring' score removes those ring positions in which agents entering the expert group before that agent scored better than its average. The reason to include the NCL A Ind Avg column in the table rather than the incremental score is to highlight that an agent's contribution to the group score is not solely dependent on its individual score but is also a function of what the group could accomplish without that agent. This fact is also clearly evident in the NCL C (NCL C is the cut-point version of the negative correlation learning method where the number of ring positions with a score higher than 95 are used as the basis for entering the expert group). In this example, the highest scoring agent had 707 ring positions with a score of 95 or higher; once those 707 ring positions were removed the next agent originally had 620 ring positions with a score of 95 or higher. However, the 6th and the 10th agents added to the group had higher original counts that the 2nd agent; what the 2nd agent had that they didn't, however, was ring positions with scores of 95 or higher that were different from the 1st agent.

4.1.2 Modified Sombrero Function Problem Space

The problem space for the heuristic simulation is extended to include a modified sombrero function as the underlying mechanism. The averages of the expected values for each agent are shown to decrease as the cycle size increase and the standard deviations increase through the

range of interest. This is caused by the agents not being able to 'step' over the monotonically decreasing sections of the problem space.

Additionally, freezing the problem space for the training and the testing phases is shown to still favor the 'random' group of experts over the individual agents who 'excel' in that very problem space which may imply that rather than 'over-training' being the cause of a random group out-performing the Top 10 group, the actual team mechanism causes the observed results

A 'Sombrero' function (sometimes also called a hat function, *sombrero* is "hat" in Spanish) is so named because when viewed in three dimensions it could be seen to represent a hat of the type historically worn by the Hollywood version of Mexican bandits. Its functional form in three dimensional space is

$$z = \frac{sin\sqrt{(x^2 + y^2)}}{\sqrt{(x^2 + y^2)}} \qquad \textbf{6}$$

The $(x^2 + y^2)$ parts of the formula create the circular part of the surface. The sine function portion of the sombrero function cycles at a rate of once every 2π. This inherently irrational nature ("irrational" as in "cannot be expressed as the ratio of two integers") of the sombrero's period makes this specification of function inappropriate for this analysis. Therefore an alternative form was developed.

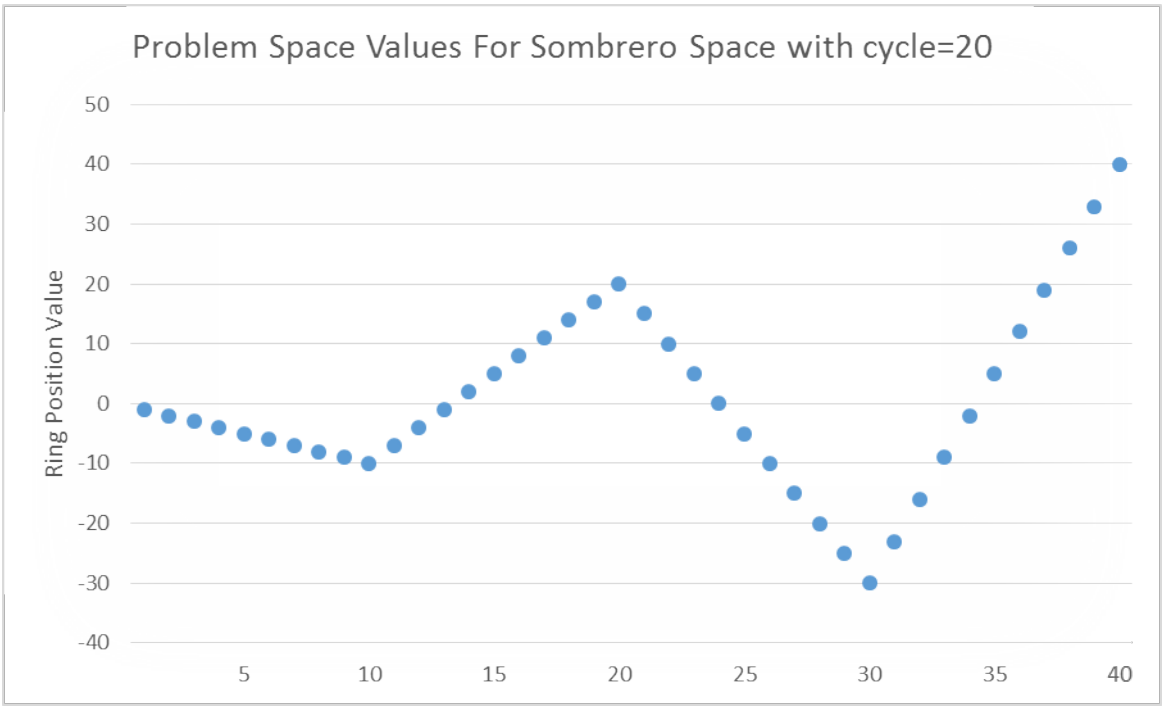

Figure 5 Modified Sombrero Function Initial Cycles

This modified sombrero function has two parameters: the first specifies the length of the cycle and the second represents the degree to which successive crests increase in value (beta for the coefficient on x term in a line connecting those crests; however is not the beta of the actual line segments which change with each change in direction). The value at 'c' (the length in the cycle) is set to 'beta'*c. The value at c/2 is –'beta'*c/2. For example if c=20 and beta=1, left most 40 positions are shown in Figure 5: first trough is (10, -10) and first crest is (20, 20). Note that the crests are on even multiples of the cycle length, c, and that troughs are at the mid-points between the crests. Additionally, note that the absolute value of the slopes increase as you move toward the right edge of the problem space, the distance between successive crests and troughs remains constant, so the increase slope means the points are more widely spread out on the line

segment. The actual slope of the line segments connecting the crests and troughs is a function of both the cycle length and the beta used.

Note that in stepping from ring position 29 to 30 the value at 30 is less than at 29 however at position 31 is the greater, so a heuristic of 1 could fail to advance but a heuristic of 2 would advance. The issue at the crests is more complicated: at 20 the next position with a higher value is at 38 (beyond 12 which is the highest valued heuristic).

Since the Heuristic Simulation's heuristics always move to the right, the modified sombrero is adjusted to be centered at the right edge of the ring (position 2,000 which is immediately followed by position 1 in the ring). Note that there is no random element in the problem space. At the far left of Figure 6, the cycling nature of the problem space can be seen whereas towards the right side of the problem space, it appears that there are 11 separate lines. Actually the top and bottom ray are separate and the interior rays are all double rays having slightly different values. The top ray represents the crest of the cycle, the next ray down shows the values for the the point just before the crest and the point just after the crest (the post crest point is slightly lower than the pre-crest point and the opposite with respect to the troughs). Since beta is 1 in this figure, the final position's horizontal value is equal to its vertical axis value.

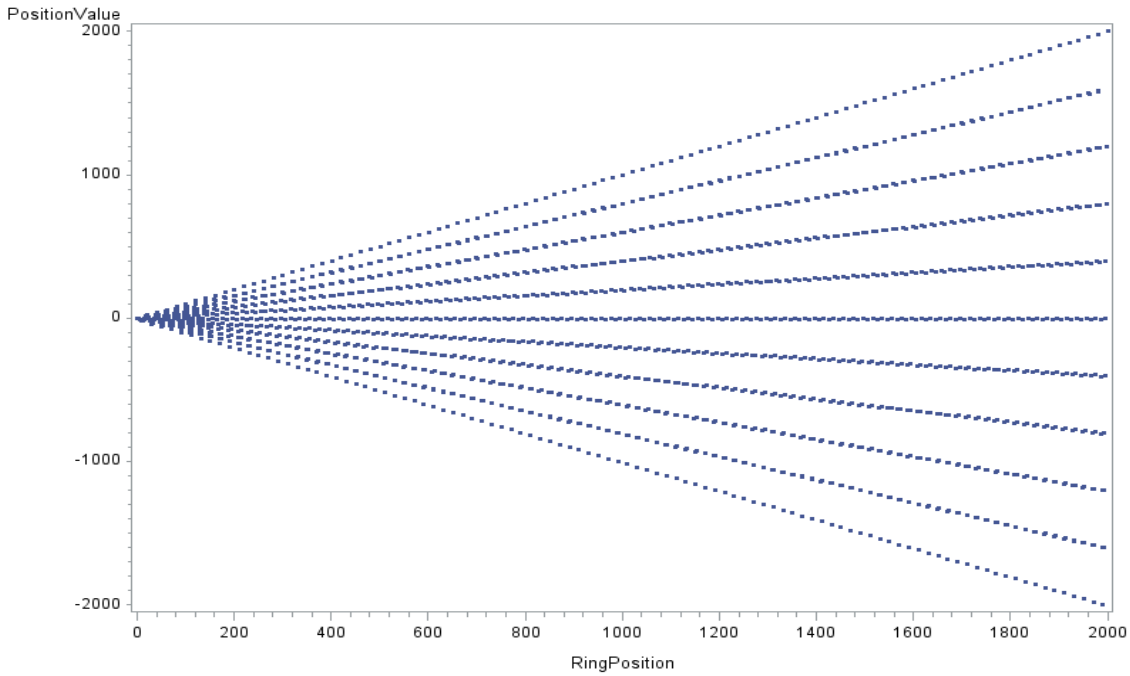

Figure 6 Modified Sombrero Function Entire Ring

The values for the last twenty values with a cycle length of 20 and beta of 1 are: position (value): 2000 (2000), 1999 (1601), 1998 (1202), 1997 (803), 1996 (404), 1995 (5), 1994 (-394), 1993 (-793), 1992 (-1192), 1991 (-1591), 1990 (-1990), 1989 (-1593), 1988 (-1196), 1987 (-799), 1986 (-402), 1985 (-5), 1984 (392), 1983 (789), 1982 (1186), 1981 (1583), 1980 (1980). The first point past the crest at 1980 is 1583 and the last value before the crest at 1979 is 1585 which appear to lie on the same ray in the above plot. The slope of the line leading to 2000 is 399 and the slope of the line leading to 1980 is 395 which is why the values are slightly different: 399 = (2000-(1990)/10 = 3990/10 and 395 = (1980-(-1970)/10 = 3550/10.

Altering the cycle size of the sombrero function has a very large effect on the average agent's score in the problem space. Table 25, below, shows the means and standard deviations over all the agents for the entire ring over the 10 cycle sizes (1,320 agents' average scores for 10

different cycle lengths = 13,200 observations): a generalized linear model shows that the cycle size predicts 93% of the variance of the agents scores; the f-statistic for the model is 20,477 $p<.0001$. The beta for this set of executions was 1, so the maximum possible score was 2,000.

The mean agent score decreases as the cycle size increases from 8 to 256 and then from 256 to 4056 (but actually any number of 4,000 will produce the same results since the resulting problem space will be a continually upward sloping line) the mean agent score increases. The standard deviation monotonically decreases from 8 on upwards. The very large values for the sombrero function are not very interesting since they describe very simple problem spaces. In general, for the interesting part of the cycle length's range, it looks as if the more rugged the problem space is (i.e., the greater number of local maxima) the higher the agents' average score.

Table 25 Cycle Effect on Agent Mean and Standard Deviation

Cycle	Mean	Standard Deviation
8	980	322
16	755	112
32	679	57
64	617	28
128	578	16
256	567	9
512	586	5
1024	646	3
2048	774	2
4056	1999	1

There is also effect on the 'best' heuristics when the cycle length changes. The following table shows the results of running a multiple comparisons Scheffe test on each of the cycle lengths. Each test consists of the 1,320 agents * 3 heuristics each for 3,960 observations per

cycle. The general model statistic, the ANOVA f test, verifies that different heuristics have different expected means; the Scheffe test then groups those agents whose means are not significantly different: Group A is the group with the highest average score, Group B the second highest, etc. The first row, cycle length of 8, is significantly different from the other rows in that the cycle length is less than the maximum heuristic and therefore once an agent gets to a local maximum it can use a heuristic equal to the cycle length to step to the next local maximum and the next until it reaches the global maximum. Starting in the second row, cycle length of 16, heuristic 12 is in Group A or Group B except for 4096; heuristic 1 is in Group A or Group B for each cycle length. The minimal significant distance between any two groups (there is up to Group F for some cycle lengths) decreases from 99 for cycle length 8 to .27 for cycle length 4096. This almost 400 fold decrease mirrors the very large difference in the standard deviations of the agents' score shown in Table 25. As the cycle length increases, there are fewer local maxima and the upward sloping and downward sloping regions become longer and longer. Once an agent gets on an upward sloping region, it will climb towards the top and as the length of that climb becomes larger and larger, its average score over the entire problem space will increase. It appears that the heuristic 12 has the highest average because it allows for longest step off a downward sloping section on to an upward sloping section. The heuristic, 1, allows to agent to reach the local maximum once it gets on an upward sloping section. This could be seen as a reflection of the experts examining the problem space with finer grained capabilities than non-experts, for the heuristic 1's importance; and the high valued (12, 11, and 10) heuristics being important to step out of regions dominated by a local maxima.

Table 26 Heuristic Groups with different cycle lengths

Cycle Length	Prob of F statistic/ R2	Minimum Significant Distance	Group A	Group B
8	<.0001/.21	99	8 (1427)	7 (1051) 1 (1038) 9 (997) 6 (969)
16	<.0001/.16	35.6	1 (864)	2 (802) 12 (787) 3 (777)
32	<.0001/.09	18.9	1 (717) 12 (704)	12 (704) 2 (693)
64	<.0001/.12	9.7	12 (641)	1 (631) 10 (622)
128	<.0001/.12	5.2	12 (592)	1 (584) 10(581) 11 (581) 2 (579)
256	<.0001/.14	2.8	12 (575)	1 (570) 11 (569) 10 (569) 2 (568)
512	<.0001/.14	1.6	12 (590)	11 (588) 1 (588) 10 (587)
1024	<.0001/.15	1.1	12 (649)	11 (647) 1 (647)
2048	<.0001/.16	.8	12 (776)	11 (775) 1 (774)
4096	<.0001/.21	.3	1 (2000)	2 (2000) 3 (1999)

Using a cycle size of 6 allows for the observation of the stepping from local maxima to the next local maxima and the doubling effect, both: the following table shows the results of a general linear model of each heuristic as the independent variable and the agent's value as the dependent value. For each agent, there are three records in the dataset, one for each heuristic.

The heuristics are represented by dummy variables. The model F's value is 64.22 (df = 11,3948) for p<0.0001, the R^2 is .15. (The number of observations is 1,320 agents times 3 heuristics = 3,960, also equals 11 + 3948 +1 from the degrees of freedom used in the F-test.) The "Effective Beta" is taken from adding the value of the intercept to the beta for that heuristic, which would be the beta if the regression was run with a 'no intercept' option.

Table 27 Heuristic Average Value Sombrero Space cycle=6

Heuristic	Effective Beta	Prob of beta=0
1	1216	0.0001
2	1028	0.0001
3	970	0.0001
4	968	0.0001
5	1117	0.0001
6	1357	0.16
7	1141	0.0001
8	996	0.0001
9	970	0.0001
10	964	0.0001
11	1058	0.0001
12	1395	0.0001 (intercept)

With a cycle length of 6, both 6 and 12 once they land on a local maximum can continue stepping on the local maximums to the global maximum, except for the next to last local maxima where 12 would step past it to the first local maxima in the ring and fail to advance at that point. The regression results above show that heuristic 6 and 12 are not significantly different from each other and that they are the two highest average values. (The effective beta column is the beta from the regression plus the intercept which is numerically the same value that is obtained when taking a simple mean for each heuristic.)

Figure 7, below, shows the average values of the different heuristics with a Sombrero problem space with a cycle length of 6 in red, with a cycle length of 4 in blue, and with a cycle length of 10 in green. Note that the cycle length of 4 has multiples at 8 and 12. The cycle length of 10 has no multiples and only one peak. Each line in addition to peaking on multiples of the cycle length also has a peak at the value 1 which always allows the agent to climb up the upward sloping sections of the problem space. As was noted above, 93% of the variance in the agents' scores are accounted for by the cycle size: this can be seen in the figure below with the length 4 cycle having a higher overall average than the length 6 cycle which is then higher than the length 10 cycle.

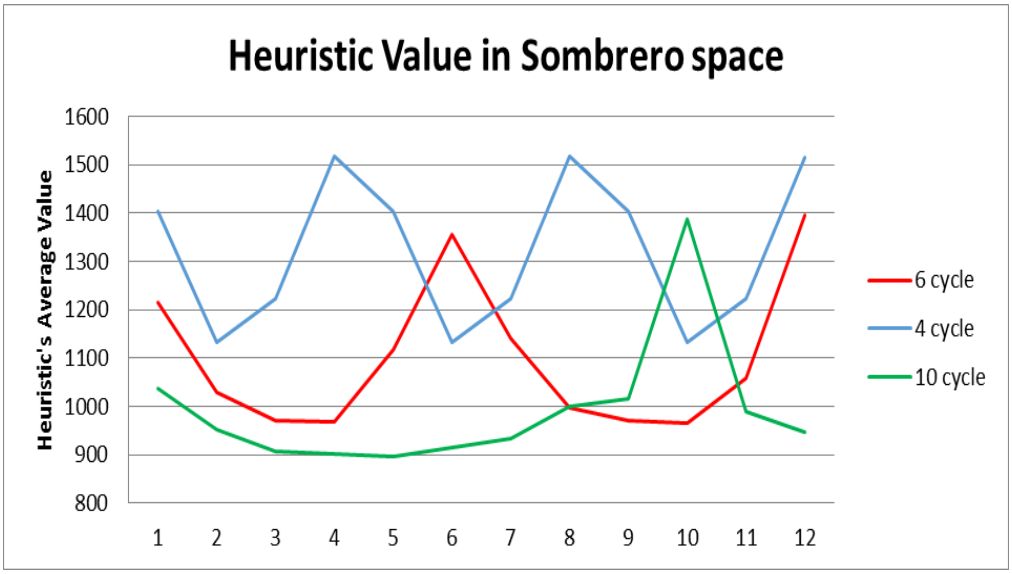

Figure 7 Average Heuristic Value Sombrero (Cycle=4, 6, and 10) Space

4.1.3 Heuristic Simulations Descriptive Statistical Analysis

A ‘basin of attraction’ in this dissertation is meant to be the area to which agents are drawn as they traverse a problem space; it could be either by some trait associated with that location in which case they are positively drawn to the basin or by some trait associated with the problem space away from that location which the agents are trying to avoid so they congregate in areas in where that negative influence is minimized. Both of these characterizations are location based; however, in a uniformly random problem space rather than focusing on a global location, smaller, local regions are the focus of interest. In a uniformly random problem space in the Heuristic simulation, there will be ring position with the highest value (there is a vanishingly small probability that this may not be a unique position which is safely ignored: the probability that the highest double precision random number in a set of 2,000 will be duplicated). Each and every agent will return this value for this starting location, since no heuristic will find a location with a higher value, and each will agent could also assign, at a minimum, three other starting locations this value, the three single step locations to the left from this location corresponding to the three heuristic each agent possesses. However, in actual fact agents often miss this. Let’s say the agent’s heuristics are 12, 6, and 1: at 12 positions away from the ring’s global maximum is will find the global maximum on the first step and never from that position, at 6 positions from the global maximum, if the value 12 steps away is higher than the position 6 steps away it will move that position and bypass the global maximum since the agent changes its base position as soon as a high valued position is found.

Obviously, this problem space position, the highest on the ring, represents a basin of attraction: all 1,320 agents report it for at least four starting positions. In one test simulation, 146 ring locations had all 1,320 agents reporting the same score for that location. The next ring

position with the highest number of agents reporting the same value had 638 agents reporting the same score, a drop off of more than 50%. Although the ring position will the global maximum will have all 1,320 agents reporting it as the value for that position, the subsequent 145 positions that had all the agents reporting the same value for that position were not the ring positions with the next highest 145 values.

The more interesting problem in this ring problem space is when very few agents report the high score for a ring position: what do they have in common? For the test simulation mentioned above when there were 146 positions for which all the agents agreed on the high score; there were 35 ring positions in which only one agent reported the highest score for that position and 71 ring positions for which two or fewer agents reported the same highest score.[18] The following table shows the distribution of the agent heuristics when only 1, 2, or 3 agents report the same highest score along with the last column which is the case when less than every agent reported the same highest score for a given ring position.

[18] In a sample running of the problem space, there were also 446 ring positions that were never reported to be return values.

Table 28 Distribution of Heuristics for selected high scoring agents

	1 agent		2 agents		3 agents		<1,320 agents	
heuristic	#	%	#	%	#	%	#	%
1	1	.95	3	1.25	6	1.12	87,667	8.13
2	0	0	6	1.87	11	2.05	94,228	8.74
3	2	1.90	7	2.80	15	2.79	91,734	8.51
4	4	3.81	2	1.87	15	2.79	84,135	7.80
5	4	3.81	3	2.18	15	2.79	85,965	7.97
6	8	7.62	27	10.90	58	10.80	94,425	8.76
7	11	10.48	22	10.28	55	10.24	94,882	8.80
8	7	6.67	14	6.54	37	6.89	86,871	8.06
9	18	17.14	32	15.58	73	13.59	89,527	8.30
10	6	5.71	34	12.46	66	12.29	89,928	8.34
11	21	20.0	36	17.76	97	18.06	91,913	8.53
12	23	21.90	30	16.51	89	16.57	86,829	8.05
Total	105		216		537		1,078,104	
Chi Square	31 p<.0001		46 p<.0001		60 p<.0001		1.5 p=.21	

Table 28, again, shows the preference for high valued heuristics with heuristics 9, 10, 11, and 12 almost always being the most frequent heuristics in the groups, the only exception being 10 in the 1 agent group. This table shows that when only 1 agent reports the highest score for a ring position (this happened in 35 ring positions) (3*21.90: the percentages reported in the table are the percentages of heuristics, since each agent has 3 distinct heuristics that number is multiplied by 3; otherwise percentage columns would sum to 300 instead of 100) 65.7% of the time that agent had 12 as one of its heuristics (randomly one would expect that number to be 25% of the time). The total of 216 for the 2 agent column reflects that 36 positions had 2 agents reporting the same high score, giving 72 agent/position combinations with each agent having 3 heuristics 3 * 72 = 216.

The final column, <1,320 agents, shows the distribution of heuristics for all agents that report the highest value associated with a ring position provided that some agents did not report that value (in this case the 146 positions for which all agents report the same high value were eliminated). The Mantel-Hanszel Chi-Square shows no significant difference between this distribution and a distribution in which each heuristic is equally likely to appear (chi^2=1.53, p=.21) whereas the Mantel-Hanszel Chi-Square tests for the other 3 columns show a significant difference from a uniform distribution of heuristics. This may be the origin of the preference which has been observed for high valued heuristics: when only a few agents are able to reach a high valued position, it might be that there is a downward sloping region prior to that high value that can only be reached from a more distant location.

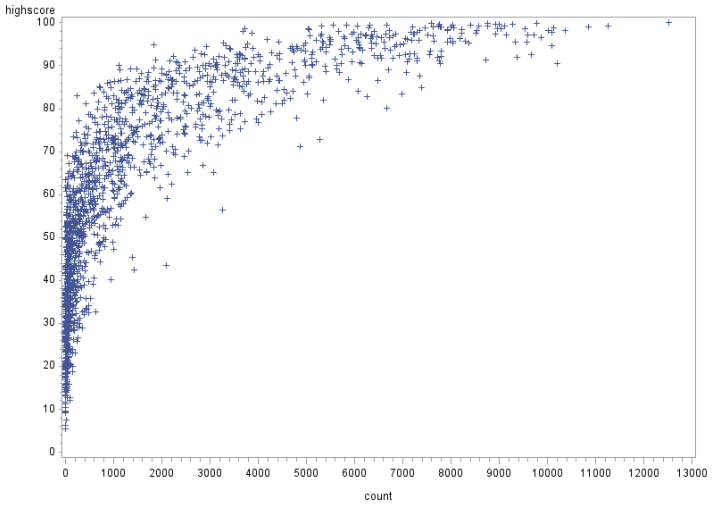

Figure 8 Frequency of distinct returned values

Figure 8 shows the distribution of returned values by the frequency of that value being returned. The vertical axis is the score associated with the ring position which ranges from 0 to

100 in the problem space (exclusive of end points) and the horizontal axis is the count of the number of times any agent returned that value from any position. There were 2,640,000 (1320 * 2000) data points used to compute the chart: the top left '+' represents the 12,524 times 99.9920, the highest value in the problem space, was returned by any agent. The far left edge of the graph has a minimum value of 6 since the agent could move to a higher position given its any of its heuristics, there are 6 permutations of those 3 heuristics all of which would not be able to move to a higher location.

An alternative point of view is to look at basins of attraction from the perspective of the problem space rather than from the perspective of the agent. Figure 9 Returned values on problem space, below, shows the distribution of maximum and minimum values returned by the 1,320 agents over the 2,000 positions on the ring. The vertical axis represents the value assigned to the ring position by the agent and the horizontal axis represents the ring position (1 to 2,000). The blue bars represent the maximum scores and the orange bars represent the minimum values returned for that ring position. Note that the blue bars tend to group together into plateaus demonstrating that the agents tend to agree on maximums in a local area whereas there is much variance in the minimums (the variance is 11 for the maximums and 536 for the minimums). The instances in the figure where the minimum equals the maximum are those cases where all the agents reported the same value for the ring position. When the average for each ring position is included (which was not done on the figure included), the average is seen, obviously, as more variable than the maximums and less variable than the minimums with a variance of 44.

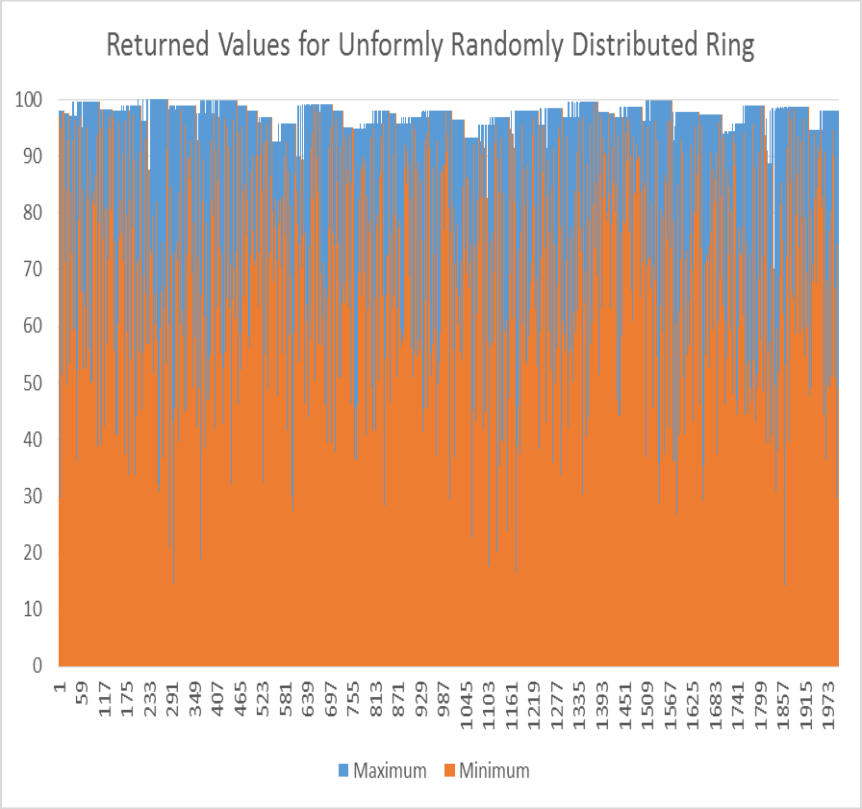

Figure 9 Returned values on problem space

An additional way to look at the distribution of scores is to plot each agent's score for each ring position, Figure 10 Problem Space, below, is a plot of each agent's value for each ring position with darker dots indicating a lower score and lighter dots indicating a higher score. Pure white represents the highest score for the entire figure and a black dot represents the lowest score for the entire plot. The vertical axis is ring position (1 to 2,000) and the horizontal axis represents agents (1,320), ordered by agents' average score over the problem space going from high (left side) to low (right).

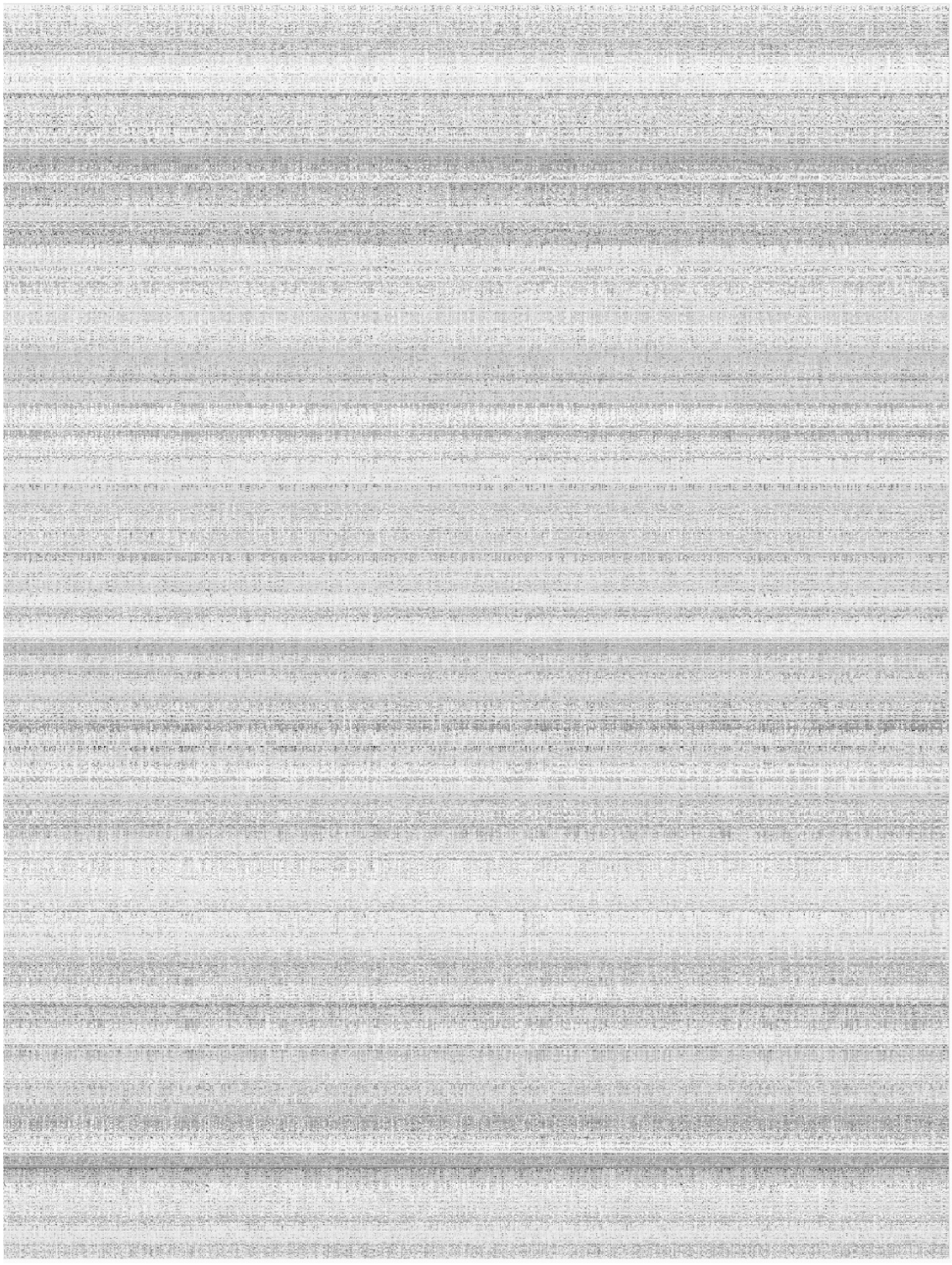

Figure 10 Problem Space

What is clear from Figure 10 Problem Space is the horizontal pattern; the ring position is what dominates the figure. The differences between the values at the different ring positions are far greater than the differences between what the different agents return as the value for a particular ring position.

In the Sombrero function problem space, the basin of attraction issue is quite different. It is a function of the cycle length if an agent has a heuristic that allows it to climb to a crest and then hop from crest to crest, it will end up at the maximum value in the problem space. For example, if the cycle size is 12, and the agent has a 12 in its set of heuristics, once it gets to a local maximum (if it gets there), it will be able to hop from local maximum to local maximum until it reaches the global maximum.

Figure 11, below, shows the values returned by all 1,320 agents. The vertical axis represents the ring position, from 1 at the top to 2,000 at the bottom and the horizontal axis represents the individual agents with agent order determined by their heuristics with 1, 2, 3 on the far right preceded by 1, 2, 4; 1, 2, 5, etc. to 12, 11, 10 on the far left followed by 12, 11, 9; 12, 11, 8, etc. Again darker indicates a lower score and lighter indicates a higher score. The Sombrero problem space associated with this figure has a cycle length of 12 and a beta value of 1. That means that the maximum value is 2,000 (would be a white spot within the figure or a white line for position 2,000 at the bottom of the figure) and a minimum value of -1,988 would be a black spot except that no agent would ever report the minimum value in the problem space as the value for the position unless its heuristics were 0, 0, 0 which is not a valid heuristic set (it will be used later to indicate an absent group member when analyzing the marginal effect of additional agents on the group's score.) Moving from the right to the left, the 10 light bands represent the agents for whom 12 is the second heuristic in their set; within each of these sections

between light bands, a single light line can be seen and that represents the agents with 12 in their third heuristic position. The area between the far right edge and the first light band represents the agents whose heuristic set starts with 1, and each successive band shows where the initial heuristic changes, for example moving from 1, 12, 11 (far left of first light band) to 2, 1, 3.

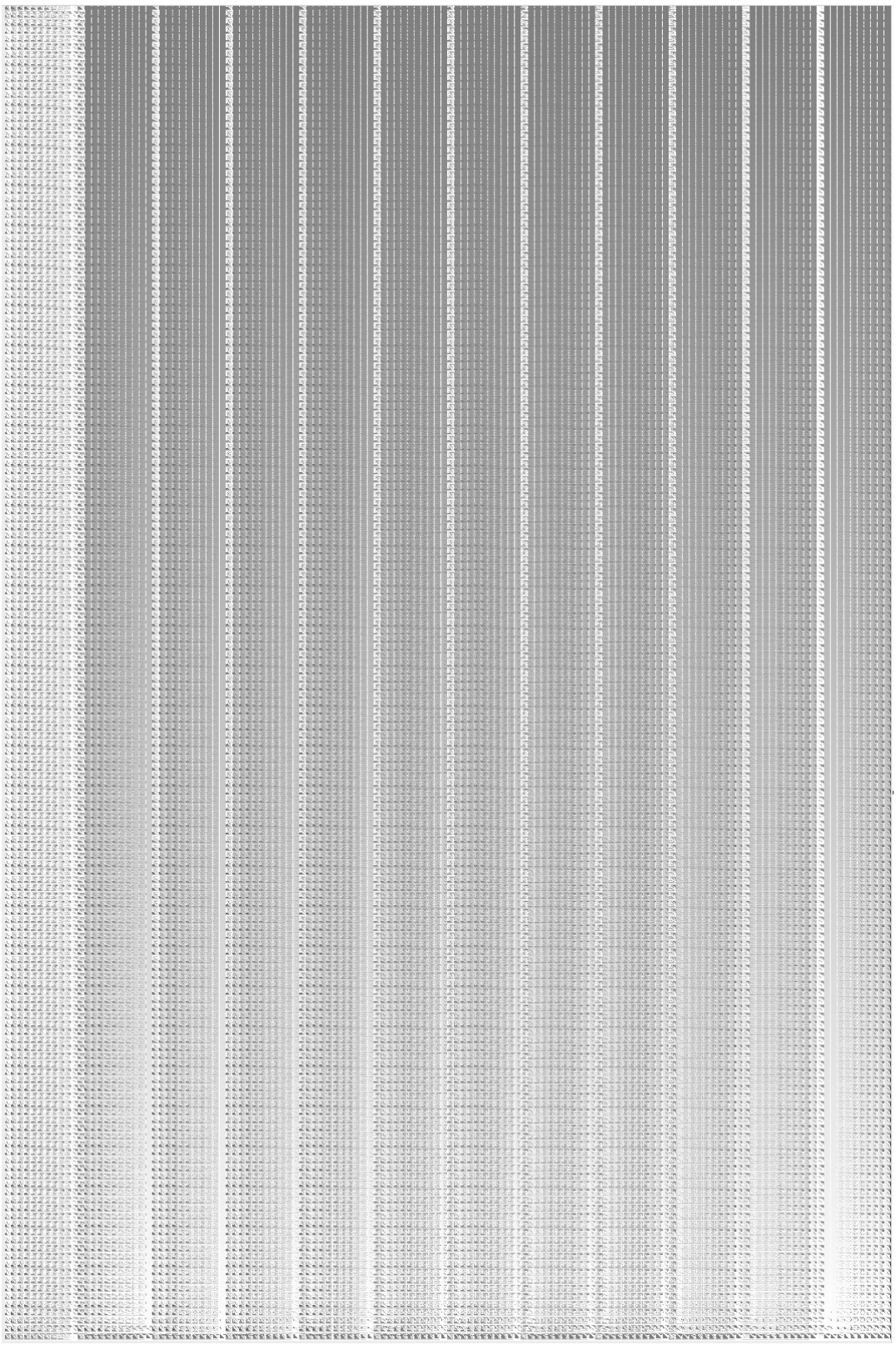

Figure 11 Sombrero Agent Space Cycle = 12 by Heuristics

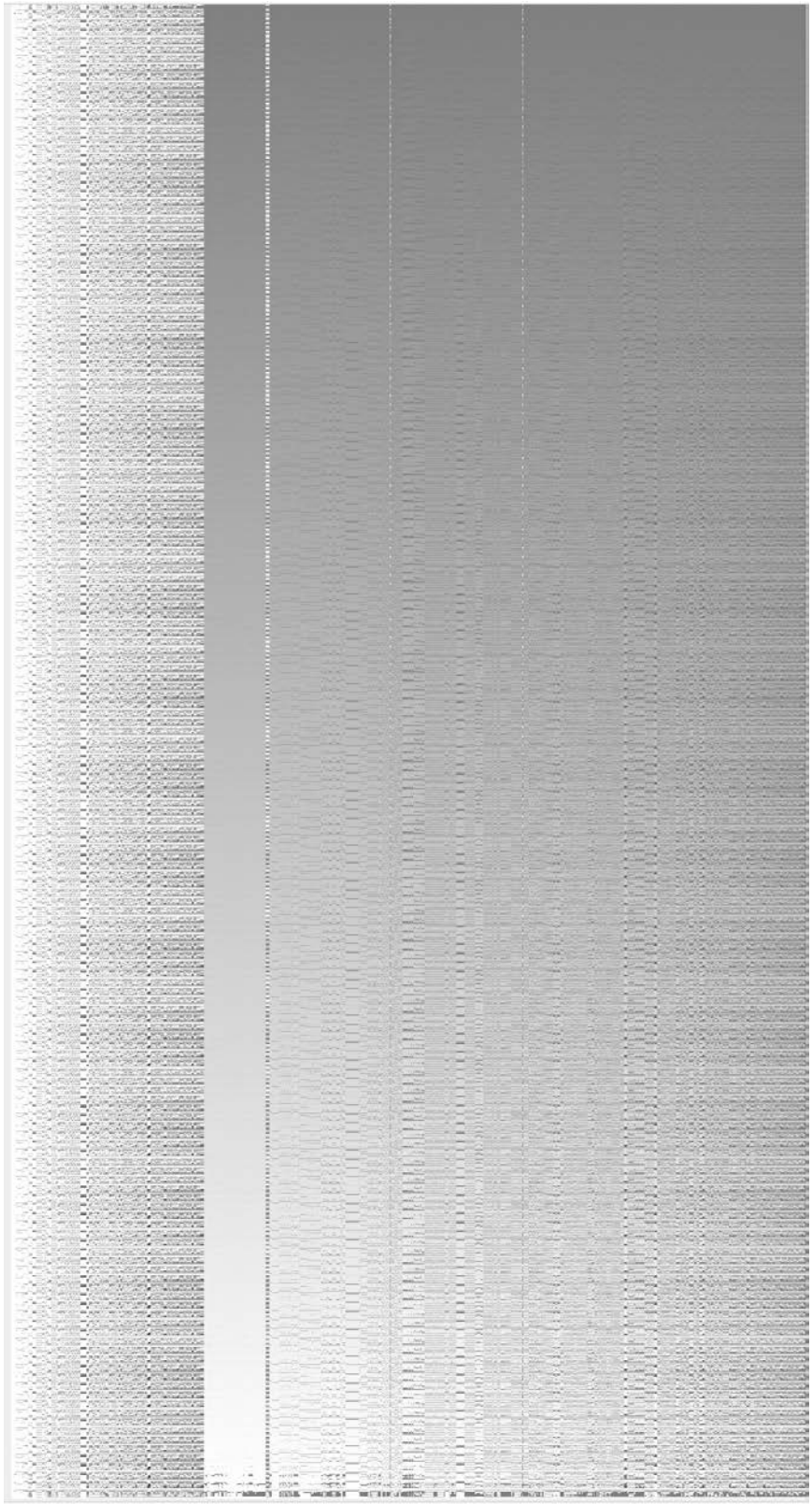

Figure 12 Sombrero Agent Space ordered by Average Score

Figure 12 represents the problem as Figure 11 except that the horizontal axis instead of being order by heuristic values is ordered by the average score that agent has for the problem space (Sombrero Problem Space with a cycle length of 12 and beta of 1). The agents are ordered from high score on the left to low score on the right. The first light grey band on the far left is a background marker to distinguish the top 6 agents who report the global maximum for all 2,000 ring positions and are consequently white lines in the figure. There is a clear line where the scores for the beginning ring positions fall; on the left of the line values are 2,000; 1,334; 669 and on the right of the line they are 32; although the difference in the agents' average score across that line only goes from 1024.6 to 1010.1. The agent's standard deviation shows how distributed the scores are.

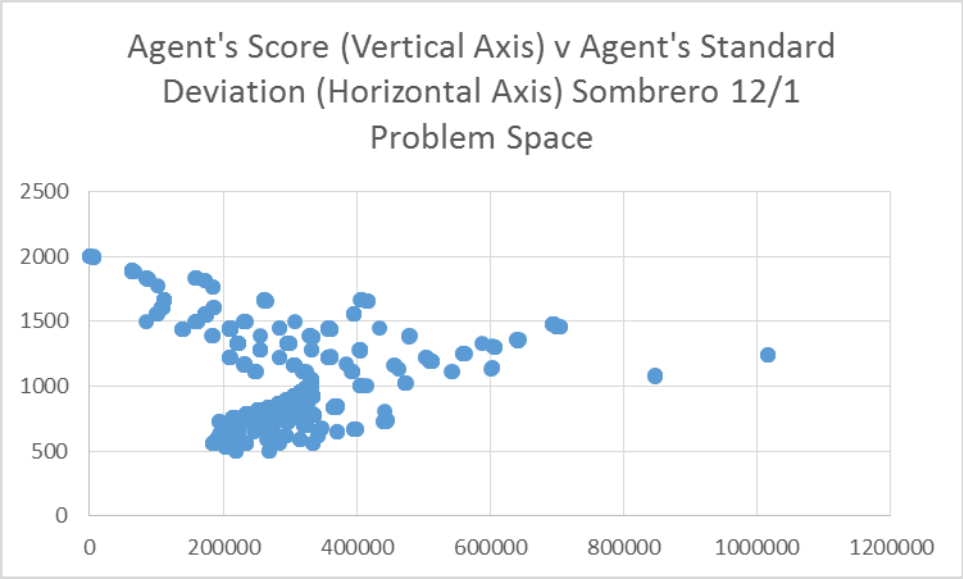

Figure 13 Agent's Score v. Agent's Variance

Figure 13 show the distribution of each agent's score and its variance with the Sombrero Problem space with cycle length of 12 and beta of 1. The extreme (high) variances belong to agents with heuristics 12, 11, 10 in their 6 different permutations. The highest being 12, 11, 10 (1,015,898) going to 10, 11, 12 (847,915) the seventh highest variance is 705,593 for 11, 9, 12. In the figure above, there are three points under the far right point and three points under the second point from the right. The lower scores (below 1,000) have less variance in their variances than the agents who score above 1,000. There are six agents with a score of 2,000 for each point on the right which gives them a variance of 0 and three with a score in the 1,995's whose variance is 5,500 to 6,500. So although the highest scoring agents have zero variance, having a score of over 1,000 increases the range of expected variance.

The other obvious artifact from Figure 12 are the three relatively narrow vertical lines that are light on the top and darker on the bottom. What distinguishes the wide light band on the left, approximately 25% of the figure, from the rest is that every agent to the left of the boundary has 12 as one of its heuristics and the right 75% does not, generally. However, those three significantly thinner lighter lines within that darker 75% also represent agents with a 12 as a heuristic. The first one has 7 agents all of which have a 12 and a 6 (the remaining heuristic is 11 once, 7 and 8 three times each). The next band is for 12, 11, 7 and the last one for 12, 11, 8. Those three lines are noticeably lighter at the top of the figure and also noticeably darker on the bottom, hence their overall average puts them in the darker 75% of the entire figure instead of in the left most 25%.

In general, no clear cut basins of attraction were found through the visual display of the individual agent's scores with the uniformly random problem space or with the modified

Sombrero function problem space. Although the modified Sombrero function problem space was significantly better organized that the Uniformly Random problem space.

4.2 FAST AND FRUGAL SIMULATIONS

The Fast and Frugal Simulation test bed has 5 cues and a random error term for each of the 15 items, called 'options' in (Luan, Katsikopoulous, & Reimer, 2012, p. 7), corresponding to cities in the original Fast and Frugal demonstration; each of the 15 items is compared with the other 14 leading to 105 pairs of comparisons (15 * 14 / 2) per iteration. The 'population' of these 'cities' is created by summing the product of a random value, U(0,1)[19], and each of the 5 cue's betas plus an error term multiplied by its beta. The value of the betas changes for the four different test environments: Large, Medium, Small, and No Difference which reflect the level of differences among the betas. In the No Difference case, all the betas are .17, in each other environment the difference between the largest and smallest beta increases from small to large.

Luan's model defines the cue weights, see Table 6 for values, to all contribute the same amount of information to the 'population' of the 15 cities. The difference is how that information is distributed with the Large difference case having approximately 10 times as much information in the first cue as in the fifth and all five cues having equal amounts of information in the No Difference case. This dissertation uses Luan's cue weights as given except in the 'perfect cues' section, page 141 and following, in which perfect cue weights are defined and then transformed into No Difference weights in a series of 20 steps.

[19] U(0,1) identifies a uniform distribution between zero and one excluding the two end points, with double precision (15 significant digits).

For each iteration, 15 ‘cities’ are created and 6 random values obtained for each, five for its cues and one as an error term and criterion scores computed for each of the cities on the basis of the actual random values. Then for each cue, the 15 values for each cue are dichotomized to 0 or 1 depending on if that particular value is above or below the median: that creates a vector of seven 1’s and eight 0’s (depending on where the median goes the 1’s signify either the larger or smaller values, it has no effect on this analysis which codes the median as 0.) At this point each ‘city’ has associated with it five cues whose value is zero or one, and a criterion score (computed from the original cue values and the error term.), and this information is then what is available to the agents to determine how to use the cues to identify which of two ‘cities’ has the larger ‘population’.

The basic simulation follows these steps.

Table 29 Fast and Frugal dichotomization simulation

Step	Activity
1	For Large, Medium, Small, and No difference: create 5 ~U(0,1) random cues plus a ~U(0,1) error term, sum the cues multiplied by the appropriate coefficient. This produces 4 sets of cues and four criterion scores.
2	Repeat Step 1 15 times for the different items
3	Create (15 * 14 / 2 = 105) sets of comparisons between items for each of the 4 environments
4	Execute the Fast and Frugal algorithm on these sets of trials; counting the number of time correct decision are made and what cue was used as the basis for that decision. The algorithm is either ‘Take the Best’ or the ‘Minimal’ selection rule.
5	Repeat steps 1-4 for each iteration of the simulation, computing an average number of correct summary and the distribution of what cue was used to make the decision.

4.2.1 The effect of the dichotomization rate

Fast and Frugal decision analysis is based on having simple questions for which there is a simple Yes/No answer. If the answer is yes for one option ('city') and no or unknown for the other, then the first option is chosen. If the answer is the same for both options, the next question is asked. In the 'take the best' algorithm, questions are asked in the order of their decreasing frequency of providing the correct response; whereas in the 'minimalist' algorithm, questions are asked in a random order[20]. The dichotomization rate is the rule for transforming a response to a question that could be a continuous value into a Yes/No response. The dichotomization rate is an artifact of the test bed; in the German cities demonstration there was, in fact, a simple Yes/No response to each question which is an important part of the underlying theory, people use the results of a single simple response to make a decision rather than a response weighted on several characteristics. Katsikopoulos (2010) does a simulation study of the Fast and Frugal decision method on both dichotomized and not-dichotomized data; however, he just uses the median as the dichotomization point without further comment.

In each simulation, with 15 values in each of four different environments (Large Difference, Medium Difference, Small Difference, or No Difference) there were 60 1st cues, 60 2nd cues, etc. To compare this to the original German cities study, the 15 different values would be represented by 15 different cities and the 5 cues would be 5 different characteristics of the cities. Since 7/15's of the cues were coded as 1 (and 8/15ths as 0), there were 28 1st cues coded as 1, 28 2nd cues coded as 1, etc. In the original study, the questions are all Yes/No questions of the form "Does the city have this characteristic?" (e.g., does it have a professional soccer team

[20] Gigerenzer's original paper (1996) also included a third algorithm that used the last cue that was used to make a distinction as the first cue.

in the premier league (Bundesliga) , does it have an inter-city train station, does it have a university, etc.); therefore in this simulation the cue values are converted to 0/1's to match the Yes/No responses of the original study. Summing over all the values and environments, however, the total number of times each cue is coded as 1 remains 28 for every simulation (i.e., in each iteration, cue 1 is coded as 1 seven times and for the four different cue environments 4 * 7 = 28). The following table shows the results of the average number of times a cue was used to make a decision (for a single iteration of the simulation) followed by the maximum number of times the cue was used and the minimum number of times the cue was used within the parentheses:

Table 30 Distribution of Decision Cues: Average (max, min)

Average (Max,Min)	Large	Medium	Small	No
Cue 1	56(56,56)	56(56,56)	56(56,56)	56(56,56)
Cue 2	26(18,28)	27(22,28)	26(18,28)	26(18,28)
Cue 3	12(5,18)	12(5,15)	12(6,16)	12(3,18)
Cue 4	6(2,12)	5(1,10)	6(0,10)	6(1,11)
Cue 5	3(0,5)	3(0,7)	3(0,5)	3(0,7)
Random	2(0,9)	2(0,8)	2(0,5)	2(0,6)
Sum	105	105	105	105

Table 30 shows very little difference in the distribution of which cue is used to make the decision across the different environments: however, what is evident is that there is no diversity in the number of decisions being made at the 1st cue level (it is always 56) and only slightly more diversity at the 2nd cue level. There were 105 (15 'cities' compared with 14 other 'cities' divided by two since 'a' compared to 'b' is the same as 'b' compared to 'a') different comparisons being

made therefore each column has to sum to 105 since if the decision was not made on the basis of any cue, it was randomly decided.

Table 31 Effect of Dichotomization

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
1	0									X	X	X	X	X	X	X
2	0									X	X	X	X	X	X	X
3	0									X	X	X	X	X	X	X
4	0									X	X	X	X	X	X	X
5	0									X	X	X	X	X	X	X
6	0									X	X	X	X	X	X	X
7	0									X	X	X	X	X	X	X
8	0									X	X	X	X	X	X	X
9	1	X	X	X	X	X	X	X	X							
10	1	X	X	X	X	X	X	X	X							
11	1	X	X	X	X	X	X	X	X							
12	1	X	X	X	X	X	X	X	X							
13	1	X	X	X	X	X	X	X	X							
14	1	X	X	X	X	X	X	X	X							
15	1	X	X	X	X	X	X	X	X							

Table 31 shows why there will always be 56 decisions being made at the 1st cue level. Assume that the agents ('cities') have been sorted by the value of their first cue, when the cue is dichotomized to 0/1 based on being equal to or less than the median (0) or greater than the median (1), the column and row heading in the table above are produced[21]. The column and row labels of 1-15 numbers represent the agents in sorted order and the 0/1's represent their recoded

[21] Note that if the median was coded as 1 instead of 0 the shaded block would be 8 high and 7 wide rather than 7 high and 8 wide, the total number of shaded cells would remain the same.

1st cue values. The shaded X's show the numbers of times one of the 15 agents has a different value for the 1st cue than the other agent. For example, the first shaded X is in row 9 column 1: when the agent with the 9th lowest value for cue 1 is compared with the agent with the lowest value for cue 1, their recoded cue 1 values will differ. The table is symmetrical around the diagonal: agent 1 being compared with agent 9 is the same as agent 9 being compared with agent 1; therefore only the lower triangle is highlighted. This rectangle is 7 by 8 which produces 56 pairs. The size of the rectangles decreases (6*9=54, 5*10 = 50, 4*11=44, 3*12=36, 2*13=26, and 1* 14=14) as the point at which the dichotomization is made decreases, which shows that using the median as the point produces the situation where the maximum number of decisions are being made at the 1st cue level. Since the 1st cue is the one with the most information in absolute terms and marginally (for large differences especially, but actually for all but the no difference environment), it biases the results towards the large difference solution making the highest possible number of their decisions where they have the highest probability making a correct decision.

Now, even though Table 31 was sorted in 1st cue order, it should be clear that how the columns and rows are arranged in the table would not influence the number of cells that contain shaded X's, they would just be randomly distributed. Now, in this simulation there are 5 such tables stacked on top of each other: each one representing one of the cues. In those cases when the decision was not made by the 1st cue, it then does matter what order the remaining cues are in. The remaining cues are randomly distributed, so although the top table represented above is in 1st cue order, the lower tables representing 2nd, 3rd, etc. cues will be randomly distributed. Even though each table has the same format of 56 shaded X's, they won't be distributed in the

same arrangement as the top (at least, there is a very, very small probability that the two sets of randomly generated cues will be distributed in identically the same order.)

There are 105 possible comparisons (15 * 14 /2) between the 15 different agents, and 56 are always resolved at the first level which leaves (105-56) 49 decisions to be made at lower levels.

Table 30 shows that the 1st cue was used exactly 56 times in all four different categories and it shows that the 2nd cue is used between 18 and 28 times for three categories and between 22 and 28 times for the remaining category averaging 26 or 27. The ratio of cases decided at the first cue to total cases is 56/105 (.533333) which when multiplied by 49 (the number of cases remaining to be decided) equals 26.13333 which is observed in the data.

In a sample of 5,000 iterations of this test bed, Table 32, below, shows the R^2's when the criterion score (the sum of the betas times the random ~U(0,1) variables) were regressed against the 1st cue only and all five cues. When all five cues are used as independent variables, the percentage of variance explained (the R^2) matches that reported in (Luan, Katsikopoulous, & Reimer, 2012)'s Table 1. Since all four classes (Large Difference to No Difference) will make 56 decisions based on the value of the first cue, this table shows that the Large Difference class by virtue of have a first clue with a high R^2 will make the correct decision more often that the other classes. However, using all five cues results in a more accurate prediction which is constant across the different environments. The betas for the criterion equation were weighted to make the same amount of variance explained by the collection of cues in the Luan's paper (2012).

Table 32 Cue1 v Cue1-5 R^2

	Cue 1 regression	Cue 1 – Cue 5 regression
Large Difference	.57	.86
Medium Difference	.36	.87
Small Difference	.22	.87
No Difference	.18	.87

The following table, Table 33, shows the effect of the dichotomization rule on the percentage of times a particular environment makes the correct decision. In the first column, the dichotomization is performed at the median which gives the maximum number of decisions to the first cue which the large difference environment is mainly based upon and consequently the large difference environment has the highest correct percentage. When the dichotomization rule is altered so that the first cue only distinguishes the 'city' with the highest value, the second cue the 'cities' with the two highest values, etc., the percentage correct drops and the other environments perform better than the Large difference environment which demonstrates that the dichotomization procedure can influence the rate at which correct decision are made independently from the distribution of the cue validity. The overall percentage of correct responses is lower with the variable dichotomization rate because fewer decisions are being made by the cues with higher information content and more are being made with lower information valued cues and finally just on the basis of a random guess.

Table 33 Percent correct with different dichotomization rules

	Dichotomization at Median	Dichotomization varying
Large	77%	65%
Medium	73%	67%
Small	70%	69%
No	69%	68%

The following table, Table 34, shows the percentage of correct responses based on the nature of the cue environment and the cue used as the top number in each cell in the table and the number of times that cue was used to determine the response as the bottom number in each cell. The data comes from a run of 5,000 iterations of the simulation. The cue validity is a function of the relationship between the cue value and the value of the variable which the cue is being used to explain. In the German Cities example, the cue "Is it the capital of the country?" had a validity of 1: the cue was only true for Berlin which has a larger population than any other city in Germany. However, the cue is only useful in distinguishing pairs of cities when Berlin is one of the cities in the pair.

Table 34 Percent correct by environment and cue and frequency cue used[22]

% correct Times used	1	2	3	4	5	Random
Large	.84 280,000	.76 130,784	.64 61,035	.59 28,373	.55 13,274	.50 11,534
Mid	.77 280,000	.73 130,749	.70 61,021	.67 28,472	.66 13,161	.51 11,597
Small	.70 280,000	.70 130,594	.71 60,928	.71 28,493	.73 13,456	.50 11.529
None	.68 280,000	.69 130,825	.70 60,808	.72 28,328	.74 13,413	.50 11,626

Each iteration of the simulation made 105 comparisons; for 5,000 iterations there were 525,000 comparisons made within each environment; the rows sum 525,000 (5,000 * 105 = 525,000 decisions and from the first row at 1st cue 280,000 + @ 2nd 130,784 + @ 3rd 61,035 + @ 4th 28,373 + @ 5th 13,274 + @ random 11,534 = 525,000). . The final column reflects that if no cue could be used to distinguish the two items in the comparison a distinction was made randomly. This random decision is totally independent of the random component used to calculate the target values. The previous section, 4.2.1, addressed why each environment had the same number of decisions made at the first cue level; what is interesting about this table is the final environment: None. In the None environment the dependent variable being estimated is the sum of six ~U(0,1) with the 5 betas for the cues all being .17 and the error term beta being .15. (The error term beta was adjusted so that the R^2 for the cues in the four different environments

[22] This table refers to which cue was used to make each decision whereas the prior table shows the overall result: .77 ~= .5333*.84 (% of times the first cue was used and was correct) + .2489 * .76 (% of time the 2nd cue was used and was correct) + … (.5333 = 280,000/525,000)

were all equal.) The fact that the percentage of times the right decision was being made increases from .68 for the first cue to .74 for the 5th cue highlights another structural artifact in the simulation.

Each cue has been recoded from a ~U(0,1) random variable into a 0/1 variable on the basis of its value being above or below the median for that iteration; therefore when cues have the same value, they lie on the same side of the median which means they are contributing a more similar amount to the item's total score than two cues on different sides of the median. Thus when the first four cues were all on the same side of the median a higher percentage of the differences in the final scores will reside in the fifth cue, which is why the fifth cue has a higher correct percentage than the fourth which is higher than the third etc. This effect isn't obvious in the Large and Medium environments because the betas decrease far more significantly, so that the 5th cue has significantly less information about the final score than the 4th cue which is less than the 3rd cue has, etc. However the final column, the decisions made on the basis of a random variable rather than a cue, all basically reflect the same ½ probability of making a binary decision correctly.

An interpretation of this effect is that when comparing two items and the main contributors to the value of those items are very similar, the contributors with a significantly lower overall influence can be used to make an effective decision. Thus when an 'expert' fails to distinguish between two items due to their being too similar; a 'generalist' may be able to correctly distinguish between the two based on contributors which by themselves only have a minimal effect on the final score, but which when the main contributors are tied provide sufficient information to make the correct decision. This marginal information becomes the distinguishing factor. In this test bed, there is a casual element, the item's score is the sum of the

random cues times the appropriate beta. So when the factors with high betas are tied, the relative value of two items' score can be distinguished by elements that provide small increments to the item's total score. But these factors can only be effective in making the distinction when the more heavily weighted factors are very similar. However, what must also be kept in mind is that (for the 'None' difference in cue betas) although the correct decision is being made at the 5th cue 74% of the time, it is only being made at that level (13,413/525,000) 2.6% of the time: decisions are overwhelming made before they reach the 5th cue: given the median dichotomization rule being used in this simulation.

Additionally note that the effect of the random factor does not change, it is approximately 50% effective in making the desired distinction. This emphasizes the point that in order for a lower valued cue to be useful in making a distinction, it must have bearing on the problem. Having a point of view that is different from the more informed points of view about a problem is only useful if the point of view is relevant to the problem and not useful if it is just randomly different.

4.2.2 Modify Distribution of cue validities

Cue validity, literally, is the number of times the cue was used to make a decision and the decision was the correct one divided by the number of times the cue was used to make a decision (the number of correct decisions plus the number of incorrect decisions). This definition merges two important aspects of the concept: it isn't just the percentage of times the cue would provide the correct response; it also includes how many times the cue is going to be used to make that decision. If a cue has a perfect correlation with the criterion variable, but it is never used to make any decisions, its validity is zero (actually undefined 0/0).

A cue's utility is a relationship between the cue's value in the set of comparisons (is it used to make a decision) and the relationship between the items' scores (does it lead to the right decision being made). The order the cues are polled influences which cues are going to be used to resolve any question; a cue at the end of the list will be polled only when all prior cues have failed to distinguish between the two items being compared and the frequency of that happening is a function of the dichotomization rate in this test bed.

In the German cities framework, "Is the city the National Capital?" has perfect validity because it is only true for Berlin and Berlin has the larger population than any other city in the collection: so when the cue is true that city also has the largest population. It is the question/response that always provides the correct answer to a different question (when answered positively). Remember the question is "Which of these two cities has the larger population?" and not "Is this city the National Capital?" In this test bed, the theory is that the higher the beta, the higher the cue's validity. In the case of Berlin, when the 83 cities' populations are regressed against the cues used, the largest estimating equation coefficient (beta) is for "Is the city the National Capital?" and it has a value of 2,836,924 (Berlin's population is 3,433,695 in this data collection) explaining 85% of Berlin's population when the intercept is included (all independent variables in the equation are binary, the dependent variable is the city's population). The R^2 for this regression is .87 which is quite similar to the .85 used in this test bed. When the populations are normalized with Berlin's population being one and all others are a proportion of Berlin's population, the beta is .83 for "National capital" with $p<.0001$, only two other clues' betas have p values below .05 with betas of .08 and .06. So the "National capital" has over 10 times the weight as the next two most useful cues[23]. However, as was discussed

[23] The second most useful cue is "Was it is site of an Exposition?" followed by the soccer team question.

previously, this question only contributes to the answer when one of cities being compared is Berlin which happens 82 times out of a potential (83*82/2) 3,403 times – not a particularly useful assay when viewed in that light.

In this test bed, the criterion score is the sum of 5 random (~U(0,1)) values times the corresponding betas. The five random values are independent which is not the case in the original German cities example[24]: even though Bonn was the (acting?) capital when it was not the largest city, being the site of an exposition and supporting an expensive soccer team are facilitated by a base population that supports such endeavors (Green Bay, of course, is a counter-example from the United States). The implications of the independence of the random values will be further discussed in the next section in which the number of cues is increased past the five used in this test bed and past the number used in the original German Cities presentation.

However, the very structure of the 'Take the best" paradigm, makes it less likely that more cues will be used to resolve a question. Previously (

Table 31), it was demonstrated that the dichotomization rate effects the percentage of comparisons that are resolved at each successive cue (when the median is used approximately 50% of the comparisons are resolved at the first cue; and each successive cue resolves approximately 50% of the remaining questions).

The highest percentage of correct responses would be whenever each cue was used it provided the correct response and that takes a little manipulation with the test bed. Rather than having the cues be distributed uniformly from zero to one and then dichotomized to zero or one

---

24

Table 8 presented the correlation coefficients among the different cues used in the original study.

when used to respond to a question, a random number (~N(0,1) normally distributed rather than uniformly distributed) was created which when less than zero lead to the cue being coded as zero and when greater than zero as one.[25] This binary value was then multiplied by the cue's beta to compute its contribution to the criterion variable. The betas were initially set such that at each level the sum of all the subsequent betas was less than the prior beta. That is to say: the beta for the first cue was .51 (all subsequent betas summed to .49), this insured that if one item in the comparison had a 1 in the first cue and zero in all subsequent cues it would still have a higher criterion score than every other possible combination, specifically the case where the first cue is zero and all subsequent cues are one. When both items in comparison have a one for the first cue, it isn't used to make the comparison; it is only when their first cue values differ that it is the deciding factor. The same is then true for the subsequent cues, the second cue's beta is greater than the third, fourth, fifth and error term betas (the error term is ~N(0,1) not uniform; however, its beta of .01 means it has no effective influence on any comparison, it only insures that no two options have the same criterion score). An example of this in the United States can be seen with paper currency: with a hundred, a fifty, a twenty, a ten, a five, a two, and a one dollar bill – if one person has a hundred and the other person doesn't (and you can only have 1 of each domination) no matter what the second person has s/he will never have more than the first, 50+20+10+5+2+1=88<100. If neither have a hundred dollar bill, and the first has a fifty dollar fill and the second doesn't, the second can never have more than the first, 20+10+5+2+1=38<50, etc.

[25] This is actually closer to the German cities example where the cues are inherently binary. Being the site of an exposition is true or false not a continuous value between zero and one. The method in the Luan study uses the continuous value of the cue times the corresponding beta to add into the criterion score, but then converts this continuous value in to a binary value based its relationship to the median value when using that cue to compare two items.

Using this cue structure, each time a cue is used, in the Take the Best paradigm, it makes the correct decision: its validity is one and the dichotomization rate insures it is used as frequently as possible. However, this does not mean that the Take the Best paradigm always makes the correct decision. When all five cues have the same value (probability of that event being approximately $.5^5$ or .03125 (1/32)), the decision was made randomly[26] which is going to be wrong 50% of the time, leading to the expectation that correct decision would be made 98.4375% (100-(.03125/2)) of the time and as the following table shows in a sample of 100 iterations correct decision were made 98.40% of the time.

Table 35 Cues used for maximum accuracy

Cue	Beta	Take The Best
1	.51	5,222 (49.73%)
2	.25	2,699 (25.70%)
3	.12	1,249 (11.90%)
4	.06	658 (6.27%)
5	.02	328 (3.12%)
Random Value	-	344 (3.28%)

The beta for the random component of the criterion score was .01 which is unrelated to the random value in the table above. The paradigm that uses a random cue to compare two items, the Minimalist, was correct 73% of the time using these betas. This highlights the fact that it isn't just the relationship between the cues and the criterion value that is important but also the order in which the cues are used to compare two items. Remember each cue always gave the

[26] This is the point at which the random variable needed to be distributed in a manner different from the binary value of the cues, since if it was distributed as a binary variable half the time the two items being compared would have the same value for the error term and the criterion scores would be equal. Implicit in the test bed is that the criterion scores differ; e.g., no two German Cities have exactly the same population.

correct response when it was used in the order prescribed by the 'Take the Best' paradigm and lead to a 98.4% correct rate for that paradigm and when used by the Minimalist paradigm, use those cues in a random order, those same cues lead to a 73% correct response rate.

Figure 14, below, shows the results of a set of simulations in which the Take the Best paradigm is compared to the Minimalist (use cues in random order) paradigm. The Minimalist paradigm is used as the number of agents making the decision increases (the Take the Best paradigm would not benefit from different group sizes since all agents would always use the same cues in the same order and thereby have the same estimate of which criterion value would be larger). The group size 0 in the figure is the 'Take the Best' agent. The horizontal axis represents steps away from the distribution of cues described in Table 35 (which is step 1) to all betas having the same value .17 in step 21 (which is the 'No difference' environment from the test bed). Each step represents a 5% change from the initial value to .17, when the initial value is larger than .17 it represents a decrease when it is smaller it represents an increase. This chart shows that as the distribution of the cues moves away from the optimal the percentage of correct responses decreases for the 'Take the Best' agent; however, what is interesting is that the last interval between step 20 and step 21 where the percentage correct falls from 86.6% to 76.7%. Step 16 is 87.8% followed by 87.5% (17), 87.5% (18), and 87.25(19); the plot clearly levels out around 16 before dropping on step 21.

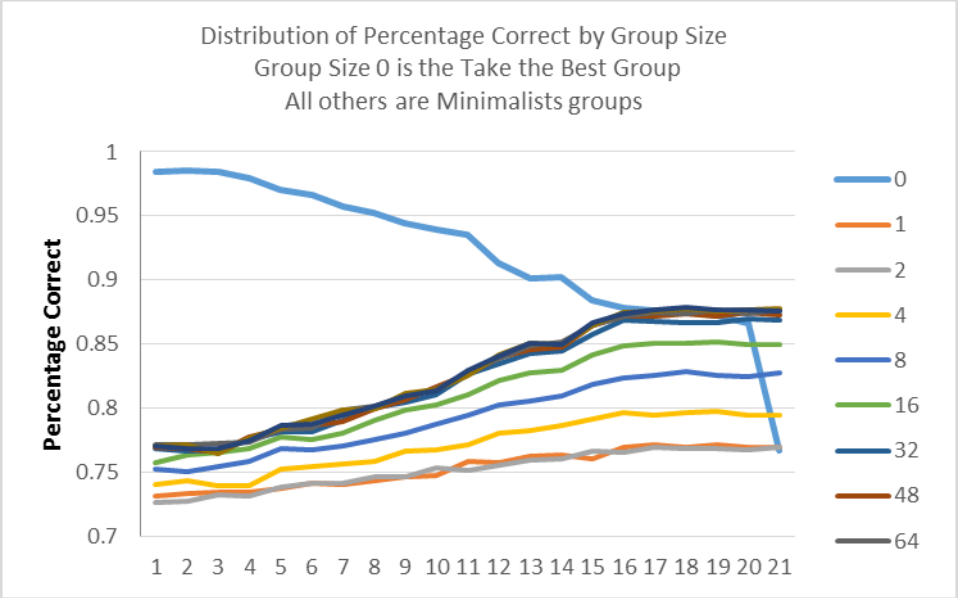

Figure 14 Distribution of Percentage Correct by Group Size

This shows that the structure of choosing cues in the order of the betas allows this paradigm to effectively extract what information is available in the cues. As is expected, when the cues all have the same value (step 21) the 'Take the Best' agent performs the same as a 'Minimalist' agent since there is no information in the cues to extract an advantage from.

Groups of "Minimalist" agents vote on which of the two items will have the higher score and if the group is tied, the decision is made randomly. Groups of 32 agents and higher have similar performance and perform similar to the Take the Best agent starting around Step 16. The different groups of agents' percentage correct plots also appear to level out around Step 16 and, of course, they stay at that level and don't fall off as the "Take the Best" agent does. Step 1 in this figure uses the 'perfect' betas described above, if the 'Large' difference betas are used instead there is a similar drop off at the final step for the 'Take the best' paradigm; the extent of this final step drop off decreases as the initial betas (the Step 1 betas) move from "Large"

difference to "Small" difference. Of course, when the initial betas are the 'No difference' betas the line is basically horizontal for the 'Take the Best' paradigm.

Another interesting feature of this figure is that a group of size two is not better than a single agent whereas by size 4 the larger group is clearly superior to the groups of size 1 or 2. The reason that the individual minimalist agent (and the groups of minimalists agents) increase from Step 1 to Step 21 is that as the explanatory power of the three bottom cues is being increased on each step and that of only the top two is being decreased, so more cues are providing more information as the chart moves to the right. The reason that the "Take the Best" line decreases is that the top two cues (which it uses for 75% of its decisions) provide less information about the criterion score in each step as the chart moves to the right.

4.2.3 Modify the number of cues

The number of cues in this section was increased to 25 with the beta weights varying from 1 to 10 in equal increments[27] with an error term with a weight of 1. The cues were distributed uniformly between -.5 and +.5. The scores for each criterion variable ranged from ~-22 to ~+22. A regression of the score on the cues provided an r^2 of .99. A ten-fold difference in the value of the cues is approximately the same as the Large Difference environment where the cue betas ranged from .37 to .04. In this simulation, 100 samples were created and each sample tested against each other (100 * 99 /2 = 4,950) for 4,950 different comparisons: using the German Cities paradigm this is equivalent to comparing 100 different cities. This simulation was

[27] For the i^{th} beta, the formula was beta(i) = (9*i)/24 + 15/24. Values are (1, 1.375, 1.75, 2.125, 2.5, 2.875, 3.25, …, 10)

executed 5,000 times and the results were that the correct decision was made 16,384,008 times (66%) and incorrect decisions made 8,365,992 (34%) times out of 24,750,000 decisions. The dichotomization rule was the cue being greater or less than zero which was an approximately 50% rather than the strict median rule (approximately because the value was a random number (~(U(-0.5,+0.5)). When compared with Table 33, which presents the correct decision rate as a function of the dichotomization rate, this is a lower rate correct than even the No difference between the cues environment when compared to the 50% dichotomization column. When compared with Table 34, which presents the correct decision rate as a function of the cue used to make the decision, this simulation does not show a pattern similar to any of the original test environments (Large cue difference to No cue difference): the highest weighted cue was used 49.99% of the time and lead to a correct choice 66.21% of the time, the next cue was used 24.99% of the time and lead to a correct choice 66.17% of the time, the next was used 12.50% of the time (66.26% correct) followed by 6.25%/66.18, 3.12%/66.24%. So the top five cues had indistinguishable correct rates.

This shows that it is not the range of the betas in the cues that leads to the differences in the correctness rates for the different environments as much as the proportional difference between successive cues. The Pearson's Correlation Coefficients between the cues and item score are represented in Table 36, below.

Table 36 Pearson's Correlation Coefficient for Cue 1 to Cue 5 and 25 cue models (p value)

	Large	Medium	Small	No	25 cues
Cue 1	.79 (<.0001)	.60 (<.0001)	.46 (<.0001)	.42 (<.0001)	.33 (<.0001)
Cue 2	.45 (<.0001)	.45 (<.0001)	.45 (<.0001)	.39 (<.0001)	.31 (<.0001)
Cue 3	.26 (<.0001)	.42 (<.0001)	.44 (<.0001)	.42 (<.0001)	.30 (<.0001)
Cue 4	.17 (<.0001)	.30 (<.0001)	.42 (<.0001)	.43 (<.0001)	.30 (<.0001)
Cue 5	.03 (=.26)	.23 (<.0001)	.38 (<.0001)	.43 (<.0001)	.28 (<.0001)

Although the 25 cue model has a higher overall r^2 (.99 v .85), each individual cue explains less of the variance in the item's score than in the 5 cue models. So when decisions are made in the 25 cue model at the level of the first cue, roughly 50% of all decisions are made at the first cue, they are made on the basis of a cue that only have a .33 correlation with the item's score leading to an overall lower percentage of correct responses. The lowest ranked cue in the 25 cue model has a correlation coefficient of .04 (<.0001) which is similar to the value of the 5th cue in the Large Difference 5 cue model. The main factor driving the ability to make the correct decision in the 5 cue models is the higher correlation between the cue values and the item's value in the cues used to make the decisions.

In the 25 cue model, the wide spread of the contributors to the item's score dilutes the explanatory power of any single cue even though the most influential cue has 10 times the influence of the least influential cue. It has only 1.04 times the explanatory power as the second cue (beta for top cue 10/ beta for next cue 9.625) whereas in the Large Difference 5 cue model the highest rank cue has 1.61 times the explanatory power (beta for cue 1 .37/ beta for cue 2 .23) as the second cue. This is true when the cues are independently distributed, which we know from the German Cities data (

Table 8) is not the case in the original real world example. In the case of a city's population, it is easy to imagine that there are fewer independent factors that influence a city's population than possible metrics. There may be mere chance historical events in some instance such as Bonn being the capital of the 'West Germany' and the Indian Packing Company's Packers remaining in the NFL without having a large population base in its home city; but, in general, either non-measured factors or an over-determined causal structure would lead to the cues not being independent.

It is possible to create groups of related cues. A group of simulations was run with 25 cues in which the cues, in groups of five, were correlated. The method for creating the correlated cues was to run a cycle of 100 items ('cities' in the German Cities paradigm) with each item having 25 cues grouped in units of 5. Each group was assigned a random value (~U(.5,1)) and a group value (5 for cues 1-5; 4 for cues 6-10; 3 for cues 11-15, etc.). This makes the first five cues have five times the weight as the last five cues when the group values are the same. Each member within the group was assigned an individual random value (~N(0,1)), this second value was divided by different amounts to control the degree of correlation among the members of the group, the individual item (a trait such as 'Hosted an exposition' in the German Cities paradigm) was then given a score which was the sum of group's value times its weight and its individual value. For example, an exemplar (a city) for cues 1-5 would have a value that was a ~U(.5,1) random number times 5 (for being in the first group) and then each separate cue would also have a ~N(0,1) random number which was divided by some discount factor added to it.

When the individual values were divided by a large number (e.g. 20), the group score dominated and the correlation coefficient was high; when the individual values were divided by smaller numbers (e.g. 1), the individual random elements had more influence and the correlation

coefficients were lower. For the criterion score, if any member of the group's individual component was greater than zero (50% of the time this is expected to be true for each individual) then the group's score was added to the criterion score. Thus if cue 1 and cue 2 had positive values the first group's score times its weighting factor was only added to the criterion score once. If none of the group's cue had positive values, that group's score was not added to the criterion score. This models the situation in which each group reflects a factor in the criterion score plus individual error terms for the separate items in the group.

In summary, the cues are combined into groups of 5 in which each group can be considered to reflect some trait, each individual element of the group of five has the same basic trait value to which an individual term is added. The criterion score is the sum of all the individual element scores (times the appropriate coefficient) plus the trait scores whenever at least one cue of that group has a positive value. For example, assume the criterion score is a composite score reflecting if an applicant should be accepted or not and assume the first trait is intelligence, the composite score could reflect several surrogates for intelligence – GPA, SAT scores, and others, all normalized in some manner. One method would be to average the 5 scores and use that; however, the method used here is to add in an amount for 'intelligence' if any of the measures indicate that this applicant has above average intelligence and then add in all the individual scores from individual instruments theoretically minus the true (common) intelligence score. So the applicant gets credit for being above average in intelligence (if any cue value is positive) and plus or minus that given how that applicant scored on each individual instrument.

The following table shows the results from a set of simulations using this structure. There were 100 items (cities) in each simulation with 25 cues for each item; this created 100 *

99/2 = 4,950 comparisons for each simulation run and there were 1,000 iterations for each simulation for a total of 4,950,000 comparisons for each column.

Each column represents how much weight each individual item kept in the final criterion score: the initial values are group score plus random numbers (~N(0,1)) then they are multiplied by .5, .2, .1 or .05 which alters how correlated the individual items in each group are. The '% correct' reflects the percentage of times the paradigm correctly identified the item with the higher criterion score. The items are correlated within their groups (cues 1-5, 6-10, 11-15, 16-20, and 21-25) since each individual item's value is the sum of the group score (5, 4, 3, 2, 1 times a random number (~U(.5,1)) for the above groups) and the individual item's random component. The correlation coefficients, correspondingly, are higher for cue 1 than cue 25 which are correlation coefficients reported in the table.

Table 37 Correlated Cues

Individual cue error term weights	1	.5	.2	.1	.05
% correct	62%	66%	72%	75%	75%
Correlation Coefficients (Cue 1 – Cue 25)	.33 - .01	.67 - .07	.93 - .34	.98 - .67	.99 - .89
R-squared	.73	.59	.53	.52	.52
Cue 1	49.99%	49.98%	49.97%	50.03%	49.97%
Cue 2	23.39%	17.45%	8.15%	4.33%	2.17%
Cue 3	11.75%	8.75%	4.03%	2.16%	1.14%
Cue 4	6.17%	5.21%	2.62%	1.40%	0.72%
Cue 5	3.41%	3.32%	1.78%	0.94%	0.51%
Cue 6	2.64%	7.66%	16.71%	20.59%	22.75%
No cue	0%	0%	.02%	.27%	.99%

The R-squared reported in Table 13 are the results from a general linear model regression of all 25 cues onto the criterion variable. Except for the .05 individual weight column, the individual betas on cues in the regression equation were all significant at the <.0001 level. In the .05 individual weight column, four of the individual cues were not significant at .05 level and only 8 were significant at the <.0001 level. The Cue 1 to Cue 6 rows show the percentage of times a decision was made at that particular cue and the No cue row shows the percentage of times that a random decision was made since all 25 cues had the same information.

The structure of this table is that the amount of randomness in the items' cue scores decreases as the columns move to the right. Methodologically, the individual cue's random component was multiplied by the individual weight before being added to the item's criterion score. Now the interesting trends within that structure is that the percentage correct increases as the columns move right from 62% to 75% while the R-squared from the regression decreases from .73 to .52. The R-squared is not directly comparable to the percentage correct; however, the trend is clear that the 'Fast and Frugal' paradigm is able to adjust for highly correlated cues whereas the regression based paradigm preforms worse as the degree of correlation increases. The regression model, however, is attempting to predict the actual criterion score whereas the 'Fast and Frugal' paradigm is only trying to choose which of two items has a higher criterion score.

The final rows of the table show how the 'Fast and Frugal' paradigm ignores highly correlated cues: as you move to the right in the table, the percentage of times the decision is based on cues 2 to 5 decreases and cue 6 which is the first cue in the second group of correlated cues increases. So as the cues are more correlated, they more frequently are dichotomized to the same value and therefore cannot be used to make a decision.

5.0 EXPLORATION/EXPLOITATION MODEL

A key issue in solving a problem is to decide where to look for a solution. Long before March's 'Exploration and Exploitation' article (1991), the problem of where to look for a solution had been addressed in a variety of fields. In Information Retrieval, Pirolli (1999) uses animal foraging theory, which of course goes back at least to the original hunters-gathers, to shed light on human information processing practices. He (Pirolli, 2007) uses Charnov's Marginal Value Theorem to identify when a forager will leave one patch and find another patch. An example could be an animal eating berries from a berry patch and the question is then at what point is it worth abandoning the current patch and looking for a new one. This is an exploitation/exploration issue: what are the conditions under which it is more profitable to exploit a known resource than to explore for new resources or in this case when does one quit exploiting this berry patch and go find another one. This is based on an assumption that the problem space is 'clumpy' meaning that solutions are not randomly distributed throughout the space but localized in patches[28].

[28] This may not always true: for example, in picking cotton one person starts at one end of the furrow and works down to the other end of the furrow since the cotton plants are 'continuous' for the entire furrow. Now, one could argue that if the goal is maximizing the amount of cotton one has in his/her sack, then a cotton picker would only pick the most easily accessible cotton bolls and would move to the next plant when the cost of extracting hard to reach cotton bolls exceeded its value. However, the cotton grower's goal may well be to get 'all' the cotton out of the field. In this case, the 'clumps' can be seen as just occurring closer together than in the case of bears and berry bushes so that the amount of energy needed to find the next 'clump' is negligible. Adamic (2007) uses network

In March's (1991) article, the issue was defined in terms of beliefs about the environment (reality). Individuals have a set a beliefs which may or may not correspond to the environment; those individuals whose belief set more closely corresponds to the environment are more successful when dealing with the environment. An institution has its own set of beliefs about the environment, its institutional code and standard operating procedures, and those beliefs are constantly being tested against the environment; when an individual in the institution has a belief set that is more successful than the institution's, the institution will recognize that fact and randomly adopt some elements of that individual's belief set. At the same time, the individuals are being acculturated into the institution and that means that they adjust their belief set to match the institution's belief set again by randomly adopting some elements of the institution's belief set. So the issue becomes what happens first: does the institution learn from the new individual entering it or does pressure to conform lead to the individual abandoning his/her (possibly more correct) beliefs before the institution can learn incorporate them into its belief set. The end result of the acculturation process is that everyone within the institution shares the same set of beliefs and the thus the institution's belief set becomes fixed even in the face of changing environmental conditions. Actually, this condition rarely happens due to turn-over, new people are constantly entering the institution bringing in their belief set. March's simulation then varies 1) the rate at which the institution absorbs high functioning individual's beliefs, 2) the rate at which individuals conform to the institution's belief set, 3) the rate at which new individuals enter the institution, and 4) the rate of environmental change.

---

analysis to analyze clustering of information in webpages with her co-author, Bhavani (2005) a CMU PhD, who initially started this analysis with his work on the distribution of information on melanoma throughout the web. Information isn't as clumpy as berries, there seems to be a dense cluster of pages that share a lot of facts and a few peripheral pages with unique pieces of information whereas each berry is fungible each piece of information may not be.

Bocanet and Ponsiglione (2012) modified March's model so that the internal environment of the institution was modelled as a Kauffman NK landscape and did not find an optimal solution to the exploitation/exploration issue. The internal environment was defined by the relationships between the individuals within the organization and the organizational code (the institution's belief set). A Kauffman NK landscape is a landscape in which there are N different parameters and the value at anyone point is determined by the relationship between K of those parameters. If K=1 then there is one global peak, as K approaches N the landscape becomes more rugged. In the berry patch model, one could imagine a landscape that had a large food source available but only if reachable from one other berry patch with a specific set of skills which itself was only reachable with a different set of skills in which case exploiting this resource would require both sets of skills.

Lazer and Friedman (2007) extended March's model focusing on a network structure among the individuals within an institution and found that a dense network lead to information being distributed more quickly but that a less dense network lead to information more aligned with 'reality' to eventually dominate: much like March's finding that when acculturation was slower, the final level of congruence with 'reality' was higher than when acculturation was faster, although in the short run the slower acculturation rate lead to worse performance than a higher acculturation rate. Their model allowed high functioning individuals to directly influence other individuals rather than requiring that their influence be mediated by the institutional code thus an individual's links to others in the institution was the key factor rather than some explicit measure of that individual's influence on the institutional code. Bocanet and Ponsiglione also allowed 'fitness' to be transferred from individual to individual bypassing the institutional code step in March's presentation.

5.1 EXLPOITATION AND EXPLORATION PATTERNS

In this simulation, exploitation is defined as searching in the near neighborhood and exploration as exploring as far from the current position as possible. In the basic framework of the Hong-Page simulation, heuristics ranged from 1 to 12 and each agent had a set of three heuristics. The process that was followed was that each agent would use the value of its heuristics to step through the problem space and whenever it found a location with a higher value would restart the process at that position and when no higher value was found after cycling through all three heuristics it would return that value as the value associated with the starting position. It would repeat this process for each of the 2,000 starting positions in the problem space. There is a critical step in that description of the agent's process that needs to be further explored.

The heuristics associated with agent have an order, first through third. If the agent applies the first heuristic and finds a position with a higher value and moves to that position, what is the next heuristic that the agent applies? The rankings of the agents depends on if the agent restarts with the first heuristic or continues with the next heuristic after finding a position with a higher value. This relationship was demonstrated by running the basic Hong-Page simulation 1,000 times and regressing the value returned for each starting position in the problem space against the values of the first, second, and third heuristics. The 1,000 regression coefficients for each of the heuristics were then compared: the average value of the coefficient for the first heuristic was -0.0045498 in one sample running whereas the regression coefficients for the second (0.0116232) and third (0.0167393) were both positive. This pattern of a negative coefficient for the first heuristic and positive coefficients for the second and third was

consistently found for all of the executions of this simulation (>10 times). T-tests verify that the means of these three sets of coefficients are different at the <.0001 level.

The dependent variable (the 'y" in algebra's y=c + mx) in this regression is the value being returned for the problem space (a higher value is better) and the values for the first, second, and third heuristics (the independent variables) are integers from 1 to 12 (inclusive). The average value for each of the heuristics is 6.5 ((1+12)/2), each of the three heuristics has the same distribution characteristics. Thus the average of the regression coefficients being negative says that the higher the value of the first heuristic the lower the score for that position on the problem space ring; whereas the positive values for the second and third heuristic say that larger heuristics lead to high scores for that ring position. Figure 15 shows the distribution of the coefficients for the three heuristics.

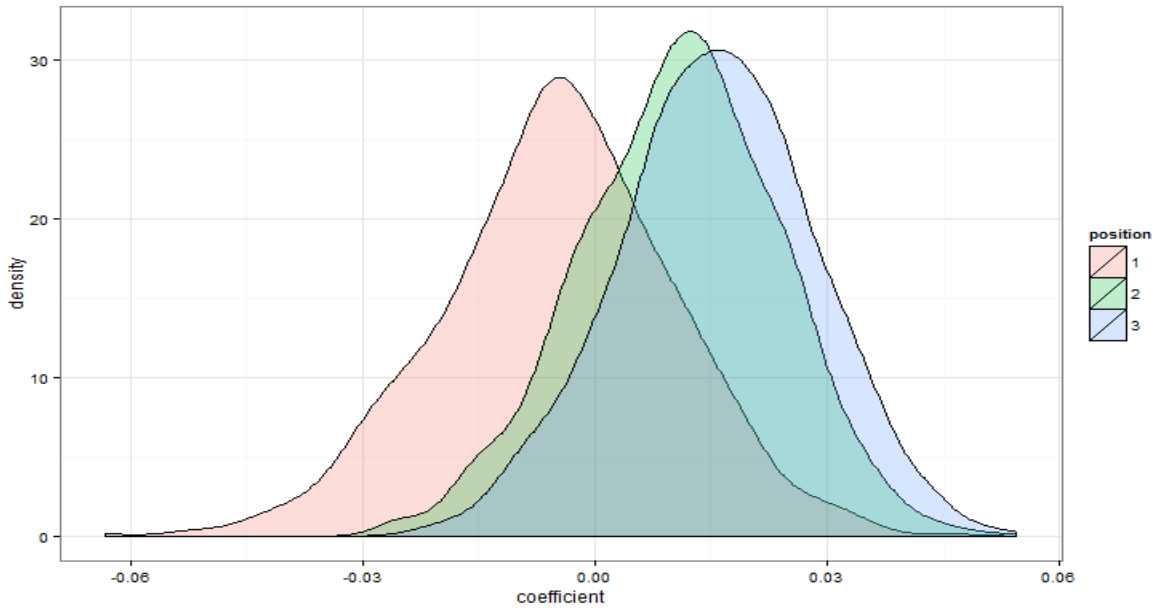

Figure 15 Distribution of Heuristic coefficients

If the implementation of the model is altered so that when an agent finds a larger position on the ring, rather than restarting from its first heuristic to search for a new value, it continues with the next heuristic, the pattern noted above changes. For example, assume the agent is starting from position 1 with (10, 1, 2) as its set of heuristics, it first looks at its current position to get its base value (let's say 50), it then checks position 11 (1 + 10: the value of its first heuristic), let's say the value at 11 is 75, then 75 becomes the base value and it checks position 21 (11 + 10). The alternative would be once the base was reset at position 11 to 75 to then check position 12 using the second heuristic. In this example, one can see that since the first heuristic is larger than the next two, the agent will never explore the positions between its starting position and 10 plus its starting position when the first step of 10 finds a higher value than the base value. If the first heuristic was 1, it would not have 'jumped' over any position that possibly held a value even higher than the one found. It would have 'exploited' the local environment first and then 'explored' more distant ring positions. This implementation leads to a different distribution of coefficients when the average agent values are regressed against the value of the heuristics. In the case above, after the agent finds a higher value at ring position 11, it starts looking for a new high ring position but instead of having (10, 1, 2) as its heuristics, the order has been rotated so that it now searches in (1, 2, 10) order. The same effect could have been accomplished by instead of always restarting with the first heuristic letting the starting number change to reflect the next heuristic in order. This rotation happens each time the agent finds a high value, but when it fails to finds any higher values and moves on to calculate the value of the next position in the ring, its set of heuristics reverts to its original order (10, 1, 2) in this instance.

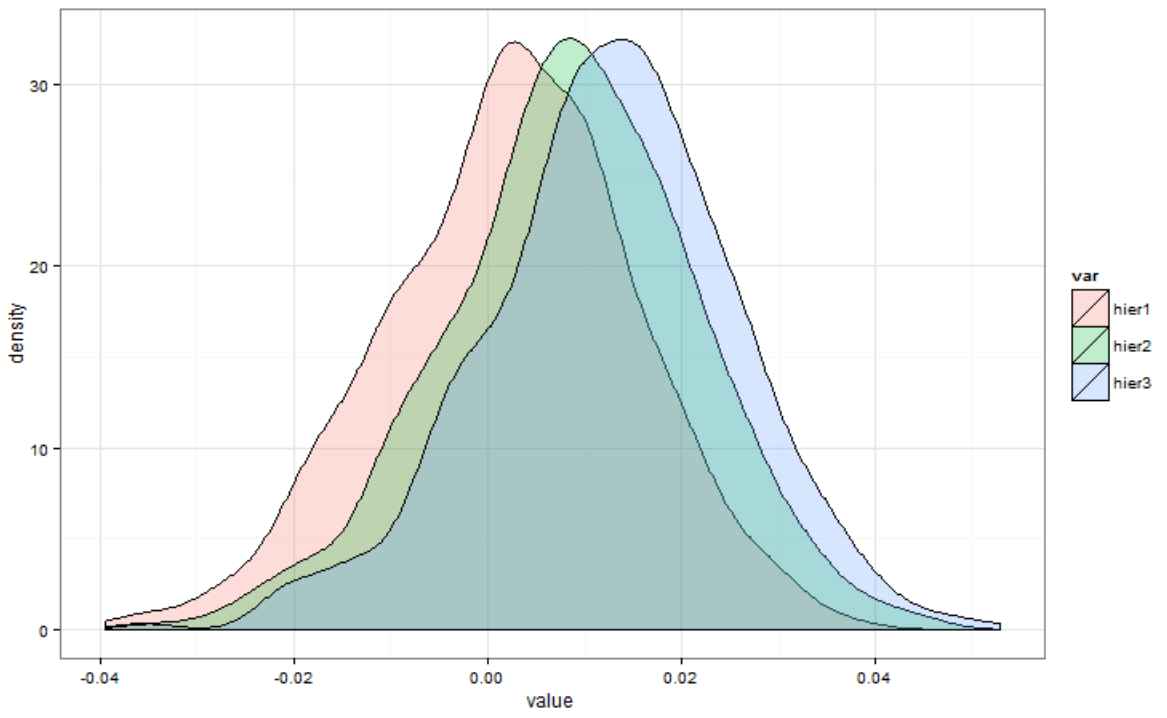

Figure 16 Distribution of Heuristics coefficient with rotation

Figure 16 still shows the average value of the 3rd heuristic coefficient being larger than the second's which in turn is larger than the first's; however, the main artifact to be observed is that the absolute value of the first heuristic is above zero in this chart whereas it was negative in the implementation without the rotation of the heuristics implemented. In this implementation of the rotation procedure, the agent restarts with its 'first' heuristic whenever it starts to evaluate a new ring position (2,000 times in each simulation). This biases the frequency with which the first heuristic is used.

The rotation implementation can be modified to eliminate that bias by not having the heuristics reset to the original order when starting at a new ring position. The following table shows the distribution of times a heuristic (by position not by value) was used to check if an

alternative position on the ring had a higher value or not if the check ended in improving the positions score. It shows that the final rotation implementation (r2) removes the bias to using the first heuristic which was only decreased by the initial rotation implementation (r1).[29]

Table 38 Heuristics (by position) used with different Rotation definitions

Heuristic Position	No Rotation (r0)	Rotation within a ring position (r1)	Rotation at all comparisons (r2)
1	2,276,644 (59%)	1,785,929 (48%)	1,244,645 (33%)
2	982,206 (26%)	1,150,000 (31%)	1,244,440 (33%)
3	574,568 (15%)	804,433 (22%)	1,244,186 (33%)

The above table comes from simulating the system one time for each rotation type, 1,320 agents, in the 2,000 position ring. When the average score for each agent is regressed against the rotation used, the model F-statistic is significant at <.0001 and the first and second rotations (r0 and r1 in the table) resulted in statistically significantly lower average scores than the third rotation method (the t-tests for the regression coefficients are significant at the <.0001 level); however, the first and second rotation methods are not different from each other at a statistically significant level. Figure 17, below, shows the mean agent value over the problem space with different rotations (r 0 – no rotation, r 1 – rotating within position, r 2 – rotating within position and between starting ring positions), different heuristic positions (p 1 – 1st heuristic, p 2 – 2sd heuristic, p 3 – 3rd heuristic), and heuristic values (1 – 12) when simulated 1,000 times. Each simulation had the normal 1,320 agents and each agent had its normal three heurisitcs; each agent then contributed three data points to this figure: its rotation pattern was set by the

[29] Actually a minimal bias still remains since the searching at position 1 always starts with the 1st heuristic.

simulation, its average agent value was determined by its average over the entire problem space, however what changes in the three data points are which heuristic is being identified (p 1, p 2, or p 3) and what is the value (range is 1 to 12) for that heuristic. The dotted lines represent no rotation, the continuous lines rotation with ring positions and the dashed lines continuous rotation throughout the problem space. The red lines represent the heuristic (horizontal axis) being in the 1st position, the golden lines the heuristic being in the second position and the green lines the heuristic being in the 3rd position.

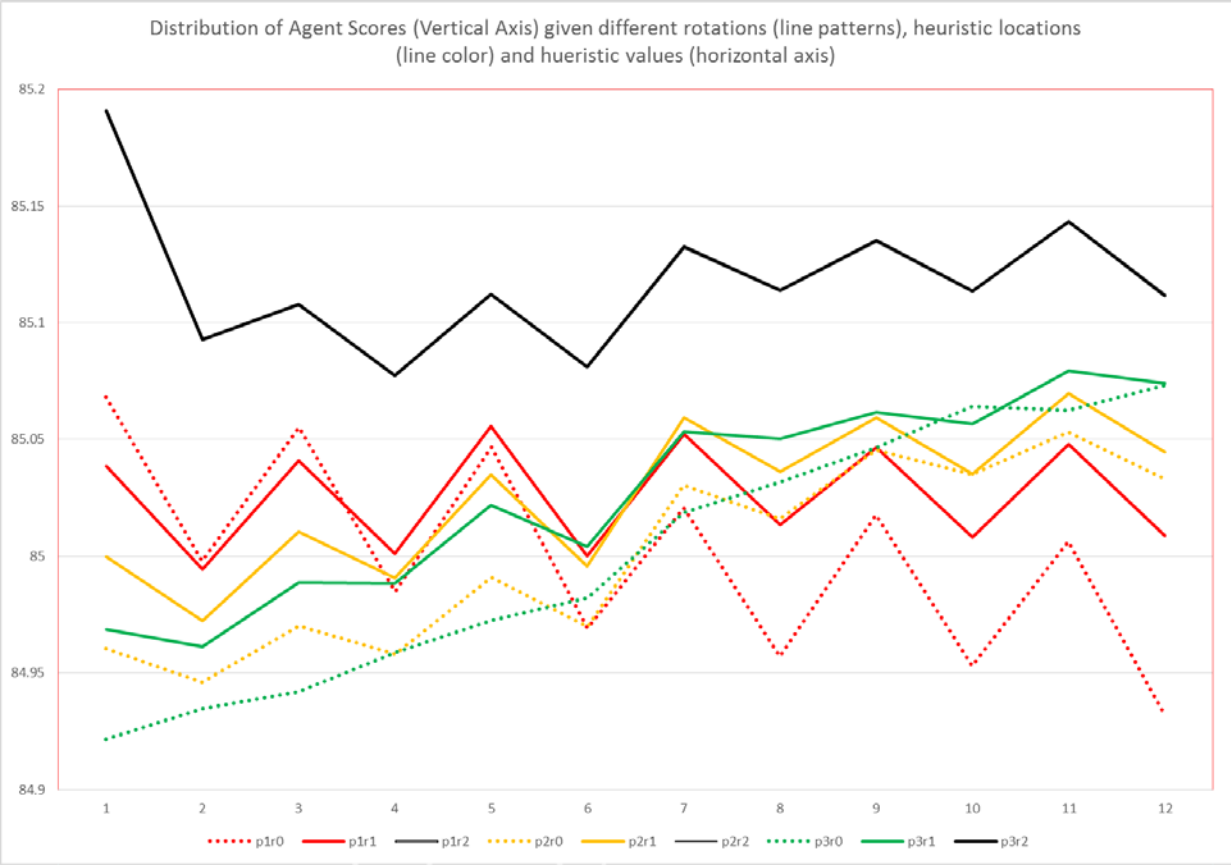

Figure 17 Distribution of Agent Scores given different parameters

The figure shows that the solid black lines (all three lines lie on top of each other), representing continuous rotation (r2), are always the highest. It also shows that the red lines are generally distinct from the green and gold lines although they cross each other at the 6/7 zone, showing that the lower valued heuristics result in high scores when they are the early in the group of three heuristics and lower scores when they are in the later part. An ANOVA shows that the continuous rotation (r2, solid black lines in Figure 17) has a larger average value than either the basic no rotation or the first rotation model ($p < .0001$) and that the first rotation (r1) model has a higher average value than the basic model (r0) ($p < .0001$). This finding demonstrates that rotation patterns contribute a significant amount of information. The 'exploration/exploitation' effect is demonstrated by 1) the order of the heuristics (1^{st} heuristic, 2^{nd}, and 3^{rd}) switching order as the value of the heuristic changes from 1 to 12 for rotation patterns r0 and r1. Rotation pattern r2, however, has virtually no heuristic order effect as seen by all three lines lying on top of each other as would be expected since it is only at ring position 1 that there is a preference for the first heuristic, each successive step to the determining the value for ring position 2,000 just depends on what the next heuristic is given the last one used. That being said, the black line does show that the ring position values for heuristics 7-12 are generally larger than those for 1-6 with the sole exception of heuristic value 1. In this case, heuristic value 1 is always associated with the highest score regardless of its position.

Table 39 Top and Bottom 10 r0 and r2 agents

r0 top				r2 top			
Heuristic 1	Heuristic 2	Heuristic 3	Value	Heuristic 1	Heuristic 2	Heuristic 3	Value
1	5	10	85.14	3	1	5	85.49486
11	2	10		1	5	3	85.49483
5	1	10		5	3	1	85.49475
1	4	5		3	1	7	85.47483
1	3	5		1	7	3	85.47482
3	4	10		7	3	1	85.47451
5	10	1		8	3	1	85.466
1	9	10		1	8	3	85.4659
3	4	9		3	1	8	85.4658
1	3	10	85.13	3	2	4	85.4502
r0 bottom				r2 bottom			
12	8	4	83.95	4	12	8	84.51
6	4	2		12	8	4	
9	6	3		8	4	12	
3	2	1		9	6	3	
12	4	8		3	9	6	
6	2	4		6	3	9	
9	3	6		6	4	2	
3	1	2		2	6	4	
8	6	2		4	2	6	
11	4	7	84.28	6	3	12	84.58

Table 39 shows the top and bottom ten agents from the previous simulation. The 'value' column represents the agent's average score for the 2,000 ring positions over 1,000 iterations of the problem space. 'r0' is the original Hong-Page rotation rule and 'r2' is the continuous rotation rule. The 'r0 top' quadrant (upper left) shows the exploration-exploitation effect just discussed. Heuristic values of 10 or larger appear 8 times and they are the 3rd heuristic in all but one case (in

which 2 heuristics of 10 of above appear, so they could not both occupy the 3rd position) and heuristic 1 appears 7 times and it is the first heuristic in 5 cases. In 7 of the ten cases, the heuristics are ordered from smallest value to highest value. In the r0 bottom (the 10 agents with the lowest scores), the opposite holds true: heuristics of 10 or above appear in three cases and in each case they are the first heuristic. Heuristic values of 1 or 2 appear in 7 times and in only one case are one of these values found prior to a larger value (when they both appear in the same case, they cannot both in the 3rd position).

In the upper right quadrant (r2 top), the agents are appears in the expected sets of three. Even more so that was expected: the heuristic 3 followed by heuristic 1 is present in the top 9 agents. The reason the values are shown in the detail presented is to highlight the fact that the groups of three score very similarly. When the top nine agents are assigned into their three groups, the ANOVA section of a Generalized Linear Model regression is significant at more than the .0001 level ($f(2,6)=41{,}812.1$; $r^2=.9999$). The equivalent regression cannot be run for this result with the r0 rotation results since there is not natural grouping of the agents into subgroups: the most obvious grouping factor would be sharing the same heuristics (1, 5, and 10) which occurs in three agents, but the other 7 don't naturally fall into two other groups. The set of bottom scoring agents has not been investigated prior to this point in this dissertation. It is interesting to note that the r0 set of the 10 worst agents have a repeating pattern, the lowest scoring 4 agents are repeated in the second set of the worst scoring agents with their heuristics in a different order. The stability of this pattern though multiple simulation runs and its possible origins are not investigated in this dissertation. Additionally each of these four combinations have the larger value equal to the sum of the smaller two which is noted here but not further explored in this dissertation.

Another interesting pattern from Figure 17 that was not commented upon was the zig-zag nature of the lines: odd value heuristics were always higher than their two adjacent even value heuristics. This pattern was consistent in multiple iterations of the simulation (the simulation being defined as the 1,000 iterations of the basic model). The table above represents the actual data used in that figure and this pattern which in the figure represents the results of all 1,320 agents is present in the top 10 and bottom 10. In the r0 top group, there are 19 odd values out of a possible 30 and in the r0 bottom there are 19 event numbers. In the r2 top there are 25 odd values and in the r2 bottom there are 23 even values.

The overall average scores for all the agents over 1,000 iteration of the problem space show that the r2 rotation has a higher average value than the r1 or r0 rotations: r2 is 85.12, r1 is 85.03, and r0 is 85.00. A generalized linear model shows that the differences between r1 and r0 with r2 is statistically significant at the greater than the .0001 level. ($f(3,3.9 \text{ million})=10204.5$; $r^2=.005$); however, the effect size is not significant. When r0 is compared to r1 and r2, r0 and r1 are statistically different at the .0001 level also.

Thus the r2 rotation method leads agents to more effectively exploit the problem space than the original r0 method. Using the original Hong-Page diversity measure, the diversity measures for the top agents with r0 is .76, r1 is .81, and r2 is .84. This extends the Hong-Page 'diversity trumps ability' finding, slightly. The Hong-Page diversity findings was that an expert group of the ten best scoring agents in a training problem space would be outperformed by a random group of agents and that the random group of agents would have higher diversity than the 'expert' group. This finding is in line with the 'greater diversity is associated with a higher score' but shows that a different mechanism for selecting the 'expert' group that picks out an

expert group with a higher individual average also has greater diversity than the original expert group. So that even among experts the diversity means better performance holds.

To review the Hong-Page diversity metric, it counts the number of times that a pair of agents have different heuristics in the same heuristic position and divides that by 3 to obtain a metric. A diversity metric of 1 means that the two agents had no heuristics in the same positions, there cannot be a diversity score of 0 since that would mean that all 3 heuristics were the same which is not allowed in the structure of the simulation, so .33 is the lowest score and that means that the two agents differ by 1 heuristic/position combination. (1, 2, 3) and (3, 1, 2) would have a diversity metric of 1 and (1, 2, 3) and (4, 2, 3) would have a diversity metric of .33.

The above figure showed that with the r2 rotation pattern, agents tended to have score similar to other agents with the same cycle of heuristics just starting the cycle at a different heuristic. This would lead one to think that a diversity metric which claimed that the three agents with the same were perfectly diverse needs to be altered for this rotation pattern. (3, 1, 5), (1, 5, 3), and (5, 3, 1) were the top three agents with very similar scores and a diversity metric should be sensitive to this similarity. To address this issue, a secondary diversity metric was created that compared pairs of heuristics and counted the number of times the pairs were different, divided by three and that the score. In the case just cited, comparing the top two agents, 3 to 1, 1 to 5, and 5 to 3 exist in each agent therefore the diversity score is zero. In comparing the 3rd to the 4th, 3 to 1 was the only pair that existed in both with (1 to 7 and 7 to 3 being the other two for the 4th and 1 to 5 and 5 to 3 in the 3rd) so its diversity metric was .66. This metric has the values 1 when no pairs match, and .66 when one pair matches, and 0 when all pairs match (two pairs matching means the third will also match, so there is no .33).

This modified diversity metric is .90 for r0, .88 for r1, and .60 for r2; thus it moves in the opposite direction as the Hong-Page diversity metric since the r2 pattern has a higher average score for the top ten agents than the r0 rotation.

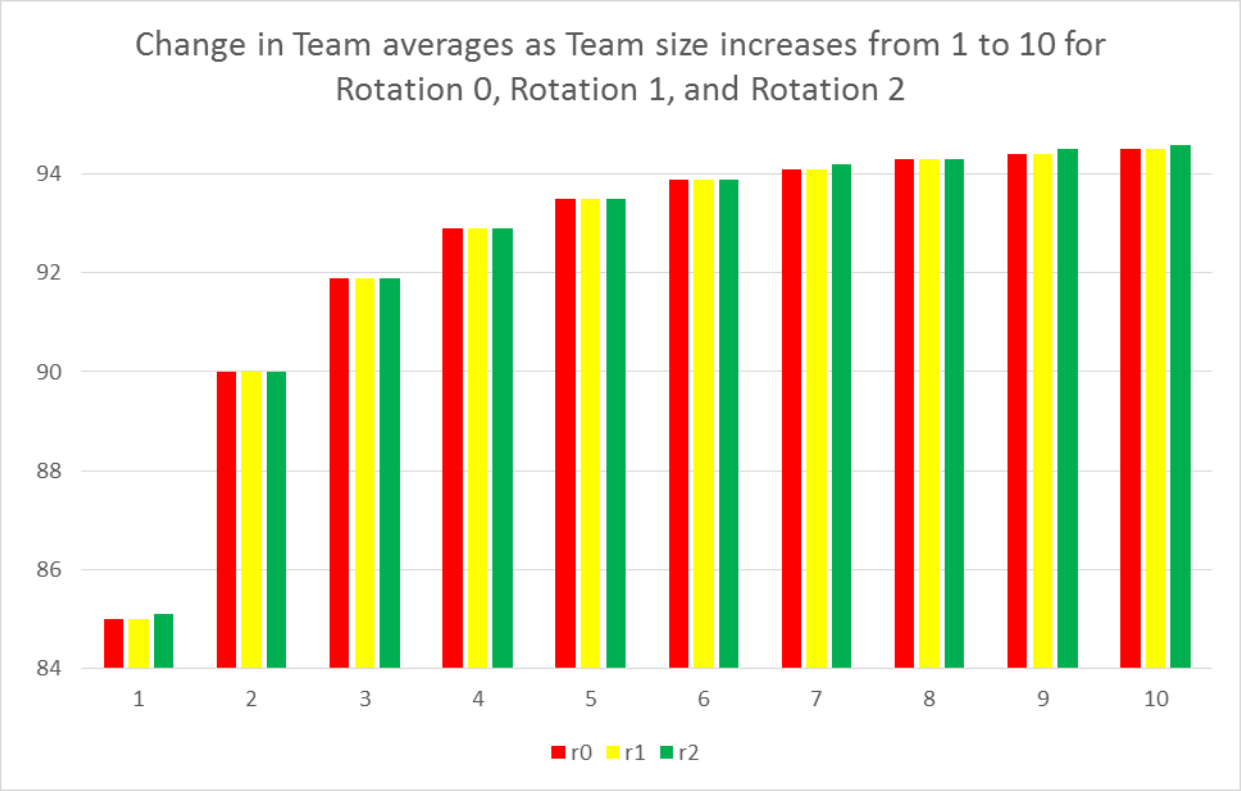

Figure 18 Group Scores (Vertical Axis) by Group Size (Horizontal Axis) and Rotation Pattern

The rotation used applies to the individual agent; however it does have an effect on the group scores when agents are combined into groups. The following simulation tested 132 groups of ten agents (all 1,320 agents were randomly placed into groups of 10) for 100 iterations of the simulation; now there was an additional factor of 10 that used to show the effect of group size, each group started with only its first member and after each iteration of the simulation the next

group member was added. These simulations were performed for the three different rotation patterns.

The figure above clearly shows that group size, which varies from 1 to 10, dominates the differences observed among groups with different rotation patterns. Although there is a very slight, but statistically significant at the $p<.0001$ level, effect caused by rotation pattern in which the continuous rotation pattern (r2) has a high average than the other two rotation patterns which are not statistically different from each other.

5.2 DECISION POINTS

When multiple agents are grouped together, a similar type of rotation issues needs to be addressed at the group level in addition to that already discussed at the individual agent level. If a group is composed of 10 agents each with three heuristics, is the group a 'super-agent' with 30 heuristics or is it a group of ten independent agents. In the case of the 'super-agent' with 30 heuristics, the rotation issue is exactly the same as for the individual agent with only 3 heuristics. Since in the Hong-Page model there are only 12 different heuristics, the 'super-agent' would have multiple copies of some heuristics. The same logic as demonstrated above would lead to high valued heuristics in the 1st position leading to a lower average score. There are $2.03 * 10^{38}$ different permutations of 3 heuristic agents into groups of 10 (1,320 * 1,319 * … * 1,310); so the same exhaustive simulation is not feasible to demonstrate this fact. However, in addition to the rotation issue from the single agent example, the question of how are the agents ordered must be addressed. This issue is not explicitly addressed in the Hong-Page simulation, since when the agents are randomly selected their order does not matter; however, when the agent group is the

ten ‘best’ agents, it is assumed that they ordered according to their average scores on the ‘training’ dataset.

At this point, if the agents are ordered then the ‘super-agent’ with 30 heuristics and the group of ten agents in a row will always yield the same value for each position on the ring: since as soon as a higher value is found, the search terminates and the base position is shifted to that position, the rotation rule then becomes operative as to what the next heuristic will be. If the rotation rule is that the agent always restarts with the first heuristic, then the ‘super-agent’ and the row of ten agents yields the same results; provided, of course, that the ten agents’ in a row heuristics are in the same order as the ‘super-agent’s’ 30 consecutive heuristics. An obvious alternative would be to let each agent determine its value for the current ring position and choose the highest one (when the ring is defined as random values uniformly distributed between 0 and 100, the probability of two positions having exactly the same value is vanishingly small). This option is quite different when applied to a single agent; if a single agent was allowed to select the heuristic which pointed to the largest value as its next step, the order of the heuristics would be rendered immaterial and each of the 6 permutations of a set of 3 heuristics would yield identical results. In the case of choosing the agent with the highest value for any ring position is just asserting that the order of the agents doesn’t matter.[30]

[30] If the picture one has of these agents is a group of investigators in a research and development facility, one can easily imagine both cases – the agents are not all equal the senior agent’s solution will be used before a superior junior agent’s or that all agents will hop on to the most promising looking development path at each point of comparison.

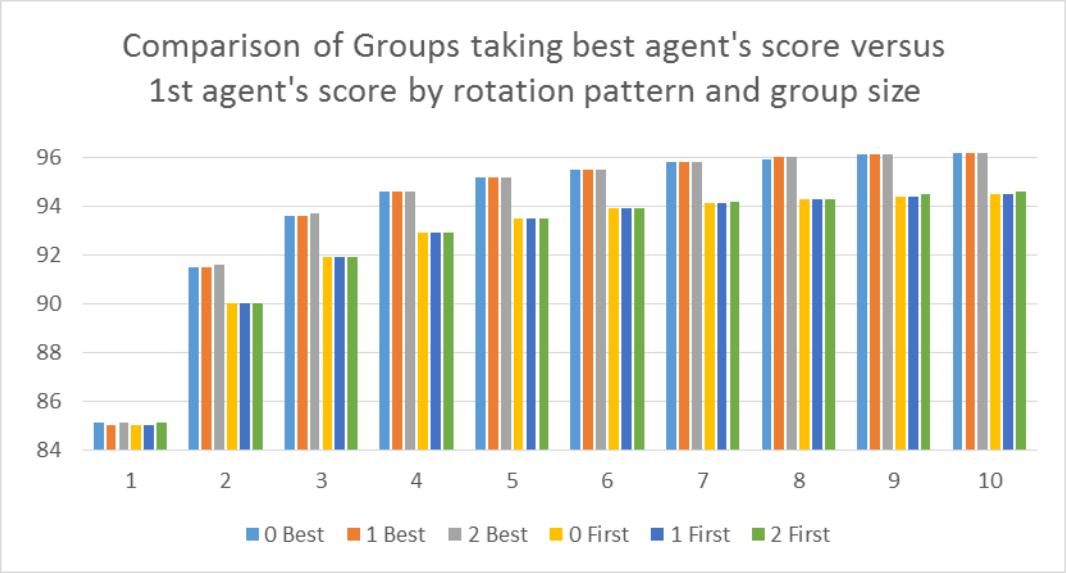

Figure 19 Best Agent's Score versus First Agent's Score

This table shows the results from simulating the group using the “Best Agent’s” score at each position in the ring rather than the first agent’s score, varying group size from 1 to 10 and altering the individual agent’s rotation pattern through the three formats discussed. ‘0 Best’ is the group with rotation pattern 0 and chooses the best agent’s score at each step in determining the value of a ring position. ‘0 First’ is the group with rotation pattern 0 and chooses the first agent’s score that is higher than the base position in determining the value of the ring position. The difference between the similar groups is very stable from group size 2 to group size 10 for rotation pattern 0, 1, and 2: the ‘Best’ group’s score is 1.6 or 1.7 points above the group that used the first agent with a higher score as the new base position. Both sets of groups are identical with the group size is 1. There is no statistical difference among the three different rotation patterns when the “Best Agent’s” score is used rather than the “First Agent’s” score.

5.3 DIVERSITY EFFECT

What the “Hidden Profiles” literature shows is that information is generally not shared when groups make decisions and consequently inferior decisions are made. Specifically, it is the information that is most widely held that influences the decision and information held by a single member of the group having the least influence. This effect can be demonstrated in this simulation by removing any heuristic that only one agent possesses. This will, obviously, happen to all three of the agent’s heuristic when the group has only one member and then probabilistically decrease as the number of agents in the group increase. This simulation is run on the 1,320 being randomly assigned to 132 groups of 10 each. Each of these groups is then sun through the problem space 10 times with 1 to 10 agents in the group each time for a total of 100 simulations for each group and 132,000 (132 groups * 10 sizes * 100 iterations) simulations for each of the three different rotation models. The group’s score is its average value over all 2,000 ring positions.

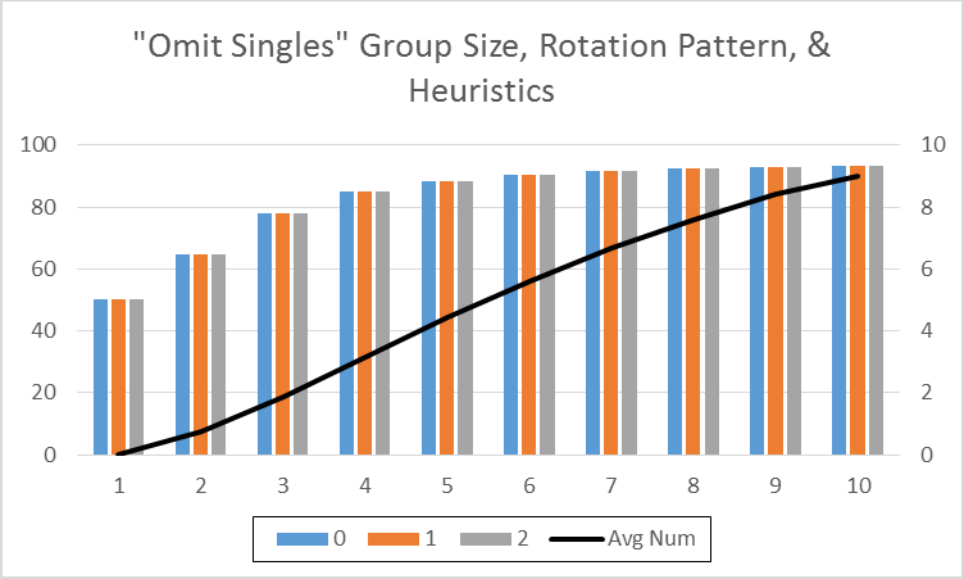

Figure 20 "Omit Singles" by Group Size, Rotation Model, and average number of heuristics per group

Figure 20 shows the results of omitting any heuristic that is only referenced by one member of the group. In groups of size 1, that, of course, means omitting all heuristics, so the average score for groups of size 1 is approximately 50 which is what the three bars show with their scale being marked on the vertical axis on the left. The black line shows the average number of heuristics being used for groups at each of the ten different sizes and its scale is represented by the vertical axis on the right. For groups of size 1, all heuristics are only used once therefore the average number of heuristics used for groups of size 1 is zero. The groups' scores at size 10 is approximately 90.3 which is approximately 96 average for "Best Agent" and 94 for "First Agent" shows in Figure 19. An ANOVA test showed that there was no difference among the three different rotation patterns for these results. An interesting finding in this chart is that the black continues to rise almost linearly after the increase in the group's scores start to level off around group size 6 at which point the average group only 5.6 out of the 12 possible

heuristics. So yes it does show that more diversity a higher score is obtained; however it also highlights that fact that the marginal benefit of increasing diversity decreases rapidly.

Alternatively to omitting single instances of heuristics, decision making in groups has been shown to settle to group's mean when the superiority of any particular decision cannot be convincingly demonstrated in real time (Gigone & Hastie, 1997). This can be simulated with the current model by using the median of the group since it reflects an actual point in the problem space whereas the mean of all the agents in the group will most likely represent a value not in the problem space. When the same simulation is run with each agent determining its decision as to the value for each point in the ring, the median of those scores can be found and that point used as the starting point for the next iteration until no improvement is found. For example, all agent when starting at one return the highest value as being the one at 1(3 times), 5(3 times), 7 (4 times) and the values for those positions are in the same order, the value at 5 will be the median and the agents will then all restart from ring position 5. When this simulated 100 times for each of the 132 groups making up groups of ten agents, the average for no rotation is 88.34, for rotation pattern 1 88.35, and for rotation pattern 2 88.36. These means are not significantly different from each other with an ANOVA test with a generalized linear model ($f(2; 39{,}497)=1.6$ $p=.20$). This score is approaching the score for the average individual acting individually (~85 from Figure 18, for example).

This mechanism, effectively, allows individual agents to use the group to remove their lowest scores and replace those lower scores with the group's median score; whereas the protocol behind the original model allowed each agent the opportunity of replacing their score with the higher score obtained by the first member with a higher score. The 'best agent' or 'best heuristic' modifications improved that model by selecting the 'best' agent when it was agent

based, or removed the agent basis and created a ‘super-agent’ with 30 heuristics and let the agent choose the heuristic with the best value to the operative heuristic. These are all, obviously, greedy algorithms; an agent only moves when it can obtain a higher score in one step. The different mechanisms merely adjust the criteria for which of the improving movements to make.

6.0 DISCUSSION

This dissertation has presented a review of several decision making simulations and their extensions to group decision processes. The parameters for the different models were explored to demonstrate the sensitivity of the results to the specific parameter values chosen.

6.1 HEURISTIC SIMULATION

The Hong-Page Simulation originally compared the top ten agents from a training phase with a random group of ten agents and demonstrated that the random group of agents outperformed the top agents. The conclusion they drew from their simulations was that the randomly chosen group demonstrated a wider diversity of heuristics than the group of the top ten agents and that, therefore, 'diversity trumped ability'. This dissertation extended the analysis of the Hong-Page test bed by altering the mechanisms that were used to select the 'experts' group and altering the problem space. It differs from the Hong-Page analysis by primarily focusing on the value of the heuristics and the distribution of those values rather than the agents' diversity score to explain variation in the groups' average score over the problem space.

6.1.1 Group Average

The mechanism used to select the group of experts was varied to allow less overlap among the heuristics its members. One mechanism was labelled 'Group Average' which, instead of using the average value for an agent over several iterations of the training problem space, used the number of times the agent ended up in the top group for that problem space (see Table 10 for a definition of how this process was accomplished.) The results of these experiments demonstrated that the group of experts with the highest average score were those agents who were chosen by identifying the agents who were most frequently the highest scoring agent in a training run (cf. Table 12). The difference between this mechanism and the Hong-Page mechanism goes back to Galton's initial articles about 'one man one vote' (1907 a). By counting the number of times an agent was among the highest scoring 10 agents, the only thing that mattered was that the agent was among the highest scoring agents and not the value of that agent's score. Galton's initial point was that the median was a better measure of a group than the average since the average was subject to be unduly influenced by an extremely low or extremely high example. This result doesn't use the median as recommended by Galton, but it removes the weight that extreme values might have which was Galton's goal. However, there is substantial literature (Soll & Larrick, 2009) recommending the average as the most efficient mechanism, which this finding appears to contradict. How frequently different agents appeared together in the 'top ten' group was investigated to identify clustering among agents: the observed co-occurrence of groups was higher than would have been expected with a random distribution of agents, but not as severe as was expected, in the 10% range rather than the 50% range which is how agents cluster within the same problem space

6.1.2 Negative Correlation Learning – Cut Point

The construction of groups was then based on the level of negative correlation between the agents in a training space. The first method was called Negative Correlation Learning Cut Point (see Table 15 for the exact description of steps involved in this process). In general, the problem space has an expected value of 50 for any particular position on the ring; however, the expected value returned by an agent for any position on the ring is in the low 80's. Therefore all agents were run through the problem space and ranked on the number of ring positions for which they returned 95 or higher. The highest scoring agent was added to the NCL C group and those ring positions for which the agent scored 95 or higher were removed from the problem space. Removal was accomplished by resetting the value of the problem space at those positions to zero, so the spaces remained for heuristics to traverse, but they would only be returned as the highest value when all other values were also zero. The logic behind this was that the group possessed the skills necessary to answer those problems and now the group needed agents who could obtain higher scores at different sites. The process was then repeated with 1 less agent, the highest scoring agent from the previous iteration, and the reduced problem space. This process continued until the group had ten members. Table 17 shows the results of one simulation. The table also shows the overlap that remained among the agents even when the process was explicitly trying to remove that overlap. The highest scoring agent had 95 on 642 of the 2,000 ring positions, the second agent when those 642 ring positions were removed had 296 positions with a score of 95 or higher; however, that agent originally had 595 ring positions with a score of 95 or higher, giving an overlap of 595-296 or 299 positions which is approximately 50% of its 95 or higher positions which is why in the preceding section a 50% overlap was expected among agents in the top ten groups.

6.1.3 Negative Correlation Learning Average

A second negative correlation based mechanism was defined labeled Negative Correlation Learning Average. In this method rather than being a function of the number of sites with a score over a cut value (95), the agents were run through the simulation and each agent's average score calculated (see Table 18 for the exact steps taken for this method.). The agent with the highest score was put into the NCL A group and the ring positions for which it obtained a score higher than its average score were removed from the ring and the process was repeated until the NCL A group had ten agents.

Table 20 showed that both of these methods produced teams with higher scores than a random team or the 'top ten' team; however, the degree of improvement was negligible, on the order of one tenth of one percent improvement, and although not part of the regression, the average value for the Average Group Size 1 group is included which is approximately midway between the random Group score and the two NCL groups.

6.1.4 Incremental agent effects among Negative Correlation Learning Models

The effect of the groups' diversity and the relationship between the training problem space and the testing problem space were also investigated. The general structure of the simulations with a randomly distributed problem space is to generate a new, completely random, problem space for every iteration. This makes the testing problem space completely independent from the training problem space. However, it is reasonable to assume that the training problem space should somehow be related to the testing problem space. In the framework of the problem, we have an instrument that measures some factor and we want to see if the group of people who score high

in that instrument consequently perform well in the testing problem space which is assumed to be similar to the training space. Note that this is not necessarily always the case, for example, take the NFL's draft: it is assumed that players in the NCAA who perform well will do so in the NFL; however, although both are playing football, there is a difference between the two environments and being successful in one is not perfectly correlated with doing well in the other. But, the most efficient training space available to the NFL is the NCAA so that is used; the other available training space is, of course, the NFL itself which then serves as input into the free agent draft system. To deal with this situation, a mechanism was developed to control the extent to which the testing problem space differed from the training problem space by specifying the probability that any particular ring position would be different for the testing phase from the training phase. These experiments showed, surprisingly, that 'top 10 expert' groups did not do particularly well even on those problem spaces identical to the training problem spaces.

The agents in these simulation are not adaptive; they have their heuristics which never change. In the negative correlated learning group selection process what becomes adaptive is the problem space. After each agent is selected to be included in the group of experts, the problem space (training space) is altered to discount those positions for which that agent performed better than the selection criteria. Thus the problem space (training space) adaption is where the intelligence resides in these simulations. However, since the problem space is random, there is not much structure to capitalize on. Nevertheless, the negative correlation learning processes did include all heuristic values into the group of experts quicker than the top ten method (see Table 22).

6.1.5 Problem Space Manipulation – Sombrero Problem Space

The second major area of investigation was manipulating the problem space. The Hong-Page problem space was a 2,000 position ring with values randomly distributed between 0 and 100. To give structure to the problem space, it was defined as a modified sombrero function which had two parameters: cycle which was the distance between successive peeks and beta which the coefficient of the line connecting successive peeks. It was actually built backwards starting at position 2,000 and working back towards position 1, so that although position 2,000 had 2000*beta as its value, position 1 did not necessarily have 1 as its value. There was no stochastic element to the problem space, therefore all iterations of simulation would always return the same values (except, of course, when the group was a random group).

Altering the cycle length from 8 to 4096 resulted in the average agent score decreasing from 8 to 256 and then increasing from that point on; the standard deviation of the agent scores decreased throughout the entire range of cycle lengths. At a cycle length of 4,056 (with any positive beta value), the problem space is a straight line from position 1 to position 2,000 and every agent marches up towards the top, not all agents arrive at position 2,000 since their last step may take them past that position so they would stop at most 9 positions away from 2,000 (agent 12, 11, 10 for example when it landed on 1,991 would remain at that position unable to get any closer to position 2,000)[31], which is why heuristics 1, 2, 3 are the associated with the highest values for cycle 4096 in Table 23. It was demonstrated that altering the cycle length changed the relative value for any particular heuristic. For example, when the cycle length was 8, heuristic 8 had the highest average value (1,427) which represented its ability to move from

[31] This is the reason the sombrero was built from ring position 2,000 backwards – to insure that the highest score was at ring position 2,000 rather than having the last cycle leave position 2,000 with some other value.

peek to peek until it reached the last upward slope. However, once the cycle length was greater than the maximum heuristic, the heuristics 1 and 12 were always in the top two groups: 1 (one) could insure that the agent could reach the local peek and 12 let the agent get onto an upward sloping section of the problem space from the furthest distance. Figure 7 showed the effect of changing the cycle length on the expected value of any individual heuristic value.

6.1.6 Basins of Attraction

The third area investigated was the existence (or lack thereof) of basins of attraction. No overwhelmingly identifiable basins of attraction were found for either the Uniformly Randomly distributed problem space or the modified Sombrero function problem space (cycle length 12, beta 1), although the Sombrero function problem space was better organized at separating out agents with a 12 as one if its heuristics. There were cases however where it failed to include agents with a 12 into the main body of agents with a 12 in their heuristic set which clearly stand out in the solution space plots. The original problem space did show how high scores grouped locally and created many plateaus which were associated with local maxima.

6.2 FAST AND FRUGAL SIMULATIONS

Luan's decision model (Luan, Katsikopoulous, & Reimer, 2012) is based on the Gigerenzer's Fast and Frugal Decision Model (Gigerenzer & Goldstein, 1996). Although it pre-dates the Hong-Page model, they have a root similarity. In the Hong-Page model, the agent applies its heuristics in sequential order and selects the first response that produces a better result. In the

Fast and Frugal model, the agent has a set of cues which have a fixed order and applies those cues sequentially until a distinction is made and then uses that distinction as the decision. Both paradigms are sequential and non-compensatory using a single factor at a time to make a decision rather than a weighting paradigm that applies the same or different weights for each que/heuristic and makes a decision based on the some combination of those weights.

Briefly, the Luan paper (2012) creates a test bed in which the five cues and an error term each have different coefficients. Fifteen examples are randomly created by sampling cues from a random uniform zero to one distribution plus the error term and creating an example value by multiplying the cue value by its coefficients and summing those values. These are equivalent to cities in the original exposition of the Fast and Frugal model. There are then 105 possible comparisons of pairs ((15 * 14)/ 2) to determine which example has the higher total score based on its cue values. The main point in the paper is that the effectiveness of the Fast and Frugal decision process lies in the relationship among the coefficient terms which is defined by four states – a large difference among the cues, a medium difference, a small difference, and no difference.

Luan defines the coefficients for each 'city's population' to be explained by the weighted cues to the same extent (approximately 85% of the variance in the population is explained by the cues) but the distribution of that information is spread out differently from the large difference environment where there is 10 times the information in the first cue as in the fifth cue to the no difference environment where there is an equal amount of information in each of the five cues. He then demonstrates that a wisdom of the crowd effect is observable in the no difference environment; whereas the Fast and Frugal paradigm is superior for the large difference environment, which has all agents making the same choices so there is, by definition, no wisdom

of the crowd effect due to no independence among the agents. This study focuses first on the dichotomization rate and creates a set of ‘perfect’ cue weights for the paradigm and finally shows the effect of ‘perfect’ cue weights moving to the no difference cue weights in 20 steps.

Table 30 and Table 31 show the effect of using the median as the dichotomization point, since in the original model the cues were facts that could be answered with a yes/no response. It maximizes the number of decisions that will be made on the basis of the first cue and when the first cue has the most information (i.e., when its coefficient is highest of all the cues in the criterion equation) then it makes the first cue the most likely to provide the correct response. Table 32 shows that using all 5 cues in a weighted regression function for making the guess about which of the two options has the larger score is always superior to only using the first cue.

The Fast and Frugal Simulation uses a dichotomization rate in order to convert the values used to create the test bed into the format used by the "Take the Best" algorithm. The dichotomization rate used is to code values greater than the median as one and values equal to or less than the median as zero. Each iteration of the simulation creates 15 exemplars with random values, ~U(0,1), generated for five cues and one error term. Therefore each of the sets of cues has seven values coded as one and eight values coded as zero. Each exemplar will then have a random assortment of zeros and ones as its cue values and its criterion score, its 'population' so to speak, will, however, be the product of the original random (zero to one) values times the associated betas plus the error term times its beta. The cues are ordered from one to five and their betas are shown in Table 6. Table 29 details the steps used for each iteration of the Fast and Frugal dichotomized simulation.

The initial consideration was to identify the effect of the dichotomization rate used. Table 30 and Table 31 show that using the median as the point at which the dichotomization

takes place produces an environment in which a maximum number of decisions are made at each individual cue level. In this case, choosing the median meant that 56 decisions were made at the 1st cue level whereas if the dichotomization rule been to only identify the highest value only 14 decisions would have been made at that level. So the dichotomization rule is optimized for the 'Take the Best' algorithm given that the cues are used in the order of their beta weights. Table 8 shows the median dichotomization rule does not correspond to the real-world test bed used for the original 'Take the Best' algorithm. Having decisions made as early as possible means that they are being made on the basis of the cue with the most information and therefore be more likely to be correct. Table 33 shows the percentage of correct decisions made in the different cue environments when the dichotomization rule changes from median to having the first cue identify the highest value option ('city with the highest population'), the second cue identifying the two high value options, the third cue the three highest, etc. Although this may seem like an ad hoc dichotomization process, it can be thought of 'what is most correlated with a city's population', then recursively adding additional criteria to increase the selected group. The first cue is correlated to the criterion, then the second cue is correlated to the first and the criterion, etc.

Table 34 shows that the percentage of times a cue informs a correct decision is a function not only its validity but also of the context in which the question is asked. In the bottom row where all cues have the same validity, the percentage of correct responses increases going from the first to the last cue due to the context in which the cue was queried. This demonstrates the fact that when the higher valued cues are tied, the marginal information in a lower valued cue can make the correct decision; whereas if the higher valued cues are able to make the decision

the decision, the lower valued cues (which will never be queried) do not have sufficient information to distinguish between the two candidates with the same degree of accuracy.

6.2.1 Luan Distribution of cue validity

Cue validity is roughly a measure of how well the cue informs a correct decision. In order for a cue to inform a decision, it has to be polled; therefore, not only is the amount of information in the cue important, its position in a series of cues is also important. In the German cities example, "Is the city a national capital?" uniquely identifies Berlin which is also the largest city in Germany. However, assume that "Was this city the home of the 1936 Olympics?" would also uniquely identify Berlin. If one of these questions was the first one asked, it would always provide the right answer when Berlin was one of the choices and not provide any information when Berlin wasn't one of the cities. If the other question was the second question, it would only be used when one of the cities wasn't Berlin and therefore it would never provide any useful information. Switch their order and the same thing happens, in this case due to their perfect correlation they are redundant questions and the second never provides additional information. As was demonstrated in Table 34, the validity of a cue is also contingent on the context; it is possible for a cue with low validity by itself becoming a highly valid que when conditioned on the responses to a set of other cues.

To analyze cue validity, the test bed was altered to create absolutely dominating cues. In this test bed, if the first cue's value was zero for option A (City A) and the first cue for option B was one, option B would always have the larger criterion score regardless of what the values for cues 2 to 5 were for option A. This then is also true for the second cue, given than the first cues match between option A and B, the second cue determines which has the larger criterion score.

There is an error term, but it is small enough that it never alters the results that one expects based on the cue values. This is the perfect scenario for 'Take the Best' since the first cue that distinguishes between two options will always do so correctly. The only time wrong decisions will be made is when all five cues match and the decision is made on the basis of a random value (this is different from the random value associated with the criterion score), and in this case the decision is expected to be incorrect half the time. When simulated 1,000 times, the correct decision was 98.40% of the time (the expected value was 98.44%). The 'Minimalist' algorithm when applied to the same test bed was correct 73% of the time, again demonstrating that the order in which the cues are used is an important aspect of their validity.

Now, allow the cues to gradually move from the maximally informative state they are in the simulation just described into a state in which each cue has the same validity (i.e., the cues' betas are all the same, equal to the no difference environment) and, as is demonstrated in Figure 14, the 'Take the Best' method drops its rate of being correct until it matches the Minimalists' groups of size 32 and above, at step 16 and then abruptly drops to rate of a single Minimalist when the cues are all equally informative. This figure showed that a group of two was not effectively superior to a single individual Minimalist agent and that the percentage of correct decisions increased with group size up to approximately 32 and then did not increase past that point. It additionally had a leveling out of rate at which group decisions were improving around step 16 which is a measure of how different the cue weights were from the absolutely dominating values to the 'No difference' values. The 'Take the Best' method was shown to be able to extract what information there was in the different cues and then when there was no difference in the amount of information in the individual cue's, it, of course, fell to equal the 'Minimalist' method

which by choosing cues randomly implicitly acts as if all cues have an equal amount of information.

A question could be raised about why the correct percentage of the group does not approach 100% that appears to be what Condorcet's Jury Theorem implies as the group size increases. This is because there is only a limited amount of information from which the agents can draw, so their decisions are not independent, when they use the same information source, they make the same decision and in that case it doesn't matter how many of the are in the group. These groups are made up of Minimalist agents which means that they are randomly choosing cues in random order on which to make their decisions and there is a limited number of random orders. There is not only a limited number of cue orderings; the dichotomization rule makes the probability of making the decision at an early cue significantly higher than a later cue which additionally restricts the range of information used for the agents to make their decisions.

6.2.2 Luan Modifying the number of cues

The number of cues was increased to 25 with betas ranging from 1 to 10 in equal increments. The range of the betas in this simulation was approximately that of the Large Difference environment where the betas ranged from .37 to .04 (also almost a 10 fold difference). Table 36 shows that even in the 'No difference' environment when there are only five cues the cues have a higher correlation with the criterion value than when there are 25 cues with a 'Large difference' beta structure. A descending linear relationship between the betas lead to the first five cues having approximately equal success in leading to correct decisions (66.21% for the first cue to 66.24% for the fifth cue). Table 36 shows that the actual correlation between the criterion score and the cue values does decrease from .33 to .28 for the first five cues, this decrease is counter-

balanced enough by other mechanisms so that the percentage of correct decisions does not decrease.

To adjust for this effect, instead of having all 25 cues be independent they were grouped into groups of five and each group of five cues had a common value summed with the cue's individual value. The ratio of the common value and the individual value was modified to represent highly correlated to minimally correlated cues. Table 37 showed the results of these simulations, when the cues were highly correlated, they dichotomized to the same side (top half or bottom half of the range of that cue value) and thus could not be used to make a decision. Again, this shows that a cue which could have a high validity score if used as the first cue could be almost meaningless when used as a second cue after a first cue with which it is highly correlated. In the far right cues, the first cue is used almost 50% of the time to make a decision and the second through fifth cues are used 2.17% to 0.51% of the time even though they contain almost identical information as the first cue.

The table shows that the percentage of correct decisions increases from 62% to 75% as the amount of "error" decreases; however, when the criterion score is regressed against all 25 cues the r-squared decreases from .73 to .52. The percentage of correct decisions increases because the cues are more correlated with the criterion score and as the table shows only one cue in each group is used to a significant degree. Notice that in the final column the percentage of times a cue is used is 49.97 for the first cue, almost a constant for all columns, but drops to 2% for the second cue to 0.5% for the fifth cue before jumping back to 22.75% for the sixth cue which is the first cue in the second group. The third row shows the correlations of Cue 1 with the criterion score and Cue 25 with the criterion score.

6.3 EXPLORATION/EXPLOTATION MODEL

6.3.1 Rotation Pattern

An analysis of the relationship between the value of a heuristic and its position in the triad of heuristics revealed that there was an interaction between the heuristics values and their positions; therefore in this section a 'rotation' parameter was developed which altered how agents in a Hong-Page simulation use their heuristics as they traverse the problem space. The original (Hong-Page) rotation model was found to associate lower average ring position scores with higher heuristics in the 1st heuristic position and higher scores with low valued heuristics in the 1st position. A new set of rotation models was developed with r0 being no rotation, r1 being rotation applied with each starting point on the ring, and r2 being continuous rotation starting from ring position 1 through to finding the final value for ring position 2,000.

Additionally, Table 38 shows that the different rotation systems lead to a substantial difference in how frequently each of the different heuristics was used in traversing the problem space with the first heuristic falling from 59% to 33% when going from the original rotation pattern to the final rotation pattern. This 59% is, of course, not associated with the 8/15's caused by the dichotomization rule which is associated with the Fast and Frugal model (8/15 = .53).

Figure 17 demonstrates the different effects caused by the three different rotation patterns, the three heuristic positions and the heuristic value on the average position score for each agent. The figure clearly demonstrates that in the rotation pattern r2 the heuristic value of 1 had the highest average score. In general the figure shows that for each of the three rotation

patterns, the average scores are reversed from heuristic value 1 (1st position having the highest average score followed by the second position and then the last position) to the heuristic value 12 in which the opposite order is observed for each rotation pattern (i.e., Agent 1,x,x will on average have a higher score than Agent 12,x,x and Agent x,x,1 will have a lower score than Agent x,x,12 – where the x's are any other heuristic values.)

This can be interpreted as an exploration/exploitation effect. The smaller valued heuristics being preferred in the initial positions can be interpreted as favoring exploiting the local situation before exploring positions farther away. The actual mechanism at work in these simulations is that when a larger valued heuristic is initially used, it may step over higher valued positions whereas when the lowered valued heuristics are initially used at least to investigate some of those positions before taking a larger step. With the bear in the berry patch simile, the bear is better off first looking at the patch it is in for more berries before heading out looking for the next berry patch. The other most obvious trait from this figure is the zigzag nature of the lines, with the odd values having higher values than the even values on either side, which is almost always true (the green dot p3r0 being the primary exception). This artifact is not addressed in this dissertation other than to note that it was observed in each simulation that lead to this figure and not just an artifact of one particular simulation (approximately four or five separate versions of this figure).

Table 39 shows the top ten agents and the bottom 10 for rotation r0 and r2. It was shown that with r2 agent scores come in groups of 3: heuristics (a,b,c) and grouped with (b,c,a) and (c,a,b) which clearly demonstrates that it is the order of the heuristics that is the diversity measure of interest rather than the heuristic value in the different positions. If r2 is used as the

rotation pattern. However, if r0 is used as the rotation pattern the actual heuristic value independent of its position appears to be a better measure.

Figure 19 shows the effect of the rotation patterns on groups ranging from groups of one agent to groups of 10 agents. From this figure, it is clear that group size is far more influential in determining the final average score for the problem space than the rotation pattern used by each individual agent. The figure shows the expected marginal improvement drop off with the improvement from an individual agent to a team of two agents matching the improvement that comes from adding the next eight agents. This is different from the Luan simulation in which groups of 1 and 2 agents were identical; since in this simulation, there is no voting to determine what the best score for each ring position, it is taken as being solely a function of the highest value found by any agent. This may appear to conflict with Figure 14 in which it was noted that groups of size 1 and size 2 were both the same: that was for the Luan model in which group members voted for their choice whereas the Hong-Page model of the group process has the group members sharing heuristics and being able to use each other gains to produce further gains.

6.3.2 Decision Points

Figure 19 shows the effect of when the decision takes place: Does the group go with the first agent to achieve a higher score or does it let each agent reach its own decision independently and then choose the best decision from the group? This figure shows the improvement that happens when the best agent score is chosen rather than the first. What is counter intuitive about these results is that as the group size increases the difference between the ‘best’ agent and the first agent with a better score does not increase, it stays stable around 1.6 to 1.7. The agents are

added into the group in order of their average score in the training space and this may influence this pattern which was not further explored in this dissertation.

6.3.3 Diversity Effect

The "Hidden Profiles" literature leads one to expect that not all heuristics available to a group would be equally used. Those heuristics which were only possessed by one agent may well never be used in the pursuit of a higher score. Figure 20 shows the result of only using those heuristics that two or more agents share as the group size increases from 1 to 10. Of course, at group size 1, there are no heuristics shared so the agents return the original value of the ring at each position which averages 50. The value returned is shown on the left axis, the right axis shows the average number of heuristics in the group. The rotation patterns are shown in different color bars, but they do not contribute any variation. The initial rate of increase in this chart is significantly slower than that in the previous figures because as additional members are added to the group, the principal effect they have is to cause another heuristic to be included if it already exists (as a singleton), unique heuristic they possess don't contribute to the group abilities. Note that although a group of ten agents with 3 unique heuristics each could have all 12 heuristics, the final average for the group of ten is around 9.

An additional loss documented during group processes comes from the group settling for the group's mean score rather than the best score. This is most evident in those cases in which it is not clear which decision is the best decision, when this is implemented in this simulation, the differences between the different rotation pattern are not statistically significant; however, the group score approaches that of the average individual. In the sample presented, the group averaged 88.35 and the average of all agents is generally approximately 85. In the 'Omit Single

instances of a heuristic' the average was in the 90's for groups of size 10, 96 for 'the best agent', and 94 for the 'the first agent'. So this process loss represents the largest loss; however, in instances in which the value of a decision cannot be known, the information required for the 'first agent', 'best agent', or even 'omit single' algorithms to work is unavailable. In this case, the Heuristic simulation's test platform operates under the same constraints as the Fast and Frugal platform with each agent voting for its decision.

7.0 CONCLUSION

This dissertation analyzed two Wisdom of the Crowd simulations and created a modified version of those two. Its principal method of analysis was to focus on the micro elements of the simulation: the actual values of the heuristics used in the Hong-Page simulation along with a finer-detailed description of the mechanisms used to select which heuristic was to be used next. In the Fast and Frugal simulations, the dichotomization rule's effect on the results was explored in detail along with the effect of correlation among the clues. Its principal finding was that an "Exploration/Exploitation" effect could be found in the Hong-Page simulation using this micro-level analysis. It was demonstrated that smaller valued heuristics used earlier in the search process produced higher results than larger valued heuristics used in those early positions. Additionally, larger valued heuristics were more valuable when they were used later in the process than were smaller valued heuristics. Smaller valued heuristics were associated with exploiting the local environment in detail; whereas the larger valued heuristics were associated with exploring a different region. This finding is original to this dissertation. This finding was the result of observing what average scores the different agents were obtaining as a function of what the value of the heuristics was and where in the set of three heuristics that value occupied. It was a wisdom of the crowd effect that was observable much like Smith's Invisible Hand: it is not an effect obvious in any particular agent but becomes clearer when they are all combined into

a ‘crowd’. It isn’t something the leads to a wisdom of the crowd effect but the outcome from that process.

The extreme value effect was explored by breaking the training dataset into groups and then identifying those agents which performed well by counting the number of groups in which they were in the top 10. Instead of using the agent’s average value of the entire 2,000 ring position, the 2,000 positions were partitioned and the top 10 agents in each partition identified. The group of experts was then determined by those agents which most frequently appeared in the various top 10 groups for the different partitions. Original to this dissertation was the finding that the best performing set of experts was the set created by making the size of the set equal to 1. So the effective mechanism for determining which agent was in the top 10 was not that agent’s average score over different simulations, but just the number of times it was in the top 10. This finding is reminiscent of Galton’s original point in the ‘how much does the ox weigh’ series of articles: It could be not using the mean across multiple simulations, the effect of extreme values is limited to only the simulation in which it occurs. If an agent is the top agent in one partition by a wide margin but mediocre in succeeding partitions, the one high value does not override the multiple mediocre values. The concept, of course, is not original to this dissertation but identifying it in the context of this wisdom of the crowds’ simulation is original to this dissertation. However, the finding that the group size of 1 was the best partition size was not stable across multiple executions of this simulation. More work is needed on this process to identify an optimal algorithm for determining the best partition size.

7.1 LIMITATIONS/DELIMITATIONS AND FUTURE WORK

A frequently mentioned issue with this type of agent based decision process is that there is no cost function. In the Hong-Page simulation, one could easily imagine that there was some cost associated with each time an agent checked the value of the ring at a different position. It could be a fixed cost or even a cost associated with a complex function taking into account the local problem space, the past successes of that agent, and purely random elements. The Fast and Frugal decision process, equally, lacks any provisions for costs. Information is, generally, not free. Organizing the cues into the optimal order may be an expensive enterprise and in an evolving problem space may have to be adjusted frequently. However, one of the potential benefits of crowd sources is that costs are so widely distributed that they may be ignorable. For example, product reviews are provided by users without cost to the product providers (or at least that is the theory, in actuality producers may manage the 'reviews' in order to make them actually more like Public Relations bulletins than independent opinions).

In order to implement a cost effect, the agents would have to be more intelligent than they are. In the simulations presented in this dissertation, none of the agents were adaptive to their environment. In the Hong-Page and Exploration/Exploitation simulations, the agents had their set of heuristics and applied those heuristics in a given order. Partially this is a result of the limited problem space in which it was possible to simulate every possible heuristic combination (agent). Since all possible agents were present, there was no issue of trying to let agents improve their scores. As the number of heuristics an agent possesses increases and the range of heuristics increases, it becomes increasingly more difficult to instantiate every possible agent and the

simulation would benefit from agents trying to find better combinations of heuristics. With three heuristics, there were 1,320 possible agents (12 * 11 * 10), if the number of possible heuristics increased to 20 there would have been (20 * 19 * 18) 6,840 possible agents and if the number of heuristics an agent possessed increased from 3 to four there would have been (12 * 11 * 10 * 9) 11,880 agents for the 12 distinct heuristics model and (20 * 19 * 18 * 17) 116,280 for the 20 distinct heuristic model. Increasing the number of distinct heuristics increases the total number faster than increasing the number of separate heuristics that an agent can possess. An additional modification could be to remove the restriction that each agent have the same number of distinct heuristics, by letting heuristics have the same heuristic in more than one position (for example 1,3,1) the agent would model an agent with only two heuristics rather than 3 which would just reflect a world in which some agents had a wider range of abilities than other agents.

As the ABC Group correctly points out in their Fast and Frugal publications, scissors have two blades that have to work together in order to cut (Herb Simon is credited with the original formulation.) Not only are the agents very limited in these simulations, but the environments are equally limited. The modified sombrero does introduce some structure into the models but, as Bocanet and Ponsiglione attempted, the creation of a problem space with more realistically tunable parameters could potentially lead to more robust generalizability of any results with Kauffman's NK fitness landscape being a frequently used format for modeling problem spaces.

A principle finding of this dissertation, that exploitation should proceed exploration, is delimited to a very artificial simulation model. Literally the findings were that when the initial search pattern was exploratory the expected score was lower than when the initial search pattern was exploitative and when the final search was exploitative it has a lower average score than

when the final search steps were exploratory. This finding is suggestive of the results of my Preliminary Examination in which I demonstrated that in a sequence of keywords describing a web site from delicious the initial keywords were the most frequently used ones and the last keyword was always the least frequently used keyword. This analysis was possible when it was conducted because tags were stored as a single character string rather than being parsed into individual tags; thus tag order was preserved. Later delicious tag handling procedures parsed the tags which, of course, made them more useful as tags but also lost some important information. The prevalence of tag information now makes that enterprise riskier since entered tags probably also strongly reflect what keywords were used to locate the item being tagged

The obvious areas to extend this work in its present format is to modify the basic paradigm to include a cost function to each agent for investigating an option. With a cost function, the basic structure of every agent checking every position may not be the most efficient search strategy. Increasing the problem space's dimensions and structuring it into a multidimensional fitness landscape would then make the simulation more amenable to adaptable agents. A simulation with intelligent adaptable agents in a complex fitness landscape is clearly the region into which this type of research could be directed.

There are several papers which claim that the Fast and Frugal paradigm proves that by applying that method agents that 'know' less function better than agents that 'know' more. This claim is 'supported' by the findings in Table 37 Correlated Cues; however, it is such a counter intuitive claim that further work needs to be published to demonstrate what the mechanisms are that underlie the trend for Fast and Frugal's percentage correct to increase while the regression's r-squared to decrease. The regression was not answering the same question as the Fast and Frugal mechanism in Table 37; therefore a comparison of how well two methods predict two

completely different events isn't sufficient. If method A is trying to predict an exact value and method B is just trying to discriminate between two examples, the question of which is 'better' needs to also answer the 'for what' question before making that determination and if only discrimination between two examples is all that is needed than one also needs to say that B does that better than A does the same thing.

This dissertation paraphrased Page (page 33), Sunstein (page 36), and Plott (page 36) among others to the effect that the wisdom of the crowd effect does not create information, it merely aggregates that information. As Page points out you don't expect a group of 5th graders to provide a good estimate of how much a fully loaded 747 weighs and adding more 5th graders to the group wouldn't, necessarily, help the estimate. All that is mathematically true is that the error from the average of the group's estimate will be smaller than the average error from the group's individual members, provided that the estimates lie of both sides of the true value and that the error is measured with a convex function (for example absolute value). Krause et al. (2011) showed that groups did well estimating the number of marbles in a jar but very poorly estimating the probability of winning the German Lotto. Sunstein (2006) also pointed out that wisdom of the crowd didn't work when it was trying to predict who President Bush would nominate as a Supreme Court Justice (Roberts) in an information market although the market mechanism works very well for other questions, pointing out that there wasn't any information to be aggregated until just before the nomination.

There is, as yet, no clear definition of those conditions under which the wisdom of the crowd effect will outperform experts – it appears that it will be a mixture of matching what information the crowd has to the nature of the problem. There are groups that would have very accurate estimates of how much a fully loaded 747 weighed and there are questions that a crowd

of 5th graders would answer very accurately. Palmer (1994) demonstrates this in his study of an artificial stock market where agents that are successful in one part of the history of the market are not successful if they are moved to a different place in time.

Gigerenzer (1996) uses Simon's simile of a pair of scissors for decision making where one blade is the environment and one blade the decision rule, they are have to work together. The simile is apt if you extend the scissors to include those that cut patterned edges instead of a straight line. In this case it is clear that there is a very wide variety of patterns that could be cut along the edge of a piece of paper but only if the two blades fit together which is not guaranteed when there are many different styles of scissor blades to choose from.

8.0 BIBLIOGRAPHY

Adamic, L. A., Suresh, K., & Shi, X. (2007). Scatter networks: a new approach for analysing information scatter. *New Journal of Physics*.

Althaus, S. (2003). *Collective preferences in democratic politics: Opinion surveys and the will of the people.* New York: Cambridge University Press.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review, 84*(2), 406-411.

Asch, S. (2003). Opinions and social pressure. In E. Aronson, *Readings about the social animal* (pp. 17-26). Macmillian.

Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2008). Results from a dozen years of election futures markets. In C. R. Plott, & V. L. Smith, *Handbook of experimental economics results* (pp. 742-751). Amsterdam: North-Holland.

Bhavnani, S. K. (2005). Why Is It difficult to find comprehensive information? Implications of information scatter for search and design. *Journal of the American Society for Information Science and Technology, 56*(9), 989-1003.

Bikhchandani, S. e. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives, 12*(3), 151-170.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 992-1026.

Bocanet, A., & Ponsiglione, C. (2012). Balancing exploration and exploitation in complex environments. *VINE, 42*(1), 15-35.

Brown, G. (2004). Diversity in neural network ensembles. Birmingham, United Kingdom: The University of Birmingham.

Caplan, B. (2001). Rational ignorance versus rational irrationality. *Kyklos, 54*(1), 3-26.

Caplan, B. (2007). *The myth of the rational voter: why democracies choose bad policies.* Princeton, NJ: Princeton University Press.

Caplan, B. (2009). Majorities against utility: Implications of the failure of the miracle of aggregration. *Social Philosophy and Policy, 26*(1), 198-211.

Chi, E., & Mytkowica, T. (2008). Understanding the efficiency of social tagging systems using information theory. *HT '08.*

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important. *Annual Conference of the British Educational Research Association.* University of Exter. Retrieved from http://www.leeds.ac.uk/educol/documents/00002182.htm

Condorect, J.-A.-N. d. (1994). *Condorcet: Foundations of social choice and political theory.* (I. McLean, & F. Hewitt, Trans.) Northampton, MA: Elgar Publishing, Inc.

Darwin, C. (1859). *On the origin of species.* London: John Murray. Retrieved from http://darwin-online.org.uk/converted/pdf/1859_Origin_F373.pdf

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy*, 135-150.

Engel, D. e. (2014, December 16). Reading the mind in the eyes or reading between the lines? *PLoS One*.

Ferandez-Abascal, E. e. (2013). Test-retest reliability of the "Reading the Mind in the Eyes" test: a one-year follow-up study. *Molecular Autism, 33*(4).

Galton, F. (1907 a, Feb 28). One vote, one value. *Science*.

Galton, F. (1907 b, March 7). Vox populi. *Science*.

Galton, F. (1907 c, March 28). The ballot-box. *Science*.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4).

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*(1), 149-167.

Guarino, A. e. (2011). Aggregate information cascades. *Games and Economic Behavior, 73*(1), 167-185.

Hackman, J. R. (2011). *Collaborative intelligence.* San Francisco: Barrett-Koehler Publishers, Inc.

Hardin, G. (1968). The tragedy of the commons. *Science*, 1243.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction.* Springer.

Hirshleifer, D., & al, e. (2003). Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management, 9*(1), 25-66.

Ho, T. K. (2002). Multiple classifier combination: Lessons and the next steps. In K. A, & H. Bunke, *Hybrid Methods in Pattern Recognition* (pp. 171-198). Hackensack, New Jersey: World Scientific Publishing.

Hong, L., & Page, S. (2012). Some microfoundations of collective wisdom. In H. Landemore, & J. Elster (Eds.), *Collective Wisdom: Principles and Mechanisms* (pp. 56-71). New York City: Cambridge University Press.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences, 101*(46), 16385-16389.

Hong, L., Page, S. E., & Riolo, M. (2012). Incentives, information, and emergent collective accuracy. *Managerial and Decision Economics, 33*, 323-334.

Jones-Rooy, A., & Page, S. E. (2012). The complexity of systems effects. *Critical Review: A Journal of Politics and Society, 24*(3), 313-342.

Kahneman, D. (2011). *Thinking, fast and slow.* New Yrok: Macmillan.

Katsikopoulos, K. V., & Schooler, L. J. (2010). The robust beauty of ordinary information. *Psychological Review, 117*(4), 1259-1266.

Keller, E. F. (2005). Ecosystems, organisms, and machines. *Bioscience, 55*(12), 1069-1074.

Keller, E. F. (2008). Organisms, machines, and thunderstorms: A hisotry of self-organization. Part one. *Historical Studies in Natural Sciences, 38*(1), 45-75.

Keller, E. F. (2009). Organisms, machines, and thunderstorms: A history of self-organization. Part two. *Historical Studies in the Natural Sciences, 31*(1), 1-31.

Kerr, N. L. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103*(4), 687-719.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-655.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456-460.

Krause, J., Ruxton, G., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology and Evolution, 25*(1), 28-34.

Krause, S., James, R., Faria, J., Ruxton, G., & Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behavior, 81*, 941-948.

Kuncheva, L. I. (2004). *Combining pattern classifiers: Methods and algorithms.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Lakhani, K. R. (2008). InnoCentive.Com (A). Cambridge, MA: Harvard Business School Publishing.

Landemore, H., & Elster, J. (2012). *Collective wisdom: Principles and mechanisms.* New York: Cambridge University Press.

Larrick, R. P., & Soll, J. B. (2006, Jan). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*(1), 111-127.

Laughlin, P. R. (1999). Collective induction: Twelve postulates. *Organizational Behavior and Human Decision Processes, 80*(1), 50-69.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly, 52*(4), 667-694.

Lehner, P. E., Adelman, L., Cheikes, B. A., & Brown, M. J. (2008). Confirmation bias in complex analysis. *IEEE Transactions on Systems, Man, and Cybernetrics - Part A: Systems and Humans, 38*(3), 584-592.

Liu, Y. (1998). Negative correlation learning and evolutionary neural network ensembles. Canbeerra, Austrialia: The University of New South Wales, Austrailian Defense Force Academy.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Acadmey of Sciences, 108*(22), 9020-9025.

Lu, L., Yuan, C., & Lauretta, M. P. (2012). Twenty-five years of hidden profile in group decision making: A meta-analysis. *Personality and Social Psychology Review, 16*(1), 54-75.

Luan, S., Katsikopoulous, K., & Reimer, T. (2012). When does diversity trump ability (and vice versa) in group decision making? A simulation study. *PLoS One, 7*(2).

Malthus, T. (1798). *An essay on the principle of population.* London: J. Johnson. Retrieved from http://www.gutenberg.org/catalog/world/readfile?fk_files= 1455152&pageno=1

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organizational Science, 2*(1), 71-87.

McKenzie, C. R. (2004). Judgment and decision making. In K. Lamberts, & R. Goldstone (Eds.), *Handbook of cognition* (pp. 321-338). Sage.

Miller, J., & Page, S. (2007). *Complex adaptive systems: An introduction to computational models of social life.* Princeton, NJ: Princeton University Press.

Mitchell, M. (2009). *Complexity: a guided tour.* New York: Oxford University Press.

Mojzisch, A., & Schulz-Hardt, S. (2010a). Process gains in group decision making: A conceptual analysis, preliminary data, and tools for practitioners. *Journal of Managerial Psychology, 23*(3), 235-246.

Mojzisch, A., & Schulz-Hardt, S. (2010b). Knowing others' preferences degrades the quality of group decisions. *Journal of Personality and Social Psychology, 98*(5), 794-808.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* New York: Cambridge University Press.

Page, S. (2007). *The Difference: How the power of diversity creates better groups, firms, schools, and societies.* Princeton: Princeton University Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96-146.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information.* Oxford (UK): Oxford University Press.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643-675.

Plott, C. R., & Chen, K.-Y. (2002). *Information aggregation mechanisms: Concept, design and implementatikon for a sales forecasting problem.* Pasadena: California Institute of Technology.

Polikar, R. (2006). Ensemble based systems in decision theory. *IEEE Circuits and Systems Magazine*, 21-44.

Polikar, R. (2012). Ensemble learning. In C. a. Zhang, *Ensemble Machine Learning: Methods and Applications* (pp. 1-34). Springer Science+Business Media.

Porter, T. M. (1986). *The rise of statistical thinking.* Princeton, NJ: Princeton University Press.

Putnam, R. D. (2007). E pluribus unum: Diversity and Community in the twenty-first century -- The 2006 Jahan Skytte Prize lecture. *Scandinavian Political Studies, 30*(2), 137-174.

Rand, A. (2005). *Atlas shurgged: (Centennial Edition).* Penguin.

Riesman, D., Glazer, N., & Denney, R. (1961). *The lonely crowd: A study of the changing American character.* New Haven: Yale University Press.

Schelling, T. C. (1978). *Micromotives and macrobehavior.* New York: Norton.

Servan-Schreiber, E. (2012a). *Intelligence is collective.* South Orange, NJ: Lumenogic, LLC.

Servan-Schreiber, E. (2012b). Prediction markets: Trading uncertainity for collective wisdom. In H. Landemore, & J. Elster, *Collective Wisdom: Principles and Mechanisms.* Cambridge University Press.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 83-104.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics, 69*(1), 99-118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review, 63*(2), 129.

Smith, A. (1776). *The wealth of nations.* London: T. Nelson and Sons. Retrieved from http://www.gutenberg.org/files/38194/38194-h/38194-h.htm

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780-805.

Strasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 67-78.

Sunstein, C. (2006). *Infotopia.* New York: Oxford University Press.

Sunstein, C. R. (2007). Deliberating groups versus prediction markets (or Hayek's challenge to Habermas). *Episteme: A Journal of Social Epistemology, 3*(3), 192-213.

Surowiecki, J. (2005). *The wisdom of crowds.* Random House Digital.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 167-171.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make ussmart: Ecological rationality. *Current Directions in Psychological Science, 16*, 167-171.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainity. *Science, 185*, 1124-1131.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129-140.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures.* Thousand Oaks, CA: Sage Publications.

Wittenbaum, G. M. (2000). The bias toward discussing shared information why are high-status group members immune. *Communication Research*, 379-401.

Wittenbaum, G., Hubbell, A. P., & Zuckerman, C. (1999). Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of Personality and Social Psychology, 77*(5), 967-978.

Woolley, A. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*, 686-688.

Woolley, A. W., Gerbasi, M. E., Chabris, C. F., Kosslyn, S. M., & Hackman, J. R. (2008). Bringing in the experts: How team composition and collaborative planning jointly shape analytical effectiveness. *Small Group Research, 39*(3).

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications.* New York: Springer Science+Business Media.