

Each day people are presented with circumstances that may require speculation. Scientists may ponder questions such as why a star is born or how rainbows are made, psychologists may ask social questions such as why people are prejudiced, and military strategists may imagine what the consequences of their actions might be. Speculations may lead to the generation of putative explanations called hypotheses. But it is by checking if hypotheses reflect encountered facts that behaviour demonstrating a true understanding results. If evidence shows a hypothesis to be false, then people should rationally abandon it, especially if there are negative consequences. The aim of this thesis is to examine how effectively people search for evidence in their hypothesis testing to test whether or not their hypotheses are true or false in competitive games. Research findings from six studies of hypothesis testing behaviour are explored. Chapter by chapter the thesis tests how everyday people, and master chess players, tackle hypothesis testing in competitive deductive tasks. Implications are discussed for aspects of general cognition: such as reasoning, social hypothesis testing and planning.

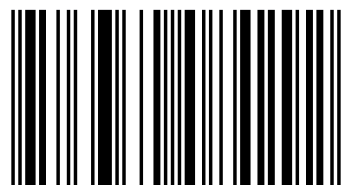


Michelle Cowley



Dr. Michelle Cowley is a Social Statistician and a Royal Statistical Society Fellow. She has been an Irish Research Council for the Humanities and Social Sciences Scholar, and a postdoctoral and visiting Katzenbach Fellow at the University of Oxford and Princeton University. She is an early-career expert on foresight and behavioural epistemology.

## Hypothesis Testing: How We Foresee Falsification in Competitive Games



978-3-330-08429-2

Cowley

LAP  
LAMBERT  
Academic Publishing

**Michelle Cowley**

**Hypothesis Testing: How We Foresee Falsification in Competitive Games**



**Michelle Cowley**

# **Hypothesis Testing: How We Foresee Falsification in Competitive Games**

**LAP LAMBERT Academic Publishing**

## **Impressum / Imprint**

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Coverbild / Cover image: [www.ingimage.com](http://www.ingimage.com)

Verlag / Publisher:

LAP LAMBERT Academic Publishing

ist ein Imprint der / is a trademark of

OmniScriptum GmbH & Co. KG

Bahnhofstraße 28, 66111 Saarbrücken, Deutschland / Germany

Email: [info@omniscryptum.com](mailto:info@omniscryptum.com)

Herstellung: siehe letzte Seite /

Printed at: see last page

**ISBN: 978-3-330-08429-2**

Copyright © Michelle Cowley

Copyright © 2017 OmniScriptum GmbH & Co. KG

Alle Rechte vorbehalten. / All rights reserved. Saarbrücken 2017

# **Hypothesis Testing: How We Foresee Falsification in Competitive Games**

**Dr. Michelle Cowley**

Royal Statistical Society Fellow (2012- )



## **Acknowledgments**

This work has been submitted as a doctoral requirement for the qualification of Doctoratus Philosophia (DPhil) while studying at the University of Dublin, Trinity College. The author was awarded a graduate scholarship from the Irish Research Council for the Humanities and Social Sciences (IRCHSS), travel funding from the Trinity Trust at Trinity College Dublin, and a Law Faculty Small Grant, University of Oxford.

The author would like to thank the School of Psychology Trinity College, the Trinity College Institute of Neuroscience, the Centre for Thinking and Language Plymouth University, and the MMR Lab at Princeton University for supporting the discussion of this research in both its graduate and preparation for publication stages. The author would like to thank the Royal Statistical Society for presenting the opportunity to work on publishing the manuscript to commence early career research development as a Royal Statistical Society Fellow Member.





**Dedicated to ... Mia Moo – Who taught me what is possible...**



## Table of Contents

	<b>Page</b>
<b>Chapter 1 Introduction</b>	
Introduction	9
Hypothesis Testing	10
The 2-4-6 Task	20
Theories of Hypothesis Testing in the 2-4-6 Task	27
Alternative Hypotheses Accounts	40
Aims of the Thesis	44
<b>Chapter 2 The Role of Alternative Hypotheses in Hypothesis Testing: An Imaginary Participant 2-4-6 Task</b>	
Introduction	49
The Imaginary Participant 2-4-6 Task	50
Experiment 1	52
Experiment 2	63
Experiment 3	74
General Discussion	81
<b>Chapter 3 The Effect of Competition on Hypothesis Testing</b>	
Introduction	85
The Philosophy of Falsification	87
Experiment 4	91
Experiment 5	99
General Discussion	107
<b>Chapter 4 Chess Masters' Hypothesis Testing</b>	
Introduction	111
Methodological Advantages of Chess Experiments	112
Theories of Chess Expertise	115
Evaluation of Chess Moves by Experts and Novices	122
Experiment 6: A Protocol Analysis	125
General Discussion	146
<b>Chapter 5 Discussion</b>	
Introduction	153

Summary of Findings and Theoretical Implications	155
General Implications for Human Learning and Cognition	164
Future Questions	168
Conclusions	172
<b>References</b>	<b>175</b>

## **Appendices**

Appendix A (i): Materials used in experiment 1	
Appendix A (ii): Recording sheets	
Appendix B: Materials used in experiment 2	
Appendix C: Materials used in experiment 3	
Appendix D: Materials used in experiment 4	
Appendix E: Materials used in experiment 5	
Appendix F: Materials used in experiment 6	
Appendix G: Set up for chess program <i>Fritz 8</i>	
Appendix H: Segmented protocols of chess players' thinking	
Appendix I: Selected problem behaviour graphs of chess players' thinking	
Appendix J: The experimenter's think-aloud script used in experiment 6	
Appendix K: The experimenter's recording sheet used in experiment 6	

## Chapter 1 – Introduction

*All...by nature desire knowledge*

– Aristotle 384 BC- 322 BC

Each day people are presented with circumstances that may require speculation. Scientists may ponder questions such as why a star is born or how rainbows are made, psychologists may ask social questions such as why people are prejudiced, and military strategists may imagine what the consequences of their actions might be. In everyday life people may wish to contemplate questions that are curious for their own sake, such as why their dog tries to eat bees when he knows they sting, or a young child may wonder what the consequences of kissing a frog might be when reading a fairy-tale at bedtime. Speculations may lead to the generation of putative explanations called *hypotheses*. But it is by checking if hypotheses accurately reflect the encountered facts that lead to a true understanding. For example, if the evidence shows a hypothesis to be untrue, then people should abandon it.

The objective of this thesis is to examine how people search for evidence to test if their hypotheses are true, and to understand how people abandon a hypothesis when evidence shows that the hypothesis is false. This type of thought is called *hypothesis testing*, and searching for evidence which shows a hypothesis to be false is called *falsification*. The literature examining how people test their hypotheses and how they reason about evidence showing that their hypothesis is false contains many contradictory findings and dissonant theories. For example, there is still no consensus about what factors may help people to falsify their untrue hypotheses, or whether they find falsification possible at all. This thesis develops two novel experimental approaches to address these issues. First, a novel component to a standard hypothesis testing task known as the 2-4-6 task is introduced. A version of the 2-4-6 task in which people must interact with an imaginary participant is created. This version of the 2-4-6 task allows the investigation of how properties of the hypothesis, properties of alternative hypotheses and competition between two people testing hypotheses, may facilitate the possibility of falsification. Second, two previously disparate cognitive domains of chess expertise and hypothesis testing are concurrently examined for links to investigate how expert knowledge may affect falsification in hypothesis testing.

A range of methodologies which are novel to the study of hypothesis testing are presented. The thesis presents not only standard reasoning tasks, but also protocol analysis to elicit the structure of what people mentally represent when testing their hypotheses. A powerful computer chess program (*Fritz 8*) is utilised, to provide an objective measure with which to compare peoples' evaluation of evidence in their hypothesis testing. Third, the two main alternative theories of hypothesis testing are tested experimentally and show that they are not fully corroborated by the data in this thesis, and what these results imply for future theories of hypothesis testing is discussed. Finally, a new theoretical component of expert hypothesis testing, based on the framework of hypothesis testing derived from the experimental findings in this thesis, is put forward.

This chapter is divided into four sections. The first section reviews the literature on hypothesis testing. The types of strategies people use to test hypotheses, the relevance of hypothesis testing to human rationality, and the functions of hypothesis testing are outlined. The second section reviews the literature on the standard hypothesis testing task used in reasoning research — Wason's 2-4-6 rule discovery task. The history of hypothesis testing through findings from this task is traced, and how different researchers' conceptions of hypothesis testing and testing strategies have changed over time are highlighted. The third section reviews the two main theories of how people test their hypotheses which were initially derived from findings in the 2-4-6 task. The chapter points out the shortcomings of these theories, and outlines questions which need to be answered by future theories of hypothesis testing. In the final section a review is presented as to how these questions can be addressed by experimental manipulations conducted by this thesis in the 2-4-6 hypothesis testing task, and experimental manipulations of this hypothesis testing is extended in more complex yet controlled domains such as chess.

## **Hypothesis Testing**

From the beginning of history humans have been driven to explain phenomena of interest occurring in their environment. For example, to explain natural incidences such as flooding, modern day scientists look towards geo-technical measurement, medieval people looked towards godly intervention, and ancient civilizations looked towards the movement of the stars in the night sky. Today, more than ever, there are countless breakthroughs in scientific endeavour that

aim to advance an increasing knowledge base. Scientists have even gone so far as attempting to understand the workings of the human organic body and the human mind. Primarily, this accumulation of knowledge owes much to the human ability to explain phenomena by generating hypotheses (e.g., Bruner, Goodnow, & Austin, 1956; Cherubini, Castelvechio, & Cherubini, 2005; Peirce, 1992). For example, the ancient Egyptians explained the flooding of the Nile by generating the hypothesis that the movement of the star Sirius triggered the rising water levels (Gooch, 1981). This hypothesis is inaccurate in light of modern day knowledge indicating that the rise in water was caused by melting snow in mountain ranges thousands of miles away. This example shows that the ability to generate hypotheses to explain phenomena is a vital part of accumulating new knowledge, but the example also hints that the ability to *test* whether or not these generated hypotheses are accurate explanations leads to a true understanding (e.g., Popper, 1959). Essentially, hypothesis testing is a major task of human thinking facilitating the comparison of internal thoughts with external facts in order to interact efficiently with the environment in a way that reflects reality (Poletiek, 2001).

### ***Hypothesis testing strategies***

For people to test if their hypothesis reflects reality they must search for evidence. To study how people search for evidence to test their hypotheses, cognitive psychologists needed to borrow the crucial concepts of *confirmation* and *falsification* from the philosophy of science (e.g., Carnap, 1950; Popper, 1959). The search for evidence that is consistent with a hypothesis and that indicates that the hypothesis is true is called confirmation, and the search for evidence that is inconsistent with a hypothesis and that indicates that a hypothesis is untrue is called falsification. One school of thought in the philosophy of science, known as the *logical positivist* school, proposed that people should follow a confirmation strategy and try to find as much evidence as possible to confirm a hypothesis (Carnap, 1950). While searching for confirming evidence may be initially useful for generating a hypothesis worthy of examination (e.g., Mynatt, Doherty & Tweney, 1978), there is a problem with this strategy. The compilation of a large number of confirming instances does not necessarily guarantee that a hypothesis is true; no matter how much evidence confirms a hypothesis, there may always be a chance that some piece of falsifying evidence will come to light (Popper, 1959; 1963). But if a major



prediction of a theory is proved false, it is made known that the theory is incorrect or at least incomplete (e.g., Tweney, Doherty & Mynatt, 1981). For this reason it was proposed that a falsification strategy which requires searching for evidence that is inconsistent with a hypothesis was better than a confirmation strategy (Popper, 1959). Confirmation by itself may lead to an accumulation of numerous hypotheses with some confirming evidence explaining the same phenomenon. In contrast falsification ensures that some of these hypotheses may be abandoned in favour of better hypotheses, ensuring a progression towards truthful explanations which can be added to our knowledge base (Kuhn, 1993). As a result science views empirical falsification as the optimal procedure for testing the truth of hypotheses, and as an essential process for the growth of scientific knowledge (Popper, 1963; Platt, 1964; Lakatos, 1970).

Popper applied his theory to everyday human thinking as well as science and he claimed that scientific knowledge is a development of common sense knowledge. Rational hypothesis testing was equated with falsification in both scientific and everyday thinking. When cognitive psychologists first examined how people tested the truth of their hypotheses, they accepted that the rational way to test hypotheses was to subject them, where possible, to falsification (e.g., Wason, 1960; Mynatt, Doherty & Tweney, 1977; 1978).

To illustrate how falsification can be better than confirmation consider the following scenario:

*You are a scientist and your job is to identify the cause of a dangerous new disease. You identify a previously unrecognized virus in tissue samples of symptomatic patients and your hypothesis is that this 'new virus' is the cause of the disease. However, other scientists have identified two viruses, including your new virus in their tissue samples. They hypothesise that it is the 'other virus' and not the new virus that is the cause. Both hypotheses have confirming evidence. A case is reported where the new virus is present and the other virus is absent. What should you conclude?*

A situation similar to this one faced scientists working on the cause of the now notorious SARS virus. They concluded that the 'new virus' hypothesis was correct. The case where the 'other virus' was absent falsified the 'other virus' hypothesis and proved that the 'new virus' hypothesis was right. The example illustrates how falsification can be vital to the discovery of truth.

Research in experimental psychology typically addresses how people can be successful hypothesis testers (e.g., Gale & Ball, 2003; 2005; Van der Henst, Rossi, & Schroyens, 2002), how they can sometimes employ faulty hypothesis testing strategies (e.g., Poletiek, 2001; 2005), and how certain prescriptive measures may or may not address these faulty strategies and improve hypothesis testing accuracy (e.g., Klayman & Ha, 1987; 1989). It has generally been accepted that where people fail to adopt a falsification strategy in hypothesis testing tasks, they fail to think rationally (Manktelow, 1999). In the following section I explore how people can test their hypotheses in a rational way, and how they may use an irrational strategy called *confirmation bias*, by seeking evidence consistent with a hypothesis and avoiding inconsistent evidence when testing the truth of their hypotheses.

### ***Confirmation bias as irrational hypothesis testing***

A central question for theories of cognition is whether human beings are rational. That is, whether or not they think about and interact with their environment in a sound and sensible manner (Nixon, 2002). For example, it has been shown that people do not always arrive at accurate deductions in their reasoning (e.g., Johnson-Laird & Byrne, 1991), or generate realistic judgments consistently (e.g., Kahneman, Slovic & Tversky, 1982), or even proceed to solve problems in an optimal way (e.g., Newell & Simon, 1972). A similar picture emerges from hypothesis testing research. Early studies of hypothesis testing showed people to be unsuccessful because they persisted in testing their hypotheses in an irrational way (e.g., Wason, 1960; Wetherick, 1962; Mynatt, Doherty, Tweney, 1977; 1978).

For example, in one of the first experimental studies of hypothesis testing participants were instructed to discover a rule the experimenter had in mind to which the number triple 2-4-6 conforms. This task is called the 2-4-6 task, where the participant is analogous to the scientist, and the experimenter's rule is analogous to the law of nature to be discovered (Wason, 1960). Participants used a biased strategy whereby they sought evidence to confirm their hypothesis about the nature of the numerical rule and to avoid falsifying evidence, even when their hypothesis was untrue. The experimenter's rule is simply 'any ascending numbers' but participants tended to focus on the salient features of the initial 2-4-6 triple and generated hypotheses such as 'even numbers ascending in twos'. They proposed triples consistent with their hypothesis such

as 10-12-14 and 16-18-20, rather than triples inconsistent with their hypothesis such as 5-10-15. If they had tested their hypothesis with at least one triple that was inconsistent with their hypothesis, such as the triple 5-10-15 which contains odd numbers, their hypothesis 'even numbers ascending in twos' would have been falsified. They would have been given the information that the triple 5-10-15 was consistent with the experimenter's rule. Participants would then know that odd numbers are consistent with the experimenter's rule and they could infer that their hypothesis containing the property of evenness was incorrect. Instead participants persisted in testing with triples that would lead to confirmation such as 10-12-14. This tendency for people to seek out information consistent with their hypotheses and avoid inconsistent information was termed *confirmation bias*. The result has been replicated many times in the 2-4-6 task (e.g., Tweney *et al.*, 1980; Gorman, Gorman, Latta, & Cunningham, 1984; Kareev & Halberstadt, 1993), and has contributed to the view that human thinking was irrational and biased (e.g., Evans, 1989; Evans, Newstead, & Byrne, 1993).

Researchers began to extend hypothesis testing experiments to more realistic tasks to investigate the implications of this irrationality. For example, confirmation bias was investigated in other laboratory tasks intended to simulate scientific discovery, notably the *artificial universe* task (Mynatt, Doherty, & Tweney, 1977; 1978). Participants were required to discover the law governing the motion of particles in an artificial universe displayed on a computer screen. Because of constraints in the set-up of the universe, most participants form the hypothesis that a triangle shape causes the particles to cease moving. In fact shape, size and location do not affect the motion of the particles and the rule to be discovered is that figures with low brightness levels have boundaries which cease particle movement. The only way for participants with the 'triangle' hypothesis to discover the rule is to try to falsify their hypothesis by firing a particle at a non-triangle low brightness figure and observe it bouncing off it, but instead participants try to confirm their 'triangle' hypotheses by firing particles repeatedly at triangles. The results observed from laboratory tasks indicate that people may be prone to a pervasive confirmation bias in their reasoning.

But are people irrational hypothesis testers only in experimental tasks? Researchers began to extend hypothesis testing to everyday thinking such as social inference (e.g., Snyder & Swann, 1978; Allport, 1979). There was a

concern that if confirmation bias was pervasive in human reasoning, then it could be responsible for the formation and maintenance of irrational beliefs such as prejudices and stereotypes (e.g., Aronson, 1999; Kruglanski & Webster, 2000). A series of social inference studies demonstrated that people do tend to seek confirmation of a hypothesis they hold about the personality of a target individual (Allport, 1979; Snyder & Swann, 1978; Nisbett & Ross, 1980). In one study a group of participants was asked to judge if another person was an extrovert, and a second group of participants was asked to judge if a person was an introvert (Snyder & Swann, 1978). They were given a list of possible interview questions and both groups tended to choose to ask the individual questions that were related to the trait they were interested in. For example, participants testing the extrovert hypothesis most often chose the question ‘what would you do if you wanted to liven things up at a party?’ However, this question is not conducive to an answer such as ‘I never try to liven things up’. Both introverts and extroverts tended to accept the premises of the question and give similar answers. If we consider the response ‘play lively music’ it could be interpreted as consistent with the idea of a ‘lively’ extrovert’s choice in music, or an introvert’s choice to rely on the music rather than themselves to liven a party up.

These examples show how hypothesis testing has a central role in human rationality. Hypothesis testing plays an important role in many aspects of human thought apart from laboratory tasks and social inference, such as in scientific discovery (e.g., Gorman, 1995a; 1995b). If we could not search for falsifying evidence to overcome untrue hypotheses our ability to interact with others and our surroundings would be hampered. The suggestion that people may have a tendency to test their hypotheses and beliefs in an irrational way by exclusively searching for confirming evidence presents us with a paradox (Poletiek, 1996). How can people be irrational hypothesis testers given the scientific and technological advancement they are capable of achieving? For example, how can we put a man on the moon if our thinking is inherently flawed (Mitroff, 1974)? One explanation is that people may be more capable of falsification than experimental studies have so far shown.

To show how falsification may be important to rational reasoning this thesis recounts several main ways in which hypothesis testing is important to human cognition. Is successful hypothesis testing then affiliated with human achievement? In the next section I describe the functions of hypothesis testing in

a broad range of domains and highlight how falsification may be important in each case.

### *The functions of hypothesis testing*

Hypotheses may start out as anticipations of future events, tentative solutions to problems we are presented with, or even guesses about occurrences around us (Bruner, Goodnow, & Austin, 1956; Poletiek, 2001). The testing of hypotheses allows people to inspect whether their hypotheses are accurate by searching for evidence that can prove their truth or falsity. Many aspects of cognition may require the testing of hypotheses and the search for falsifying evidence to test if these hypotheses are good ones. For example, experts who are defined as people who tend not to make as many errors in their thinking in their domain of expertise, may need to search for falsifying evidence to test whether their hypotheses are correct, especially when they are generating ideas to create new knowledge or extend an initial knowledge base (e.g., Einhorn & Hogarth, 1978). Planning a course of action may require searching for evidence to show how a plan may lead to negative consequences, in other words how a plan may be falsified by an opponent responding to a plan in a way that was not anticipated (e.g., Cowley & Byrne, 2004; Johnson-Laird & Byrne, 1991). Solving a problem may require searching for evidence that a potential hypothetical solution works out by searching for the possible ways a solution may not work out (e.g., Newell & Simon, 1972; Gobet, 1998; Gobet *et al.*, 2004), and remembering hypotheses that previously turned out to be incorrect may help people learn from experience (e.g., Roese & Olson, 1995). In the following sections how hypothesis testing may be required in several areas of cognition will be outlined in more detail including: expert thinking; planning; learning from experience and problem solving.

### *Expert thinking*

Much of expert thinking proceeds by hypothesis testing. Experts must generate and test hypotheses to advance understanding in many domains. For example, scientists generate hypotheses to discover new knowledge (Gorman, 1995a; Mitroff, 1974; Tweney, 1989), and must subject these hypotheses to experimentation in order to discriminate between hypotheses as the best explanations of the data (e.g., Fugelsang, Stein, Green & Dunbar, 2004; Kuhn, 1993). Legal experts such as criminal psychologists and the police must

generate hypotheses to detect motives and suspects for acts of crime (e.g., Britton, 1997), and subsequently evaluate the evidence to ascertain criminal responsibility (e.g., Wagenaar, Van Koppen & Crombag, 1993). Medical experts generate hypotheses to understand the causes of disease in order to develop cures for illnesses (e.g., Christensen-Szalansky & Bushyhead, 1981), and they must discriminate between relevant and irrelevant symptoms to diagnose illness (e.g., Koriati, Lichtenstein & Fischhoff, 1980).

An important role for an expert working in any field is the generation of new knowledge. For example, a scientist working at the current boundaries of existing knowledge in a domain may make a significant new discovery thereby generating new knowledge (e.g., Dunbar, 2000). It is possible that when the testing of a hypothesis leads to an encounter with evidence to prove that the hypothesis is false, it may lead to the generation of new knowledge (Popper, 1963). In scientific terms a falsification of theory is termed a refutation (Kuhn, 1993). When theories are refuted either an alternative theory which explains the result is accepted as superior, or an alternative theory is developed which can explain the falsifying result (e.g., Wason & Johnson-Laird, 1972; Kuhn, 1993). A theory is revised to incorporate the new result rather than abandoned altogether (Howson & Urbach, 1993; Klayman & Ha, 1989; Kowslowki, 1996), or occasionally the refutation is labelled as an anomaly until a viable alternative theory is generated (Kuhn, 1993; Koslowski, 1996).

Refutations are generated by rival theorists (e.g., Mitroff, 1974; Kuhn, 1993), and to safeguard against many refutations being labelled anomalies by scientists who disagree with one another it is important to test hypotheses with specific alternatives in mind (e.g., Platt, 1964). For example, successful hypothesis testers who use falsification to overcome hypotheses which are untrue, consider at least one alternative hypothesis in rule discovery tasks (e.g., Klayman & Ha, 1989). Identifying falsifying evidence indicates what is wrong with a hypothesis or theory (e.g., Fugelsang *et al.*, 2004). Falsification drives hypothesis revision because it hints at what should be incorporated into the hypothesis. When we encounter inconsistent evidence relevant to a current state of knowledge we may update our knowledge by revising it to include the new piece of information (e.g., Gardenfors, 1988; Harman, 1986). Falsification in expert thinking ensures that theories which have outlived their usefulness are either improved or abandoned in favour of theories which offer better explanations (Popper, 1963).

### *Planning*

Hypothetical thinking entails the prediction of some future event, and an important type of reasoning which requires hypothetical prediction is planning (e.g., Craik, 1943; Gazzaniga, Ivry, & Mangun, 1998). There may be many situations in which it is helpful to predict the consequences of a hypothesized plan of action, for example, in interactions with an opponent or collaborator, in political or social engagement, in games such as tic-tac-toe or poker, or in cases of military strategy (e.g., Mallie, 2001). When people plan for the future they may attempt to predict the most likely outcome given some scenario (Giroto, Legrenzi, & Rizzo, 1991), which helps them to test whether their hypothesized plan will lead to their desired goal (e.g., Camerer, 2004).

It may be helpful to consider the possible alternative ways a plan may be falsified, in other words how a plan may go wrong, for example, by an opponent responding to a plan in a way that was not anticipated (e.g., Cowley & Byrne, 2004; Hedden & Zhang, 2002; Zhang & Hedden, 2003). For example, military strategists must evaluate hypotheses about different possible courses of action to establish the best plans of action in war and peace (Beavor, 1998). Consider the Russian victory at the battle of Stalingrad (1942- 1943) which thwarted the German advance on the eastern front in World War II. This success rested on a tactic that falsified the planned German invasion. Instead of fighting the German army head on as expected, the Russian side retreated from the borders of Poland to Stalingrad where they unexpectedly started to fight. The Germans simply continued with their plan and sent a wave of reinforcements. Meanwhile, the Russian forces had amassed an army in excess of one million men outside Stalingrad and then encircled over 250,000 German troops (Beavor, 1998). If the German side had foreseen this counter attack by exhaustively searching for all alternative plans at the Russians disposal, they may have foreseen the falsification of their plan and stopped investing German troops in a lost cause.

### *Learning from experience*

Falsification and the consideration of alternative hypotheses may also be directly related to how people learn from their mistakes. Much of reasoning concerns the consideration of alternative possibilities such as how things could have worked out better (e.g., Roese & Olson, 1995; Walsh & Byrne, 2001). The

consideration of alternative possibilities may play a role in more general facets of thinking such as imagining alternative possible worlds when thinking creatively (e.g., Byrne, 2005), and thinking about alternative causes of an event (e.g., Goldvarg & Johnson-Laird, 2001). When falsification results in error it may provoke the generation of an alternative possibility in which a person imagines a way that things could have worked out better (e.g., Roese & Olson, 1995). For example, when a person makes a mistake they may imagine how a past event might have been avoided (e.g., Mandel & Lehman, 1996), or what a better solution to a problem might have been (e.g., Anzai & Simon, 1979). In other words when people remember how the hypothetical possibilities they had imagined turned out to be false they may be able to avoid investing in similar false hypotheses in the future.

### *Problem solving*

When a hypothesis is a proposed solution to a problem a number of alternative paths may need to be tested in order to find the optimal solution. Problems which require an action require consideration of a number of possible alternative paths towards solution (e.g., Mynatt, Doherty & Dragan, 1993; Simon & Hayes, 1976). For example, chess problems require one to choose a move among many possible alternatives at each stage in the game, and to predict a number of possible opponent counter moves to each move (e.g., Nunn, 1999; Newell & Simon, 1972; Newell, 1990). The Tower of Hanoi problem requires the movement of little discs from one wooden peg to another according to rules about what size disc can be placed on top of another on each peg (Anzai & Simon, 1979). Problems in real life may involve the consideration of possible alternative solutions and the weighing up of potential positive and negative outcomes to each alternative. Each alternative solution path is tested by evaluating whether or not the solution achieves the desired goal (e.g., Newell, 1990; Newell & Simon, 1972). Often there are a number of sub-goals which need to be achieved in order to reach an overall goal when solving a problem (e.g., Simon & Hayes, 1976; Kotovsky, Hayes, & Simon, 1985; Klahr & Dunbar, 1988). For example, completing an undergraduate degree before you complete a PhD. But how do people select a path that will lead them towards a solution? Searching for evidence for (confirming evidence) and against (falsifying evidence) one hypothesised solution over another may help people to evaluate which solution path is best (e.g., Cowley & Byrne, 2004).



These examples testify that a major function of hypothesis testing is the discovery of untrue hypotheses by falsification, whether they are expert hypotheses such as scientific hypotheses, hypotheses concerning plans of action, or hypotheses concerning the solutions to a problem. Yet the research on hypothesis testing has shown that people find falsification difficult and often find confirmation useful, whether they are NASA Apollo mission scientists (Mitroff, 1974), scientific discoverers such as Alexander Graham Bell (Gorman, 1995a), or participants who confirm early on in their hypothesis testing in complex laboratory experiments (e.g., Mynatt, Doherty & Tweney, 1978). In the next section the literature questioning the universal notion that falsification is the optimal strategy in hypothesis testing is reviewed. How people can find confirmation useful under some circumstances is examined. That there is a difference between confirmation and confirmation bias and how this difference has been poorly distinguished in previous research is put into context. How different conceptions of what constitutes confirmation and falsification, and how these strategies have been measured in different ways over time is discussed. To show how these different conceptions of hypothesis testing have contributed to contradictions in the literature, I now trace these different conceptions of hypothesis testing strategies, and the debate about whether falsification is the optimal strategy, using findings from the most widely used hypothesis testing task—the 2-4-6 task.

### **The 2-4-6 Task**

The 2-4-6 task is the main task used in hypothesis testing research. Participants must discover a rule an experimenter has in mind that the number triple 2-4-6 conforms to. They generate their own number triples with sets of three numbers. For each triple they are told ‘yes’ or ‘no’ by the experimenter as to whether or not it conforms to the rule or not. Each triple is taken to be a test of the participant’s hypothesis about what the rule is. For example, participants tend to focus on the salient features of the initial 2-4-6 triple and generate hypotheses such as ‘even numbers ascending in twos’. To test this hypothesis they propose triples such as 10-12-14 and 16-18-20. For each one of these triples they receive a ‘yes’ response from the experimenter. But the experimenter’s rule is the deliberately general rule, ‘any ascending numbers’ (Wason, 1960). Hence, a test triple such as 10-12-14 receives a ‘yes’ because it is consistent with the rule to be discovered (‘any ascending numbers’) as well as the hypothesis under test

(‘even numbers ascending in twos’). Participants generate test triples until they think they have discovered the rule and they then announce what they think it is.

Typically participants tend to test triples such as 10-12-14 and compile confirming ‘yes’ responses and announce an incorrect rule such as ‘even numbers ascending in twos’. Only 21% of participants tend to announce the correct rule first time round (e.g., Wason, 1960; Tweney *et al.*, 1980; Gorman, Gorman, Latta, & Cunningham, 1984; Gale & Ball, 2003; 2005). In the first 2-4-6 study successful participants tended to have what was termed a higher *eliminative-index*, that is, they tested their hypothesis with at least one inconsistent triple such as a triple with odd numbers 5-10-15. This inconsistent triple falsified their hypothesis ‘even numbers ascending in twos’ because when they were told that odd numbers were consistent with the experimenter’s rule then they knew that their hypothesis containing the property of evenness was incorrect.

***The logic of hypothesis testing: Forty years of misdiagnosis in the 2-4-6 task?***

Initial classifications of hypothesis testing as confirming and falsifying tests were equated with consistent tests (tests that were consistent with the participant’s hypothesis) and inconsistent tests (tests that were inconsistent with the participant’s hypothesis) (Wason, 1960). But Wetherick (1962) argued against this division of test instances into confirming and falsifying based on whether a test was consistent or inconsistent with the hypothesis under test. Instead, a four-way classification was suggested where confirmation and falsification were split into two different strategies based not only on whether participants expected instances to be consistent with their hypothesis *but* also whether participants *intended* instances to be consistent with the experimenter’s rule. For instance, when a participant’s hypothesis is ‘numbers ascending in twos’ and they generate the test triple 3-5-7, it is clear that 3-5-7 is consistent with the participant’s hypothesis because it ascends in twos. But this test is only a confirming test if the participant *intends* the triple to confirm by also expecting it to be consistent with the experimenter’s rule. If the participant expects a ‘yes’ from the experimenter, they are attempting to confirm their hypothesis, and they expect their hypothesis is correct. But if the participant expects a ‘no’ from the experimenter, they are attempting to falsify their hypothesis as they expect their hypothesis is incorrect.

The same is true for inconsistent tests. For example, when a participant's hypothesis is 'numbers ascending in twos' and they generate the test triple 5-10-15, it is clear that 5-10-15 is inconsistent with the participant's hypothesis because it is not ascending in twos. However, it is a falsifying test only if the participant intends the triple to conform to the experimenter's rule. If the participant expects a 'yes' from the experimenter, then they expect a triple that is inconsistent with their hypothesis to be consistent with the experimenter's rule, therefore they expect their hypothesis to be incorrect. But, if the participant expects a 'no' from the experimenter, then they are in fact attempting to confirm. They expect that the triple 5-10-15 is neither inconsistent with their hypothesis 'numbers ascending in twos' nor with the experimenter's rule. The inconsistent test in this instance is intended to provide confirmation. Inconsistent tests can be intended to either confirm or falsify. The classifications are:

- (i) *consistent-confirming*: the triple is consistent with a participant's hypothesis and is expected to conform to the experimenter's rule
- (ii) *consistent-falsifying*: the triple is consistent with a participant's hypothesis but is expected not to conform to the experimenter's rule
- (iii) *inconsistent-falsifying*: the triple is inconsistent with a participant's hypothesis but it is expected to conform to the experimenter's rule
- (iv) *inconsistent-confirming*: the triple is inconsistent with a participant's hypothesis and it is expected not to conform to the experimenter's rule

Later theorists also considered the above system to be the best method for classifying confirming and falsifying hypothesis tests in the 2-4-6 task (e.g., Poletiek, 1996), but the terminology used to describe this classification has changed (Klayman & Ha, 1987; 1989). A test triple that is consistent with a hypothesis test (i and ii) is renamed a *positive test*, because it is a positive instance of the hypothesis, so 3-5-7 is a positive test of the hypothesis 'numbers ascending in twos'. A test triple that is inconsistent with a hypothesis test (iii and iv) is renamed a *negative test*, because it is a negative instance of the hypothesis, so 5-10-15 is a negative instance of the hypothesis 'numbers ascending in twos'. Positive and negative tests have been split into confirming and falsifying sub-classifications: positive confirming (i); positive falsifying (ii); negative falsifying (iii); negative confirming (iv).

The important point is that although this classification has now been accepted as the best way to classify confirming and falsifying hypothesis tests in recent years, earlier research on the 2-4-6 task tended to rely only on the distinction between positive and negative tests to tell confirming and falsifying hypothesis testing apart (e.g., Gorman, Gorman, Latta, & Cunningham, 1984; Tweney *et al.*, 1980; Kareev & Halberstadt, 1993). For example, when researchers tried to improve participant's ability to falsify in the 2-4-6 task by instructing them to falsify, they based their instructions on the concept of confirmation as a positive test and falsification as a negative test (e.g., Gorman, Gorman, Latta, & Cunningham, 1984; Gorman & Gorman, 1984). Their analysis did not record participants' intention to confirm or falsify, and as a result confirmation and falsification may have been confused in many studies (see Klayman, 1995; Poletiek, 2001 for review). The critical point is that a test is considered to be a confirming test when it is intended to confirm a hypothesis. Likewise, a test is considered to be a falsifying test when it is intended to falsify a hypothesis. To clarify I present an example of each test type in a Table on the next page which reflects this current method for classifying the logic of hypothesis testing in the 2-4-6 task. I use one of the hypotheses typically generated in the 2-4-6 task (and which I will use in the experiments later)—'even numbers ascending in twos':

Table 1.1: Categorising confirming and falsifying test types in the 2-4-6 task for the hypothesis ‘even numbers ascending in twos’.

<b>Test triple</b>	Is the test triple a positive or negative test?	Does the person intend the test to confirm or falsify?	<b>Confirming or falsifying Test</b>
<b>8-10-12</b>	positive	confirmation expected	<b>Confirming</b>
<b>24-26-28</b>	positive	falsification expected	<b>Falsifying</b>
<b>5-10-15</b>	negative	falsification expected	<b>Falsifying</b>
<b>23-25-27</b>	negative	confirmation expected	<b>Confirming</b>

In the 2-4-6 task the terminology of hypothesis testing has not only reflected how confirmation and falsification have been measured, but how hypothesis testing has been labelled over time. Falsification has been termed *disconfirmation* in some contexts. For example, in case studies of scientific discovery falsification was conceptualised as evidence proving a theory to be untrue at a *micro level* in a simple experiment, and disconfirmation was conceptualised as evidence proving a theory to be untrue at a *macro level* in a series of experiments (e.g., Gorman, 1995a; 1995b). Concepts of hypothesis testing in the 2-4-6 task have been concerned with how confirmation and falsification should be labelled given the processes underlying each strategy. For example, there is the suggestion that confirming is not a conscious process but the result of a preconscious bias to attend to information that is positive rather than negative, such as attributing more relevance to triples leading to ‘yes’ than ‘no’ responses from the experimenter in the typical 2-4-6 task context (Evans, 1989). Recent work on the 2-4-6 task has called into question the notion of such a positivity bias. Participants were found to be able to use triples that generated feedback of ‘no’. When participants were told there were two rules to be discovered they used these negatively labelled triples just as effectively as triples that generated ‘yes’ feedback. The consideration of dual goals, that is, the aim to discover two rules, may play a more important role in hypothesis testing than any type of preconscious attending to positive information (Gale & Ball, 2003). If positivity bias plays a role in hypothesis testing, then it is a small role.

In order to summarise the different ways researchers have defined hypothesis testing strategies over the last forty-five years in the 2-4-6 task, the following table is outlined:

Table 1.2: The different ways hypothesis testing strategies have been conceptualised in hypothesis testing research over the past forty-five years.

---

**Term and Definition {main author(s)}**

---

*Severity of test* Severity of test is a philosophical term used to refer to falsification. A hypothesis tester should test their hypothesis as severely as possible. In other words, they should choose a test that can result in the strongest possible evidence against a hypothesis. This type of hypothesis testing was termed falsification (Popper, 1959).

*Falsification* Falsification became the favoured scientific and psychological term used to refer to the severity of test as outlined above. Falsification has tended to be associated with the search for evidence to show a hypothesis to be untrue (e.g., Wason, 1960).

*Confirmation bias* Confirmation bias is a tendency to search for evidence that is consistent with a hypothesis and avoid inconsistent evidence (e.g., Wason, 1960).

*Positive and negative test strategies* A triple that is consistent with a hypothesis is a *positive test* of that hypothesis. For example, the triple 8-10-12 is generated when the hypothesis is ‘even numbers ascending in twos’ because it contains the target properties of evenness and ascending in twos. A triple such as 5-10-15 is inconsistent with the hypothesis because it does not contain these target properties and it is called a *negative test*. Participants may have a tendency to test cases that have the property of interest rather than those that do not have the property in the 2-4-6 task, that is, they have a tendency to follow a positive test strategy of testing positive instances, which does not necessarily constitute a bias in all reasoning contexts (Klayman & Ha, 1987).

*Intentional confirmation and falsification* Confirmation and falsification depend on whether a test is consistent with a hypothesis *and* on whether it is intended to confirm or falsify. Participants' tendency to generate triples that are consistent with a currently held hypothesis may not constitute a bias, because they may not 'expect' a consistent triple to result in a confirming response from the experimenter. Participants must *expect* a confirmation for it to constitute a confirmation bias. Participants must expect a falsification for it to constitute a falsification (Wetherick, 1962).

*Positivity bias* One claim is that human reasoning is biased towards attending to positive instances of a current representation at a preconscious level (Evans, 1989). This tendency to attend to positive instances corresponds to a bias towards attending to positive instances of a current hypothesis and selecting these positive instances as tests of that hypothesis which may be symptomatic of confirmation bias.

*Disconfirmation* One purpose is that there are two levels of hypothesis testing. At the micro-level hypothesis testing corresponds to individual tests, for example one experiment may falsify a hypothesis, but at the macro-level hypothesis testing corresponds to a series of tests, for example a series of experiments which lead to disconfirmation of a theory (Gorman, 1995a).

---

Next the thesis will discern ways of discriminating between confirmation bias, non-biased confirmation, and falsification. First, it is suggested that a test can be considered an instance of *confirmation bias* in the following circumstances when a hypothesis is untrue (Cowley & Byrne, 2004; 2005):

- (i) when participants indicate in their responses that they intend their test to result in confirmation of their hypothesis, *even though* falsifying evidence is available (in line with Wetherick, 1962, Klayman & Ha, 1987; Poletiek, 1996);
- (ii) *or* when participants evaluate the result of a test as confirming their hypothesis when the test result objectively falsifies their hypothesis.

Second, it is suggested that a test can be considered an instance of *falsification* in the following circumstances when a hypothesis is untrue:

- (i) When falsifying evidence is available to the participant, and participants indicate in their responses that they intend their test to result in

falsification of their hypothesis (in line with Wetherick 1962; Klayman & Ha, 1987; Poletiek, 1996);

(ii) *or* when participants evaluate the result of a test as falsifying their hypothesis when the test result objectively leads to falsification.

Third, it is suggested that a test can be considered an instance of non-biased confirmation in the following circumstance:

(i) When the hypothesis is true or of exceptional quality such that there is very little falsifying evidence to search for, and a hypothesis test, even though it is intended to falsify, may result in confirmation (e.g., Poletiek, 1996; 2005). In other words when a person seeks to falsify their hypothesis as much as possible in order to identify falsifying cases, if any exist. But these severe tests in fact lead to confirmation of a hypothesis. In this case the hypothesis is confirmed but not in a biased way (Cowley & Byrne 2005).

It is important to note the distinction between the process of a test choice and the outcome of a test choice when we refer to confirmation and falsification in the above examples. For example, when a hypothesis is generated it may actually represent the true state of affairs. To test this hypothesis a person may generate a test with the intention to falsify it, but because the hypothesis is in fact true the test outcome can only confirm the hypothesis regardless of the process the person has used (in the experimental chapters I detail this point further). In the next section two main theories which have been developed to explain the findings observed in the 2-4-6 task are detailed. The main tenets of each theory, and how the factors pertinent to these main tenets affect hypothesis testing are described. I consider the shortcomings of each theory and explain how the experimental designs employed in this thesis test the main tenets of each.

### **Theories of Hypothesis Testing in the 2-4-6 Task**

I will now outline two main theories of hypothesis testing developed from findings in the 2-4-6 task. The first theory proposes that people find falsification difficult if not impossible in the 2-4-6 task, and that confirming and falsifying are one and the same process (Poletiek, 1996; 2001; 2005). I will refer to this theory as the uniformity theory of hypothesis testing in the 2-4-6 task because it proposes that confirming and falsifying testing are the same process. The second theory proposes that hypothesis testing is constrained by the mathematical structure of the hypothesis testing task at hand (Klayman & Ha, 1987). Klayman



and Ha suggest that people find it difficult to falsify, not because they find falsification impossible, but because their tendency to use positive tests is not conducive to falsification due to the mathematical constraints in the standard 2-4-6 task (Klayman & Ha, 1987). I will refer to this theory as the mathematical relationship theory of hypothesis testing in the 2-4-6 task. Next I detail each theory in turn and point out the main tenets of each.

### ***The uniformity theory (Poletiek, 2001)***

Are people able to perform two distinct types of hypothesis tests, that is, confirming and falsifying tests? In other words when people generate a hypothesis do they feel able to control the hypothesis testing process in order to bring about a confirming or falsifying result? Or do people perform just one type of test that will either lead to a confirming or falsifying outcome depending on the quality of the hypothesis rather than their own test choice (Poletiek, 1996)? Experimental studies in the psychological literature have shown that people are rarely capable of intentionally bringing about a falsifying result (e.g., Wason, 1960; Mynatt *et al.*, 1977; Tweney *et al.*, 1980; Gorman, Gorman, Latta, & Cunningham, 1984).

Recent evidence from the 2-4-6 task indicates that people may find falsification difficult if not impossible, and in performing a test people cannot sensibly intend to confirm or falsify (Poletiek, 1996, Experiment 1). To test hypotheses, people perform a test, and the test will either confirm or falsify a hypothesis depending on the quality of the hypothesis initially generated (Poletiek, 1996, Experiment 2). In other words participants cannot deliberately intend to falsify or control test outcomes in order to falsify a hypothesis; hypothesis testing is simply experienced as performing a test and therefore confirmation and falsification are the same strategy (Poletiek, 2001; 2005). In one experiment participants were explicitly instructed to falsify their hypothesis in the standard version of the 2-4-6 task (Poletiek, 1996, Experiment 1). The rule to be discovered was the ‘any ascending numbers’ rule and participants were instructed to generate their ‘best guess’, that is, their hypotheses about what the rule might be. Participants typically focused on the salient features of the 2-4-6 triple and generated hypotheses pertaining to ‘evenness’ or ‘ascending in twos’ (e.g., Cherubini *et al.*, 2005). The only type of hypothesis test a participant can use to intentionally falsify a hypothesis such as ‘even numbers ascending in twos’ in the standard 2-4-6 task is a *negative falsifying test* triple;

such as 5-10-15 which they then *expect* to lead to falsification (See Table 1.1). If they receive the feedback that the triple is consistent with the experimenter's rule, then they know their hypothesis 'even numbers ascending in twos' is untrue because 5-10-15 contains odd numbers so the rule cannot pertain to even numbers only. Poletiek claims that people cannot generate these test triples with the intention of getting a falsifying test result.

The ability to generate negative falsifying tests is pivotal to the debate about whether people can falsify in a useful way. Participants were given instructions either to 'test', 'confirm', or 'falsify' (Poletiek, 1996). For the 'test' and 'confirm' conditions, the majority of tests fell into the positive confirming category (86% and 80% respectively), and few tests fell into the negative falsifying category (0% and 3% respectively). Participants in the 'falsify' condition were instructed to 'try to test in such a way as to get your hypothesis about the rule rejected' (Poletiek, 1996; p.454). The majority of tests in this condition fell into the two confirming categories, the positive confirming and negative confirming categories (32% and 54% respectively). (See Table 1.1). Although the participants who were instructed to falsify proposed test triples that were negative tests, they in fact intended these tests to confirm. It was concluded that people do not seem to be able to make sense of falsification because they expect their test result to confirm their hypothesis regardless of the tests they proposed. Poletiek (1996) points out that people do not appear to be able to intentionally perform negative falsifying tests and therefore they find falsification an impossible hypothesis testing strategy to conduct. The claim is made that negative confirming tests may have been misidentified as falsifying tests in early studies of hypothesis testing claiming people could sometimes falsify (Poletiek, 1996, p. 448; see Gorman, Gorman, Latta, & Cunningham, 1984; Gorman & Gorman, 1984). People find falsification impossible because their hypothesis represents their 'best guess' about the truth, that is, the hypothesis incorporates all the information they had access to when it was generated and therefore they cannot know where to find falsifying information. Poletiek explains that negative tests are a first reflex to make a mismatch between the hypothesis and test item when participants are instructed to falsify, because participants appear to have little insight into their test choices.

Poletiek also reasons that people simply perform a test which should not be considered something that more or less corroborates *or* falsifies the hypothesis, but something that uniformly more or less confirms the hypothesis. She argues

that participants do not consider the test outcome to be a consequence of the test choice as their lack of expectation to falsify means they believe they cannot control the test outcomes by choosing a particular testing strategy beforehand. In other words confirmation and falsification are experienced as a uniform process by participants, that is, they are experienced as the process of carrying out a hypothesis test. Poletiek (1996) argues that there is a paradox in hypothesis testing: how can people falsify their hypotheses given that they incorporate all information at the time of hypothesis generation? To generate a falsifying test requires searching for information that has been left out at the time of hypothesis generation (Poletiek, 1996; 2001; 2005). In other words, how can people know where to find falsifying evidence if they have used all the evidence at their disposal to generate the hypothesis?

On the surface this claim may make intuitive sense. However, an important criticism is that participants may not have been given adequate opportunity to show they could intentionally falsify. First, participants were requested to generate three test triples in each condition which is very few in comparison with the previous literature allowing the generation of a minimum of fifteen triples (See Klayman & Ha, 1989), or up to forty five minutes of testing (e.g., Wason, 1960). Second, the results section of the experiment reports statistical analyses for the first test triple only, the remaining two triples were excluded, suggesting that the uniformity theory was initially developed from a small data set of ninety-four triples (ninety-four people generated one triple each) (Poletiek, 1996, p.455). A clear problem is that negative tests do not tend to appear until at least after the first three test triples (Klayman & Ha, 1989), and attempts at falsification may occur at a later stage in the hypothesis testing process (e.g., Mynatt, Doherty, & Tweney, 1978). The conclusion that people find it impossible to intentionally falsify may be an artifact of the limited opportunity and subsequent analysis of the data in the experiment. Accordingly, the main tenets of Poletiek's uniformity theory are summarized in Table 1.3 below. I suggest experimental tests that may falsify the theory by showing hypothesis testers can experience falsification as possible and as distinct from confirmation.

Table 1.3: Tenets of the uniformity theory (Poletiek, 1996; 2001)

---

**Tenet 1:** Falsification is impossible because it presupposes that people know where to find information to intentionally falsify a hypothesis.

*Criticism 1:* Falsification may be possible. For example when testing somebody else's as opposed to one's own hypothesis people may have information that will help to generate a falsifying test with the aim of falsifying that test. Or people can intentionally generate tests inconsistent with a current hypothesis (i.e. negative tests) and expect them to result in falsification (i.e. negative tests). Given the opportunity to test more than three triples, people may begin to use these negative falsifying tests.

**Tenet 2:** Falsification is indistinguishable from confirmation and they are the same process, because the strongest attempt at falsification of a hypothesis results in the most convincing type of confirming evidence should that attempt to falsify fail.

*Criticism 2:* The strongest attempt at falsification may lead to the most convincing type of confirming evidence should that attempt to falsify fail. Even though hypothesis testers may choose the same test, for example a negative test in the standard 2-4-6 task, an objective hypothesis tester may intend it to falsify, whereas a biased hypothesis tester may intend it to confirm. The process of confirmation and falsification may be distinct (See Table 1.1).

**Tenet 3:** A result of a hypothesis test may be as much a consequence of the quality of the hypothesis under test, as of any specific strategy employed by the hypothesis tester.

*Criticism 3:* The result of the hypothesis test may be a consequence of the quality of the hypothesis under test, but if hypothesis quality is responsible for the test result it implies that people do not have an active role in hypothesis testing. But it may be possible for individuals to be active. Consider an experiment in which two conditions are compared when the hypothesis being tested in each condition is equally untrue and an additional factor leads to falsification in one condition and not in the other.

---

These criticisms are addressed in the experimental chapters. In each case an experimental test is created using materials derived from the 2-4-6 paradigm that corresponds to each criticism and tests whether the main tenets of the uniformity theory hold (see Chapter 2 and Chapter 3). To address criticism 1 and 3 the tests seek to discover whether or not people generated different types of hypothesis tests depending on whether they themselves or somebody else owned the same hypothesis. To address criticism 2 and 3 whether or not the intention to confirm or falsify differed depending on whether participants considered an opponent hypothesis tester or not is differentiated. I turn now to examine the second main theory of hypothesis testing in the 2-4-6 task—the mathematical relationship theory.

***The mathematical relationship theory (Klayman & Ha, 1987)***

The second hypothesis testing theory I describe posits that it is the mathematical relationship between the hypothesis under test and the rule to be discovered that is the main factor affecting hypothesis testing in the 2-4-6 task (Klayman & Ha, 1987). The central idea is that the type of mathematical relationship between the hypothesis under test and the rule to be discovered constrains hypothesis testing. In the standard 2-4-6 task it is the relationship that requires the most difficult type of falsifying test which helps participants to overcome their untrue hypotheses. That is, there are two types of falsifying tests, but it is the falsifying test that people find most difficult and do not need to generate as often in hypothesis testing that is required in the standard 2-4-6 task.

Consider the situation in the standard 2-4-6 task when the participant's hypothesis is 'even numbers ascending in twos' and the properties of evenness and ascending in intervals of two are embedded in the experimenter's rule 'any ascending numbers'. 'Any ascending numbers pertains to any numbers that increase by any interval. This relationship corresponds to the typical relationship that occurs in the 2-4-6 task. This typical relationship is an 'embedded' relationship because the participant's hypothesis is embedded within the truth (i.e. embedded within the experimenter's rule). This embedded relationship is one of five possible relationships that can occur given variations of what the experimenter's rule and the participant's hypothesis could be. Klayman & Ha (1987, 1989) suggest that this embedded relationship is the most difficult for participants, because it is the only relationship that requires them to discover that their hypothesis is incorrect by generating a *negative test* that leads to

falsification, whereas positive tests which participants may find easier to generate can lead to falsification in several of the other relationship types including another type of embedded relationship.

For example, when the experimenter's rule is 'any ascending numbers' and the participant's hypothesis is 'even numbers ascending in twos' the triple 3-5-7 is a negative test because it is not an instance of 'even numbers ascending in twos' as it contains odd numbers. When the researcher replies 'yes' indicating that a triple with odd numbers is consistent with the experimenter's rule, the hypothesis pertaining to evenness is falsified.

Klayman and Ha point out that this relationship is not representative of the majority of hypothesis testing situations that can occur, and people tend to test their hypotheses using positive tests, which are more effective at producing falsification in the other hypothesis testing situations. Consider a medical professional researching the cause of a birth defect such as *Spina bifida*. Spina bifida is a neural tube defect which means the spinal cord has not developed fully during the early stages of pregnancy and most babies with spina bifida have difficulties walking. Medical professionals (Molloy & Scott, 2001) hypothesized that the main factor leading to the birth of a baby with Spina bifida was genetic. Researchers hypothesized that the birth of a baby with Spina bifida might be genetically linked because they noticed a high incidence rate of babies being born with Spina bifida in Ireland and Scotland. Having family ancestors from indigenous populations in Ireland, Scotland and parts of the UK suggests that part of one's genes comes from the *Celtic* gene pool and researchers hypothesized that the birth of a baby with Spina Bifida was genetically linked to having genes from this Celtic gene pool. Breakthrough research chose to examine blood samples and family history data collected in genetic studies of the Irish population. The choice of an Irish test population for examination was a positive test of the hypothesis that the main factor leading to the birth of a Spina bifida baby was genetically linked to the Celtic gene pool. This positive test led to a theory of genetic predisposition as a major cause of neural tube defects. If a significant pattern of neural tube defects was not observed in relationships among families from the Celtic gene pool then the scientists could conclude that their genetic predisposition hypothesis was incorrect.

Consider if the researchers had chosen to focus on a group that was not composed of the hypothesized risk factors, such as a population from the African gene pool which would qualify as a negative test of their hypothesis.

They would have found close to zero percent cases of Spina bifida and their search would not yield new information because it would be like the proverbial search for a needle in a haystack (Klayman & Ha, 1987). This example shows that it may be often more useful to examine positive instances from the group composed of the hypothesized risk factors in scientific research. For this reason Klayman and Ha suggest that people have a tendency to engage in a general *positive test strategy* because they are familiar with the usefulness of engaging in hypothesis testing in real world examples similar to the Spina bifida example. People persist in testing their hypotheses using positive tests because they are familiar with the success of using positive tests. Yet the traditional 2-4-6 task does not allow a positive test to lead to falsification and successful discovery of the experimenter's rule which we will describe in detail shortly. People must use a negative test to falsify, which Klayman and Ha argue is much more difficult given the propensity people have to engage in a positive test strategy. Klayman and Ha (1989) predict that people may be able to falsify using a negative test when they are given information to help them infer what the relationship between their hypothesis and the true rule is, for example, when the relationship between their hypothesis and the truth is made explicit to them by presenting people with an alternative hypothesis.

In what follows I illustrate in more detail how the usefulness of positive and negative tests depends on the relationship between the hypothesis and the experimenter's rule. Two types of embedded relationship that may occur in the 2-4-6 task are outlined. I show how a negative test leads to falsification in one relationship (the typical 2-4-6 task situation), and how a positive test leads to falsification in the other (the situation akin to the Spina bifida example).

### *Embedded hypotheses and falsifying test types*

The embedded situation described above is one of five possible relationships that can exist between a participant's hypothesis and the experimenter's rule (Klayman & Ha, 1987). The relationships concern how much the participant's hypothesis and the experimenter's rule overlap with one another. I will focus on three relationships that are relevant to our discussion of hypothesis test effectiveness and embedded relationships in the 2-4-6 task (See Figure 1.1).

There are two embedded situations. Critically, a negative falsifying test is best in the first situation, and a positive falsifying test is best in the second. The first embedded relationship is the one characteristic of the 2-4-6 task where the

experimenter’s rule applies to ‘any ascending numbers’ and it overlaps any triples that are even and/or ascend in twos such as when the participant’s hypothesis is ‘even numbers ascending in twos’. This relationship is illustrated in Figure 1.1(a).

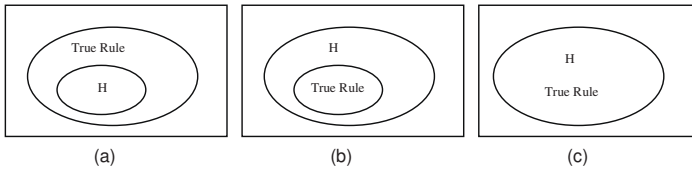


Figure 1.1: Embedded relationships between a participant’s hypothesis (H) and the experimenter’s rule (True Rule).

The only way to intentionally achieve falsification in this relationship is to use a negative falsifying test. For example, consider a participant who generates the triple 5-10-15 (which is a negative test as ascending in five or odd numbers is inconsistent with the hypothesis ‘even numbers ascending in twos’), and they expect it to be consistent with the true rule. They will receive a ‘yes’ from the experimenter, because 5-10-15 is consistent with ‘any ascending numbers’, and so they can infer that the hypothesis about ‘evenness’ and/or ‘ascending in twos’ cannot be true.

Consider on the other hand a participant who tries to intentionally falsify by generating a positive falsifying test such as 24-26-28 (which is a positive test because it is consistent with the hypothesis ‘even numbers ascending in twos’, *but* they expect it to be inconsistent with the true rule). Perhaps the rule only corresponds to triples ending in the digits 2, 4, 6, such as 2-4-6, 22-24-26 etc. This time when a ‘yes’ is received from the researcher they may not infer that the hypothesis pertaining to properties of ‘evenness’ and/or ‘ascending in twos’ is untrue. Although the positive test was intended to falsify, it cannot. It is consistent with the hypothesis and the true rule.

In the second type of embedded relationship it is possible to falsify with a positive falsifying test when the participant’s hypothesis overlaps the experimenter’s rule, for example, when the hypothesis is ‘numbers ascending in twos’ and the experimenter’s rule is this time ‘even numbers ascending in twos’. (See figure 1.1 b). (The relationship is akin to the Spina bifida example). This



time the true rule 'even numbers ascending in twos' is embedded within the hypothesis 'numbers ascending in twos'.

Consider when a participant generates the triple 3-5-7 (which is a positive test as it is consistent with the hypothesis 'ascending in twos'), and they intend it to falsify because they expect it not to be consistent with the experimenter's rule. This time they receive a 'no' from the researcher, because 3-5-7 contains odd numbers. They can infer that their hypothesis is falsified and it does not correspond to the experimenter's rule because it may pertain to numbers with the property of 'evenness'.

Now consider a participant who tries to falsify by generating a negative test such as 5-10-15. They may intend this test to falsify by expecting it to be consistent with the experimenter's rule. This time when they receive a 'no' from the researcher they may not infer that the hypothesis 'numbers ascending in twos' is untrue. The triple is both inconsistent with their hypothesis and the true rule and so cannot discriminate between them (Klayman & Ha, 1987).

The third situation is when the hypothesis is the same as the experimenter's rule. where the hypothesis 'any ascending numbers' completely overlaps the true rule 'any ascending numbers'. (See figure 1c). When a participant generates a positive test triple such as 24-26-28, even if it is intended to falsify it receives a 'yes' response. This leads to ambiguous confirmation because 'any ascending numbers' contains an infinite number of triples that can be confirmed. And negative test triples such as 6-4-2 receive a 'no' response (because descending numbers are not consistent with the true rule 'any ascending numbers'). When a descending triple receives a 'no' it does not help the participant infer that their hypothesis 'any ascending numbers' is certainly the true rule. It is not possible to be certain that the hypothesis is the truth, but it is still possible to attempt to falsify it. Each failed attempt to falsify indicates that a hypothesis is at least close to the truth (Popper, 1959).

Klayman and Ha point out that the attempt to confirm and the attempt to falsify may be distinct hypothesis testing strategies rather than a uniform test strategy in the 2-4-6 task. They distinguish between positive or negative test strategies, and they suggest that participants generally 'choose' to generate positive test strategies rather than negative test strategies (Poletiek, 2001). They also indicate that people may find falsification possible using a negative test strategy under certain circumstances such as when they are aware of what the

relationship between the hypothesis and the rule to be discovered is (Klayman & Ha, 1989).

However, their account does suggest that the hypothesis tester largely plays a passive role in hypothesis testing because they suggest the mathematical relationship between the hypothesis and the truth (in this case the experimenter's rule) is the major factor controlling how effective a participant's choice in hypothesis testing is. This suggestion is akin to the view of the uniformity theory which proposes that hypothesis quality creates a passive role for the hypothesis tester (e.g., Poletiek, 1996). In other words, both the mathematical relationship theory and the uniformity theory suggest the hypothesis tester is largely at the mercy of the properties of their hypothesis (i.e., the quality of the hypothesis and the relationship between the hypothesis and the truth), post-hypothesis generation. If this assertion is true it implies that the discussion of hypothesis testing as being biased or rational may be redundant because people may not be in a position to actively pursue a confirming or falsifying strategy, and research should start to focus more on a previous stage in the process such as the hypothesis generation stage. An important objective for this thesis is to investigate whether people choose their tests in a way that reflects an active role for a hypothesis tester.

In Table 1.4 below I summarise the main tenets of Klayman and Ha's mathematical relationship theory. Accordingly, I suggest experimental tests that may falsify the theory by showing that hypothesis testers can experience falsification as possible even in the hypothesis testing situation they find most difficult: when their hypothesis is embedded within the true rule.

Table 1.4: Tenets of the mathematical relationship theory (Klayman & Ha, 1987; 1989)

---

**Tenet 1:** There is a tendency to test instances that are consistent with a hypothesis. This tendency is called a *positive test strategy* and it is usually a helpful strategy in hypothesis testing, such as in the Spina bifida example. In the 2-4-6 task a positive test strategy is not the same as confirmation bias. Even if the participant intends their positive test to lead to falsification, it can only lead to confirmation of their incorrect hypothesis. The relationship between the hypothesis and the true rule constrains the effectiveness of the positive test

strategy. Only negative tests can falsify in the mathematical relationship standard of 2-4-6 task; but participants find it difficult to disengage from positive testing and that is why they are not successful.

*Criticism 1:* People sometimes successfully discover the rule in the 2-4-6 task. Perhaps there are conditions under which people readily follow a negative test strategy, for example, when they are competing with an opponent. The finding that participants disengage from positive to negative tests in the embedded relationship typical of the 2-4-6 task, with or without knowledge of task constraints, would indicate that the mathematical relationship alone cannot predict hypothesis testing—people can play an active role in hypothesis testing. For example, when people consider an alternative hypothesis in addition to the initial hypothesis they may generate negative tests.

*Tenet 2:* People often do not know when a positive test strategy is wise and when it is not.

*Criticism 2:* If people can show that they know when it is wise not to use a positive test strategy, and rely on a negative test strategy instead, then they do know when a positive test strategy will not work. For example, when people test a hypothesis belonging to somebody else that they know is untrue, they may rely on negative tests.

*Tenet 3:* The mathematical relationship between the hypothesis and the experimenter's rule affects how useful positive and negative test strategies are.

*Criticism 3:* The mathematical relationship only affects whether or not a positive or negative test will lead to a confirming or falsifying outcome. The relationship does not affect the part of the process where people intend to falsify or confirm. For example, when people compete with an opponent hypothesis tester, they may attempt to falsify their hypotheses by generating negative tests, but they may actually intend these negative tests to confirm.

---

*Negative tests and embedded relationships in the real world: Prejudiced beliefs*  
A critical question to ask is whether the ability to falsify in an embedded relationship using a negative test is important to hypothesis testing in the real world? To address this question it is essential to understand the zeitgeist in which the 2-4-6 task was designed. While the initial experiments on

confirmation bias in psychology were being carried out in reasoning, social psychology in post-holocaust America was attempting to understand the conditions under which prejudiced behaviour could lead to harmful consequences such as genocide (see Gross, 1999; Aronson, 1999). While social psychologists were examining how external factors such as being ordered to perform an action could be used to justify harming another person (e.g., Milgram, 1963/1974), little was known about the internal factors used to justify this behaviour, such as perceiving a person as being less than another human being in order to maintain a type of prejudiced thinking.

Popper (1959; 1963) included a moral force to his argument which advocated that falsification was a way of revising prejudiced thinking, especially in light of the way Jews were perceived during the holocaust (e.g., Popper, 1959, see also 1992). The idea was that a prejudiced belief is untrue and it is often embedded within the truth because it may often be possible to find cases to confirm a prejudiced belief about a group of people. That is, a prejudiced belief may consist of a smaller set of positive instances confirming a belief which is untrue of the total population of the group against which the prejudice is targeted. However, a negative instance which would falsify it exists outside of that set. To illustrate how falsification by negative testing may play a role in the revision of a prejudiced belief, consider the example provided by Anne Frank in her famous diary:

“...Jews are regarded as lesser beings. Oh, it’s sad, very sad that the old adage has been confirmed for the umpteenth time: ‘What one Christian does is his own responsibility, what one Jew does reflects on all Jews’.”

(22<sup>nd</sup> May, 1944, p. 302)

If someone held the prejudiced belief that Jews were lesser beings they may cite cases which are consistent (positive instances) with this belief, such as a person with a criminal record who was also Jewish, and avoid any inconsistent instances (negative instances) which exist outside of their collection of confirming evidence. But a falsifying case is available in the very human story of Anne Frank, and one falsifying case can prove that this prejudiced belief is false. Likewise, that the Frank family received help from non-Jews falsifies the persistence of the belief that all Christians are prejudiced against Jews. The standard version of the 2-4-6 task, when the participant’s hypothesis is

embedded within the true rule, is analogically equivalent to this prejudiced belief (Wason, 1960). When the participant's hypothesis is 'even numbers ascending in twos' they may search for an infinite number of positive cases such as 10-12-14, 16-18-20, and avoid negative cases outside of the set of even triples ascending in twos. For example, the triple 5-10-15 would also receive a 'yes' response from the researcher which will show that the truth does not only concern triples that are even and ascend in twos.

As a result people should seek to challenge the hypotheses which they *believe* to be true to prevent the maintenance of beliefs with potentially harmful consequences (Popper, 1959; Wason, 1960). A distinctive feature of irrationality is to maintain beliefs with consequences inconsistent with the facts (Johnson-Laird, Girotto, & Legrenzi, 2004), and ultimately the search for evidence in hypothesis testing is the search for truth, and the truth should be objective and not prejudiced (Popper, 1963).

### **Alternative Hypotheses Accounts**

One way to begin to overcome a prejudiced belief is to consider the alternative, for example, that Jewish people are people and all peoples deserve equal standing. Perhaps the consideration of an alternative hypothesis may help people to search outside of their untrue hypothesis in order to discover that their belief is prejudiced or even that their hypothesised rule is incorrect (e.g., Wason & Johnson-Laird, 1972). Presently there is a collection of novel experimental findings relating to the role alternative hypotheses have in hypothesis testing. There is also a collection of explanations for how the consideration of an alternative may help the discovery of the truth. In what follows these accounts are outlined to show how the consideration of an alternative hypothesis may in fact play a pivotal role in revealing falsifications and the discovery of the truth in hypothesis testing. An advantage of addressing the role of alternative hypotheses in hypothesis testing is pointed out, namely, that alternatives may go hand in hand with falsification in the discovery of the truth. Finally, I make suggestions about how the inclusion of these accounts may be advantageous to our present understanding of hypothesis testing.

The first experimental result to highlight the facilitating role of considering alternative hypotheses in rule discovery was the DAX-MED experiment carried out in the 2-4-6 task (Tweney *et al.*, 1980). In this version of the 2-4-6 task participants are told that there are two rules to be discovered; a DAX rule and a

MED rule. The DAX rule was the standard ‘any ascending numbers’ rule and the MED rule was ‘any other number sequence which does not ascend’. Participants were instructed to generate number triples and the researcher responded with the feedback ‘DAX’ or ‘MED’ rather than ‘yes’ and ‘no’ respectively. Participants discovered the rule ‘any ascending numbers’ significantly more frequently than the usual 21% rule discovery rate; 60-80% of participants tend to discover the rule in DAX-MED manipulations, even though they have generated the same number of test triples (e.g., Valle-Tourangeau, Austin & Rankin, 1995; Wharton, Cheng & Wickens, 1993; Gale & Ball, 2003; 2005).

Yet little is known about how considering alternative hypotheses facilitates rule discovery. Initial DAX-MED experiments found that participants simply generated more triples in the DAX-MED feedback condition than in the standard ‘yes’-‘no’ feedback condition and it was proposed that increased rule discovery rates were due to the additional information obtained from testing additional triples (e.g., Wharton, Cheng & Wickens, 1993). This *information quantity* hypothesis was ruled out by the results of an experiment which allowed participants to generate a fixed number of triples (15 triples in each condition). In a standard ‘yes’-‘no’ feedback condition and a DAX-MED feedback condition it was found that rule discovery remained elevated in the DAX-MED condition and at the typical ~20% in the standard condition (Vallee-Tourangeau *et al.*, 1995).

A second explanation was that people generally have a bias to process information with a positive label as opposed to a negative label, for example ‘yes’ versus ‘no’ or ‘right’ versus ‘wrong’, and that the labels DAX and MED allow the processing of more triples (e.g., Evans, 1989). This *positivity bias* hypothesis was ruled out by the results of an experiment with two groups of participants who were asked to either discover one rule or two rules, and in half of each group the linguistic labeling for feedback was either ‘DAX’-‘MED’ or ‘fit’-‘does not fit’ (Gale & Ball, 2003). Regardless of feedback participants discovered the rule more often when they were instructed to discover two rules rather than one. The rule discovery rate was 73% when participants were instructed to discover two rules and were given ‘fit’-‘does not fit’ feedback; participants appear to be able to process information with a negative label (Evans, 1989).

A third explanation is that the DAX-MED manipulation induces a mental

representation which requires less effort to switch between two alternative hypotheses and test them both simultaneously. That is, the two hypotheses are complementary to one another; one is 'ascending' and one is 'not ascending' (Wharton, Cheng, & Wickens, 1993). This *complementarity* hypothesis was questioned by the results of an experiment that found similar rule discovery rates in a condition with feedback inducing non-complementary representation by labeling triples 'DAX'-'MED'-'DAX or MED', and in the standard DAX-MED condition (Vallee-Tourangeau *et al.*, 1995).

Oaksford & Chater (1994) suggest that participants may find the consideration of an alternative hypothesis useful because it helps them to decide which information to include or exclude from their hypothesis. For example, if the hypothesis under test is 'even numbers ascending in twos' and the experimenter's rule is 'any ascending numbers' participants may generate the alternative hypothesis 'numbers ascending in twos' which excludes the property of evenness (see also Farris & Revlin, 1989). Participants may then generate the triple 5-7-9, which ascends in twos, and ascends but is not even. If participants receive a 'yes' response from the experimenter for the triple 5-7-9, then they can revise their hypothesis to 'numbers ascending in twos' by excluding the property of evenness (see also Gale & Ball, 2005).

This varying of one property at a time may help participants to identify what is wrong with a hypothesis. Oaksford and Chater's account assumes that participants need to generate a specific alternative at the outset in order to discover what should be excluded from a hypothesis. But, why would participants with the hypothesis 'even numbers ascending in twos' choose an alternative hypothesis which would bring them towards the truth such as 'numbers ascending in twos' (the true rule is 'any ascending numbers')? What prevents them from generating an alternative which would bring them away from the truth such as 'even numbers ascending in twos ending in the digits 2, 4, 6' (see Klayman & Ha, 1989). This time a triple with the properties of evenness, ascending in twos and not ending in the digits 2,4,6 such as 28-30-32 would show that the alternative hypothesis was incorrect. In fact the DAX-MED manipulation is largely successful when participants are not required to generate either a specific hypothesis or alternative hypothesis at the outset. It is possible that a specific, that is, an explicit alternative is not always needed to generate a falsifying test. That is, participants may implicitly represent the concept that another alternative exists and that it may offer a better explanation than their

hypothesis.

A similar phenomenon may occur in science. For example, Kuhn (1962) suggests that the competitive endeavour of science presents readily available alternatives to scientists such as two alternative teams of scientists battling for a discovery. For example scientists may try to falsify alternative theories belonging to rival scientists (e.g., Gorman, 1995a; 1995b). In this way alternative hypotheses are generated by others, and previous research has had very little to say about how competition between hypothesis testers may facilitate falsification of others' hypotheses or exacerbate confirmation of one's own hypothesis.

Despite the wealth of findings from research on hypothesis testing in the 2-4-6 task there are still several unanswered questions. In particular we still know very little about how participants mentally represent hypotheses, even though mental representation is critical to theories of reasoning (e.g., Johnson-Laird & Byrne, 1991). In Chapter 2 we will pay particular attention to the way people consider alternative hypotheses. In Chapter 3 we will examine how competing with another hypothesis tester may lead to the consideration of alternative hypotheses, and in Chapter 4 we will take a closer look at how people mentally represent their hypotheses. Presently, we do not know how active the role of the participant as hypothesis tester is, even though an active role is vital to proving that reasoning can be biased (e.g., Kahneman, Slovic, & Tversky, 1982). In chapters 2 and 3 particular attention is paid to how people can play an active role in the process of hypothesis testing in the 2-4-6 task, both in terms of their test choice and their intention to use their test to confirm or falsify. We also know little about how prior knowledge may affect hypothesis testing, even though an understanding of how knowledge interacts with the task at hand is essential to theories of expert reasoning (e.g., Chase & Simon, 1973a; 1973b; Newell, 1990). In Chapter 4 I pay particular attention to how domain knowledge may affect the selection of hypothesis tests, and how the evaluation of the test result may have been underestimated as an important stage in hypothesis testing. In addition we know little about how factors such as competition may affect hypothesis testing, even though factors which require interaction with another person can play a critical role in belief bias in social contexts (e.g., Kruglanski & Webster, 2000). In Chapter 3 therefore we will examine how competition with another hypothesis tester may affect hypothesis testing. In the next section I outline the aims of the thesis in more detail, and introduce the two novel



approaches used in this thesis to address these aims.

### **Aims of the Thesis**

The aims of this thesis was to discover whether hypothesis falsification is consistently possible in human thinking, whether there are factors which facilitate hypothesis falsification, and whether hypothesis falsification is important to expert thinking. Two novel experimental approaches are introduced in the thesis to address these aims. The first novel experimental approach brings an interactive component to the standard 2-4-6 task by introducing an imaginary participant to the task. This concept allows us to address whether alternative hypotheses and competition may play a role in hypothesis testing and hypothesis falsification. The second novel experimental approach brings together two previously disparate domains of cognitive research, namely, the domains of hypothesis testing and chess problem solving. This approach allows us to address the role that domain knowledge and mental representation may play in hypothesis testing and hypothesis falsification.

### ***Hypothesis testing in an imaginary participant 2-4-6 task***

I chose the 2-4-6 task as our first approach because the task has been a standard test bed for theoretical development in hypothesis testing over the last forty-five years (see Poletiek, 2001 for a review). In this review of the task I showed that it presents a well-defined hypothesis testing situation for which the relationship between a hypothesis and the available evidence is precisely worked out (e.g., Klayman & Ha, 1987). In addition it has a well-defined logic for classifying confirming and falsifying hypothesis testing (e.g., Poletiek, 1996; Wetherick, 1962). The well-defined nature of the task allows us to bring an interactive component to the standard 2-4-6 task by introducing an imaginary participant to the task, and allows us to directly compare our results with existing theories. Participants are asked to test an imaginary participant's hypothesis in some of the experiments I report (see Chapter 2). In others participants are told that an imaginary participant (i.e., an opponent) is also testing their hypothesis (see Chapter 3). The concept of introducing an imaginary participant allows us to directly address the questions of competition and the active role of the hypothesis tester.

In Chapter 2, a series of experiments designed to examine how the consideration of an alternative hypothesis affects how people test the hypotheses

are reported. Participants are told what Peter hypothesizes the experimenter's rule to be and they must generate number triples to test Peter's hypothesis. In a series of three experiments I vary the properties of the alternative hypothesis to investigate how hypothesis quality, knowledge of hypothesis quality, explicit and non-explicit alternatives, and misleading alternatives affect falsification and rule discovery.

In Chapter 3, the role of competition in hypothesis falsification in a series of two experiments that compare hypothesis testing between hypotheses owned by the self and by the imaginary participant are examined. In this series of two experiments a comparison of hypothesis testing between self owned hypotheses when participants competed with an opponent, or did not compete with an opponent, is made. The implications of the findings of the two series of experiments for the two main alternative theories of hypothesis testing are examined in detail.

### ***Chess masters' hypothesis testing***

In Chapter 4 experimental data on hypothesis testing in chess problem solving is presented. I chose the domain of chess as our second approach to address further the theoretical shortcomings in hypothesis testing research. Hypothesis testing of chess players has not previously been investigated, and the chess domain is akin to both controlled laboratory tasks such as the 2-4-6 task and the complexity of realistic contexts (DeGroot, 1965; Newell & Simon, 1972). Chess has also been the standard testbed for theorising about the nature of domain knowledge (e.g., Chase & Simon, 1973a; 1973b), and chess players must play moves so good that they may not be refuted (falsified) by their opponent (Saariluoma, 1995), yet we do not know how chess players identify falsification of their own or their opponent's planned moves. Studies of chess have contributed substantially to our understanding of human cognition (e.g., Newell & Simon, 1972; Gobet & Simon, 1996a; 1996b). Chess findings provide strong external validity and have contributed to explanations of expertise in non-game domains such as physics (Larkin, McDermott, Simon, & Simon, 1980; Chi, Glaser, & Rees, 1982), computer programming (McKeithen, Reitman, Rueter, & Hirtle, 1981), and music (Sloboda, 1976). Chess provides a neat micro-world of cognition in which to investigate different aspects of thinking with unparalleled precision (Newell, 1990). For example, chess research employs protocol analysis which allows reliable recording of what is currently mentally

represented while solving a problem (Ericsson & Simon, 1993), and chess players use fluent algebraic notation when thinking aloud to describe the precise square coordinates where chess pieces are imagined to move to (e.g., DeGroot, 1965; Nunn, 1999). An experimental analysis of hypothesis testing comparing chess players with different levels of expertise allows us to address the question of how domain knowledge affects hypothesis falsification. The use of a protocol analysis methodology allows us to address the question of how hypothesis testing is mentally represented, how tests are searched for and how they are evaluated (see DeGroot, 1965).

I conceptualize each move a chess player considers as a hypothesis, and the opponent moves they consider in response to each move as tests of the hypothesis. An error in chess is a *refutation*, that is, a disproof that a move under consideration for play will work out in a player's favour. Instead, an opponent move will lead to a worsening of the player's position. To secure the best possible result in chess, players must play moves that are so good they cannot be refuted (Saariluoma, 1995). Tests of these moves may be either confirming (the opponent's move fits in with the player's plan, that is, it leads to an improvement in the player's position), or they may be falsifying (the opponent's move refutes the player's plan, that is, it leads to a worsening of the player's position). That is, the player's conception of the board problem is a 'working hypothesis' (De Groot, 1965, p. 395). In this thesis I suggest that expert chess players may be better players than novices precisely because they are better at anticipating how their plans may be falsified by the opponent's counter moves. The reason put forward is that experts consider moves for play that are less likely to be refuted because they have accessed them from their domain knowledge which is superior to that of novices (e.g., Chase & Simon, 1973a; 1973b; Gobet & Simon, 1996a; 1996b; 1996c; Gobet, 1998; Gobet *et al.*, 2004).

In Chapter 4 an experimental analysis of chess masters' hypothesis testing enables us to examine the implications hypothesis falsification may have for expert thinking, problem solving, planning, and learning from experience. To this end the thesis examines how expert knowledge affects hypothesis falsification by comparing the hypothesis testing of master and novice chess players. I suggest that hypothesis testing is an important component of chess expertise that will enable theories of cognitive expertise to move forward from pattern recognition frameworks.

In Chapter 5, the final chapter, the implications of the findings from the 2-4-

6 experiments in this thesis and hypothesis testing in chess expertise for the two main theories of hypothesis testing, the uniformity theory and the mathematical relationship theory are discussed. How these experiments challenge the views that falsification is impossible; that confirmation and falsification constitute the same strategy; and that the hypothesis tester's role is passive and dependent on the quality of the hypothesis and the relationship between their hypothesis and the truth is argued. The discussion emphasises how domain knowledge is important to hypothesis falsification, and future theories addressing the role that alternative hypotheses and domain knowledge may play to explain the findings of the thesis and lead to a more complete theory of hypothesis testing are surmised. Let us now turn to examine the role of alternative hypotheses, competition, and expert knowledge in hypothesis testing and hypothesis falsification.



## Chapter 2 The Role of Alternative Hypotheses in Hypothesis Testing: An Imaginary Participant 2-4-6 Task

*Scientific observation is not merely pure descriptions of separate facts. Its main goal is to view an event from as many perspectives as possible.*

– A. R. Luria, (1987, p. xv)

The aims of this chapter are twofold. The first aim is to describe an adapted version of a reasoning task that has been the test bed for theories of hypothesis testing for over forty years: the 2-4-6 rule discovery task (Wason, 1960; 1968). The second aim is to report a series of experiments examining hypothesis testing in the new version of the task.

In the standard 2-4-6 task participants are asked to discover a rule a researcher has in mind that the number sequence 2-4-6 conforms to. Participants are instructed to generate hypotheses about what the experimenter's rule might be. They test the truth of their hypotheses by generating hypothesis tests consisting of number sequences with sets of three numbers (referred to as triples from this point onwards). As we saw in Chapter 1, participants have been found to overwhelmingly confirm their hypotheses and rarely falsify them even though they are inaccurate (e.g., Wason, 1960; Tweney *et al.*, 1980; Klayman & Ha, 1989; Poletiek, 1996). The aim in this chapter is to investigate if participants can falsify an untrue hypothesis under some conditions. One central implication that has been drawn from experimental work in the 2-4-6 task is that people find it difficult if not impossible to falsify their hypotheses (Poletiek, 1996; 2001).

Consider the real world situation where a teacher corrects a student's incorrect hypothesis by providing a falsifying example to it. For example, when teaching someone how to play chess they may play a move and ask "is that a good move?" Suppose the move is actually a bad move. The teacher can provide the student with a falsifying example to show them why it is a bad move by showing them how an opponent piece may move to ruin their plan. This example contains properties that are akin to hypothesis testing research which indicates falsification may be possible. For example, it has been suggested that falsification may be possible if the relationship between the hypothesis and the experimenter's rule is made explicit to the participant (Klayman & Ha, 1987; 1989). Falsification may also be possible when a participant approaches the task as a group problem solving task (Gorman *et al.*, 1984), or when a participant

evaluates negative tests previously generated by other people and can understand that these tests falsify a hypothesis (Kareev & Halberstadt, 1993).

How does the chess example correspond to these findings and the possibility that people are able to falsify in the 2-4-6 task? Consider when a participant knows the answer to the 2-4-6 problem and they also know what another participant's incorrect hypothesis is. Can a participant who knows the answer provide a falsifying triple to another participant that will help them realise that his/her hypothesis is incorrect? In the next section this question is applied to a 2-4-6 situation in which participants must generate falsifying tests of someone else's hypothesis to demonstrate how they are incorrect in their hypothesis testing. This variation of the task introduced participants to an imaginary participant called Peter, and participants are asked to test Peter's hypothesis.

### **The Imaginary Participant 2-4-6 Task**

The first aim of this chapter is to describe this new imaginary participant 2-4-6 task. The imaginary participant 2-4-6 task is identical to the standard 2-4-6 task except that the participant is given a hypothesis to test that belongs to 'Peter'. In the standard 2-4-6 task participants are instructed to discover a rule the number triple 2-4-6 conforms to. Participants generate hypotheses about what the rule might be, and they generate triples to test whether or not their hypothesis is the rule. Participants in the following experiments were given a hypothesis to test. They were told 'a participant called Peter was asked to discover a rule an experimenter had in mind that the number triple 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was 'even numbers ascending in twos'. Participants are then told to go about testing if Peter's rule is the experimenter's rule by generating your own number triples with sets of three numbers... to test the rule'.

This new imaginary participant 2-4-6 task has a number of advantages. First, by giving participants specific hypotheses to test we can control how representative the hypothesis is of the truth, that is, how high or low-quality the hypothesis is. We can control what hypotheses the participants test. Second, this control of hypothesis quality in addition to the well-defined logic of the 2-4-6 task allows precise comparisons of hypothesis testing in this imaginary participant task with hypothesis testing in the previous literature (e.g., Klayman & Ha, 1989; Poletiek, 1996). Third, we can control what alternative hypotheses the participant may consider alongside the initial hypothesis (Peter's hypothesis,

that is), by giving participants the alternative hypothesis. Fourth, we can introduce competitive factors to a laboratory task by asking people to consider someone else's hypothesis. In this way the introduction of an imaginary participant may add to the ecological validity of the classic version of the task (see Halberstadt & Kareev, 1993).

The second aim of this chapter is to describe the first experimental study of hypothesis testing carried out in this imaginary participant 2-4-6 task. The chapter reports a series of three experiments that focus on the facilitation of hypothesis falsification in hypothesis testing, that is, searching for evidence that refutes a hypothesis. The results are reported in the context of the two main alternative theories of hypothesis testing derived from findings in the 2-4-6 task. As we saw in Chapter 1, the uniformity theory posits that people find falsification to be psychologically difficult if not impossible (Poletiek, 1996). People experience a confirming and falsifying hypothesis test as simply a test, that is, hypothesis testing is a uniform process (Poletiek, 2001; 2005). The second theory posits that the mathematical relationship between the hypothesis under test and the rule to be discovered constrains how effective different types of hypothesis tests are in the 2-4-6 task (Klayman & Ha, 1987). The central idea is that the mathematical relationship between the hypothesis under test and the rule to be discovered in the 2-4-6 task has tended to be the relationship that requires the most difficult type of falsifying test. There are two types of falsifying tests, but it is the falsifying test that people need to generate less often in hypothesis testing in general and find most difficult that is required in the standard 2-4-6 task, that is, a negative falsifying test.

In this chapter the main tenets of each of these two theories are tested in three imaginary participant experiments. In Experiment 1 the possibility that participants can falsify using this negative falsifying test consistently when they must is tested, that is, when the hypothesis they are testing is an inaccurate and therefore a low-quality hypothesis. In Experiment 2 the conditions in which participants can falsify by examining properties of the quality of the alternative hypothesis are further investigated. In Experiment 3 a further property of the alternative hypothesis, that is, how explicit the hypothesis is and how this explicitness affects falsification is examined in more detail.



### **Experiment 1: Falsifying a low-quality hypothesis**

This experiment aims to investigate if people find falsification consistently possible when they must falsify a hypothesis when the hypothesis is low-quality. A low-quality hypothesis refers to a hypothesis that does not accurately represent the truth; it is either untrue or untrue in parts. To show that participants find it possible to consistently falsify, the evidence must show that participants exhibit insight into the implications of particular test choices, by understanding how they will be interpreted by the participant (Peter) who owns the incorrect hypothesis. In other words, the participant who knows the answer to the 2-4-6 task must be shown to *intend* to falsify by indicating that they expect a falsification. For example, participants who know the solution to the 2-4-6 problem (any ascending numbers) may know that Peter's hypothesis is incorrect (even numbers ascending in twos), and they may generate a falsifying test (5-10-15) and expect it to falsify (because they know what the rule is). In this experiment I ask two questions related to this example: (1) Does the quality of the hypothesis affect the ability to falsify; and (2) can people consistently falsify a hypothesis belonging to someone else when they know it is a low-quality and untrue hypothesis?

First, does the quality of a hypothesis under test affect the extent people can confirm or falsify? For example, hypotheses of a higher quality may be less likely to result in refutations (i.e. falsifications) than low-quality hypotheses. The implication is that the quality of the hypothesis under test may determine the availability of confirming or falsifying evidence and hence the extent people can confirm or falsify. It follows if people can falsify then it will be in a situation where to falsify will be of help, for example, when testing a low-quality hypothesis. By falsifying, the participant can conclude that the hypothesis is not true and abandon it. To this end it is constructive to compare participants testing a low-quality hypothesis and compare their hypothesis testing to participants testing a high-quality hypothesis (a high-quality hypothesis refers to a hypothesis that is representative of the truth).

Second, can participants consistently generate negative falsifying tests? Poletiek (1996) identified a major problem for hypothesis falsification. She suggested that many experiments tended to equate falsifying instructions with negative testing and as a result it was not clear if people were aware they were generating tests that could objectively lead to falsification. Negative confirming tests could have been scored as falsifying tests mistakenly in these studies

(Wetherick, 1962). A way of tackling this problem is to measure falsifying tests by taking the participant's *intention* to falsify into account (as outlined in Table 1.1 Chapter 1). A test must not only be an inconsistent and therefore a negative instance of a hypothesis, but it must also be expected by the participants to result in falsification. The participant must indicate that they expect a falsification before the experimenter replies whether the test actually falsifies the hypothesis or not. This method is presently considered the best way of scoring hypothesis tests (e.g., Poletiek, 1996; Rossi *et al.*, 2004; Wetherick, 1962). This shows it is important to how the tenets of the two main theories are tested because it may indicate that people can differ in their processing of hypothesis tests.

As we saw in Chapter 1, Poletiek (1996) investigated whether people found it possible to intentionally falsify their hypotheses by expecting their negative tests to result in falsification. In the standard 2-4-6 task participants had to come up with their best guess about what the experimenter's rule was. They were presented with the number triple 2-4-6 and the experimenter's rule was the usual 'any ascending numbers' rule. There were three conditions: participants were given instructions to "test" or "confirm" or "falsify". For the "test" and "confirm" conditions, the majority of tests fell into the positive confirming category (86% and 80% respectively), and the least amount of tests fell into the critical negative falsifying category (0% and 3% respectively). Participants in the "falsify" condition were instructed to "try to test in such a way as to get your hypothesis about the rule rejected" (Poletiek, 1996; p.454). The majority of tests in this condition fell into the two confirming categories, the positive confirming and negative confirming categories (32% and 54% respectively). Although the participants instructed to falsify proposed test triples that were negative tests, they in fact intended them to confirm. It was concluded that people do not seem to be able to make sense of falsification; people do not seem to find falsification to be a possible strategy. Participants did not appear to have insight into the implications of their test choice; they did not realise that their negative tests resulted in falsification.

This experiment tests whether people can in fact make sense of falsification under some circumstances. The situation where a participant must test the imaginary participant Peter's hypothesis is examined in order to show him that his hypothesis is a low-quality one and compare this to the situation where a

participant must test the imaginary participant Peter's hypothesis in order to show him that his hypothesis is a high-quality one.

Hypothesis quality was defined in terms of how closely the hypothesis corresponded to the correctness of the researcher's rule. That is, the hypothesis quality was based on the number of numerical properties that corresponded to the correctness of the researcher's rule. For example, when the hypothesis was 'any ascending numbers' it was 100% correct, because it corresponded perfectly to the researcher's rule 'any ascending numbers'. When the hypothesis was 'numbers ascending in twos' it was half correct (50%), because 'ascending numbers' was one of two numerical properties that corresponded to the researcher's rule. When the hypothesis was 'even numbers ascending in twos' it was one third correct (33%), because 'ascending numbers' was one of three numerical properties that corresponded to the researcher's rule. This measure is a crude measure, but it was the logical criterion available given the constraints that: (i) the same embedded relationship between the hypothesis under test and the alternative hypothesis must be used; and (ii) approximately equivalent interval decreases in hypothesis quality should occur (see Klayman & Ha, 1989).

The experiment makes the prediction that it may be possible to generate falsifying tests in this imaginary participant 2-4-6 task because participants are presented with the solution to the problem (they are given the correct alternative hypothesis 'any ascending numbers') and this alternative indicates that the hypothesis under test (Peter's hypothesis is 'even numbers ascending in twos') is incorrect, because the participant knows it is the actual rule to be discovered, much in the same way a teacher falsifies a student's inaccurate hypothesis. Counter to Poletiek (1996; 2001), I predict participants may generate tests to consistently falsify Peter's hypothesis in order to show him he is incorrect and that the quality of the imaginary participant's hypothesis may affect hypothesis testing.

## **Method**

### ***Materials and design***

One group of participants were told that Peter hypothesised that the rule was 'even numbers ascending in twos' (low-quality hypothesis), and another group were told that Peter hypothesized that the rule was 'any ascending numbers' (high-quality hypothesis, which is in fact the researcher's rule). Hypothesis

quality in this instance refers to how closely the hypothesised rule fits the experimenter's rule. The experimenter's rule was the standard 'any ascending numbers'. Half of the participants in each case of hypothesis quality were given knowledge about hypothesis quality by being told the solution to the 2-4-6 problem. They were given the following additional sentence: "The experimenter's rule was in fact 'any ascending numbers'". Participants were assigned at random to one of four groups (known low-quality, unknown low-quality, known high-quality and unknown high-quality,  $n = 16$  in each). Appendix A provides the full instructions given to each group. Participants in all four groups were instructed to generate number triples of their own. They had to test Peter's hypothesis in such a way as to help him discover if his rule was the experimenter's rule. To illustrate the task, the example of the known low-quality condition is provided:

"In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. The experimenter's rule is in fact 'ascending numbers'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to.

When you feel highly confident that you have helped Peter discover that his rule is not the experimenter's rule, and not before, you are to write down "Peter now knows his rule is not the experimenter's rule". You are to write this under your most recent number sequence. The researcher will then write whether or not you are correct beside your announcement."

In the unknown low-quality condition participants were given Peter's untrue hypothesis 'even numbers ascending in twos' to test and they were not told anything about the experimenter's rule to indicate which quality Peter's

hypothesis was. In the known high-quality condition participants were given Peter's true hypothesis 'any ascending numbers' which was the experimenter's rule and they were told it was the experimenter's rule. Lastly, in the unknown high-quality condition participants were again given Peter's true hypothesis 'any ascending numbers' and they were not told anything about the experimenter's rule to indicate which quality Peter's hypothesis was.

Each participant was given a three-page recording sheet which had 5 columns. Appendix B provides a copy of the recording sheet. The first column was labelled 'number sequence', in which participants wrote their number triple, e.g., 2, 4, 6, and the second column was labelled, 'reasons for choice', in which participants provided a justification for their triple. The next two columns required participants to provide 'yes' or 'no' answers to the questions, 'do you expect it to conform to Peter's rule?', 'do you expect it to conform to the experimenter's rule?'. The fifth column was headed 'Feedback from the experimenter: does your number sequence conform to the experimenter's rule'. Feedback was given in the form of a 'y' for a yes and an 'n' for a no as to whether or not the generated number sequence conformed to the experimenter's rule. There were 18 rows in the recording sheet. There were also spare sheets for participants to insert as many tests as they considered necessary. The dependent measures were: the number of triples generated by the participants; the percentage of confirming and falsifying triples; the percentage of positive and negative tests; the number of correct announcements of the rule achieved by the participants at the end of the task (in the two conditions in which the participants did not know what the rule was).

### ***Participants and procedure***

The participants were 64 members of the psychology department's participant panel, that is, members of the general public recruited through national newspaper advertisements, who were paid a nominal fee (8 euro) and undergraduate students who participated for course credits. The 50 women and 14 men were aged from 15 years to 73 years, with a mean age of 35 years. No participants had taken courses in the philosophy of science.

Participants were tested individually and in small groups of up to four individuals. The experimenter read the instructions aloud to participants (and the participant could re-read the instructions by themselves if they wished). The participants were told that they could take as long as they needed to complete

the task. Most participants took approximately fifteen minutes to complete the task.

## Results and discussion

Next I report the number of triples generated by the participants, the percentage of confirming and falsifying triples, the percentage of positive and negative tests, and the number of correct announcements of the rule achieved by the participants.

### *Number of triples*

Participants generated 343 number triples, and an average of 5.36 number triples per participant. The mean number of triples that were generated for each condition is illustrated in Table 2.1. Reliably more triples were generated for high-quality hypotheses than low-quality hypotheses (6.28 versus 4.44, Mann-Whitney  $U_{32,32} = 363$ ,  $Z = -2.025$ ,  $p = .043$ , two-tailed). There was no difference in the triples generated for hypotheses for which the quality was known or unknown (5.40 versus 5.31, Mann-Whitney  $U_{32,32} = 463$ ,  $Z = -.666$ ,  $p = .505$ , two-tailed).

Table 2.1: The number of triples generated in each condition of Experiment 1.

	<i>Known</i>	<i>Unknown</i>	<i>Total</i>
High-quality	6.56	6.00	6.28
Low-quality	4.25	4.63	4.44
<b>Total</b>	5.40	5.31	5.36

There was no difference between the number of triples generated for the low-quality hypothesis for which the quality was known or unknown (4.25 versus 4.63, Mann-Whitney  $U_{16,16} = 91.5$ ,  $Z = -1.401$ ,  $p = .171$ , two-tailed). And there was no difference between the number of triples generated for the high-quality hypothesis for which the quality was known or unknown (6.56 versus 6.00, Mann-Whitney  $U_{16,16} = 117.5$ ,  $Z = -.400$ ,  $p = .696$ , two-tailed).

The results show that the fewest number of triples were not generated when participants tested a low-quality hypothesis, and knew they tested a low-quality hypothesis. This result suggests that when participants know a hypothesis is

untrue, they test as much as when they do not know whether a hypothesis is true or not.

The results also show that participants did not generate more test triples for high than low-quality hypotheses, and that the highest number of triples was not generated by participants who tested a high-quality hypothesis and knew it was a high-quality hypothesis. The result suggests that when a hypothesis is high-quality, people do not necessarily assume the best way forward is to confirm the hypothesis as much as possible.

### ***Correct announcements***

Participants' announcements of Peter's rule as being either correct or incorrect in the unknown conditions were scored by allocating a score of one for a correct announcement and a score of zero for an incorrect announcement. The percentages of correct announcements were 100% for the high-quality unknown condition and 56% for the low-quality unknown condition, and this difference was reliable, ( $\chi^2 = 10.667 (1), p = .001$ ). This result may not be surprising because participants in the high-quality unknown condition are given the researcher's rule as the hypothesis to test. As in real life where one scientist may test a significantly higher quality hypothesis than another scientist, it was instructive to compare correct announcements between participants who tested the hypothesis that was the researcher's rule, with participants who tested the hypothesis that was not the researcher's rule, when neither group knew what the researcher's rule was. Achieving falsification was not possible when the hypothesis is high-quality: even if the participant expects a falsification it will result in confirmation because unknown to them the hypothesis is the experimenter's rule. At the end of the testing session they will have accumulated many instances of confirmation and no falsifying evidence is available. They can correctly announce that Peter's hypothesis is the experimenter's rule. When the hypothesis is low-quality more than half of the participants tested the hypothesis in such a way as to conclude that Peter's hypothesis was not the experimenter's rule. Falsifying evidence is available when a hypothesis is low-quality, and more than half of the participants came across at least one instance of falsification. They can correctly announce that Peter's hypothesis is not the experimenter's rule.

***Hypothesis quality and knowledge of hypothesis quality and hypothesis testing***

Triples were scored as one of the four test types outlined in Table 1.1 and the results are illustrated in Table 2.2. Overall, more confirming triples were generated for testing high-quality (90%) than low-quality hypotheses (40%), and this difference was reliable, ( $\chi^2 = 86.087 (1), p < .0001$ ). Somewhat more falsifying triples were generated for low-quality (60%) than high-quality hypotheses (10%), although this difference was not reliable, ( $\chi^2 = 10.442 (1), p = .165$ ).

Table 2.2: The percentage of confirming and falsifying triples generated for high and low quality hypothesis when quality type was known or unknown.

	<i>Known</i>	<i>Unknown</i>	<i>Total</i>
	Confirm	Confirm	Confirm
	<b>Falsify</b>	<b>Falsify</b>	<b>Falsify</b>
High-quality	100 <b>0</b>	80 <b>20</b>	90 <b>10</b>
Low-quality	10 <b>90</b>	70 <b>30</b>	40 <b>60</b>
<b>Total</b>	<b>55 45</b>	<b>75 25</b>	<b>65 35</b>

*\*Note: The percentage of falsifying triples is presented in bold.*

The percentage of falsifying triples is the mirror image of the percentage of confirming triples. For high-quality hypotheses confirming triples were generated by participants who knew they were testing a high-quality hypothesis (100%) than those who did not (80%), and this difference was reliable, ( $\chi^2 = 4.308 (1), p = .038$ ). More falsifying triples were generated by participants who did not know they were testing a high-quality hypothesis (20%) than those who did know (0%), and this difference was reliable ( $\chi^2 = 21.895 (1), p < .01$ ), (although this p value may be elevated because zero cases were present in all cells for the known condition). The result suggests that even when a hypothesis corresponds to a true state of affairs, participants cannot be certain that it does and so they will still attempt to falsify the high-quality hypothesis in the unknown condition. In this way the *knowledge* that the hypothesis under test is a good one (by telling participants what the experimenter's rule is) affects confirming and falsifying in addition to the effect of hypothesis quality.



For low-quality hypotheses more confirming triples were generated by participants who did not know they were testing a low-quality hypothesis (70%) than participants who did know (10%), and this difference was reliable ( $\chi^2 = 34.322$  (1),  $p < .0001$ ). Critically, participants who knew they were testing a low-quality hypothesis falsified more often (90%) than those who did not know (30%), and this difference was reliable, ( $\chi^2 = 18.325$  (1),  $p < .0001$ ). This is a novel result and does not corroborate the theory that people cannot make sense of falsification (Poletiek, 1996; 2001). Participants found it possible to intentionally generate falsifying tests for Peter's low-quality hypothesis when they knew it was low-quality.

#### ***Four types of hypothesis test***

When participants tested Peter's low-quality hypothesis 'even numbers ascending in twos' and they knew that the experimenter's rule was 'any ascending numbers', they generated the critical negative falsifying test 90% of the time as Table 2.3 shows. All falsifying triples fell into the negative falsifying category in this condition. Not even one positive falsifying test was generated. Every participant in this condition generated at least one negative falsifying test and announced that Peter should know from the evidence they gathered that his hypothesis is incorrect. Participants found it possible to consistently falsify when the hypothesis was low-quality (Poletiek, 1996), even though the relationship between the hypothesis and the experimenter's rule required the most difficult type of falsifying test (Klayman & Ha, 1987).

Table 2.3: Percentages of confirming and falsifying positive and negative test types generated in Experiment 1.

		<i>Low-quality</i>		<i>High-quality</i>	
		Known	Unknown	Known	Unknown
<i>Confirming</i>	Positive	6	61	86	72
	Negative	4	9	14	8
<i>Falsifying</i>	Positive	0	8	0	14
	Negative	90	22	0	6

Negative falsifying triples were generated more often by participants who knew they were testing a low-quality hypothesis (90%) than in any other condition, that is, when they did not know they were testing a low-quality hypothesis (22%), when they tested a high-quality hypothesis and did not know it was high-quality (6%), or when they tested a high-quality hypothesis and did know it was high-quality (0%,  $\chi^2 = 46.938$  (21),  $p = .0005$ ). Overall more negative falsifying tests were generated for low-quality hypotheses (56%) than for high-quality hypotheses (3%,  $\chi^2 = 24.737$  (7),  $p = .0005$ ). Overall the generation of negative falsifying tests did not differ between conditions where the quality of the hypothesis was known (45%) and when it was unknown (14%,  $\chi^2 = 8.18$  (7),  $p = .159$ ).

Even participants who tested Peter's low-quality hypothesis and did not know it was low-quality generated more negative falsifying tests (22%) than is usual in the 2-4-6 task. For example, 6% of tests were negative falsifying tests in Poletiek's falsifying condition (1996). Simply testing someone else's hypothesis may help people to falsify, and I return to this point later in the context of successful science in the next chapter.

Positive confirming triples were generated less often by participants who tested a low-quality hypothesis and did know it was low-quality (6%), compared to when participants knew they were testing a high-quality hypothesis (86%), or when they did not know they were testing a high-quality hypothesis (72%), or when they tested a low-quality hypothesis and did not know it was low-quality (61%,  $\chi^2 = 63.161$  (33),  $p = .0005$ ). Overall reliably more positive confirming tests were reliably generated for high-quality hypotheses (79%) than for low-quality hypotheses (34%,  $\chi^2 = 32.732$  (11),  $p = .0005$ ). Overall the generation of positive confirming tests did not differ between conditions where the quality of the hypothesis was known (46%) than when it was unknown (67%,  $\chi^2 = 11.34$  (11),  $p = .208$ ). There were too few negative confirming and positive falsifying tests in the data set to justify a statistical analysis (Hollander & Wolfe, 1999).

### ***Summary***

The results show that falsification is consistently possible in a situation where people must falsify. They can falsify when the hypothesis under test is a low-quality hypothesis and it belongs to someone else, in this case the imaginary participant Peter. This result does not corroborate the view that people cannot make sense of falsification (Poletiek, 1996; 2001). Participants falsified in the most difficult situation for a hypothesis tester in the 2-4-6 task; they falsified using a negative test when the untrue hypothesis is embedded within the truth. People can generate a negative falsifying test in an intentional and consistent manner in a situation that demands it, that is, when the hypothesis under test is embedded within the true rule. Overwhelmingly they generated negative falsifying tests, that is, a triple that was not consistent with the hypothesis under test ('even numbers ascending in twos'), but which was expected to be consistent with the experimenter's rule ('any ascending numbers'). For example, one participant generated the negative falsifying test 3-5-7 that is not consistent with the 'even numbers...' part of Peter's hypothesis, but they expected it to be consistent with the experimenter's rule 'any ascending numbers'. They received a 'yes' from the researcher and announced that Peter will know his rule is not the experimenter's rule, because he will know that the odd numbers are consistent with the experimenter's rule. This result does not corroborate the view that people do not know when a negative test is wise and when it is not (Klayman & Ha, 1987). The presentation of an alternative hypothesis may help

participants to infer what type of relationship exists between the hypothesis and the truth and they can understand when a negative test can falsify; they can infer that a positive test strategy is unwise in this situation (Klayman & Ha, 1987).

The results also show that the quality of the hypothesis affects hypothesis testing. Participants falsified Peter's hypothesis more often when it was untrue than when it was true regardless of whether or not they knew what the quality of the hypothesis was. When a hypothesis is low-quality there is more falsifying evidence available to people than when a hypothesis is high-quality. Yet the results showed that hypothesis quality alone cannot explain falsification in this experiment because participants falsified the low-quality hypothesis more when they knew that it was low-quality than when they did not know. One important implication that can be drawn from this finding is that the properties of the hypothesis which are generated, such as the quality of the hypothesis, cannot by themselves explain the hypothesis testing strategies people adopt. Other factors, such as the quality of the alternative hypothesis considered alongside the initial hypothesis may help explain the hypothesis testing strategies people adopt, and I return to the factor in the next experiment.

In the next experiment the effect that different types of alternative hypotheses have on hypothesis testing is examined with a focus on the quality of the alternative hypothesis. How the consideration of alternative hypotheses which are either higher or lower quality than Peter's hypothesis can lead to hypothesis falsification using negative falsifying tests will become apparent.

## **Experiment 2: Quality of Alternative Hypotheses**

The aim of this experiment was to determine what properties of an alternative hypothesis facilitate high levels of negative falsifying tests. In Experiment 1 in the known low-quality condition participants were presented with Peter's hypothesis 'even numbers ascending in twos' and were presented with the alternative hypothesis 'any ascending numbers' and they knew that this alternative hypothesis was the experimenter's rule. Participants were instructed to test Peter's hypothesis in such a way as to show him if his rule was the experimenter's rule.

The hypothesis 'even numbers ascending in twos' is typically generated in the 2-4-6 task, and people usually do not falsify it successfully. What was it that helped participants to falsify the 'even numbers ascending in twos' hypothesis in the known low-quality condition in Experiment 1?

Participants were told what the experimenter's rule was '...you know that the experimenter's rule is any ascending numbers'. This sentence introduces several factors which may explain the resulting high levels of hypothesis falsification. First, this sentence provides participants with an alternative hypothesis to consider alongside the initial hypothesis. This alternative hypothesis *is the correct rule* ('any ascending numbers') and is higher quality than the hypothesis under test ('even numbers ascending in twos'). Second, the sentence provides participants with the *knowledge that Peter's hypothesis is untrue*, because participants are told that the alternative ('any ascending numbers') is in fact the experimenter's rule. Third, the fact that this alternative is simply *higher quality* than the hypothesis under test may help participants to falsify; the alternative may not necessarily have to be the correct rule to promote falsification. This experiment examines the role each one of these factors may play in hypothesis falsification. I ask whether the knowledge of hypothesis quality affects hypothesis falsification, and whether the alternative hypothesis needs to be very high-quality (the actual experimenter's rule) in order for participants to falsify Peter's hypothesis.

First, did people falsify because they *knew* that the alternative 'any ascending numbers' was the experimenter's rule? If they had not known this alternative was the experimenter's rule would they have falsified just as much? In Experiment 2 two conditions similar to those employed in Experiment 1 are included to examine these properties further, that is, the known low-quality condition from Experiment 1 in which participants know that the alternative hypothesis 'any ascending numbers' is the experimenter's rule are compared to a condition in which participants do not know that the alternative 'any ascending numbers' is the experimenter's rule. In each condition participants test Peter's low-quality 'even numbers ascending in twos' hypothesis. The aim is to replicate the finding from Experiment 1. If the knowledge that the alternative hypothesis ('any ascending numbers') is the experimenter's rule helps participants to falsify better, then we can conclude that the knowledge of hypothesis quality plays a role in hypothesis falsification.

Second, did people falsify because the alternative 'any ascending numbers' was in fact the experimenter's rule? If they had considered an alternative that was higher quality than Peter's hypothesis 'even numbers ascending in twos', and yet was not the experimenter's rule, such as the alternative 'numbers ascending in twos', would they have falsified as much? And if participants had

considered an alternative hypothesis that was lower quality than Peter's hypothesis, such as 'even numbers ascending in twos ending in the digits 2, 4, 6' would they have been able to falsify Peter's hypothesis? Perhaps the consideration of an alternative hypothesis that is lower quality than the hypothesis being tested would constrain falsification and rule discovery (Klayman & Ha, 1989).

In Experiment 2 participants are presented with alternative hypotheses of different quality in two further conditions. I define the hypothesis quality by how close it is to the correctness of the experimenter's rule 'any ascending numbers'. Participants were given an alternative hypothesis alongside the initial hypothesis that belongs to Peter. Peter's hypothesis was the low-quality hypothesis 'even numbers ascending in twos'. One condition as we have seen received an alternative of high-quality (You know another participant called James hypothesised that the experimenter's rule is 'any ascending numbers'), which I attributed to another participant so that I could manipulate the quality of the hypothesis. Participants in the third condition received an alternative of medium quality (You know another participant called James hypothesised that the experimenter's rule is 'numbers ascending in twos'), and in the fourth condition participants received an alternative of low-quality (You know that another participant called James hypothesised that the experimenter's rule is 'even numbers ascending in twos that end in the digits 2,4,6'). The known low-quality condition from Experiment 1 is included as a control condition with which to compare hypothesis falsification resulting from the consideration of a high-quality, medium quality, and low-quality alternative hypothesis. (The measure of hypothesis quality used in this experiment was adapted from Klayman & Ha, 1989). The alternatives have the same embedded relationship between the hypothesis under test and the alternative hypothesis in each case, and approximately equal interval decreases in hypothesis quality (Klayman & Ha, 1989).

The prediction is that the higher the quality of the alternative hypothesis the more negative falsifying tests participants would generate. The idea is that the higher quality alternative presents the hypothesis tester with an explicit set of possibilities from which to generate negative falsifying tests. For example, when participants test Peter's hypothesis 'even numbers ascending in twos' and they consider the alternative 'numbers ascending in twos', they can generate the triple 3-5-7 from the possibilities that this alternative may make explicit to them

but which are not explicit in Peter's hypothesis. In addition the consideration of an alternative may help people infer what the relationship between Peter's hypothesis and the true rule is. Consider that the higher quality alternative hypotheses used in this experiment are not only higher quality but they are *more general* than the hypothesis under test. Peter's hypothesis is embedded within the alternative, and so they may generate negative instances more often (Klayman & Ha, 1987). The prediction is that the lower the quality of the alternative hypothesis the fewer negative falsifying tests participants would generate. The idea is that the lower quality alternative constrains the explicit set of possibilities from which to generate negative falsifying tests. For example, when the lower quality alternative is 'even numbers ascending in twos ending in the digits 2,4,6' participants may generate the triple 22-24-26. When they receive a 'yes' response from the experimenter they may conclude that the alternative is the rule, or they may then generate a triple from the possibilities made explicit in Peter's hypothesis such as 14-16-18 and when they receive a 'yes' they may then conclude that 'even numbers ascending in twos' is the rule. It may be difficult for participants to discover the true rule because they may persist in generating triples consistent with both of these two low-quality hypotheses. In sum I predicted that alternative hypotheses, particularly higher quality alternatives, facilitate the generation of negative falsifying tests and hence rule discovery.

## **Method**

### ***Materials and design***

Participants were randomly assigned to four conditions (three experimental conditions and one control condition,  $n = 16$  in each). In each condition they were given a low-quality hypothesis belonging to the imaginary participant Peter: 'even numbers ascending in twos'. Participants were then given another piece of information in the form of an alternative hypothesis. In the three experimental conditions participants were given one of three alternative hypotheses (high, medium and low-quality) to consider alongside the initial hypothesis that belonged to Peter. In the control condition they were given the alternative: 'in fact you know the experimenter's rule is 'any ascending numbers' (this is the replication of the known low-quality condition in Experiment 1). In the first experimental condition (high-quality alternative) participants were given a high-quality alternative hypothesis that was the

experimenter's rule, but they did not know it: 'you know that another participant called James hypothesised that the experimenter's rule was any ascending numbers'. In the second experimental condition (medium quality alternative) participants were given the medium quality alternative that was not the experimenter's rule but which was higher quality than Peter's hypothesis: 'you know that another participant called James hypothesized that the experimenter's rule was numbers ascending in twos'. In the third experimental condition (low-quality alternative) they were given a low-quality alternative that was lower quality than Peter's hypothesis: 'you know that another participant called James hypothesized that the experimenter's rule was even numbers ascending in twos that end in the digits 2,4,6'. The instructions for the second experimental condition are given below to illustrate (see Appendix B for the instructions for the control condition and the other two experimental conditions):

“In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. You know that another participant called James hypothesised that the experimenter's rule was 'numbers ascending in twos'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the experimenter's rule, and not before, you are to write down 'Peter now knows his rule is the experimenter's rule' or 'Peter now knows his rule is not the experimenter's rule'. You are to write this under your most recent number sequence and raise your hand. The experimenter will then write whether or not you are correct beside your announcement.”



### ***Participants and procedure***

Forty eight participants completed the task (one was excluded because she said she was familiar with the task). Most participants were undergraduate students and some were individuals from the general population. The age of the participants ranged from 16 to 49 years. The mean age was 22 years, and there were 33 women and 14 men who took part. No participants had taken courses in the philosophy of science.

### **Results and discussion**

I report the results of the following dependent measures: the number of triples generated by participants; the number of correct announcements achieved by the participants; confirming and falsifying triples; and positive and negative confirming and falsifying triples.

#### ***Number of triples***

A total of 245 triples was generated with a mean of 3.83 triples per participant. A mean of 3.38 triples was generated in the control condition when participants knew the alternative ‘any ascending numbers’ was the experimenter’s rule. A mean of 4.06 triples was generated in the high-quality alternative condition when participants considered the alternative ‘any ascending numbers’. A mean of 4.31 triples was generated in the medium quality alternative condition when participants considered the alternative ‘numbers ascending in twos’. A mean of 3.56 triples was generated in the low-quality alternative condition when participants considered the alternative ‘even numbers ascending in twos’. (See Table 2.4) Somewhat fewer triples were generated by participants in the control condition who knew that ‘any ascending numbers’ was the experimenter’s rule ( $M = 3.38$ ) than participants in the high-quality alternative condition who did not know it was the experimenter’s rule ( $M = 4.06$ ), but this difference was not reliable ( $\text{Mann-Whitney}_{16,16} = 96.5$ ,  $Z = -1.204$ ,  $p = .118$ ). Somewhat fewer triples were generated in the control condition ( $M = 3.38$ ) than in the medium quality alternative condition ( $M = 4.31$ ), but this difference was not significant ( $\text{Mann-Whitney}_{16,16} U = 87.5$ ,  $Z = -1.563$ ,  $p = .064$ , two-tailed). There was little difference in the mean number of triples generated in the control condition ( $M = 3.38$ ) and in the low-quality alternative condition ( $M = 3.56$ ), and the difference was not reliable ( $\text{Mann-Whitney}_{16,16} U = 114.00$ ,  $Z = -.536$ ,  $p = .308$ ). There was no difference for the number of triples generated in the high-quality alternative

condition ( $M = 4.06$ ) and in the medium quality alternative condition ( $M = 4.31$ , Mann-Whitney<sub>16,16</sub>  $U = 111.00$ ,  $Z = -.658$ ,  $p = .268$ ). There was a marginal difference in the number of triples generated for the medium quality alternative condition ( $M = 4.31$ ) and in the low-quality alternative condition ( $M = 3.56$ , Mann-Whitney<sub>16,16</sub>  $U = 88.5$ ,  $Z = -1.522$ ,  $p = .069$ ). These results imply that the quality of the alternative hypothesis may sometimes affect the number of tests participants generated when testing a low-quality hypothesis. There was a small indication that participants could have a tendency to test fewer triples in the control condition because they are sure that Peter's hypothesis is untrue and that their test falsifies his hypothesis; they know what the experimenter's rule is as Table 2.4 shows. This result replicates the same finding reported in Experiment 1. And there was a small indication that participants could have a tendency to test fewer triples in the low-quality alternative condition than in the other experimental conditions, perhaps because the consideration of an alternative that is even lower quality than their hypothesis may constrain their ability to generate other possible test triples or alternatives.

Table 2.4: The number of triples generated in for each type of alternative hypothesis quality in Experiment 2.

<i>Condition</i>	Control	High-quality	Medium quality	Low-quality
<i>Number of Triples</i>	3.38	4.06	4.31	3.56

Participants test more when the alternative hypothesis is higher quality than the hypothesis under test and they do not know that the alternative is higher quality.

***Correct announcements and rule discovery***

The prediction was made that as the quality of the alternative hypothesis decreased that the number of correct announcements would decrease, that is, the number of participants who would announce that Peter's low-quality

hypothesis ‘even numbers ascending in twos’ was incorrect would decrease. (The control condition is not relevant to this section because participants know what the experimenter’s rule is. I compare the three experimental conditions only). The alternative hypothesis may present the participant with an explicit set of possibilities from which to generate falsifying tests. If participants are using the alternative hypothesis to generate test triples such as 5-11-22 when they consider the alternative hypothesis ‘any ascending numbers’ they cannot falsify Peter’s hypothesis ‘even numbers ascending in twos’ when they receive a ‘yes’ from the experimenter, but they may conclude that the alternative hypothesis is the experimenter’s rule.

Participants in the high-quality alternative condition announced that Peter’s hypothesis was not the rule almost as often (81%) as participants in the medium quality condition (94%), but less often when they were presented with the low-quality alternative hypothesis (69%). This difference was not significant, ( $\chi^2 = 3.282 (2), p = .097$ , two tailed), as Table 2.5 shows:

Table 2.5: The percentages of participants who correctly announced that Peter’s hypothesis was incorrect, and the percentages who subsequently discovered the experimenter’s rule.

	<i>High-quality</i>	<i>Medium quality</i>	<i>Low-quality</i>
<b>Correct announcement</b>	81	94	69
<b>Rule discovered</b>	50	44	12

Participants discovered what the experimenter’s rule was more often in the high-quality alternative condition (50%) and in the medium quality alternative condition (44%), than in the low-quality alternative condition (12%,  $\chi^2 = 5.647 (2), p = .03$ ). The result implies that even when one of the hypotheses under consideration is correct participants may not always discover that it is correct (50%). Moreover, even when participants consider a medium quality alternative hypothesis they can sometimes discover the rule (44%). Participants

who considered a lower quality alternative hypothesis rarely discovered the rule (12%). The implication is that it is not enough to consider two alternative hypotheses to discover the rule; discovery may depend on considering at least one good quality hypothesis (Tweney *et al.*, 1980).

***Alternative hypothesis quality and hypothesis testing***

I predicted that participants would falsify more when the alternative hypothesis was higher quality, and that as the alternative decreased in quality participants would confirm more. Although participants confirmed somewhat more in the low-quality condition (67%), compared to the medium quality condition (52%) and in the high-quality condition (49%), the differences were not reliable, ( $\chi^2 = 13.017$  (12),  $p = .184$ ). As predicted participants falsified more in the high-quality condition (51%), and in the medium quality condition (48%), than in the low-quality condition (33%,  $\chi^2 = 20.323$  (10),  $p = .013$ ) as Table 2.6 shows:

Table 2.6: The percentage of confirming and falsifying triples when the alternative hypothesis was high-quality, medium quality and low-quality.

	<i>High-quality</i>	<i>Medium quality</i>	<i>Low-quality</i>
<b>Confirming</b>	49	52	67
<b>Falsifying</b>	51	48	33

The results imply that as the quality of the alternative hypothesis decreases the amount of falsification decreases. High-quality alternative hypotheses facilitate falsification of low-quality hypotheses.

***Four types of hypothesis tests***

Participants falsified reliably more often with negative falsifying tests in the high-quality condition (42%), and in the medium quality condition (48%), than in the low-quality condition (23%,  $\chi^2 = 22.167$  (10),  $p = .007$ ), as Table 2.7 shows. Participants confirmed somewhat more often with positive confirming tests in the low-quality condition (44%), than in the medium quality condition

(20%), or in the high-quality condition (23%), but this difference was not reliable, ( $\chi^2 = 7.725$  (10),  $p = .328$ ) as Table 2.7 shows:

Table 2.7: The percentages of positive and negative confirming and falsifying triples generated when the alternative hypothesis was high-quality, medium quality and low-quality.

		<i>High-quality</i>	<i>Medium-quality</i>	<i>Low-quality</i>
<i>Confirming</i>	Positive	23	20	44
	Negative	26	32	23
<i>Falsifying</i>	Positive	9	0	10
	Negative	42	48	23

A similar amount of negative confirming was observed in the high-quality condition (26%), as in the medium quality condition (32%), and in the low-quality condition (23%,  $\chi^2 = 6.686$  (10),  $p = .378$ ). There were too few positive falsifying tests in the data set to justify a statistical analysis (See Siegel and Castellen, 1994). The results imply that the quality of the alternative hypothesis does not have a strong effect on the amount of negative falsifying triples. Regardless of the quality of the alternative hypothesis, negative falsifying tests were generated.

***Knowledge of alternative hypothesis quality***

Participants generated falsifying tests when they knew the alternative hypothesis was the experimenter’s rule (61%) and when they did not know (51%), and this difference was not reliable, ( $\chi^2 = 7.244$  (6),  $p = .15$ ). The result that the majority (61%) of the tests were falsifying when participants knew the alternative was the experimenter’s rule replicates our finding in Experiment 1, although the effect in this experiment was not as large. In this experiment the

control condition was compared with the high-quality condition to examine the effect of knowing the alternative is the experimenter's rule, 'you know that the experimenter's rule is any ascending numbers' versus not knowing the alternative is the experimenter's rule, 'another participant called James hypothesized that the experimenter's rule is any ascending numbers'. In Experiment 1 it was found that 90% of the triples generated by participants were negative falsifying tests when they knew the alternative 'any ascending numbers' was the experimenter's rule and they tested Peter's hypothesis 'even numbers ascending in twos' in such a way as to show him his hypothesis was not the experimenter's rule. The alternative hypothesis they were presented with was not only the experimenter's rule, that is, the correct rule, but participants also *knew* it was the experimenter's rule. The result suggests that falsifying a low-quality hypothesis depends somewhat on the knowledge that the alternative hypothesis is the experimenter's rule, but also on the consideration of an alternative that is very high-quality.

Participants confirmed somewhat less often when they knew the alternative hypothesis was the experimenter's rule (39%) than when they did not know (49%), but this was not reliable, ( $\chi^2 = 3.352$  (5),  $p = .323$ ). Negative falsifying tests were generated somewhat more often when participants knew the alternative was the experimenter's rule (61%) than when they did not (42%), but this difference was not reliable, ( $\chi^2 = 6.819$  (6),  $p = .169$ ). Positive confirming tests were generated as often when participants knew the alternative was the experimenter's rule (33%) than when they did not know (23%,  $\chi^2 = .400$ , (4),  $p = .491$ ). More negative confirming tests were generated when participants did not know the alternative was the experimenter's rule (26%) than when they did know (6%), and this difference was marginally reliable, ( $\chi^2 = 7.133$  (4),  $p = .065$ ). The knowledge that the alternative hypothesis is the experimenter's rule has a small effect on hypothesis testing, but the consideration of a higher quality alternative hypothesis may be the clearest predictor that people will falsify a low-quality hypothesis.

### **Summary**

The results of the experiment show that more falsifying test triples were generated to test Peter's low-quality hypothesis 'even numbers ascending in twos', when an alternative hypothesis that was higher quality than Peter's hypothesis was considered. Negative falsifying test triples were generated most

often, and the consideration of a higher quality alternative hypothesis facilitated this effect. The results do not corroborate the view that falsification is impossible; people can falsify when they consider a higher quality alternative hypothesis (Poletiek, 1996). However, the higher quality alternative may have given participants information about what the relationship between Peter's hypothesis and the truth might be. That is, the higher quality hypotheses were not only higher quality but they embedded Peter's hypothesis (Klayman & Ha, 1987). (We will turn to examine if the alternative needs to embed Peter's hypothesis in order for participants to falsify in the next experiment, Experiment 3). There were marginally more correct announcements that Peter's hypothesis was incorrect when the alternative hypothesis was high or medium quality than low-quality, but the rate of rule discovery did not differ between participants' consideration of a high or medium quality alternative. The consideration of an alternative that is higher quality than the hypothesis being tested may be as useful as the consideration of the truth itself in order to generate falsifying triples. The consideration of an alternative hypothesis that is lower quality than the hypothesis under test may help participants to falsify, but this falsification may not help them to discover the truth.

### **Experiment 3: Explicit and non-explicit alternative hypotheses**

The aim of this experiment was to investigate whether an alternative hypothesis facilitates falsification by presenting participants with an explicit set of possibilities from which to generate the falsifying tests. For example, when participants are asked to test Peter's hypothesis 'even numbers ascending in twos', and they are presented with an alternative hypothesis to consider such as 'numbers ascending in twos', they may generate a test triple such as 3-5-7. This triple is one of many possible triples which are included in a set of triples which ascend in twos (Klayman & Ha, 1987). When a participant considers the alternative they may generate the 3-5-7 test triple because it is consistent with one of the hypotheses they mentally represent; 'numbers ascending in twos' (e.g., Johnson-Laird, 1983; Wason & Johnson-Laird, 1972). Even though the hypothesis they are testing is 'even numbers ascending in twos', they may identify that the two alternatives differ because Peter's hypothesis contains even numbers and the alternative does not (see Gale & Ball, 2005). By generating a triple from the set of possibilities made available by considering an alternative, participants are able to generate a test such as 3-5-7 that is inconsistent with

Peter's hypothesis 'even numbers ascending in twos'. Perhaps participants can generate negative tests of hypotheses by considering the possibilities made explicit to them by an alternative hypothesis. Key theories of reasoning make similar predictions about deductive inferences. For example, the search for explicit alternative scenarios that could show a premise to be false helps people to check the validity of their deductions (e.g., Johnson-Laird & Byrne, 1991; 2002).

In this experiment we will test whether an explicit alternative hypothesis is required in order for people to consistently falsify a hypothesis. The previous experiments found that falsification was facilitated by the consideration of an alternative hypothesis, but each alternative hypothesis presented to participants referred to specific numerical properties such as 'ascending in two' and 'any ascending numbers'; the alternative stated explicit numerical properties. Yet when people consider two alternative hypotheses labelled DAX and MED the rate of rule discovery dramatically improves, even though these labels do not present explicit alternative hypotheses to participants (e.g., Tweney *et al.*, 1980; Wharton, Cheng, & Wickens, 1993).

This time we examine whether an explicit alternative hypothesis is essential to falsification and subsequent rule discovery. Conditions in which participants consider either an explicit alternative hypothesis or a non-explicit alternative hypothesis are compared. Participants test the imaginary participant Peter's low-quality hypothesis 'even numbers ascending in twos'. A non-explicit hypothesis of the form 'James hypothesized that the experimenter's rule was *something else*' was introduced. This condition was compared to one in which the explicit hypothesis was 'James hypothesized that the experimenter's rule was *any ascending numbers*'. (I chose this explicit hypothesis from Experiment 2 because it leads to approximately 50% rule discovery rates, implying that rule discovery in this condition is neither too easy nor too difficult. Differences that could be explained by task difficulty rather than the explicitness of the alternative hypothesis were not the object of the investigation (see Kerlinger, 2000). Both of these conditions were compared to a condition where no alternative hypothesis was given.

### ***Hypothesis explicitness and the implications for theories of hypothesis testing***

When participants test Peter's hypothesis 'even numbers ascending in twos' and they consider a non-explicit alternative, the alternative does not give them any



information about the mathematical relationship between the hypothesis under test and the rule (Klayman & Ha, 1987). If participants still find it possible to falsify using negative falsifying tests when they consider a non-explicit alternative, then the theoretical prediction that people can falsify in the 2-4-6 task embedded relationship only when they can infer what that relationship is, cannot be a complete explanation (Klayman & Ha, 1987; 1989).

Furthermore the theoretical view that people find it difficult to intentionally falsify a low-quality hypothesis because there is no new information available to them, cannot offer a complete explanation either (Poletiek, 1996; 2001). A non-explicit hypothesis does not give participants new information, but it may encourage them to search for new evidence by either generating their own negative tests or alternative hypothesis. Any theoretical account that proposes people search for properties common to two explicit alternative hypotheses in order to generate a falsifying test, also tells us little about how participants can generate negative falsifying tests when the alternative is non-explicit (Oaksford & Chater, 1994; Gale & Ball, 2005).

## **Method**

### ***Materials and design***

Participants were randomly assigned to three conditions ( $n = 16$  in each). In each condition they were given a low-quality hypothesis belonging to the imaginary participant Peter: 'even numbers ascending in twos'. In the first condition they were given an explicit alternative hypothesis: 'Another participant called James hypothesised that the experimenter's rule was any ascending numbers'. In the second condition they were given a non-explicit alternative: 'Another participant called James hypothesised that the experimenter's rule was something else'. In the third condition they were given no alternative at all. The instructions for the non-explicit condition are given below:

"In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter's rule was: 'even numbers ascending in twos'. You know that another participant called James hypothesised that the experimenter's rule was 'something else'.

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the experimenter's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the experimenter's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the experimenter's rule, *and not before*, you are to write down 'Peter now knows his rule is the experimenter's rule' or 'Peter now knows his rule is not the experimenter's rule'. You are to write this under your most recent number sequence and raise your hand. The experimenter will then write whether or not you are correct beside your announcement."

### ***Participants and procedure***

Forty eight participants completed the task. They were undergraduate students who gained course credit for their participation. Their age ranged from 17 to 49 years and the mean age was 21 years. There were 33 women and 15 men who took part. No participants had taken courses in the philosophy of science. The recording sheet and procedure were the same as in Experiment 1.

### **Results and discussion**

The results present the following dependent measures: the number of triples generated by participants; the number of correct announcements achieved by the participants; confirming and falsifying triples; and positive and negative confirming and falsifying triples.

#### ***Number of triples***

A total of 222 triples were generated. A mean number of 4.63 triples were generated per participant. There was no difference in the number of triples generated when the alternative was explicit ( $M = 4.75$ ) and non-explicit ( $M = 4.81$ , Mann-Whitney<sub>16,16</sub>  $U = 122.5$ ,  $Z = -.210$ ,  $p = .417$ ). There was no difference in the number of triples generated when the alternative was non-explicit ( $M = 4.81$ ) and when there was no alternative ( $M = 4.31$ , Mann-

Whitney<sub>16,16</sub>  $U = 109$ ,  $Z = -.7224$ ,  $p = .469$ , two-tailed). And there was no difference in the number of triples generated when the alternative was explicit ( $M = 4.75$ ) and when there was no alternative ( $M = 4.31$ , Mann-Whitney<sub>16,16</sub>  $U = 106$ ,  $Z = -.840$ ,  $p = .401$ , two-tailed). The result implies that neither the consideration of nor the explicitness of an alternative hypothesis, affects how much people test their hypothesis.

### ***Correct announcements and rule discovery***

Participants announced correctly that Peter's low-quality hypothesis 'even numbers ascending in twos' was not the experimenter's rule somewhat less often when they were presented with the explicit alternative (69%), than the non-explicit alternative (81%), or no alternative (81%), but this difference was not reliable ( $\chi^2 = 0.943$  (2),  $p = 0.312$ ) as Table 2.8 shows. The rate of correctly announcing that Peter's hypothesis is not the experimenter's rule appears to be elevated in this experiment compared to the previous experiment. Nonetheless the first condition (50% discovered the rule) replicates the result of the same condition in Experiment 2 (50% also discovered the rule), suggesting there were no new extraneous variables.

Participants in this experiment were asked what they thought the experimenter's rule was once they announced Peter's low-quality hypothesis 'even numbers in twos' was incorrect. The rate of rule discovery was highest when participants considered the explicit alternative 'any ascending numbers' (50%), than the non-explicit alternative 'something else' (31%), or when there was no alternative (19%), and this difference was reliable ( $\chi^2 = 5.101$  (2),  $p = .039$ ). (See Table 2.8).

Table 2.8: The percentages of participants who correctly announced that Peter’s hypothesis was incorrect and the percentages who subsequently discovered the experimenter’s rule.

	<i>Explicit</i>	<i>Non-explicit</i>	<i>No alternative</i>
<b>Correct announcement</b>	69	81	81
<b>Rule discovered</b>	50	31	19

The results suggest that the discovery of the rule appears to depend on the consideration of an explicit high-quality alternative hypothesis. Falsification and the consideration of an alternative that is both explicit and high-quality may go hand in hand to facilitate rational hypothesis testing (Wason & Johnson-Laird, 1972; Kuhn, 1996).

***Confirming and falsifying***

More confirming triples were generated when the alternative was explicit (57%), than when it was the non-explicit (47%), or when there was no alternative (46%), but this difference was not significant ( $\chi^2 = 28.374 (16), p = .058$ ) as Table 2.9 shows:

Table 2.9: The percentages of confirming and falsifying triples when the alternative hypothesis was explicit, non-explicit and when there was not alternative.

	<i>Explicit</i>	<i>Non-explicit</i>	<i>No alternative</i>
<b>Confirming</b>	57	47	46
<b>Falsifying</b>	43	53	54

There was no difference in the amount of falsifying triples generated when the alternative was explicit (43%), than non-explicit (53%), and when there was no alternative (54%,  $\chi^2 = 10.044$  (16),  $p = .216$ ). It is not clear from this result if the consideration of explicit and non-explicit alternatives help people to falsify. Falsifying was found in each condition even when there was no alternative, however, it is possible that simply considering someone else's hypothesis helps a participant to falsify (I will return to this point in the next chapter).

**Four types of hypothesis tests**

More positive confirming tests were generated when the alternative was explicit (37%), than when it was non-explicit (22%), or when there was no alternative at all (27%), but this was not reliable, ( $\chi^2 = 11.379$  (12),  $p = .249$ ). There was no difference in the amount of negative confirming tests generated when the alternative was explicit (20%), than when it was non-explicit (25%), than when there was no alternative (19%,  $\chi^2 = 9.128$  (8),  $p = .166$ ), as Table 2.10 shows).

Table 2.10: The percentages of positive and negative confirming and falsifying triples generated when the alternative hypothesis was explicit, non-explicit, and when there was no alternative.

	<i>Explicit</i>	<i>Non-explicit</i>	<i>No alternative</i>
<i>Confirming</i>			
Positive	37	22	27
Negative	20	25	19
<i>Falsifying</i>			
Positive	8	8	12
Negative	35	45	42

There was no difference in the amount of positive falsifying tests generated when the alternative was explicit (8%), than when it was non-explicit (8%), or when there was no alternative (12%, the number of cases was not large enough to carry out a reliable chi-square test). There was no difference in the amount of negative falsifying tests generated when the alternative was explicit (35%), than when the alternative was not explicit (45%), or when there was no alternative (42%,  $\chi^2 = 12.875$  (14),  $p = .268$ ). The results imply that the consideration of an alternative need not necessarily be explicit in order to falsify using a negative falsifying test.

### ***Summary***

The explicitness of the alternative hypothesis did not affect the number of test triples generated; even when no alternative hypothesis was present there were not fewer triples than when an alternative was present (e.g., Vallee-Tourangeau *et al.*, 1995). The explicitness of the alternative did not affect the generation of falsifying tests, or negative falsifying triples (Oaksford & Chater, 1994). Nor did the presence of an alternative hypothesis affect the generation of these falsifying tests as just as many of them were generated when there was no alternative hypothesis. These results do not corroborate the view that participants can only falsify when an alternative hypothesis is made explicit to them. Participants can falsify when the alternative is non-explicit and does not give them any information about the mathematical relationship (Klayman & Ha, 1987). In other words, participants may be able to make alternative possible hypotheses and falsifying triples explicit for themselves. Although the generation of negative falsifying tests did not depend on the consideration of an explicit alternative, it is possible that participants subsequently fleshed out the non-explicit alternative to generate their own explicit alternative (e.g., Byrne, 2005). The falsification of the hypothesis ‘even numbers ascending in twos’ when there was no alternative can be considered elevated when compared to previous research, and it does not corroborate the view that people find it impossible to falsify (Poletiek, 1996). Possibly testing someone else’s hypothesis facilitates falsification and I address this possibility in the next chapter. The results imply that falsification *and* the consideration of alternative hypotheses that are higher quality than the hypothesis under test, may go hand in hand in discovering the truth in hypothesis testing (Wason & Johnson-Laird, 1972). The falsifying test is only any good if it leads to the endorsement of an explicit alternative that is

higher quality than the quality of the hypothesis under test. For example, in scientific reasoning a theory is sometimes falsified, but unless there is an explicit alternative theory to explain the falsifying result, the falsification remains an anomaly until such a time as a new theory is generated (see Kuhn, 1993).

## **General Discussion**

The experiments reveal that people find it consistently possible to falsify an incorrect hypothesis that is typical of the standard 2-4-6 task. Experiment 1 reports the novel result that people find it possible to consistently generate a negative falsifying test (Poletiek, 1996). Participants falsified Peter's hypothesis 'even numbers ascending in twos' when they considered an alternative hypothesis that told them what the experimenter's rule was. They overcame their tendency to test a hypothesis with positive tests when it was more accurate to test with negative tests; they expected their negative tests to falsify (Klayman & Ha, 1987). This result is a novel one (e.g., Wason, 1960; Wetherick, 1962; Tweney *et al.*, 1980; Klayman & Ha, 1989; Poletiek, 1996).

The chapter examined this finding by separating out a number of different factors that could have been responsible for the falsification observed in Experiment 1, in experiments 2 and 3. In Experiment 2 it was discovered that participants did not necessarily need to know that the alternative 'any ascending numbers' was the experimenter's rule in order to falsify. Counter to our predictions participants falsified Peter's hypothesis 'even numbers ascending in twos' as often when they considered the alternative 'any ascending numbers', and did not know it was the experimenter's rule. They also falsified as often when the alternative was higher quality than Peter's hypothesis, even though it was not as high in quality as the experimenter's rule. Counter to our predictions participants discovered the rule as often when the alternative was higher quality, regardless of whether it was the experimenter's rule or not. When the alternative was lower quality than Peter's hypothesis it led to falsification, but participants were not able to use this falsification to discover the experimenter's rule. Considering an alternative hypothesis that is higher quality than the hypothesis under test can reliably lead to falsification and rule discovery, but considering an alternative that is lower quality does not tend to lead to rule discovery, even though it led to falsification. The major implication of Experiment 2 is that higher quality alternative hypotheses are important in hypothesis falsification,

and while lower quality alternatives may help people to falsify, considering lower quality alternatives may reliably hamper rule discovery.

In Experiment 3 we found that participants falsified Peter's hypothesis 'even numbers ascending in twos' and they announced that Peter's rule was not the experimenter's rule as often when the alternative was explicit and non-explicit, and when there was no alternative. But participants reliably discovered the rule more often when the alternative was explicit than non-explicit, than when there was no alternative at all. This finding suggests that the consideration of an explicit hypothesis and falsification may go hand and hand in discovering the truth in hypothesis testing. While falsification is sufficient to announce that a hypothesis is untrue, an explicit alternative hypothesis that explains the falsifying result is necessary for truth discovery.

The results do not corroborate the mathematical relationship theory that asserts people have a tendency to engage in a positive test strategy in the hypothesis testing situations they encounter. In our experiments participants knew when it was accurate to test a hypothesis with a negative test; when they considered an alternative hypothesis they could often reliably generate negative tests and they reliably expected them to falsify (Klayman & Ha, 1987). The prediction that participants need to know what the mathematical relationship between the hypothesis and the truth is in order to generate negative tests was not supported by our results. Participants generated negative tests and expected these tests to falsify when they considered a non-explicit hypothesis telling them nothing about what the relationship between the hypothesis and the rule was (Klayman & Ha, 1989).

The results do not corroborate the prediction that people find falsification impossible; participants not only generated negative tests but they expected these negative tests to falsify. They showed that they understood the implications of their test choice by predicting that Peter would know from their negative falsifying tests that his hypothesis was incorrect (Poletiek, 1996).

The consideration of an explicit alternative hypothesis led not only to falsification but to the discovery of the rule reliably more often than when the alternative was non-explicit or low-quality. Falsification and the consideration of a good quality explicit alternative may go together in the discovery of the truth in hypothesis testing (e.g., Wason & Johnson-Laird, 1972; Kuhn, 1993). The major implication of the results is that a hypothesis testing theory must take



account not only of the role of alternative hypotheses, but how people reason with alternative hypotheses in order to generate falsifying instances.

We have not yet investigated what researchers refer to as objective falsification, that is, falsifying your own hypothesis (e.g., Poletiek, 2004). There is a possibility that it is somehow easier to falsify someone else's hypothesis rather than one's own, and all of our participants falsified Peter's hypothesis in this series of experiments. Given that much of real life scientific hypothesis testing proceeds by investigators attempting to falsify someone else's hypothesis, it may be easier to do so because one has not invested any cognitive effort in generating the hypothesis initially. In the next chapter I turn to an investigation of whether factors such as hypothesis ownership affect hypothesis testing and hypothesis falsification. We have learned that the consideration of alternative hypotheses is important, so let us turn to address whether the competition between hypothesis testers, each with their own alternatives helps people to falsify.

### Chapter 3      The Effect of Competition on Hypothesis Testing

*Alice generally gave herself very good advice...and once she remembered trying to box her own ears for having cheated herself in a game of croquet she was playing against herself.*

– Lewis Carroll (Alice in Wonderland; 1994/1866, p. 9)

Explanations of hypothesis testing in psychology have tended to focus on falsification as the *best way* to test a hypothesis (e.g., Tweney *et al.*, 1980; Gorman & Gorman, 1984; Klayman & Ha, 1987). However, some explanations have tended to focus on falsification as not only the best way to test a hypothesis, but as the way we *should* test our own hypotheses (Popper, 1959; Wason, 1960; Poletiek, 1996; 2001; 2005). In fact, early research considered falsification of one's own hypotheses, as opposed to the falsification of any hypothesis, as falsification in its truest form (e.g., Popper, 1959; Wason, 1960; Tweney *et al.*, 1980).

Chapter 2 showed that people could falsify; participants falsified an untrue hypothesis belonging to an imaginary participant called Peter. This falsification of the imaginary participant's hypothesis was higher in our experiments under all conditions than has previously been found in the literature (e.g., Wason, 1960; Tweney *et al.*, 1980; Klayman & Ha, 1989; Poletiek, 1996). One possible explanation is that people have a tendency to falsify hypotheses belonging to others rather than their own hypotheses. People may be able to use falsification in a rational way to disprove untrue hypotheses, but they may tend to falsify untrue hypotheses belonging to other people more than their own untrue hypotheses.

One reason that people might have a tendency to falsify other people's hypotheses is because their experience of hypothesis testing proceeds by testing competing hypotheses. For example, scientists may often proceed by attempting to confirm their own hypotheses and falsify other scientists' hypotheses (e.g., Mitroff, 1974; Gorman, 1995a; Fugelsang *et al.*, 2004). Falsification of a theoretical program is more likely to come from a scientist working outside the program than from a scientist working within the program (e.g., Kuhn, 1993).

The suggestion that scientists falsify other scientists' theories and the finding that people can falsify an imaginary participant's hypothesis, may provide an additional way of looking at hypothesis testing; competition between hypothesis

testers may affect falsification. People may be rational hypothesis testers when a hypothesis belongs to someone else, as the results of the experiments in Chapter 2 testify. The experiments in this chapter aim to examine whether participants can be rational hypothesis testers by falsifying their own hypotheses under competitive conditions. The aim is to discover what competitive factors help people to falsify their own hypotheses in the 2-4-6 task. Two experiments that test two key competitive factors; hypothesis ownership and the consideration of an opponent hypothesis tester were designed.

The first competitive factor examined is *hypothesis ownership*, which is the subject of Experiment 4. Whether participants falsify someone else's hypothesis, that is, an imaginary participant's hypothesis, more than their own is tested. Participants are simply presented with the imaginary participant 'Peter's hypothesis: even numbers ascending in twos' in one condition, and presented with 'Your hypothesis: even numbers ascending in twos' in another condition. To eliminate the possible confounding effect of personal investment participants are not asked to generate their own hypotheses, they are simply told that a hypothesis belongs to them. They are instructed to test if the hypothesis is the experimenter's rule. The experiment controls for other relevant factors in that the hypotheses are equally untrue and they are not presented with any other conditions to help them to falsify (e.g., an explicit alternative hypothesis such as in the experiments in Chapter 2). I predict that participants will falsify Peter's hypothesis more than their own, given that falsification was consistently higher in the imaginary participant experiments than in hypothesis testing research (e.g., Poletiek, 1996), and there is little evidence to suggest that people falsify their own hypotheses to date (e.g., Poletiek, 2005).

The second competitive factor examined is contending with *an opponent hypothesis tester*. Experiment 5 takes a look at whether participants falsify their own hypothesis more under conditions in which they are told to consider an imaginary opponent hypothesis tester who is also testing their hypothesis, than when they are not told anything about an opponent. What we are discovering is whether or not the awareness of an opponent testing their hypothesis would promote falsification. For example, it has been shown that when people direct their reasoning towards some goal, such as creating a new scientific theory, they may expend more cognitive effort, attend to information more carefully and process it more deeply (e.g., Kunda, 2000; Kruglanski & Webster, 2000). The extra cognitive effort may help them make other alternative theories more

explicit, or help them check that they have not overlooked an important point. What I suggest is that the situation of a hypothesis tester in competition with an opponent hypothesis tester may be akin to theoretical competition in scientific discovery; the scientist must attempt to check whether the opponent can falsify their hypothesis and whether they can falsify the opponent scientist's hypothesis (e.g., Kuhn, 1993).

Before moving on to the experiments and the explanation of how they test theories of hypothesis testing, let us consider the philosophical distinction between falsifying hypotheses belonging to oneself and hypotheses belonging to another, and how this distinction may help us to understand the effect of competition on hypothesis testing.

### **The philosophy of falsification: Falsifying one's own hypotheses**

To understand what is meant by falsification as the best way to test a hypothesis, consider Popper's black swan example, which I adapt here: Imagine we are testing if the hypothesis 'all swans are white' is true. Imagine we find one white swan, and then another white swan, and then another. Now, imagine we find a black swan. The hypothesis 'all swans are white' is falsified and cannot be true because this one black swan shows that not all swans are white. We should search for the black swan, that is, we should look for evidence which has the potential to falsify hypotheses. If we happen upon an example counter to what a hypothesis would predict, then we can say the hypothesis is untrue. In this case we can conclude that the hypothesis 'all swans are white' is untrue. Even if we search for many confirming examples of white swans, we cannot say all swans are white because one black swan may yet be discovered.

But there is an additional reason why we should search for the black swan, which proposes we should search for the black swan because we should seek to challenge our own hypotheses which we *believe* to be true (Popper, 1959; Wason, 1960). Trying to confirm again what we already believe can lead to the maintenance of incorrect ideas, such as those concerned with prejudiced stereotyping (e.g., Snyder & Swan; 1978). Consider our hypothesis when we exchange the word 'swans' for a group of people, such as an ethnic minority, and exchange the word 'white' for a word with a negative connotation. If we only search for members of the group that confirm again the negative hypothesis, and do not seek to challenge it by searching for a member of the group who does not fit the negative hypothesis, we are maintaining a prejudiced

belief. This example illustrates hypothesis testing in a social domain where we should seek to challenge hypotheses, but especially our own (e.g., Aronson, 1999).

Traditionally there have been two classes of explanation offered to understand hypothesis testing, and why people tend to engage in confirmation bias. One explanation of why people tend to confirm their own hypotheses and falsify the hypotheses of others could be motivational. People do not like the discomfort of being incorrect and are directed towards confirming biased hypotheses further, because it reduces this feeling of discomfort (e.g., Festinger, 1957; Aronson, 1999). The other explanation of why people tend to confirm their own hypotheses is cognitive. People find it easier to confirm, because confirming is easier to attend to, search for, and recall because it is consistent with information that is currently being mentally represented (e.g., Kunda, 2000). The suggestion that motives affect reasoning is controversial. A major criticism of the motivational view is that all research proposed to demonstrate motivated reasoning can be reinterpreted in entirely cognitive, non-motivational terms (Nisbett & Ross, 1980). Furthermore, no theory has clearly addressed a cognitive mechanism through which motivation may operate (for discussions see Evans, 1989; Koslowski, 2000; Sperber & Mercier, 2010).

Yet there is a suggestion that when reasoning is directed by a goal, for example, a goal to create a new scientific theory or to achieve a strategic advantage in a game of strategy, then people may expend more cognitive effort (e.g., Kunda, 1987). This is in line with the cognitive view that extra expended cognitive effort can explain goal directed reasoning (e.g., Kunda, 2000). Cognitive effort can refer to deeper processing of information, attending more carefully to information, searching more for information, maintaining more information in working memory, and manipulating more information in working memory (e.g., Eysenck & Keane, 2000). When people are competing in a hypothesis testing situation, such as with an opponent scientist or military strategist, they may expend more cognitive effort checking that their hypothesis is a good one, so that they can achieve their goal. I turn now to describe how extra cognitive effort in reasoning may help people detect falsification.

### ***Cognitive effort in competitive hypothesis testing***

Consider once again the hypothesis ‘all swans are white’. Imagine two scientists are given this hypothesis to test. One scientist holds that the hypothesis is true,

and the other scientist holds that the hypothesis is not true. Each scientist should look for falsifying cases which are non-white, such as black swans. But the scientist who holds that 'all swans are white' is true will do the very same thing as the scientist who does not hold this hypothesis as true, albeit in hopes of not finding them! (De Groot, 1969).

Now consider that the two scientists are opponents. Consider the scientist who holds that the hypothesis 'all swans are white' is true. This scientist knows that the opponent scientist who holds that the hypothesis 'all swans are white' is not true, will also be testing the hypothesis. This awareness of an opponent hypothesis tester may prompt the scientist to anticipate how their own hypothesis may be falsified, even though they may hope not to find any falsification. People may expend more cognitive effort to ensure their hypothesis cannot be falsified; they may expend more effort by raising their level of strategic sophistication in order to prevent an opponent gaining an advantage (e.g., Camerer, 2004), and accordingly they may expend more effort in order to falsify an opponent's hypothesis. There are many competitive circumstances in real life where it is advantageous to falsify somebody else's hypothesis, for example in political debating (Einhorn & Hogarth, 1978), legal inquiry (e.g., Crombag *et al.*, 1993), or even in familial argumentation (e.g., Laing, 1971). The implication is that people may expend more cognitive effort when testing somebody else's hypothesis.

There are several ways in which cognitive effort may be understood in hypothesis testing. First, the generation of negative tests may be more difficult than generating positive tests due to general difficulties with negation in human reasoning (e.g., Johnson-Laird & Byrne, 1991; Evans, Newstead & Byrne, 1993; Evans, 1989). The critical 2-4-6 task situation when the typical participant's hypothesis is 'even numbers ascending in twos' and the experimenter's rule is 'any ascending numbers' requires the generation of a negative test in order to discover that the hypothesis is untrue (Wason, 1960). People have a tendency to engage in a positive test strategy by generating positive tests of a hypothesis (Klayman & Ha, 1987). People generate positive tests not only because they are familiar with some hypothesis testing situations in which positive tests are useful, but they tend to use positive tests when task demands are high suggesting that a positive test strategy is easier than a negative test strategy (e.g., Evans, 1989; Evans, Newstead, & Byrne, 1993). When people compete with an opponent hypothesis tester the competition may help them create other salient

alternative possibilities, or the competition may facilitate the extra cognitive effort needed to use negation to falsify an opponent's hypothesis.

Second, people tend to think of few possibilities in their reasoning because their working memory is limited (Johnson-Laird & Byrne, 2002). When people test hypotheses they often represent only one hypothesis at a time in working memory (Mynatt, Doherty, & Dragan, 1993), but when they compete with an opponent hypothesis tester they may represent two possibilities; their own hypothesis and the opponent's hypothesis. These possibilities may not necessarily correspond to false possibilities, but two possibilities that may be true (e.g., Tweney *et al.*, 1980; Johnson-Laird & Byrne, 1991). Competitive hypothesis testing may provide a forum in which people can consider two possibilities, their own hypothesis and their opponent's hypothesis, and the difficulty of representing two possibilities by oneself and falsifying one's own hypothesis may be slightly less. Third, competition may help participants to be better at making possible alternative hypotheses explicit for themselves in their mental representations of hypotheses. In Chapter 2 we saw how the consideration of explicit alternative hypotheses helped participants to falsify. One way an explicit alternative helped was by presenting participants with an alternative set of possibilities from which to generate test triples. In the experiments in this chapter explicit alternative hypotheses are not presented to participants; they are presented with one hypothesis to test only. Perhaps with competition participants may understand that there are alternative hypotheses that an opponent hypothesis tester may be considering. Even though the alternatives belonging to an opponent are non-explicit, the competition may prompt participants to flesh out these properties to consider what the opponent's alternatives might be.

The major implication is that not knowing what the mathematical relationship between the hypothesis and the true rule is, cannot entirely explain why people do not generate negative tests (Klayman & Ha, 1987). Under competitive circumstances participants in the 2-4-6 task may be able to make other alternative hypotheses explicit for themselves in order to falsify using a negative test. People can find it possible to falsify in competitive circumstances (Poletiek, 1996).

The following experiments examine the effects of competition on hypothesis testing. I examine whether people attempt to falsify someone else's hypothesis in Experiment 4, and investigate whether people may have a tendency to falsify

hypotheses belonging to an imaginary participant. I examine whether people falsify their own hypotheses more when they know an opponent is also testing their hypothesis in Experiment 5. The implications that the results have for the two main alternative theories of hypothesis testing; the uniformity theory (Poletiek, 1996, 2001) and the mathematical relationship theory (Klayman & Ha, 1987) are drawn out. The remainder of the chapter attends to how the results point to major shortcomings of these two accounts and indicate the foundations along which a new theory of hypothesis testing should be built. Let us turn now to Experiment 4 to examine the effect of hypothesis ownership on hypothesis falsification.

#### **Experiment 4: Hypothesis ownership**

In Chapter 2 we showed that participants could consistently falsify Peter's hypothesis in the imaginary participant 2-4-6 task. Participants falsified Peter's hypothesis consistently more than is usual for participants testing their own hypotheses in the 2-4-6 task literature (see Cowley & Byrne, 2005; 2015). This falsification of the imaginary participant's hypothesis was more consistent than has been previously reported in the literature, regardless of whether the participants were presented with an alternative hypothesis or not (see Poletiek, 2001 for a review). In this experiment we will examine if ownership of a hypothesis is a factor that affects falsification.

#### **Mathematical explanations of hypothesis testing**

In Chapter 1 we described the theory of hypothesis testing in the 2-4-6 task proposed by Klayman & Ha (1987). They proposed that the mathematical relationship between the participant's hypothesis and the truth (the experimenter's rule) was one of the main factors predicting hypothesis testing success in the 2-4-6 task. There are five possible relationships, but the one that corresponds to the 2-4-6 task is when the participant's hypothesis is embedded within the truth (i.e., the experimenter's rule). Recall that the relationship relevant to our imaginary participant experiments is when the participant's hypothesis is 'even numbers ascending in twos', and its properties of evenness and ascending in twos are embedded within the experimenter's rule 'any ascending numbers'. The 'any ascending numbers' rule applies to numbers ascending in any order, not only ascending in twos. Klayman & Ha (1987) suggest that this embedded relationship is the most difficult for participants,



because they can only discover that their hypothesis is incorrect by generating a negative test that leads to falsification (See Table 1.1 in Chapter 1 for a full description of the logic of the 2-4-6 task).

Consider how 3-5-7 is a negative test because it is inconsistent with the hypothesis 'even numbers ascending in twos'. 3-5-7 is not an instance of 'even numbers ascending in twos' as it contains odd numbers. If the experimenter replies 'yes' to indicate that a triple with odd numbers is consistent with the rule, (because the rule is 'any ascending numbers'), then the hypothesis pertaining to evenness is falsified. Klayman and Ha point out that this relationship is not representative of the majority of hypothesis testing situations that can occur, and people tend to test their hypotheses using positive tests, which are more effective at producing falsification in the other hypothesis testing situations. Consider again the example of the medical professionals researching the cause of a birth defect such as a *Spina bifida* as outlined in Chapter 1. The hypothesis was that the main factor leading to the birth of a baby with Spina bifida was genetic; a mother who was related to the *Celtic* gene pool in Ireland and the British Isles was more at risk of having a baby with Spina bifida. The researchers carried out a positive test of their hypothesis by examining blood samples and family history data from people in the Irish population. This positive test led to a theory of genetic predisposition in the cause of neural tube defects such as Spina bifida. If the researchers had chosen to focus on a negative test of their hypothesis such as a population from the African gene pool, they would have found close to zero percent cases and their search would be like the proverbial search for a needle in a haystack (Klayman & Ha, 1987).

Klayman and Ha suggest that people engage in a general positive test strategy because they are familiar with the usefulness of positive tests in real world examples. People persist in testing their hypotheses using positive tests in any hypothesis testing situation they encounter because they are familiar with the success of positive tests.

But participants need to generate a negative test if their hypothesis turns out to be untrue and embedded within the truth, whether the hypothesis belongs to an imaginary participant in a laboratory task or whether it is a prejudiced hypothesis belonging to the pages of history as our Anne Frank example in Chapter 1 testified. Hence, they predict that people can only falsify using a negative test when they are given extra help. For example, people can generate

negative tests when the relationship between their hypothesis and the truth is made explicit to them, such as by presenting them with an alternative (Klayman & Ha, 1989).

In the first condition participants are instructed to test ‘Peter’s hypothesis: even numbers ascending in twos’. In the second condition participants are instructed to test ‘Your hypothesis: even numbers ascending in twos’. The experimenter’s rule is ‘any ascending numbers’. I predict that testing someone else’s hypothesis enables people to consider multiple possibilities such as alternative hypotheses more readily, and participants can falsify the imaginary participant’s hypothesis more than their own.

To test the mathematical relationship theory I controlled for several factors. First, the hypotheses are equally low-quality, so any differences observed in testing behaviour cannot be explained by the quality of the hypothesis and the amount of available evidence for participants (Klayman & Ha, 1987). Second, the mathematical relationship between the hypothesis under test and the experimenter’s rule is the same in each condition, so any differences observed in hypothesis testing between the two conditions cannot be explained by the mathematical relationship theory either (Klayman & Ha, 1987). Third, the participants in both conditions were given the hypothesis for testing, they did not generate the hypotheses. Testing identically untrue hypotheses that they did not generate themselves, ensures that an explanation of personal investment at the hypothesis generation stage cannot predict any differences observed in hypothesis testing (e.g., Kunda, 1987; 2000).

I predicted that participants will not only generate more negative falsifying tests of an imaginary participant’s hypothesis than their own, but that they will expect these negative tests to falsify. I predict that hypothesis ownership is a competitive factor that affects hypothesis falsification.

### ***Materials and Design***

Participants were randomly assigned to two conditions ( $n = 16$  in each): in one condition the low-quality embedded hypothesis was identified as belonging to the “imaginary participant” Peter (Peter’s hypothesis is ‘even numbers ascending in twos’) and in the other the identical hypothesis was identified as belonging to the participant (Your hypothesis is ‘even numbers ascending in twos’). Participants were not made aware of what the experimenter’s rule was.

Crucially, the relationship between the hypothesis and the true rule was identical. The instructions were as follows:

“In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the experimenter’s rule was: “even numbers ascending in twos”.

Your aim is to go about testing if Peter’s rule “even numbers ascending in twos” is the experimenter’s rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind.

You should try to go about testing if Peter’s rule “even numbers ascending in twos” is the rule the researcher has in mind by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if Peter’s rule is the experimenter’s rule, *and not before*, you are to write down “Peter’s rule is the experimenter’s rule” or “Peter’s rule is not the experimenter’s rule”. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

The words ‘your’ and ‘you’ replaced the words ‘Peter’s’ and ‘Peter’ respectively for the condition where the hypothesis belonged to the participant themselves.

### ***Participants and procedure***

Thirty two people who were members of the general public volunteered and were paid a nominal fee of 8 Euro per hour. There were 23 women and 9 men and their age ranged from 20 years to 75 years, with a mean age of 51 years. No participants had taken courses in the philosophy of science. Participants were tested individually. The testing session lasted approximately 20 minutes.

### **Results and discussion**

#### ***Number of triples***

In total, 184 triples were generated, with a mean of 5.75 triples per participant. A similar number of test triples were generated for the imaginary participant’s

hypothesis ( $M = 6.0$ ) and for participants' own hypothesis ( $M = 5.5$ ), Participants did not generate more tests of the hypothesis belonging to the imaginary participant Peter significantly more than their own hypothesis.

**Positive and Negative tests**

Participants generated more negative tests of Peter's hypothesis (46%) than of their own hypothesis, but this difference was not reliable (25%,  $\chi^2 = 10.492$  (6),  $p = .052$ ). Participants generated fewer positive tests of Peter's hypothesis (54%), than of their own hypothesis (75%,  $\chi^2 = 18.619$  (9),  $p = .015$ ). The result that participants generated more negative tests and fewer positive tests for the imaginary participant's hypothesis than their own, may imply that people put more cognitive effort into testing somebody else's hypothesis than their own.

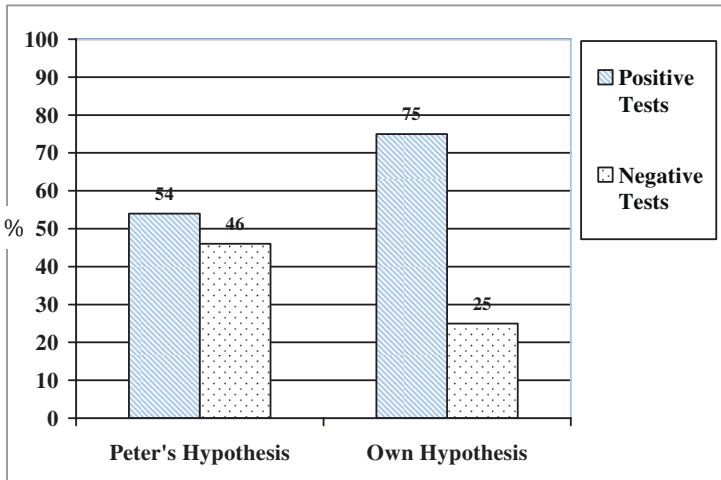


Fig.3.1.: Percentages of positive and negative tests generated in Experiment 4.

Participants were not presented with an explicit alternative hypothesis in either condition. Previous explanations have predicted that extra information such as an alternative hypothesis, is essential if people are to generate negative tests (Klayman & Ha, 1989). The result that participants generated negative tests readily for the imaginary participant's hypothesis does not corroborate this explanation (Klayman & Ha, 1989). Participants generated more negative tests of the imaginary participant's hypothesis than their own, even though the hypothesis quality was the same (Poletiek, 1996) and the mathematical

relationships between the hypothesis and the rule were identical in each case (Klayman & Ha, 1987). The only difference between conditions is that the hypothesis is said to belong to Peter in one condition and to the participant in the other. The experiment shows that hypothesis ownership can affect the generation of hypothesis test types.

### ***Falsification and confirmation***

Did participants intend to use their positive and negative tests differently to test the imaginary participant's hypothesis than their own? Participants generated more negative falsifying tests, when the hypothesis belonged to Peter (32%) than when the hypothesis was their own (7%), but this difference was not reliable, ( $\chi^2 = 2.667$  (4),  $p = .307$ ) as Table 3.1 shows. Participants intended their positive tests to falsify Peter's hypothesis a similar amount to their own hypothesis (8% vs 9% respectively,  $\chi^2 = 3.143$  (3),  $p = .185$ ). Overall participants tested Peter's hypothesis less with falsifying tests (40% vs 60%), although the difference was not reliable ( $\chi^2 = 5.25$  (5),  $p = .193$ ). Hypothesis ownership does not have a significant effect on the generation of negative tests which are expected to falsify.

Participants expected their positive tests to confirm reliably more often when the hypothesis was their own than when the hypothesis belonged to Peter (67% vs 46%,  $\chi^2 = 17.571$ (1) = 9,  $p = .02$ ). Participants expected their negative tests to confirm as often when the hypothesis was their own as when it belonged to Peter (17% vs 14%,  $\chi^2 = 4.4$ ,  $df = 5$ ,  $p = .246$ ). Overall participants tested their own low-quality hypotheses with confirming tests (84%) more than the imaginary participant Peter's (60%), but this difference was not significant ( $\chi^2 = 15.067$  (10),  $p = 0.06$ ). The results show that confirming your own hypothesis may be easier than confirming someone else's. Hypothesis ownership affects the generation of confirming hypothesis tests even though the relationship between the hypothesis and the experimenter's rule was not made explicit to participants (Klayman & Ha, 1987), nor was there an explicit alternative presented to participants (Klayman & Ha, 1989; Oaksford & Chater, 1994; Wason & Johnson-Laird, 1972). See Table 3.1.

Table 3.1: Percentages of hypothesis test types generated in each condition of Experiment 4.

	<i>Peter's hypothesis</i>	<i>Own hypothesis</i>
<i>Confirming</i>		
Positive	46	67
Negative	14	17
<i>Falsifying</i>		
Positive	8	9
<b>Negative</b>	<b>32</b>	<b>7</b>

***Using falsification to abandon low-quality hypotheses***

An important question is how participants used the falsifying and confirming test triples to reach the discovery whether they thought the low-quality hypothesis they were testing was the rule or not. Participants announced whether the hypothesis ‘even numbers ascending in twos’ was the experimenter’s rule when they finished testing. More participants abandoned the low-quality hypothesis when they finished testing Peter’s hypothesis (62%) than when they finished testing their own (38%), and fewer participants decided to abandon the low-quality hypothesis when they finished testing their own hypothesis (25%) than when they finished testing Peter’s (75%), and this result was reliable ( $\chi^2 = 4.571 (1), p = .016$ ), as Table 3.2 shows.

Table 3.2: Percentages of abandoned and endorsed hypotheses in each condition of Experiment 4.

	<i>Peter's hypothesis</i>	<i>Own hypothesis</i>
<i>Abandoned hypothesis</i>	62	25
<i>Endorsed hypothesis</i>	38	75

The result suggests that people not only find falsification to be possible but they also find it to be useful: they can use it to abandon untrue hypotheses.

### **Summary**

The results of this experiment suggest that the role of the hypothesis tester is not totally constrained by the mathematical properties of the problem (Klayman & Ha, 1987). The effect of hypothesis ownership on the generation of positive and negative tests, and the intention to turn these tests into confirming or falsifying ones, shows that hypothesis testing cannot be completely explained by the mathematical relationship between the hypothesis and the evidence. In other words the hypothesis tester has their own active role both in the selection of hypothesis tests and in the interpretation of the test result, and this role cannot be completely explained by the constraints placed upon the hypothesis tester by the mathematical properties of the problem.

In sum, participants tend to confirm their own hypothesis reliably more than someone else's hypothesis and they can somewhat falsify to test untrue hypotheses belonging to someone else, while they do not seem to falsify to test their own equally untrue hypotheses. The implication is that the results indicating higher levels of falsification than is usual in the 2-4-6 literature in these experiments are partly due to the testing of someone else's hypothesis, namely the imaginary participant Peter. The remaining question is can people falsify their own hypotheses? The next experiment addresses this question by introducing a previously unexplored factor in hypothesis testing in the 2-4-6 task—direct competition with an opponent.

### **Experiment 5: An opponent hypothesis tester**

In Experiment 4 we showed that competition affected hypothesis testing. Hypothesis ownership affected not only (1) the generation of negative tests which can falsify a hypothesis, but (2) how readily participants abandoned an untrue hypothesis. In this experiment I examine whether the consideration of an opponent hypothesis tester affects hypothesis testing.

In the introduction to this chapter it was proposed that people might have a tendency to falsify in competitive situations because hypothesis testing in realistic settings may proceed by testing other people's hypotheses or interacting with an opponent hypothesis tester. Scientists may often proceed by attempting to confirm their own hypotheses and falsify other scientists' hypotheses (e.g., Mitroff, 1974; Gorman, 1995a; Fugelsang *et al.*, 2004). Legal experts need to not only compete with opposition barristers, but to ensure that the grounds on which they base their legal arguments are irrefutable (e.g., Britton, 1997; Canter, 2000). Military strategists must engage with opposition forces, and ensure that they consider each possible alternative at the disposal of the opponent to their hypothesized plans of action (e.g., Mallie, 2001). Falsification is more likely to come from a scientist who is not working within the program than from a scientist working within the program (e.g., Kuhn, 1993).

For these reasons it was suggested that the consideration of an opponent hypothesis tester may prompt participants to expend more cognitive effort when testing their hypotheses. Participants were predicted to generate negative falsifying tests more readily when they consider an opponent hypothesis tester who is also testing their hypothesis, than when they do not consider an opponent.

Competition may help participants to be better at making possible alternative hypotheses explicit for themselves in their mental representations of hypotheses. The increased cognitive effort may help participants to consider more than one explicit hypothesis in working memory (e.g., Mynatt, Doherty, & Dragan, 1993; Byrne, 2005). Chapter 2 showed how the consideration of explicit alternative hypotheses helped participants to falsify. One way an explicit alternative helped was by presenting participants with an alternative set of possibilities from which to generate test triples. In the experiments in this chapter explicit alternative hypotheses are not presented to participants; they are presented with one hypothesis to test only. Perhaps with competition participants may understand



that there are alternative hypotheses that an opponent hypothesis tester may be considering. Even though these alternatives belonging to an opponent are non-explicit, the competition may prompt participants to flesh out these properties to consider what the opponent's alternatives might be.

If the consideration of an opponent hypothesis tester affects hypothesis testing, then there are several implications for theories of hypothesis testing. I outline these implications in the next section.

### ***Implications for the uniformity theory (Poletiek, 1996, 2001)***

In this experiment we will examine whether participants falsify their own hypothesis more readily when they consider an opponent hypothesis tester, than when they do not consider an opponent hypothesis tester.

In the introduction chapter the two main alternative theories of hypothesis testing were described; the uniformity theory (Poletiek, 1996; 2001) and the mathematical relationship theory (Klayman & Ha, 1987). Neither theory makes a prediction about the effects competition may have on hypothesis testing.

The main tenet of the uniformity theory is that people experience a hypothesis test as one and the same process regardless of whether the test leads to a confirming or falsifying result (Poletiek, 2001; 2005). The uniformity theory predicts that competitive factors should have no effect on this process. Whether one hypothesis testing situation is competitive and another is not, the hypothesis tests people generate should not differ. Positive and negative tests should be generated as often whether an opponent hypothesis tester is present or not. And these tests should be expected to lead to confirming and falsifying results as often whether an opponent hypothesis tester is present or not.

The main tenets of the mathematical relationship theory are that people do not know when it is wise to generate negative tests of a hypothesis. They follow a positive test strategy unless they are presented with an explicit alternative hypothesis, to show them that the relationship requires the generation of a negative test of the hypothesis (Klayman & Ha, 1987; 1989). The mathematical relationship theory predicts that competitive factors should have no effect on how 'wisely' people generate their hypothesis tests. Positive and negative tests should be generated as often whether an opponent hypothesis tester is present or not, unless the opponent makes a higher quality alternative hypothesis explicit to them.

### ***Materials and design***

Participants were randomly assigned to two conditions (n = 16 in each): in both conditions the low-quality embedded hypothesis was identified as belonging to the participant (Your hypothesis is ‘even numbers ascending in twos’). In the experimental condition participants were given additional information about an opponent hypothesis tester (‘However, an opponent called Peter is also testing ‘even numbers ascending in twos’’). Participants were not made aware of what the experimenter’s rule was. Crucially, the relationship between the hypothesis and the true rule was identical. The instructions for the opponent condition were as follows:

“In a previous study investigating human thinking you were a participant who was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. You hypothesised that the experimenter’s rule was: “even numbers ascending in twos”.

Your aim is to go about testing if your rule “even numbers ascending in twos” is the experimenter’s rule. However, an opponent called Peter is also testing “even numbers ascending in twos”. You must discover if “even numbers ascending in twos” is the experimenter’s rule before he does.

You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if your original rule “even numbers ascending in twos” is the experimenter’s rule, and not to test any new ideas of your own that you think the experimenter’s rule might be.

Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if your rule is the experimenter’s rule, *and not before*, you are to write down “My rule is the experimenter’s rule” or “My rule is not the experimenter’s rule”. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.”

### ***Participants and procedure***

Thirty two people who were members of the general student population in Trinity College, University of Dublin volunteered, and were given a minor

reward of one bar of chocolate. There were 26 women and 6 men whose ages ranged from 18 years to 27 years, with a mean age of 20 years. No participants had taken courses in the philosophy of science. Participants were tested individually or in a group of up to three people.

## **Results and discussion**

The results for the following dependent measures are reported: the number of triples; the number of positive and negative tests; the percentage of falsification and confirmation; and whether or not the untrue hypothesis is abandoned at the end of the task.

### ***Number of triples***

In total 147 triples were generated, with a mean of 4.6 triples per participant. A similar number of test triples were generated when participants were told there was an opponent hypothesis tester ( $M = 4.75$ ) and when they were not ( $M = 4.44$ ), (Mann-Whitney  $U = 125.00$ ,  $Z = -.118$ ,  $p = .906$ , two-tailed). The consideration of an opponent hypothesis tester did not encourage participants to test their hypothesis with more tests than when there was no opponent.

### ***Positive and negative tests***

Participants generated more negative tests when there was an opponent (62%) than when there was no opponent (21%), and more positive tests when there was no opponent (79%) than when there was an opponent (38%); this pattern was reliable, ( $\chi^2 = 4.5$  (1),  $p = .038$ , two-tailed) as Fig. 3.2 shows.

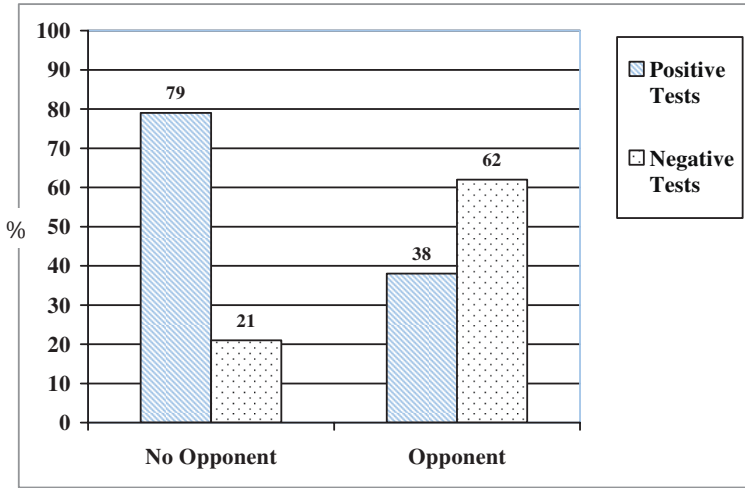


Fig. 3.2.: Percentages of positive and negative tests generated by participants for their own hypotheses when an opponent hypothesis tester was absent or present.

The result that participants generated more negative tests and fewer positive tests when there was an opponent may imply that people put more cognitive effort into testing their hypothesis when they are competing with an opponent hypothesis tester.

***Falsification and confirmation***

Do participants intend to use their positive and negative tests differently to test their hypothesis when there is an opponent? Overall participants generated a similar amount of confirming tests whether or not there was an opponent hypothesis tester (91% vs 88%), but the types of confirming tests differed as Table 3.3 shows. Participants generated more positive confirming tests when there was no opponent (75%) than when there was an opponent (37%), although the difference was not reliable, ( $\chi^2 = 8.067 (7), p = .164$ ). (This raises the possible question of power because there is a 40% difference and the p value is not significant. In fact the reason for this non-significance is a result of the degrees of freedom that are sometimes elevated because participants do not generate the same number of triples in each case of the chi square matrix;

Hollander & Wolfe, 1999). The difference suggests that participants may confirm less readily when they are aware of an opponent also testing their hypothesis.

Participants who considered the opponent hypothesis tester generated negative tests reliably more than participants who did not consider an opponent, but they intended the negative tests to confirm. Participants in the opponent condition expected their negative triples to be inconsistent with the experimenter’s rule, thereby confirming their hypothesis. Participants in the opponent condition generated these negative confirming tests (54%) reliably more often than participants in the no opponent condition, (13%,  $\chi^2 = 11.4$  (6),  $p = .039$ ). The opponent condition was the only condition in which such a high number of negative confirming tests were generated in any of the 2-4-6 experiments reported in the thesis.

Table 3.3: Percentages of hypothesis test types generated in each condition of Experiment 5.

	<i>No opponent</i>	<i>Opponent</i>
<i>Confirming</i>		
Positive	75	37
Negative	13	54
<i>Falsifying</i>		
Positive	4	1
Negative	8	8

Participants generated too few falsifying tests to warrant a statistical analysis. However, they did generate the same amount of negative falsifying tests whether an opponent hypothesis tester was present or not (8% in each case). They generated a small amount of positive falsifying tests irrespective of whether an opponent hypothesis tester was present (1% vs 4%). In both conditions the participants tested their hypothesis that belonged to themselves rather than an imaginary participant, and the rates of falsification were low. For

example, participants tested their own hypothesis with falsifying tests when an opponent hypothesis tester was not present only a small amount of the time (12%). This result replicates the previous experiment (Experiment 4) which showed that participants did not falsify their own hypotheses (16% were falsifying tests).

The introduction of an opponent hypothesis tester affected the types of hypothesis testing; participants changed the type of confirming tests they performed, from positive confirming tests when there was no opponent, to negative confirming tests when there was an opponent. This result does not corroborate the uniformity theory which predicts that people experience hypothesis testing as one and the same process regardless of other factors (Poletiek, 2001; 2005). Hypothesis testing cannot be explained by mathematical theories positing that the relationship between the hypothesis under test and the experimenter's rule constrains hypothesis testing; the relationships were identical in both conditions, and participants generated different types of tests (Klayman & Ha, 1987). Instead, participants may be expending more cognitive effort when they consider an opponent; they generate more negative tests.

### *Using falsification to abandon low-quality hypotheses*

Somewhat more participants abandoned the low-quality hypothesis when there was an opponent (56%) than when there was no opponent (38%), and somewhat fewer participants endorsed the low-quality hypothesis when there was an opponent (44%) than when there was no opponent (62%), but the difference was not reliable, ( $\chi^2 = 1.129$  (1),  $p = .288$ , two-tailed), as Table 3.4 shows. This result may indicate that participants are somewhat better able to successfully discover that their hypothesis is low-quality and untrue when an opponent hypothesis tester is introduced. (However, this result represents the fact that three more people abandoned the untrue hypothesis when there was an opponent than when there was no opponent, which can only suggest a tentative conclusion for this result). By generating negative tests, participants did in fact receive falsification even though they did not expect it; they expected their negative tests to confirm but instead received falsification from the experimenter who replied 'yes' to their negative tests, when they expected a 'no' (refer to Table 1.1 in Chapter 1).

Table 3.4: Percentages of abandoned and endorsed hypotheses in each condition of Experiment 5.

	<i>No opponent</i>	<i>Opponent</i>
<i>Abandoned hypothesis</i>	38	56
<i>Endorsed hypothesis</i>	62	44

In the opponent condition five participants (36%) maintained that their untrue hypothesis was the experimenter’s rule even though they had falsifying evidence to tell them otherwise. This result indicates that there may be a bias not only in the search for tests of one’s hypothesis, but also in the interpretation of the test result. Previous research has tended to exclusively focus on bias in the search for hypothesis tests rather than on the interpretation of falsifying evidence once it is found (Poletiek, 2005; Howson & Urbach, 1993). Theories of belief revision predict that when people encounter evidence inconsistent with their beliefs they either choose to disbelieve their belief or the evidence encountered (e.g., Gärdenfors, 1988); in this experiment some participants chose to disbelieve the falsifying evidence.

**Summary**

Competition affects hypothesis testing. Hypothesis ownership and the consideration of an opponent hypothesis tester affected hypothesis testing in several ways. First, participants generated negative tests of a hypothesis more readily when it belonged to someone else, and second when they considered that an opponent hypothesis tester was testing their hypothesis. In this experiment participants generated negative tests of their hypothesis when they considered an opponent, but they did not expect these negative tests to falsify. The results suggest that the process people use to generate a hypothesis test in a competitive situation may be different than in a non-competitive situation. They may be expending more cognitive effort in competitive situations in order to generate negative tests or in order to generate more possibilities. This result does not

corroborate the uniformity theory that predicts that people use the same process to test a hypothesis regardless of other factors (Poletiek, 2001; 2005). Whether they are able to generate their own explicit alternative hypotheses in competitive circumstances in order to falsify remains to be seen, but it is reasonable to suggest that competition affects the generation of negative tests in our experiments through increased cognitive effort. I turn now to a general discussion of the experimental results reported in Experiment 4 and 5.

### **General discussion**

The two experiments show that competition affects hypothesis testing. Experiment 4 found that hypothesis ownership affected hypothesis testing. An untrue hypothesis was falsified reliably more readily when it belonged to somebody else (the imaginary participant Peter), than when it belonged to the participant. This falsification was used to abandon the untrue hypothesis more readily when it belonged to somebody else than when it belonged to the participant. Experiment 5 showed that the introduction of an opponent hypothesis tester affected hypothesis testing. Negative tests with the potential to falsify a hypothesis were reliably and more readily generated to test an untrue hypothesis when the participant was told an opponent was also testing the hypothesis than when there was no opponent. This result suggests that the participants tested their hypothesis more rigorously when there was an opponent, perhaps attempting to anticipate how the opponent might falsify their hypothesis. Yet they may have wished that this anticipated falsification would not occur. In other words when participants are aware that they are competing with an opponent in this hypothesis testing situation, they more readily attempt to see how they might go wrong, but even so are unlikely to admit that they might actually go wrong. Over one third of these participants ignored falsifying evidence once they received it, and refused to abandon their untrue hypotheses.

The results have several implications for current theories of hypothesis testing. First, the effect of competition in hypothesis testing corroborates the separation of falsification into falsifying one's own hypothesis, and falsifying someone else's hypotheses (Poletiek, 2005). Each of the three imaginary participant 2-4-6 experiments in Chapter 2, and the hypothesis ownership experiment in this chapter shows that people can consistently falsify hypotheses belonging to somebody else, but rarely their own hypotheses. In fact, when participants considering an opponent in Experiment 5 generated a test that



would objectively falsify their own hypothesis they rarely expected that it would.

The debate about falsification has tended to assume that people find falsification to be difficult if not impossible (e.g., Poletiek, 1996), or possible under certain circumstances such as in group problem solving (e.g., Gorman *et al.*, 1984). One way the experiments in this chapter bring understanding to this contradiction in the literature, is by showing that competition can explain the results achieved from both perspectives. That is, when people test their own hypotheses falsification is very difficult, and when people test hypotheses when interacting with others falsification is possible; people can expend more cognitive effort to help them falsify.

Second, the results of both experiments have implications for mathematical theories of hypothesis testing. The mathematical relationship between the hypothesis and the experimenter's rule, and the quality of the hypothesis was identical in each condition of each experiment. Yet reliably different hypothesis testing resulted. The implication is that theories founded on the mathematical relationship between the experimenter's rule and the hypothesis (Klayman & Ha, 1987), or on the likelihood of obtaining falsifying evidence given the quality of the hypothesis (Poletiek, 2001; Howson & Urbach, 1993), have little to say about how competition may affect hypothesis testing.

The explanation for how competition may affect hypothesis testing was mentioned in the introduction to this chapter. In short, participants may try harder, that is, they expend more cognitive effort in generating more possibilities or negative instances when testing hypotheses belonging to somebody else, or when interacting with an opponent who is also testing their hypothesis. I suggest that extra cognitive effort affects a complete sequence of cognitive events involved in falsification (e.g., Kunda, 2000). A search is conducted in order to generate an alternative hypothesis, or a property relevant to numbers might be retrieved from limited domain knowledge people have about numbers (see Johnson-Laird, 1983; Johnson-Laird & Byrne 1991; Byrne, 2005 on the representation of mental models and searching for alternatives in reasoning tasks). A number triple that is not an instance of the hypothesis under test is generated from the product(s) of this search. Participants may expect the triple to result in a falsification, and when it does they may switch their attention to the hypothesis currently being represented (see Mynatt, Doherty, & Dragan, 1993 on attention switching in hypothesis testing). They must detect the

inconsistency between the hypothesis they are mentally representing and the falsifying triple (see Elio & Pelletier, 1997; Byrne & Walsh, 2005 on reasoning with inconsistency). Finally, the participant may update the truth status of the hypothesis to indicate that it is untrue.

In short, there may be many different stages and ordering of stages involved in falsifying a hypothesis; especially when no explicit alternative hypothesis is presented to participants. Extra cognitive effort may help explain how people can falsify a hypothesis belonging to someone else more easily than their own, in the absence of explicit alternative hypotheses being presented to them. Competition may prompt people to represent the hypothesis testing situation more explicitly, to consider more than one possibility in order to generate their own alternative hypotheses, to maintain negative test results in working memory, and search more for negative falsifying instances.

In conclusion, the imaginary participant experiments in this chapter point out that competition affects hypothesis testing. This is a novel result and it does not corroborate the tenets of the two main alternative theories of hypothesis testing, which importantly were also developed from findings in the 2-4-6 task. Yet we need a more detailed analysis of how alternative hypotheses, explicit representation of alternative possibilities, competition and cognitive effort relate to one another in order to create a more detailed picture of hypothesis testing. In the next chapter such a detailed analysis of hypothesis testing is carried out. One of the most extensive protocol analyses of hypothesis testing undertaken in the literature is produced in order to draw a more complete picture of hypothesis testing. I will focus on the hypothesis testing of chess masters.



## Chapter 4 Chess Masters' Hypothesis Testing

*Chess inculcates certain virtues such as foresight, patience, and the ability to accept the consequences of one's decisions.*

—Benjamin Franklin (*US President*, 1786; cited in Hartston & Wason, 1983, p. 7)

A fundamental type of thought is the prediction of events ( Craik, 1943). When people generate a hypothesis they may consider the implication of the hypothesis assuming it is true, or they may attempt to predict the most likely outcome given some scenario (e.g., Fugelsang *et al.*, 2004; Girotto, Legrenzi, & Rizzo, 1991). For example, people may anticipate a possible outcome when planning a future action or choice that they will make. This chapter examines how people test their hypotheses when they are planning a course of action. Understanding how people test their hypotheses about plans of action may help in understanding how people commit errors, learn from past mistakes and develop better planning for the future (e.g., Roese & Olsen, 1995; Vygotsky, 1986).

To explicate these thought processes further in a controlled and precise way I chose to focus on plans of action in chess players' hypothesis testing. Thinking in chess may consist of exploring different alternative hypotheses about what the best move is, in order to achieve a desired goal (Newell & Simon, 1972). Chess players must play moves that are so good that they may not be refuted (Saariluoma, 1995), implying that it is helpful to consider the possible alternative ways a plan may be falsified. In other words, how might a plan go wrong, for example, by an opponent responding to a plan in a way previously not anticipated? I chose the domain of chess because studies of chess have contributed substantially to the understanding of human cognition (e.g., Newell & Simon, 1972; Gobet & Simon, 1996a; 1996b; 1996c), and chess provides a neat micro-world of cognition in which to investigate different aspects of thinking with unparalleled precision (DeGroot, 1965; Newell, 1990).

The chapter focusses on how chess players anticipate an opponent's response to the moves they consider for play by suggesting that chess players evaluate moves they plan to play by searching for confirming evidence that their plan will work or falsifying evidence that their plan will not work. This investigation of hypothesis testing in chess emphasizes a question about chess playing

originally identified by DeGroot (1965), who published the first studies of chess during the 1940's (translated to English in the 1960's). How do chess players anticipate the ways a plan might be falsified in order to avoid making mistakes? The question has been neglected since and studies of chess have focused on other factors, such as memory for chess positions (e.g., Chase & Simon, 1973a; 1973b). DeGroot pointed out that the choice of a good move rests partly on foreseen possibilities for action and on the evaluation of their foreseen results (1965).

To address this question this chapter will present the first exploration of chess players' hypothesis testing. It will provide one of the largest analyses of chess players' thinking in the past forty years (DeGroot, 1965; Newell & Simon, 1972). Protocols from ten players who thought aloud while choosing moves for play for six positions each were selected. Sixty protocols were analysed (minus three that were inaudible, see Experiment 6 for details) and a comparison of experts to non-experts (5 master level players and 5 novices) was conducted. The protocols were selected from a larger, previously unanalysed set of data (Cowley, 2002). The data were analysed by mapping out all move sequences considered by each player for each position, and by using one of the most powerful chess programs in the world (at the time of the study, 2002) to objectively evaluate the outcomes of move choices (*Fritz 8*). The prediction was that expert chess players would attempt to falsify, by anticipating opponent moves that would ruin their plans more readily than novices, because experts may be better at detecting mistakes and recovering from them than novices.

In what follows I will describe in detail why chess presents us with an ideal domain in which to investigate hypothesis testing, by outlining the methodological advantages of the chess domain. Two main sorts of theories of chess expertise are detailed; chunking theories and search theories, and a new theoretical component of chess expertise based on a hypothesis testing framework derived from the experiments reported in this thesis is put forward. The experimental analysis of chess expertise is outlined and the implications that a hypothesis testing theoretical component of chess expertise may have for cognitive theories of how people test their hypotheses is discussed.

### **Methodological Advantages of Chess Experiments**

Studies of chess have contributed substantially to understanding human cognition, including how people solve problems (Newell & Simon, 1972), how

people chunk knowledge together to remember it (Chase & Simon, 1973a; 1973b), and how people develop their expertise (DeGroot, 1965). Chess findings provide strong external validity; results observed in chess experiments have contributed to explanations of expertise in non-game domains such as physics (Larkin, McDermott, Simon, & Simon, 1980; Chi, Glaser, & Rees, 1982), computer programming (McKeithen, Reitman, Reuter, & Hirtle, 1981), and music (Sloboda, 1976). Accordingly, I will suggest that findings concerning chess players' hypothesis testing, and chess masters' hypothesis testing in particular, may lead to the generation of similar explanations of expert thinking in non-game domains, such as scientific thinking (e.g., Fugelsang *et al.*, 2004). The findings may also have implications for understanding thinking in domains analogous to chess where interaction with a collaborator or an opponent is required, for example, in expert military strategy (Mallie, 2001).

Another advantage is that expertise in chess can be defined objectively and categorized relative to other experts by an established standard rating system. This rating system ensures that we have a relatively accurate system for classifying levels of expertise, more so than in some domains where it is difficult to classify expertise, such as in leadership or creative domains (e.g., Gardner, 1983; Charness, 1991). For example, the Elo rating system in chess classifies the level of expertise that players have obtained (Elo, 1978). Elo systems calculate an expected playing strength rating on the basis of competitive tournament games. The stronger the player one can defeat the more points one's rating is increased, and the weaker the player one is defeated by the more points one's rating is decreased. Ratings vary between approximately 1000 for an absolute novice and over 2800 for the world champion.

A further advantage is that chess has a 'unique' notation for describing what a player is currently representing while thinking about what move to play. Each player is fluent in this notation, and can verbalise what they are thinking about, with this verbalisation having minimal interference with thought processes (Ericsson & Simon, 1993). The availability of this notation allows an experimenter who is also fluent in it to transcribe a player's thinking aloud, and map each and every move a player verbalises. The notation is called *algebraic notation* and I describe it in detail in the next section.

### ***Algebraic Notation in Chess***

Chess is a board game where the overall goal is to move your pieces around in a

purposeful manner, in order to checkmate the opponent by attacking the opponent's king piece, and eliminating all the possible ways the opponent king can escape your attack. Chess thinking may consist of exploring different alternative paths or 'moves' in a 'problem space' (Newell & Simon, 1972). The problem space consists of the initial problem state, that is, the start of the game; intermediate problem states, for example, capturing an opponent piece, and the end state, checkmate. Progress from state to state is achieved through operators and in chess the operators are the ways the chess pieces are allowed to move. For example, a bishop operates diagonally backwards and forwards and captures on the square on which it lands for any one move. At the beginning of a game of chess the two players have equal numbers of pieces and theoretically almost equal chances of a win. Algebraic notation means that chess operators can be formalized mathematically and players' verbalizations of their thinking can be recorded in the common short-hand code used to refer to moves. Consider a simple example of algebraic notation with reference to Figure 4.1 below.

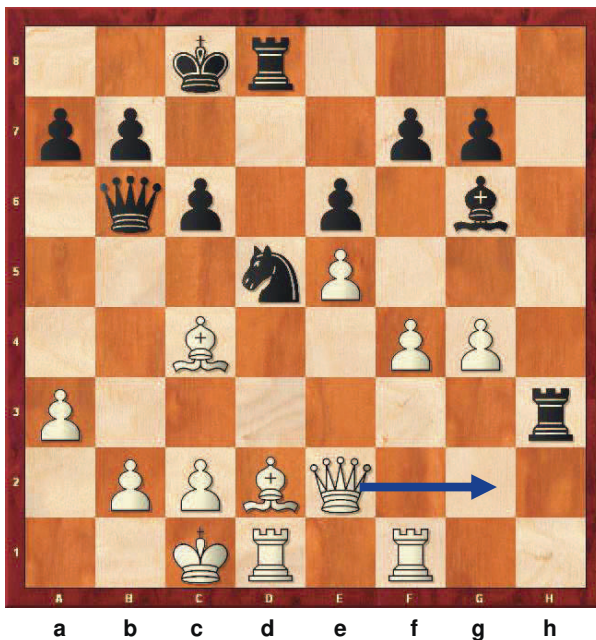


Figure 4.1: A representation of a chess board middle game, in which it is white to play.

Suppose a player faced with this board is invited to think-aloud. The player says “If I play Qg2”, which means the player is thinking about moving the queen (Q) to the square called g2. The arrow indicates that the square that the queen is perceived to move to. Each square on the chessboard has an algebraic coordinate. Starting from the left hand corner and moving horizontally to the right, each square (x8) is labelled with a letter as Figure 4.1 shows. The letters range from a - h. Again starting from the left hand corner but this time moving vertically upwards, each square (x8) is labelled with a number. The numbers range from 1 - 8. So, the coordinate of the first square in the left-hand corner is known as a1. The square to a1’s immediate right is b1 and so on. There are sixty-four coordinates altogether. The algebraic notation provides chess players with the means of recording where pieces have moved in a game of chess. The pieces have their own symbols. The symbols are King = K; Queen = Q; Rook = R; Knight = N; Bishop =B; Pawn moves are recorded merely as coordinates (e.g. b4 = a pawn has moved to the b4 square). For example a move could be Ne4 (the knight has moved to the e4 square). Typically chess players do not verbalise one move in isolation, such as Ne4 or Qg2. They often generate a sequence of moves of the form “If I play Qg2, then you might play Nd7, and if you play Nd7 then I’ll play...”. Next I turn to the most influential experimental studies of chess players’ thinking, which used this notation. First, the theoretical debate about what constitutes chess expertise is described: the ability to search through more moves than one’s opponent to find a good move, or the ability to reproduce good moves from large repositories of chess knowledge built up with extensive amounts of practice. An experimental analysis of hypothesis testing in chess playing is conjectured to help to resolve this debate. Finally, how this exploratory analysis may elicit a new hypothesis testing component of chess expertise is discussed, and the implications it may have for cognitive theories of expertise are outlined.

### **Theories of Chess Expertise**

The theoretical debate exists between chunking theories and search theories and each are described in turn.



### *Chunking Theories*

Chunking theories of chess expertise focus on how experts store their knowledge of a domain, that is, how smaller related pieces of knowledge can be grouped together into larger pieces of knowledge called chunks (Chase & Simon, 1973a; 1973b). The chunking theory of expertise was developed initially from De Groot's early examination of 22 chess players (6 grandmasters, 4 masters, 7 experts and 5 strong club players). The players were requested to think-aloud while deciding what move they would play next (DeGroot, 1965). Their protocols showed that they considered the same number of moves, and searched to the same depth in move sequences. (Search depth refers to how many moves ahead a player thinks about from the current position the player is thinking about. A player who thinks about what move he or she will play and what move the opponent will play in response, and then what move he or she will play in response to that, has searched to a depth of three moves. Each move is usually called one ply, and so the player's search depth is 3 ply.) But when players were shown board positions for two to fifteen seconds, and asked to recall where the pieces had been, the masters' recall was better (93%) than the recall of the experts and the strong players (72% and 51%). The result has been replicated (Chase & Simon, 1973a; 1973b; Gobet, de Voogt, & Retschitzki, 2004). It has been taken to indicate that the masters' greater playing experience leads to greater familiarity with chess configurations likely to occur in a game (DeGroot, 1965).

In a study of three chess players (a master, a class A expert - which is a strong club player with a US rating between 1900 and 2100, and a beginner), the participants were presented with normal chess board positions and random positions (Chase & Simon, 1973a; 1973b; 1973b). In the random positions, the pieces were scrambled to achieve a random pattern unlike any that would have occurred in a real chess game. Each of the three players showed poor recall for the random positions, worse even than the beginner's recall for normal positions (less than 17%). The result corroborates the idea that experts may be better at remembering information pertinent to their domain of expertise because they have experience in that domain. Moreover, experts may 'chunk' information more effectively than novices (Chase & Simon, 1973a; 1973b). In one task, two chessboards were placed side by side. Positions were set up on the left board and the 3 players were required to reconstruct the pieces on the empty chessboard on the right. They were allowed to glance back and forth between

the position and the empty board and they were video-taped. When players' placed pieces together on the empty board within a glance (without looking back at the position), it was taken to indicate that these pieces had been 'chunked' together in memory. Two pieces placed one after another rarely exceeded two seconds. A glance was assumed to establish the boundary of a chunk and pieces placed within a limit of two seconds were assumed to belong to the same chunk. Chunks were found to be related to one another in a number of different ways including spatial proximity, colour, similarity, attack or defense relations, pawn chains, and castled king formations. The average size of chunks in recall correlated with the 3 players' level of skill. The master placed 2.5 pieces on average within a glance, the class A expert 2.1 pieces, and the beginner 1.9 pieces. The master also encoded the chunks in shorter glances suggesting that experts may be faster encoders of domain specific knowledge. Accordingly, the ability of experts to recall briefly presented board positions may rely on better pattern recognition (more information encoded within a glance and faster encoding times). Further support for this chunking model of expertise was provided by an early computational model called MAPP (Memory Aided Pattern Perceiver) (Simon & Gilmarin, 1973). The program held what was considered to be a large number of chunks at the time and it encoded board positions into its short-term memory by recognizing chunked patterns. Two versions of the program were generated and the one with more chunked patterns performed better (1114 versus 894 patterns). According to the chunking theory each of these patterns can quickly be identified and can be accessed in a players' long-term memory through a label (Frensch and Sternberg, 1991, p.356). This label matching process activates retrieval of information about that chunk (Saariluoma, 1995). According to Chase and Simon (1973) chunks reduce the cognitive load in working memory. Since beginners do not typically have access to a large store of cognitive chunks they must try to recall the 24 individual pieces and their locations. The brief presentation of five seconds assumes that only short-term or working memory is being used, and typically it has been found to have a capacity of 7 items plus or minus 2 (Miller, 1956; Baddeley, 1999). Therefore beginners find it difficult to store up to 24 pieces in working memory, whereas masters can store a chunk containing, for example, 3 pieces as one item in working memory, increasing their working memory capacity for briefly presented positions. In random positions there are no chunks available

for recognition and so the skilled player's performance is no better than a novice (Chase and Simon, 1973).

In fact, experts may be able to access a considerable amount of information, more than an isolated chunk (Gobet, 1996). Board position knowledge may be stored in schematic retrieval structures or *templates*, that is, an abstraction of a set of similar positions and general plans and possible moves that follow from them (see Gobet, 1998; Gobet *et al.*, 2004). Templates are more general than any specific type of position. Support for the view comes from studies of chess masters' memory. Gobet & Simon (2000; for a review see Gobet *et al.*, 2004) compared five masters with eighteen experts and seven class A players. The players reconstructed a rapidly presented board position within a two second interval. Masters remembered and reconstructed the locations of more pieces from rapidly presented board positions than the experts and class A players. In addition, the sizes of the largest chunks the masters remembered were larger than the others (15-16 for masters, 9-10 for experts, 4-5 for class A players). For example, a chess player may be familiar with an opening known as the 'Sicilian Dragon' in which the pawn formation is shaped like the profile of a Chinese dragon spanning the width of the board (Gobet, de Voogt, & Retschitzki, 2004; Nunn, 1999). Their template may contain information about the general structure of the positions that arise from this opening and general plans that follow on from achieved positions. A master who is familiar with the Sicilian Dragon opening would be able to recall a briefly presented Sicilian Dragon position in chunks that contain more than 3 pieces. For example the seven pawns in the pawn structure can be encoded in a single chunk. A position template is structurally different to a set of chunks because of its integration of an entire position. Knowledge structures may become more integrated with increasing levels of expertise.

Accordingly the chunking and template theories predict that experts rely on domain knowledge to deal with routine problems with which they are familiar. These theories provide convincing explanations for how experts choose a good move, that is, a move that improves their position and does not lead to a worsening of their position. But why do experts sometimes make a bad move, that is, a move that worsens their position? Expertise is often associated with a lack of error but experts do make mistakes (Green & Gilhooly, 1992). Experts have not always been experts and they continue to have experience of making mistakes, as losses at the highest levels of competition testify. Perhaps chunking

and template theories are better at predicting performance for routine problems, which experts are familiar with. What happens when experts are confronted with a problem that does not map neatly onto their knowledge?

### *Search Theories of Chess Expertise*

In an early investigation of how experts deal with novel problems, 3 experts and 3 novices were asked to choose moves in randomized chess positions, where the pieces had been scattered about the board, until few regularities of chess configurations were present (Holding & Reynolds, 1982). The experts chose better quality moves for play. How did the experts choose better moves for play if they could not use their superior knowledge? The experts may be able to use their superior search capabilities to choose better moves (Holding & Reynolds, 1982; Holding, 1985). The implication was that deeper search may explain how masters anticipate their opponents' moves, prepare a counter-move in advance, and thereby choose better quality moves. Experts have been found to search a larger problem space, that is, they consider more moves than novices, when the board positions are more complex (Gobet & Campitelli, 2002). But experts have only been found to search deeper than novices when the problems they are presented with are not routine and do not demand reproducing solutions already known.

Overall reproductive-based theories, such as the chunking and template theories appear to explain more of the data on experts' skill than search-based theories. The overwhelming support for the view comes from evidence that experts' first moves (when no search has taken place) are of a consistent, 'good enough' quality. Klein, Wolf, Militello & Zsombok (1995) found that the protocols generated by 8 highly skilled players (with a rating range between 1700 to 2150) contained first verbalized moves that were good enough for play as rated by an independent grandmaster. Moreover in 'blitz' chess, in which each player is allotted five or ten minutes for an entire game, moves must be made quickly. Players spend about an average of 5 seconds per move in blitz chess compared to an average of 180 seconds per move in standard chess. Despite very little time for extensive search processes, performance in blitz chess shares 81% variance with players' ratings based on their performance in standard games of chess (Burns, 2004; Chabris & Hearst, 2003). Likewise, the performance of world chess champion Kasparov did not decrease when he played a simultaneous exhibition against a team of strong chess experts (the

Israeli national team with players of grandmaster strength) in which he had little time between choosing moves for different positions (Gobet & Simon, 1996a; 1996b; 1996c). Perhaps experts may search more deeply than novices only to solve problems with which they are unfamiliar or which are novel in some way.

But how informative is depth of search as a measure of the key difference between experts and novices? Consider players thinking about a move and thinking about the move their opponent may make in response. There may be many possible moves the opponent could make. One player may think immediately about the possible move the opponent could make that would *refute* the player's plan. The player's search is just two moves deep (their own and their opponent's move), but it reveals a crucial refutation of their plan. Another player may instead fail to think immediately about the refuting move of the opponent. Instead the player may think of some of the possible moves the opponent could play that will not refute the player's move, and may even lead to an advantage for the player. The player's search may be deep but confined to opponent moves that *confirm* their plan. A third player may think of just one or two moves the opponent could make, and each may be moves that would fit with the player's plan. The player's search is not deep and it is confined to confirming moves. A quantitative comparison of the depth of search based on the surface structure is not informative: the first player searched to just two moves deep (but found a potential refutation), the second player searched four or five moves deep (but found no refutation), and the third player searched to the same depth as the first player (but found no refutation). An important difference may be present in the deeper structure of the search. An expert may think about the strongest move an opponent can play - a refuting move, and so reduce the need to examine a large search space. An expert may be distinguished more by the selection of *what* to investigate than by the depth of calculation (DeGroot, 1965; Kotov, 1971).

### ***Hypothesis Testing in Chess Expertise***

The aim in this chapter is to examine how chess experts err, and in particular how they detect error and recover from it, a question never before addressed. There are hints in the literature that expert chess playing requires more than extensive and accurate memory retrieval, as experts do not automatically accept the first move they generate when they have time. Instead they appear to search through a problem space as revealed by their think-aloud protocols (DeGroot,

1965; Newell & Simon, 1972). Why do experts focus their efforts on search given that their first moves generated from their stored knowledge tend to be good enough (Gobet & Campitelli, 2002)? I propose that sometimes expert knowledge may not be perfect, or that the knowledge experts retrieve from long-term memory may not match the current position best. To avoid choosing moves that lead to error as a result of inaccurate retrieval, chess experts may evaluate the moves they consider, and their evaluation can usefully be conceptualized as a form of hypothesis testing. I suggest that expert chess players may tend to try to detect and avoid error by engaging in hypothesis testing, that is, they may attempt to falsify their move plan, especially when they have time to check the accuracy of their thinking. In other words expert knowledge may help people to falsify *their own* hypotheses. The account implies that the purpose of expert search is to detect where an opponent can refute a move planned for play, to identify errors. Choosing a move in chess may depend on accessing a large repository of domain knowledge about possible moves for play, including possible opponent moves. All players may try to think about their opponent's response to their move, but a key difference between expert and novice search may lie in the anticipation of how opponent moves can have negative, rather than positive, consequences for a planned move.

DeGroot (1965) suggested four stages to choosing a move in a game of chess, corresponding to a *progressive-deepening* strategy (De Groot, 1965). The four stages are orientation, exploration, elaboration, and proof. During the orientation phase the player adjusts to the position, weighs up the number and placement of pieces for the player and the opponent, and identifies strengths and weaknesses of the position. The function of the phase is the formation of a specific problem conception that may include a selective set of possible board plans that correspond to a set of considered moves each aimed at attaining a specific board goal. There may be a preference or 'favourite' within this set of moves. These plans, considered moves and anticipated solution are partly tentative and hypothetical. The player's conception of the board problem is a 'working hypothesis' (De Groot, 1965, p. 395). During the second phase of exploration, the player begins to explore possibilities in the problem space. During the third phase of elaboration, the player continues the exploration in more detail and may repeat some investigations. Finally, during the proof phase, the investigations are evaluated.

This thesis conceptualises each move a chess player considers to be a

hypothesis, and the opponent moves they consider in response to each move as tests of the hypothesis. These tests may be either confirming (the opponent's move fits in with the player's plan, that is, it leads to an improvement in the player's position), or they may be falsifying (the opponent's move refutes the player's plan, that is, it leads to a worsening of the player's position). Let us now turn to examine experimentally expert and novice chess players' evaluations of the moves they considered in response to normal and random board positions. Evaluation is conceptualised as a hypothesis testing process.

### **Evaluation of Chess Moves by Experts and Novices**

So, the experiment was designed to test whether there were differences in the types of evaluations generated by experts and novices when they thought about what move they would like to play in a given position. I will now outline the aims of the experiment and its method, but let us turn to the summary of the results from the previously reported retrospective evaluations measured in the experiment in the first instance (i.e., Cowley, 2002). The report of the previously unanalyzed protocol data collected in the experiment will follow.

The experiment examines how hypothesis testing could be conceptualized as the evaluation of moves considered for play by chess players (Cowley & Byrne, 2004, 2016 for details). Ten experts and ten novices (including the full population of chess masters living in Ireland at the time) took part. The aim of the experiment was to measure a) retrospective evaluations of chosen moves, and b) think-aloud protocols when choosing moves. The experiment and the results for the retrospective evaluations are reported in Cowley (2002) as a BA research project, and while I cannot report those data on retrospective evaluations in full here, I will outline the main results and methodology. The results showed that expert chess players (including master and non-master level experts) chose a higher quality move for play than novices at the end of their thinking time in the normal positions. While experts tended to retrospectively evaluate a greater proportion of the move sequences they thought about as leading to a negative outcome, novices tended to retrospectively evaluate a greater proportion of the move sequences they thought about as leading to a positive outcome. This result suggests that experts chose higher quality moves for play because they tended to anticipate the opponent moves that led to a negative outcome for their board position. Novices may have chosen lower quality moves for play because they tended to anticipate how their moves could

lead to a positive outcome even when they objectively led to a negative outcome. Experts thought about how their plans could be falsified, but novices thought only about how their plans could be confirmed.

The focus in this chapter is to examine this result in greater detail by analysing the previously unanalysed and unreported protocol analyses. First I summarise the methodology of the retrospective evaluations reported in Cowley (2002) in some detail given that the protocol analysis is based on a subset of the participants who carried out the retrospective evaluations, and the materials and design were the same. I will summarise its methodology, and describe how the result led to the protocol analysis reported in this chapter.

### ***Retrospective evaluation***

In the earlier retrospective evaluations experiment reported in Cowley (2002), there were 20 participants. The 20 participants (19 men and 1 woman) were registered members of the Irish Chess Union. They were experienced novices (mean rating of 1509) and experts (mean rating 2240). The expert group included experts from different Elo categories of expertise, including one grandmaster (Irish Elo >2500) two international masters (Irish Elo > 2300), three Fide masters (Irish Elo > 2200, i.e. International Chess Federation masters), and four initial category experts (Irish Elo > 2000).

The materials were six board positions, three normal and three random (as well as an initial practice position). The board positions were chosen from games in chess periodicals. They were middle game positions with 22-26 pieces to ensure complexity and to rule out the chances that the masters had seen them before. The positions were chosen with the assistance of a chess expert (who was not a participant in the study). Importantly, they were 'equality outcome' positions, where there were equal chances with best play for both black and white pieces (see Appendix F). This constraint ensured that there would be no obvious confirming or falsifying move sequences. These properties were selected to ensure complexity and to rule out the likelihood that the masters would have seen them before. Although they were equality outcome positions they were not 'quiet' positions such as those where very few possible moves are available to the players and they inevitably lead to a draw. Instead, they were selected to be in the style of De Groot's (1965) position A where the position is 'lively' with many possible alternative moves to choose from. Any colour preference of player's to play with white or black was controlled by balancing



the number of times they played the white or black side of the positions.

Random board positions also contained twenty-two to twenty-six pieces but the pieces had been scattered. They were randomized until positions were achieved to satisfy the constraints that neither king was in check, no pawn occupied the first or eighth rank, and they were equality outcome. They were also designed and checked by an independent chess expert and the chess program *Fritz*. These random board positions with equality outcomes were designed specifically for this experiment. The first two constraints make it possible to play and choose a move in randomized positions (Holding & Reynolds, 1982). See Figure 4.1 for an example of a chess position used (position 1).

The participants' task was to, "choose a move you would play in the way you are used to going about choosing a move in a real game". They were given instructions to think-aloud, and their verbalizations were recorded by dictaphone (See Appendix J for the experimenter's think-aloud script). Moves examined by the player during think-aloud are verbalized using algebraic chess notation. These moves were recorded by the author who is fluent in algebraic notation (See Appendix K for the experimenter's recording sheet).

Three minutes thinking time was allotted for choosing a move as it is just over the average time per move in tournament play. Exposure for each board position was timed using a standard tournament chess clock, each clock was set at three minutes and when the clock's flag fell participants were told that their time was up. To accurately access hypothesis testing we also needed participants to provide us with an evaluation of each move sequence that they examined. However, spontaneous evaluation in chess has a low probability of verbalization (Newell & Simon, 1972). Accordingly I used a combined methodology of think-aloud followed by retrospective evaluation (See Appendix J for the experimenter's retrospective evaluation script). Verbalized move sequences were recorded not only by dictaphone but also by the experimenter (the author) in algebraic notation concurrent with think-aloud. The experimenter asked the participants for their evaluation of each move sequence, by first saying back the move sequence immediately after each chess problem to reduce retrospective error and interference (Ericsson & Simon, 1993). The participants were then asked to evaluate each move sequence as having led to a positive, negative or neutral outcome for their positions. When players evaluated a move sequence as leading to a positive outcome it was scored as confirmation. When players

evaluated a move sequence as leading to a negative outcome it was scored as falsification, and when players evaluated a move sequence as leading to a neutral outcome it was scored as neutral.

The findings showed that novice chess players evaluated their move sequences as leading to a positive outcome more readily than experts, even though they chose lower quality moves to play than the experts at the end of their thinking time. Experts evaluated their move sequences as leading to a negative outcome more readily than novices, and they chose higher quality moves to play than the novices. The results suggest that experts could falsify their hypothesized moves by evaluating them as leading to a negative outcome, and then they chose higher quality moves than the novices. Experts can falsify their own hypotheses to abandon low-quality hypotheses (Popper, 1959).

However, the players' retrospective evaluations are subjective, for example, a player may evaluate a move sequence as leading to a positive outcome but objectively it leads to a negative outcome. The player may have anticipated only how opponent counter-moves to their chosen move lead to a more positive outcome for them, and they may not have anticipated how opponent counter-moves would lead to a negative outcome. Alternatively the player may have anticipated opponent counter-moves that would lead to a negative outcome, but they may have misinterpreted them as leading to a positive outcome.

In the protocol analysis reported in this chapter I aimed to discover whether players' subjective evaluations corresponded to objective evaluations. One of the most powerful chess programs was used to achieve objective evaluations of the verbalized chess moves (*Fritz 8*). First, the players' think-aloud protocols were transcribed, and then the problem behaviour graphs were created based on the transcriptions. Each possible move mentioned by each chess player was mapped out to create a problem behaviour graph, and the properties of the problem behaviour graphs were analysed to discern the role of mental representation in chess players' hypothesis testing. I now turn to the experiment.

### **Experiment 6: A protocol analysis of hypothesis testing by masters and novices**

The experimental analysis of the think-aloud protocols reported here is in full and contains previously unreported data (see Cowley, 2002, Cowley & Byrne, 2004). I selected a subset of the protocols including the protocols collected in the retrospective evaluation experiment summarised in the previous section. I

explored a new and more complete experimental and theoretical analysis of chess players' hypothesis testing. It is instructive to focus on the master level players (for comparison with masters studied in the chess literature previously) and to this end I selected the think-aloud protocols of five master level players (i.e. 1 Grandmaster, 2 International Masters, and 2 Fide International Chess Federation Masters) who took part in the earlier experiment, and compared them to the think-aloud protocols of five novice chess players, chosen at random from the full sample of novices. First I describe the aims of the protocol analysis and its method. Then I provide the first report of these data.

### ***Objective evaluation of chess moves: Fritz 8***

*Fritz 8* is a chess program. It plays chess to the standard of the world's current top five grandmaster level players, and recently drew in a match with the world chess champion Vladimir Kramnik (Bahrain, 2003). In analysis mode, *Fritz 8* can produce an evaluation of a position after a chess player plays a move. The evaluation tells a player whether a move will objectively lead to a positive outcome, a negative outcome, or a neutral outcome regardless of what the player may subjectively wish the outcome to be (see Appendix G for the setup functions in the use of the program *Fritz 8*). In fact as Table 4.1 illustrates, there are nine types of hypothesis test in chess based on whether a subjective evaluation of a move leads to a positive, negative or neutral outcome; and whether the evaluation of the move objectively leads to a positive, negative or neutral outcome.

As Table 4.1 shows there are two types of *confirmation* in chess, (i) objective confirmation in which a player evaluates a move sequence as leading to a positive outcome and objectively it leads to a positive outcome ('+/+' in Table 4.1), and (ii) confirmation bias in which a player evaluates a move sequence as leading to a positive outcome and objectively it leads to a negative outcome ('+/-'). There are also two types of *falsification* in chess, (i) objective falsification in which a player evaluates a move sequence as leading to a negative outcome and objectively it leads to a negative outcome. In this case the player has anticipated the opponent counter-moves to his chosen move that lead to a negative outcome ('-/-' in Table 4.1), and (ii) falsification bias, in which a player evaluates a move sequence as leading to a negative outcome but it objectively leads to a positive outcome for them ('-/+').

Table 4.1: The nine possible hypothesis types based on the subjective and objective evaluations of move sequences

<i>Retrospective evaluation by chess player</i>	<i>Objective evaluation by Fritz</i>		
	Positive (+)	Negative (-)	Neutral (=)
Positive (+)	+/+	+/-	+/=
Negative (-)	-/+	-/-	-/=
Neutral (=)	=/+	=/-	=/=

Key: '+' refers to a positive evaluation, '-' to a negative one, '+/-' means the player's evaluation was positive and the program's evaluation was negative.

They may not have anticipated how opponent counter-moves would lead to a more positive outcome for them. Finally there are two types of *neutral evaluation* in chess, (i) objective neutral evaluation in which a player evaluates a move sequence as leading to neither a positive nor a negative outcome and objectively it leads to this outcome ('='); (ii) neutral bias in which a player evaluates a move sequence as leading to neither a positive nor a negative outcome. However, the player may not have anticipated how opponent counter-moves would lead to a more positive outcome or a more negative outcome ('=/' or '=/-'). Players may also evaluate a move sequence as positive when it is in fact neutral, a positive bias similar to confirmation bias ('+/' or '=+') or they may evaluate it as negative when it is in fact neutral, a negative bias similar to falsification bias ('-/' or '='). Neither of these evaluations leads to error in chess however and so I do not include them in the initial analysis of confirmation and falsification biases, although I do include them in a second broader analysis of positive and negative testing.

### ***Protocol analysis in chess and the creation of problem behaviour graphs***

The aim in the experimental analysis I report here is to map each possible move mentioned by each chess player so that their move sequences can be inputted into the chess program *Fritz 8*. I wished to compare the *Fritz 8* objective evaluation with the subjective evaluation generated by the player to calculate the

amount of objective falsification (-/-), and the amount of biased confirmation (+/-) etc. To derive the objective measures from *Fritz 8* requires a detailed protocol analysis of think-aloud scripts. Protocol analyses of think-aloud scripts are time-consuming. Studies of chess have tended to rely instead on accuracy and time to recall board positions, or quality of final move choice (e.g., Holding and Reynolds, 1982). Studies that have used protocol analysis in problem solving tend to rely on very small sample sizes of just a few participants (e.g., Newell and Simon, 1972), although one chess study included a protocol analysis of five of the world's top grandmasters at that time (De Groot, 1965). I wished to carry out a protocol analysis on a reasonably large sample of players and so I selected a subset of the think-aloud scripts provided by ten of the players in the experiment reported in Cowley (2002), which reports the retrospective evaluations but not the protocol analysis.

First, our aim was to transcribe the taped think-aloud protocols for ten players, for six positions each (three of the positions were normal chess board positions, and three were random board positions). The transcribed think-aloud protocols were *segmented* into episodes, move by move (see Appendix H for all of the segmented protocols). Fifty-seven (3 protocols were inaudible including one FM normal board position and two GM random board positions) *problem behaviour graphs* were constructed using Newell and Simon's guidelines for the normal chess board positions (see Appendix I for a subset of the problem behaviour graphs generated by the experimenter). To illustrate I present a small section of a master's problem behaviour graph in Figure 4.2, and I explain each corresponding move in the text that follows.

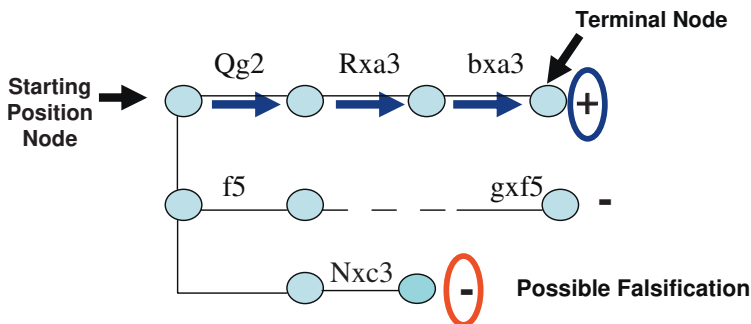


Figure 4.2: A section of a problem behaviour graph constructed from the chess master's protocol.

Each line across represents a move sequence. The order of search is from left to right, then down. Each circle (i.e. node) represents a new position following a move made in the problem space. For example, a player says Qg2 and it means the player is describing the possibility of moving his queen to the g2 square. Then the player says Rxa3 which refers to the player describing this move as a possible reply from the opponent to his Qg2 move (The opponents rook moves to the a3 square, and the x refers to the fact that the rook captures a piece, in this case a pawn). The next move bxa3 means the player says his pawn on the b-file can move to a3 to capture the opponent's rook piece. The plus sign shows that the player evaluated this move sequence as positive for the player, in their retrospective evaluation. The minus shows a negative evaluation. The dashed line indicates that moves have not been specified for either the self or opponent. I label these *skipped* moves. For example, in the second move sequence, the participant first says f5 as a move for the player which means a pawn moves to the f5 square (note that pawn moves do not have the p in front of them). Then the player says gxf5 which means the pawn on the g file of the board captures a pawn on the square f5. This second move is only possible for the player and not for the opponent. Hence, the player has generated a move sequence describing only his own moves and has omitted any description of opponent moves. As the problem behaviour graph shows, I incorporated the retrospective evaluation (positive, negative, or neutral) with the think-aloud move sequences. A subset of the problem behaviour graphs created for masters and novices think-aloud protocols of normal and random positions are presented in Appendix H.

To secure objective evaluations, *Fritz 8* was used to evaluate the chess position occurring at the final move (i.e. terminal node) of each sequence (each line). Use of *Fritz 8* was essential to enable us to identify move sequences that would genuinely be positive for a player, and to discriminate them from move sequences that a player identified as positive for them but which could ultimately end negatively for them if played. Thus confirmation bias as a move sequence that was evaluated as leading to a positive outcome when it in fact leads to a negative outcome could be identified. Likewise, I was able to ensure that move sequences evaluated by a participant as negative were objectively instances of falsification. In other words, we can now discriminate clearly between the nine types of hypothesis test presented in Table 4.1.

The problem behaviour graphs allowed us to measure several properties of

chess masters' hypothesis testing. The first property of interest is the sorts of hypothesis tests that chess players generate. The aim was to identify the confirming and falsifying hypothesis tests. A second property of interest is the quality of the first verbalised move. A third property of interest is the number of individual moves. A fourth property of interest is the search depth and the fifth property of interest is the search breadth (I define these measures shortly). The sixth property of interest is the different sorts of articulated move sequence: complete, skipped moves, base skip move, and ambiguous. The final property of interest is the point at which skipped moves occur.

For normal board positions I predicted that chess masters would use falsification to detect moves they examined for play to which an opponent could reply with a refutation (the -/- cell). The prediction is that domain knowledge facilitates hypothesis falsification, and expert hypothesis testing proceeds by seeking falsifying instances of a hypothesis (see DeGroot, 1965; Gobet, 1998; Gobet *et al.*, 2004). Access to expert knowledge may help chess masters to consider a larger number of possible alternative moves for both themselves and opponents than novices (see Johnson-Laird & Byrne, 1991). Further, chess masters may be better at representing their opponent's moves explicitly than novices; they are better than novices at detecting an opponent move that falsifies a hypothesized plan. Cognitive expertise may help chess masters to consider more possible alternative moves because they can represent their hypothesis testing more easily due to many years of practice (e.g., Ericsson & Kintsch, 1995). Novices were predicted to exhibit more evidence of confirmation bias suggestive of how people have been found to test their hypotheses in the standard hypothesis testing literature (e.g., Wason, 1960; Tweney *et al.*, 1980; Mynatt, Doherty, & Tweney, 1978; 1979; Mynatt, Doherty, & Dragan, 1993; Poletiek, 1996). Novices were predicted to be less able to detect moves to which an opponent could reply with a refutation, and only see how these moves could lead to good outcomes. For random board positions, where the pieces have been scrambled about the board coordinates in order not to match expert knowledge, masters were predicted to be no longer be able to falsify. They should perform similarly to novices because their expert knowledge would no longer be able to help them to falsify their own hypotheses.

## **Method**

### ***Participants***

Protocols from 10 of the 20 chess players who took part in the experiment reported in Cowley (2002) were chosen for analysis. I selected 5 master level players: 1 Grandmaster, 2 International Masters, and 2 Fide Masters. The selected experts conformed to the following criteria: they had an international master title of Fide Master, International Master, or Grandmaster; they had represented Ireland at international level at a Chess Olympiad team event (Chess Olympiads are the equivalent of Olympic games for chess, and are independently run by the world governing chess federation, *FIDE*); they had an ICU rating above 2250, and they were active players at the time of the study. I compared them to 5 novice chess players chosen at random from the sample of ten novices reported in Cowley (2002). The minimum and maximum ratings for this sample of novices were 1265 and 1511 respectively ( $M = 1413$ ).

### ***Materials, design and procedure***

The between participants factor was the level of expertise (master or novice), and the within participants factor was the type of board position (normal or random) and the design was a 2x2 mixed design. Participants were given six board positions, three normal and three random positions as described earlier (see Appendix F for the six position diagrams).

Each participant received three normal board positions and three random board positions. The board positions were presented in randomized order. Each participant was also presented with a practice position to help them become familiar with thinking aloud before they moved on to the six experimental board positions. The procedure and the description of data collection is outlined on pages 148-150.

### ***Protocol analysis procedure***

The previously unanalysed think-aloud protocols were first transcribed for each board position considered by the five novices and masters ( $n = 57$ , 3 protocols were inaudible including 2 random board positions from the grandmaster, and 1 random board position from a Fide Master). Then the protocols were broken down into segments according to guidelines indicating that each move, comment about the nature of the position, and implicit signifiers of evaluation should be put into a different segment (Newell & Simon, 1972). An example of



a segmented protocol is given in Table 4.2 below:

Table 4.2: An example of a segmented expert protocol. This protocol corresponds to the fourth master problem-behaviour graph in Appendix F.

---

Grandmaster (participant 4), normal position 1, black

1. So, I'm black in this position.
2. It's some kind of Alekhine Defense,
3. Caro-Kann something in that line,
4. black to play
5. and white has immediate threat f5
6. at the same time all of my pieces seem to be ok,
7. except maybe for the Bg6.
8. So I now need to find a way to stop f5,
9. I have semi-open file h,
10. which could be leading somewhere,
11. em I can stop f5 by playing Ne7
12. but that could lead to Bb4
13. threatening to,
14. well obviously wanting to get the knight
15. in or planting the bishop on d6,
16. probably need to play c5,
17. which I can do if I have to if...
18. or next... (*time up*)

---

These segmented protocols were then worked through move by move and segment by segment alongside a chess board to map each move verbalized by a player for a board position. The number of segments for normal board positions ranged between 32 to 74 segments for masters and 27 to 48 segments for novices. The length for random board positions ranged between 30 to 64 segments for masters and 24 to 54 for novices. Problem behaviour graphs were constructed from each protocol indicating the order of moves and move sequences considered with their retrospective evaluations (see Appendix I for a subset of problem behaviour graphs generated).

## Scoring

The data were scored in relation to the primary measure, the sorts of hypothesis tests, falsification or confirmation, that players carried out. It was also scored in relation to six search behaviour measures including, (i) first move quality, (ii) number of individual moves, (iii) search depth, (iv) search breadth, (v) and the different sorts of articulated move sequence: complete, skipped moves, base skip move, and ambiguous; (vi) point at which skipped moves occur.

To illustrate the scoring, I present two problem behaviour graphs, one constructed for a grandmaster and one for a novice for position 1, when it is black to play in Figure 4.3 and Figure 4.4. A visual comparison of the grandmaster graph to the novice graph, illustrates their differences. The most obvious difference is the size. The master's graph has more moves than the novice's. I will explore these differences in more detail shortly.

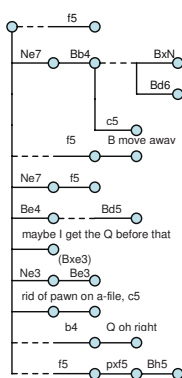


Figure 4.3: Grandmaster (participant 4), normal position 1, black

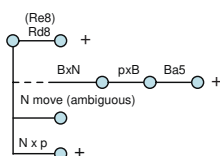


Figure 4.4 Novice (participant 12), normal position 1, black

First, I will present a step-by-step interpretation for the grandmaster's graph and outline what I mean by different types of moves sequences using examples from the graph (e.g., what I mean by a complete, skipped, baseskip or ambiguous move sequences). Search is from left to right and then down. The circles represent nodes and each node indicates where the representation of the position has changed due to the imagination of a move that has taken place. The top-left-hand node represents the starting position. This problem behavior graph corresponds to the think-aloud protocol I presented in Table 4.2. The graph shows that the Grandmaster's first move for board position 1 (where he had the black pieces and it was black to move) is indicated as 'f5' in the graph. The 'f5' move refers to a pawn being imagined to move to the f5 square. The full verbalization was 'and white has immediate threat f5' (segment 5 in Table 4.2) and the pawn move is for the opponent (white). It is clear that it is an opponent move because it is a move that is only possible for white. The dashes preceding f5 on the graph indicate that a move has been skipped, that is, the Grandmaster skipped his own turn to move in his verbalization. In the graphs a dash (---) represents a move for the opponent or for the self was not articulated (a 'skipped' move). The next verbalized move is segment 11, 'em, I can stop f5 by playing Ne7'. Ne7 refers to the black knight moving to the e7 square. Then the move Bb4 is verbalized for white to follow this move: 'but that could lead to Bb4'. This verbalization is followed by skipping a move for the self and instead verbalizing two possible moves for white after Bb4. Notice the dashes on the problem behavior graph to indicate the skip. Both of the possibilities considered are connected to what the grandmaster thinks white's plans are following his Bb4 move. First, 'well obviously wanting to get the knight', which refers to the white bishop that is now poised to capture his black knight on e7. This is indicated as BxN on the graph, where the 'x' is the symbol used to indicate 'capture'. Second, 'or planting the bishop on d6' refers to the other possibility for white of playing his bishop to the d6 square (Bd6).

In the graphs ambiguous move sequences are recorded using the player's verbalization, for example, 'R moves' implies the Rook moves, but we do not know where it moves to so it is labeled 'ambiguous' (for an example of an ambiguous move sequence, see master's problem behavior graph example 3, position 2, normal board position, fourth move sequence down in Appendix F). Repeated move sequences are annotated with 'repeat'. The end nodes or terminal nodes are annotated with a plus, minus or equals sign. A plus sign (+)

indicates that the player provided the retrospective evaluation that the position led to a positive outcome; a minus sign (-) indicates that a player evaluated it as leading to a negative outcome, and an equals sign (=) indicates that a player evaluated it as leading to a neutral outcome.

## **Results and discussion**

The results for the following are reported: the nine types of hypothesis test; the quality of the first verbalised move; the total number of moves; the search depth; the search breadth; the types of move sequences; and skipped moves along a move sequence.

### ***Hypothesis tests in complete move sequences***

Only completely articulated move sequences (sequences in which the players articulated every imagined move of their own and their opponent) can be reliably inputted to the computer program *Fritz* for objective evaluation. 50% of move sequences were articulated completely by masters and novices. The mean number of different types of hypothesis tests are presented in Table 4.3.

The hypothesis test type '-/-' indicates that the player and *Fritz* both evaluated a move sequence as leading to a negative outcome as Table 4.1 showed. It is the strongest falsification type of the nine tests. Masters falsified more often than novices for normal positions (1.07 vs 0.40, Mann-Whitney  $U_{5,5} = 4.00$ ,  $Z = -1.844$ ,  $p = .0325$ ), and not for random positions (0.83 vs 0.87, Mann-Whitney  $U_{5,5} = 9.00$ ,  $Z = -.752$ ,  $p = .226$ ). The result corroborates our suggestion that masters try to falsify their hypotheses. It also suggests that when domain knowledge is eliminated masters do not falsify more than novices.

The '+/-' test type corresponds to the strongest type of confirmation bias as Table 4.1 showed. The player evaluates a move sequence as leading to a positive outcome and *Fritz* evaluates it objectively as leading to a negative outcome. There were no differences in frequency of confirmation bias in masters and novices for the normal positions (0.87 vs 0.53, Mann-Whitney  $U_{5,5} = 8.50$ ,  $Z = -.854$ ,  $p = .197$ ), or the random positions (0.33 vs 0.42, Mann-Whitney  $U_{5,5} = 7.5$ ,  $Z = -1.195$ ,  $p = .116$ ).

Table 4.3: The mean number of different types of hypothesis tests for the complete move sequences for the normal and random board positions by the masters and novices in Experiment 6.

<i>Nine test types</i>	<b>Normal</b>		<b>Random</b>	
	Master	Novice	Master	Novice
+/+	0.60	0.27	0.67	0.27
<b>-/- (falsification)</b>	<b>1.07</b>	<b>0.40</b>	<b>0.83</b>	<b>0.87</b>
=/=	0.53	0.13	0.08	0.00
<b>+/- (confirmation bias)</b>	<b>0.53</b>	<b>0.87</b>	<b>0.42</b>	<b>0.33</b>
+/=	0.47	0.60	0.17	0.20
=-/	0.13	0.20	0.17	0.27
-/+	0.13	0.07	0.42	0.13
-/=	0.33	0.13	0.17	0.07
=/+	0.13	0.20	0.08	0.20
Total	3.92	2.87	3.01	2.34

*\*Note: The falsification and confirmation bias hypothesis tests are represented in bold.*

Falsification and confirmation bias were the two most frequently occurring of the nine test types, as Table 4.3 shows, and they account for almost half of the total number of hypothesis tests in the complete move sequences. I also collapsed the nine test types into three broad test types: objective tests, positive bias tests and negative bias tests as Table 4.4 shows).

Objective tests are tests for which there was a match between the players subjective evaluation and *Fritz's* objective evaluation (the '+/+', '-/-' and '=/=' tests). The category includes the falsification test. Masters carried out more objective tests than novices for the normal positions (6.0 vs 2.4, Mann-Whitney  $U_{5,5} = 5.00$ ,  $Z = -1.596$ ,  $p = .055$ ), but not for random positions (3.8 vs 3.4, Mann-Whitney  $U_{5,5} = 10.5$ ,  $Z = -.426$ ,  $p = .335$ ). The result again shows that masters were better at objectively evaluating their move sequences than novices, but only when they could rely on their domain knowledge.

Table 4.4. The mean number of positive, negative and objective tests by the masters and novices for normal and random board positions (with standard deviations in parentheses) in Experiment 6.

	<i>Masters</i>	<i>Novices</i>
<i>Normal board positions</i>		
Objective	6.0 (3.00)	2.4 (1.82)
Positive	3.2 (1.10)	5.0 (2.74)
Negative	1.8 (1.64)	1.2 (0.84)
<i>Random board positions</i>		
Objective	3.8 (1.79)	3.4 (1.52)
Positive	1.8 (1.64)	2.4 (0.89)
Negative	1.8 (1.92)	1.4 (1.14)

*Note: Objective tests include the falsification tests, and positive tests include the confirmation bias tests.*

Positive bias tests were tests for which the players were more positive than Fritz (the '+/-', '+/=', and '=/' tests). The category includes the confirmation bias test. There were no differences in frequency of positive bias tests between experts and novices for the normal positions (5.0 vs 3.2, Mann-Whitney  $U_{5,5} = 7.5$ ,  $Z = -1.085$ ,  $p = .139$ ), or random positions (1.8 vs 2.4, Mann-Whitney  $U_{5,5} = 9.0$ ,  $Z = -.759$ ,  $p = .224$ ).

Negative bias tests were tests for which the players were more negative than Fritz (the '-/+', '-/=', and '=/' tests). Masters and novices carried out the same amount of negative tests for the normal positions (1.8 vs 1.2, Mann-Whitney  $U_{5,5} = 10.5$ ,  $Z = -.434$ ,  $p = .332$ ), and for random positions (1.8 vs 1.4, Mann-Whitney  $U_{5,5} = 12.0$ ,  $Z = -.108$ ,  $p = .457$ ).

Overall, the primary difference in hypothesis testing between masters and novices is that masters were better at objectively evaluating their moves than novices, and in particular they were better at accurately falsifying their hypotheses than novices, when they had domain knowledge.

### ***First move quality***

Masters and novices differed somewhat in the quality of the first moves they verbalized as a possible move for them to play, as evaluated by *Fritz 8*. The *Fritz* evaluations correspond to how much of a pawn may be won or lost if a hypothesized move is played, for example  $-.99$  indicates the player will lose almost a full pawn,  $.99$  indicates that the player will win almost a full pawn. In chess terms a pawn is a considerable loss, particularly if one is losing a pawn's worth every other move. Masters' first moves tended to be of a somewhat better quality than novices for normal positions, as judged by *Fritz* ( $-.070$  vs  $-.440$ ) but the difference was not significant (Mann-Whitney  $U_{5,5} = 6.5$ ,  $z = -1.257$ ,  $p = .104$ ). Masters' first moves tended to be of a similar quality to novices for random positions ( $-6.47$  vs  $-5.05$ , Mann-Whitney  $U_{5,5} = 11$ ,  $z = -.313$ ,  $p = .754$ , two-tailed). Earlier findings (De Groot, 1965; Gobet & Campitelli, 2002) suggest that the first move is generated by masters directly from their long-term knowledge, for example by accessing a template (Gobet, 1998; Gobet *et al.*, 2004). In this experiment the masters did not generate significantly better first moves than the novices in the normal positions. This suggests that hypothesis testing may often help masters think about the position further in order to generate better moves for play. Further searching may help them to falsify their first move, if it is not a good move. Hypothesis falsification may help masters to avoid playing moves that would lead to error, even though they generated the moves.

### ***Number of moves***

In the protocol analysis I was able to calculate the average number of *individual* moves per position. Each move sequence contains different numbers of individual moves, for example one move sequence may contain one move whereas another may contain eight individual moves. Masters generated more individual moves than novices in normal positions ( $25.40$  vs  $16.80$ ) although the difference is marginal (Mann-Whitney  $U_{5,5} = 4.0$ ,  $Z = -1.776$ ,  $p = .076$ , two-tailed), and they did so in random positions too ( $31.56$  vs  $18.27$ , Mann-Whitney  $U_{5,5} = 2.0$ ,  $Z = -2.193$ ,  $p = .028$ , two-tailed) as Table 4.5 shows. The result suggests that masters can generate a larger search space than novices.

Table 4.5. The mean ply depth and ply breadth of move sequences, and the mean number of individual moves in the generated move sequences in Experiment 6 (standard deviations in parentheses).

	<i>Masters</i>	<i>Novices</i>
<i>Normal board positions</i>		
Individual moves	25.40 (6.757)	16.80 (9.001)
Ply depth	3.30 (0.961)	2.61 (0.621)
Ply breadth	2.20 (1.523)	1.60 (1.230)
<i>Random board positions</i>		
Individual moves	31.57 (4.806)	18.20 (7.596)
Ply depth	4.13 (1.181)	2.94 (0.520)
Ply breadth	1.93 (0.864)	1.00 (0.848)

### ***Search depth***

Search depth refers to the number of moves a player thinks about in advance when choosing a move for play. These moves combined together form move sequences. The first move in the sequence is one move deep. The second move is two moves deep. Search depth is measured by the number of moves that indicates the length of a move sequence. Each move in the sequence is given a measure of one ply. For example the search depth of a move sequence that is three moves long is termed a ‘3 ply’ move sequence. Consider problem behavior graph example 6 in Appendix I. In the third move sequence down the novice first considers playing his pawn to the f4 square (f4; 1 ply). Then he thinks about his opponent responding to this move by capturing this pawn on f4 with a pawn placed on an e-square (exf4; 2 ply). Now he thinks about another move for his opponent immediately (indicating a skipped move for himself; 3ply), and he thinks about the opponent playing the queen to the e2 square capturing one of his pawns and lining up for an attack on his king putting him in check (Qxe2+; 4 ply). This move sequence is four moves deep and is called a 4 ply move sequence. Notice that skipped moves are given one ply also in order to represent the search depth as accurately as possible. I calculated search depth for all move sequences, whether they were complete move sequences (every move was



articulated precisely along the sequence), skip move sequences (at least one move for the self or opponent was not articulated along the sequence), base skip move sequences (the first move of the sequence was not articulated) or ambiguous move sequences (the coordinates of at least one move were not articulated).

Masters and novices searched to the same depth in normal positions (3.3 vs 2.61, Mann-Whitney  $U_{5,5} = 7.0$ ,  $Z = -1.149$ ,  $p = .251$ , two-tailed), although masters searched to a somewhat greater depth than novices in random positions (4.14 vs 2.94, Mann-Whitney  $U_{5,5} = 4.0$ ,  $Z = -1.776$ ,  $p = .076$ , two-tailed), as Table 4.5 shows. Previous studies have not found a difference in search depth. For example, DeGroot (1965) did not find that masters searched more than lesser experts (in a study of five grandmasters, five experts, and five category A players), and the result has been replicated (e.g., De Groot & Gobet, 1996; Gobet, 1998; Gobet *et al.*, 2004). However, a recent study has shown that masters ( $n = 2$ ) search deeper in board positions that are highly complex, that is, positions that can occur in real games and require at least 23 look-ahead moves to solve (Gobet & Campitelli, 2002). The current study provides a detailed comparison of search between normal and randomized chess positions comparing a large sample ( $n = 10$ ) of master and novice players, with a large skill difference between the two groups. It replicates previous findings of no search depth difference for normal positions, but suggests that masters may be able to search deeper in the random positions.

### ***Search breadth***

Search breadth refers to the situation in which more than one possible move has been considered at the same place in a move sequence for either the opponent or for the self. In other words, a move sequence ‘broadens out’ where two or more distinct alternative move sequences follow on from the same first moves. Consider problem behaviour graph example 2 in Appendix I. In the seventh move sequence down this master first considers playing his pawn to the e3 square (e3). Then he thinks about his opponent responding to this move by playing a knight to e6 (Ne6). Now he thinks about playing two alternative moves in response to his opponent’s knight move: queen to the c2 square (Qc2) or queen to the d5 square (Qd5). Both of these queen moves are the third move of the sequence. On the graph this broadening (or branching) is represented by inserting the second alternative move directly underneath the first resembling a

fork shape. The number of prongs of the 'fork' represent the number of moves imagined possible at the same stage in a sequence. I calculated search breadth for all move sequences.

Masters searched as broadly as novices for normal positions (2.2 vs 1.6, Mann-Whitney  $U_{5,5} = 9.0$ ,  $Z = -.760$ ,  $p = .459$ , two-tailed) but they searched somewhat more broadly than novices for random positions (1.93 vs 1.00, Mann-Whitney  $U_{5,5} = 4.5$ ,  $Z = -1.735$ ,  $p = .083$ , two-tailed) as Table 4.5 shows. The result suggests that when masters are confronted with a novel problem in their domain (random chess positions) they may be able to rely on superior search driven processes.

### *Types of move sequences*

As reported above, 50% of move sequences were completely articulated sequences. I identified three other types of move sequence from the problem behavior graphs. As Table 4.6 shows, the four types of sequences are complete, skipped, base skip, and ambiguous sequences and I describe each in turn. Masters and novices produced the same amount as each other of each type of move sequence for the normal and random positions, as Table 4.6 shows.

Half of all move sequences were *complete move sequences* where every move for the player and his or her opponent was articulated. For example, the problem behavior graph in Appendix I, example 2 (International Master, position 1 with white to move) provides a good illustration of a complete sequence in the seventh move sequence down. The master thinks about moving his pawn to the e3 square (e3), then he thinks about his opponent playing his knight to the e6 square (Ne6), and then he thinks about responding to this opponent move by playing his queen to the c2 square (Qc2).

Then the sequence ends and is evaluated (negatively). There are no moves left out of the sequence for either the player or his opponent. Masters and novices generated the same number of complete move sequences for normal positions (4.00 vs 3.14, Mann-Whitney  $U_{5,5} = 7.0$ ,  $Z = -1.16$ ,  $p = .246$ , two-tailed) and for random positions (2.90 vs 2.67, Mann-Whitney  $U_{5,5} = 12.0$ ,  $Z = -.105$ ,  $p = .917$ , two-tailed). The results indicate that masters and novices were equally able to articulate their own and their opponent's moves.

Table 4.6: The mean number of different sorts of move sequences generated by the five masters and five novices in Experiment 6 for the normal and random board positions (standard deviations are in parenthesis).

<i>Move sequence</i>	<i>Complete</i>	<i>Incomplete</i>	<i>Base skip</i>	<i>Ambiguous</i>
<i>Normal positions</i>				
Master	4.00 (1.13)	2.20 (1.02)	1.67 (0.53)	0.33 (0.34)
Novice	3.13 (0.96)	1.53 (1.41)	1.14 (0.51)	0.60 (0.69)
<i>Total</i>	<i>3.57 (1.09)</i>	<i>1.87 (1.21)</i>	<i>1.40 (0.56)</i>	<i>0.47 (0.53)</i>
<i>Random positions</i>				
Master	2.90 (1.84)	2.76 (0.70)	1.30 (0.77)	1.37 (0.71)
Novice	2.67 (1.25)	1.60 (0.72)	1.20 (0.81)	1.53 (0.51)
<i>Total</i>	<i>2.79 (1.49)</i>	<i>2.18 (0.91)</i>	<i>1.25 (0.75)</i>	<i>1.45(0.59)</i>

*\*Note. Means are based on 15 cases (five players by three board positions) except for the masters' random positions which were based on 12 cases because of a recording error.*

The next most common sort (25%) were *skipped move sequences* where an essential move was not mentioned by the player at some point in the move sequence. These incomplete skipped move sequences had from one to four skips in the sequence. For example, as Appendix I, example 3 shows in the fifth move sequence down the master thought first about moving his queen to the e5 square (Qe5). He then thinks about moving his knight to e4 directly after his queen move and therefore skips a move for his opponent to respond to his queen move. This skip is represented as three dashes between the first and third moves. The fourth move in the sequence is where he thought about the opponent capturing his knight that has just moved to e4. He thinks about his opponent capturing his knight with a bishop (Bxe4). The final and fifth move in the sequence is his response to this capture. He thinks about capturing the opponent bishop that has just taken his knight with a pawn. The move sequence is now terminated. Masters and novices generated the same number of skipped move sequences for normal positions (2.2 vs 1.53, Mann-Whitney  $U_{5,5} = 8.0$ ,  $Z = -.946$ ,  $p = .344$ , two-tailed), and masters generated more than novices for random

positions (2.76 vs 1.60, Mann-Whitney  $U_{5,5} = 3.0$ ,  $Z = 2.015$ ,  $p = .044$ , two-tailed). The results indicate that masters and novices sometimes have difficulty in articulating their own and their opponent's moves.

A third common sort of move sequence (19%) was *base skip sequences* where the first move or 'base move' of the sequence was not mentioned. For example, as Appendix I, example 2 shows in the first move sequence of the graph this master thought first about the opponent moving a knight to the e6 square (Ne6) even though it is the master's turn to move. He has skipped his move and it is represented as three dashes at the start of the move sequence. The opponent Ne6 move is placed as the second move in the sequence. The third move in the sequence is where he thinks about moving his queen to the e3 square in response to his opponent's move. The fourth move in the sequence is where he thought about the opponent responding by moving his queen to the b5 square (Qb5). He thinks immediately about the next opponent move in the sequence rather than his own. He thinks about the opponent moving his rook to the e8 square (Re8). The move he skipped for himself is the fifth move in the sequence, and is represented by three dashes between the fourth and sixth moves in the sequence which both belong to the opponent. The move sequence is now terminated. Masters and novices generated the same number of base skip move sequences for normal positions (1.67 vs 1.14, Mann-Whitney  $U_{5,5} = 5.5$ ,  $Z = -1.49$ ,  $p = .136$ , two-tailed) and for random positions (1.3 vs 1.2, Mann-Whitney  $U_{5,5} = 11.5$ ,  $Z = -2.10$ ,  $p = .834$ , two-tailed). The results show that masters and novices sometimes do not articulate their first moves, and this difficulty occurs for both normal and random positions.

Finally, some move sequences were ambiguous (6%), where the move sequence could not be interpreted. For example a player said 'then the rook moves over there...', and it is not possible to determine the precise coordinate the player has in mind for it to move to. Masters and novices generated as many ambiguous sequences for normal positions (0.33 vs 0.60, Mann-Whitney  $U_{5,5} = 10.5$ ,  $Z = -.435$ ,  $p = .663$ , two-tailed) and for random positions (1.37 vs 1.53, Mann-Whitney  $U_{5,5} = 11.5$ ,  $Z = -.211$ ,  $p = .841$ , two-tailed).

Overall masters' and novices' think-aloud protocols of their move sequences reveal completely articulated sequences, skipped move sequences, base skip sequences and ambiguously articulated sequences. Masters and novices did not differ systematically in the frequency of each sort of sequence.

### *Hypothesis tests and skipped move sequences*

For the normal board positions, players failed to verbalize a move in 25% of move sequences (N = 56); some move sequences had more than one such 'skip' and so there were 75 skips overall. For the random board positions, players failed to verbalize a move in 26% of sequences (N = 57), but some move sequences had more than one skip, and there were 161 skips in total. Masters and novices produced a similar number of *individual skips* in normal positions (3.064 vs 1.932, Mann-Whitney  $U_{5,5} = 7$ ,  $Z = -1.160$ ,  $p = .246$ , two-tailed) and in random positions (7.466 vs 4.2, Mann-Whitney  $U_{5,5} = 5$ ,  $Z = -1.581$ ,  $p = .114$ , two-tailed).

Masters and novices skipped their own moves as often as each other for normal positions (.40 vs .20, Mann-Whitney  $U_{5,5} = 7.5$ ,  $Z = -1.107$ ,  $p = .268$ , two-tailed); they also skipped their opponent moves as often as each other (2.67 vs 1.73, Mann-Whitney  $U_{5,5} = 7.5$ ,  $Z = -1.051$ ,  $p = .293$ , two-tailed). It is notable that they also skipped their own moves for random positions as often as each other (3.1 vs 2.0, Mann-Whitney  $U_{5,5} = 6.5$ ,  $Z = -1.261$ ,  $p = .207$ , two-tailed) and they skipped their opponent moves as often as each other (4.37 vs 2.20, Mann-Whitney  $U_{5,5} = 5$ ,  $Z = -1.586$ ,  $p = .113$ , two-tailed). Masters and novices tended to articulate their own moves more than their opponent's moves, which may reflect a sort of confirmation tendency.

To test whether it is a confirmation bias, I inputted to *Fritz* the first few moves of a skipped move sequence until the position was reached at which the first skipped move occurred. I carried out the analysis only for the normal positions, because for the random positions the large number of skips precluded inputting them. At this point (first skip node) I sought an objective evaluation of the position from *Fritz*. Masters and novices sometimes began to skip at a point in the move sequence when the evaluation of the node was negative (4.4 vs 2.4, Mann-Whitney  $U_{5,5} = 324.5$ ,  $Z = -1.083$ ,  $p = .279$ , two-tailed); they also sometimes began to skip at a point in the move sequence when the evaluation of the node was positive (2.2 in each case). The *Fritz* evaluation for the first skipped move was somewhat more negative for masters than novices (-.6167 vs .1935) although the difference was not reliable ( $t = -1.620$ ,  $df = 54$ ,  $p = .111$ , two-tailed). The tendency to skip opponent's moves may simply reflect a richer representation of own hypotheses and plans.

### ***Summary***

The main result of the experiment is that masters objectively evaluated their moves more accurately than novices, and in particular they falsified accurately more often than novices for normal positions, and not for random positions. The result corroborates the prediction that masters try to falsify their hypotheses. It also suggests that when domain knowledge is eliminated masters do not falsify more successfully than novices.

The analysis also explored the differences in the search structure of master and novices hypothesis testing, and provided the following results:

- (i) Masters and novices did not significantly differ in the quality of their first moves.
- (ii) Masters generated reliably more individual moves than novices.
- (iii) Masters searched to the same depth as novices for normal positions, but to a marginally greater depth for random positions.
- (iv) Masters searched to the same breadth as novices for normal positions, but to a marginally greater breadth for random positions.
- (v) Masters' and novices' think-aloud protocols contained completely articulated sequences, skipped move sequences, base skip sequences and ambiguously articulated sequences. Masters and novices did not differ systematically from each other in the frequency of each sort of sequence.
- (vi) An analysis of the skipped move sequences shows that masters and novices reliably tend to articulate their own moves more than their opponent's moves. Masters and novices sometimes began to skip at a point in the move sequence when the evaluation of the node was negative and sometimes when it was positive.

Let us turn now to a discussion of the implications of the results. First, I will discuss how the results on search differences between masters and novices may contribute to our understanding of hypothesis testing and theories of reasoning. Second, I will discuss the implications of the results for theories of cognitive expertise, and the development of expertise.

### **General Discussion**

The results of the experimental analysis of chess masters hypothesis testing corroborate the prediction that masters try to falsify their hypotheses. They also

indicate that when domain knowledge is eliminated masters do not falsify more successfully than novices. The results are suggestive that domain expertise may help experts to choose better quality moves, and the process by which it may help them is through falsification.

There was no difference in the quality of their first moves. Masters generated more individual moves than novices. They searched to the same depth as novices for normal positions, but to a somewhat greater depth for random positions. They also searched to the same breadth as novices for normal positions, but to a somewhat greater breadth for random positions. The results are suggestive that masters may have superior search skills than novices for dealing with novel problems in their domain.

Masters and novices' did not differ systematically from each other in their ability to articulate their move sequences. The novices in this study were experienced novices, to ensure they were fluent in 'chess notation'. Half of all masters and novices' think-aloud protocols contained completely articulated sequences, and the remainder contained skipped moves, that is, moves that were not articulated, either during the sequence (skipped move sequences) or at the outset (base skip sequence). Masters and novices both tended to articulate their own moves more than their opponent's moves.

### ***Implications for Theories of Cognitive Expertise***

Is chess expertise comparable to other domains of expertise? Chess is adversarial and falsification in chess is a result of considering opponent moves. There are immediate consequences of committing an error: you will lose the game. Science too is a competitive enterprise and refutations may often come from competing scientists and laboratories (Kuhn, 1993; Mitroff, 1974; Gorman, 1995a). In fact, findings from the domain of chess have an excellent external validity (e.g., Larkin *et al.*, 1980; Lesgold, Rubinson, Feltovich, Glaser, Klopfer, & Wang, 1988; Sloboda, 1976).

A hypothesis testing model implies that masters sometimes consider move sequences that lead to error. The idea that expert knowledge sometimes contains errors, or that the retrieval of knowledge sometimes results in errors implies that experts must be able to detect errors that could result from such inaccurate knowledge or inaccurate retrieval. Although expertise may be reproductive in routine problems (e.g., Ericsson, Krampe, & Tesch-Romer, 1993; Gobet 1998), masters still search beyond the first move considered and the purpose of this

search requires explanation (Charness, 1991).

The addition of a hypothesis testing component to chess expertise provides the explanation that although masters may rely on templates stored in memory, the plans or moves stored with the template may not fit with every position encountered. A master may avoid error by searching to check the alternative moves retrieved as possible plans. The previous analyses of the search process in chess have focused on surface structures of the search tree, such as ply-depth and ply-breadth but these features may be only peripherally related to processing differences in search between masters and novices. A more important feature of the search structure may be the evaluation of move sequences, which may contribute to the selectivity of search.

The problem of selectivity in search is relevant for chess playing programs in artificial intelligence research (e.g., Hsu, Campbell & Tan, 1997). Chess playing programs still search extensively through the problem space when choosing a move for play. For example, the *Deep Blue* program that defeated the reigning world chess champion Kasparov considered 90 billion moves at each turn, at a rate of 9 billion per second (Eysenck & Keane, 2000). Computer chess programs today still do not perform to world champion standard without employing extensive search. The processes of chess playing programs at this level of expertise do not yet match the types of processes relied on by human world champion chess players (Hsu, Campbell, & Hoane, 1995).

Evaluative knowledge, including the goal to search for negative evaluation when searching through a problem space, as proposed by our hypothesis testing model, creates a highly selective search strategy in two ways. First, the ply depth is shortened by searching for opponent moves that falsify a plan as early in the move sequence as possible. This strategy eliminates redundant alternative lines where a poor quality opponent move that leads to a good outcome is considered even though an opponent move that falsifies it is available. If a poor quality opponent move is considered in a move sequence instead of a falsifying opponent move the player may have to search deeper or broader before they find evidence to act as a 'stop rule' for their consideration of this particular move sequence. Masters' search is more selective because it is limited largely to finding opponent-falsifying moves. Second, masters' stop rule for search is to stop when the opponent's falsifying move leads to a negative evaluation, that is a worsening of the masters' position. For example, masters skip opponent moves when they think-aloud at the point at which the move sequence begins to



lead to a negative outcome, according to *Fritz*. A crucial principle for the improvement of chess playing programs might be to simulate the development of the human player's goal hypothesis by employing a cyclic, progressive-deepening strategy based on hypothesis testing which acts as goal feedback. The program could carry out investigative tests and collect the test results in terms of whether a hypothesis is confirmed or falsified (that is, a board goal is obtainable or not).

Masters and novices both showed some evidence of confirmation bias: they sometimes thought about how their planned move sequences could lead to an advantage for their position when objectively it led to error. The type of confirmation bias exhibited was of the +/- test type, that is, the players either did not see that the opponent moves would lead to a disadvantage or they mentally minimized the negative impact of the opponent moves and found justifications for playing the moves anyway. It is important to note that there were very few of the '+/+' test types, which shows that the players were not simply searching for opponent moves that made their plans work out (if they were then the objective evaluation of the move sequence by *Fritz* at the terminal node would also often be positive). The confirmation shown by players is more likely a bias, that is, a failure to detect that the opponent move is negative for them, rather than a direct search strategy.

Masters may consider possible moves that they have not retrieved from template knowledge but that they have constructed 'on the spot' to explore new possibilities and so update their knowledge. This strategy may ensure they discover novel moves that are better than any previous moves considered best in a given position. Masters search even after they have generated a good move in normal chess. They may create new knowledge from old knowledge, by using smaller units of chunked knowledge stored lower down the hierarchical architecture than templates, to generate exploratory moves (Anderson, 1983). The structure of expert knowledge in a template system can be adaptive and promote the development of new knowledge structures if an evaluation mechanism can operate, that is, a hypothesis testing process mediated by evaluative knowledge. The conjecture provides the germ of an explanation for how knowledge may advance in a domain, and how masters deal with novel problems within the domain of chess.

### ***Some Implications for Cognitive Theories of Reasoning***

The experimental analysis of the chess masters' hypothesis testing shows that chess masters are capable of falsifying their hypotheses. They thought about how their opponent might refute their plan in move sequences that objectively led to error in normal chess. Falsification in this expert domain appears to be possible, useful and rational (see Popper, 1959; 1963; Kuhn, 1993; and Poletiek, 1996). Falsification in thinking requires thinking about negative instances. A negative instance may be counter to or inconsistent with a mental representation currently under consideration. People tend to have difficulty in representing negation, and as a result their thinking may display a bias (e.g., Evans, Newstead, & Byrne, 1993; Newstead *et al.*, 1992). It may require more working memory resources to maintain it (e.g., Evans, 1989). People may have a tendency to represent possibilities that are consistent and true rather than false (Johnson-Laird & Byrne, 2002). To represent negation they may have to think about alternative possibilities as well as the negative one (Johnson-Laird & Byrne, 1991). In this experiment, masters' ability to falsify is greater for normal positions than random ones, suggesting that falsification is mediated by domain knowledge. Their stored domain knowledge may contain some counterexamples that are typical in specific positions. Masters may actively seek counterexamples in their reasoning about the possibilities for play (see Byrne, 2005; Byrne, Espino & Santamaria, 1999; Johnson-Laird & Byrne, 1991; 2002). Indeed, the ability to falsify, and consider what may be false or negative may be part of what makes an expert, and helps them to avoid making mistakes. Novices' hypothesis testing may be similar to people's everyday reasoning when they do not have expertise in a domain. The implication is that people in general may find it easier to reason with positive information, such as true possibilities, and find it difficult to reason with negative information, such as false possibilities, and the results of this experimental analysis are consistent with this suggestion (Johnson-Laird & Byrne, 1991; 2002).

In addition, chess is an action domain (Mynatt, Doherty & Dragan, 1993). Players may generate a counter move to an anticipated counter move by a process of strategic reasoning at progressively deeper levels (e.g., Camerer, 2004; Hedden & Zhang, 2002; Zhang & Hedden, 2003; 2003). Action domains require consideration of a number of alternative paths towards solution, and there is much speculation on how the consideration of alternatives plays a role

in more general facets of thinking such as counterfactual possibilities in imaginative thought (e.g., Byrne, 2005), thinking about alternative causes of an event (e.g., Goldvarg & Johnson-Laird, 2001), and thinking about how things could have worked out better when one makes an error (e.g., Roese & Olson, 1995). The results of this experimental analysis imply that it may be necessary to switch attention in the search space from one move sequence to another, in order to evaluate alternative possible plans of action. As experts have better problem representations due to chunked or template structured knowledge, it may be easier to switch between alternatives and to maintain the results of previous tests (Ericsson & Kintsch, 1995). Novices on the other hand may find it difficult to disengage from a current line of investigation and may find switching to another line of investigation difficult (e.g., Cowley, 2002; Cowley & Byrne, 2004). It may also be difficult for novices to retrieve the results from earlier tests, and the same may be true for peoples' reasoning in general. When people do not have expertise in a domain they may find it difficult to generate and represent many needed alternative possibilities, because they do not have the knowledge to generate an alternative (e.g., Cowley & Byrne, 2005; 2015), or because they do not have the practice built up that allows people to mentally represent more information when they have domain expertise (e.g., Baddeley, 1999; Ericsson & Kintsch, 1995).

A further important feature prominent in the problem behaviour graphs of chess thinking was where a player 'skipped' a move. A skip represents a move in the sequence that was not verbalised for either the opponent or the player. In other words these skips may represent where a move has not been explicitly represented as have other verbalised moves in the sequences. The skips appeared to occur most often for opponent moves, for both masters and novices. This may imply that it is more difficult to represent an opponent's plan or an alternative plan to one's own in an explicit way. Theories of reasoning suggest that mental representation may be explicit or not explicit, given that people may reason using mental representations (e.g., Craik, 1943; Johnson-Laird, 1983). Our detailed chess problem behaviour graphs may testify that chess players may reason by constructing mental representations that are explicit in some respects (see DeGroot, 1963; Newell & Simon, 1972).

This experimental analysis has shown that not only is the consideration of multiple alternative possible moves important to hypothesis falsification, but that it is important to represent these alternative possible moves as explicitly as

possible. Chess masters falsified their hypotheses, by searching for alternative moves an opponent could play and these alternative moves were explicitly represented. Expert knowledge prompts the falsification of one's own hypotheses, regardless of the availability of falsifying evidence (Klayman & Ha, 1987). Chess masters must consider an opponent, but it is not clear whether this competitive variable is affecting how they test their hypotheses apart from their access to large repositories of domain knowledge. This experimental analysis shows that expert knowledge is the main factor facilitating falsification in chess masters' hypothesis testing. People can experience falsification as a possible and rational strategy; chess masters use hypothesis falsification to avoid error in their thinking (Cowley & Byrne, 2004; 2016).



## Chapter 5 - Discussion

*It is the mark of a strong education to be able to entertain a thought without accepting it*

—Aristotle 384BC-322BC

The aim in this thesis was to examine the factors which help people search for evidence to falsify untrue hypotheses. The chapter summarises the findings from the thesis, and focuses on how the work contributes to the literature on hypothesis testing, and what the results may mean in the broader context of human learning and cognition.

Two new experimental approaches to study how people test the truth of their hypotheses were adopted. First, a new version of a standard reasoning task was adapted to the research: an imaginary participant to the 2-4-6 task was introduced. Rather than participants testing their own hypothesis, which is usual in the 2-4-6 task, I designed a version of the task which required participants to think about an imaginary participant's hypothesis—the imaginary participant 2-4-6 task. I conducted two series of experiments to investigate not only competition with opponent hypothesis testers, but how the consideration of alternative hypotheses affected hypothesis falsification. Second, I extended the investigation of hypothesis testing to a new domain to investigate how expert knowledge affected hypothesis falsification; I compared the hypothesis testing of master and novice level experts in the domain of chess. Third, a range of methodologies was drawn on in these experimental investigations of hypothesis testing in chess and in the imaginary participant 2-4-6 task. Additional recording measures in our pen and paper tests for the 2-4-6 task were included. Participants were asked not only whether they intended their hypothesis test to lead to a confirming or falsifying result, but whether they considered their test to be consistent (a positive test) or inconsistent (a negative test) with their hypothesis. Both measures are used to categorize hypothesis testing in the 2-4-6 task. In previous studies the researcher decides whether a test is a positive or negative test, even though the experiments in this thesis found that participants may sometimes not conform to the researcher's prescriptions (e.g., Wason, 1960; Gorman & Gorman, 1984; Kareev & Halberstadt, 1993). Fourth, for the first time in this research domain protocol analysis was used to examine the

mental representations people rely on when they test their hypotheses; what chess players mentally represented while they searched for evidence to test if their hypothesized good moves were objectively good moves was tested. Finally, I used a powerful computer chess program to measure whether or not chess players correctly interpreted the evidence they found, which consisted of move sequences they had generated, with corresponding evaluations of outcome for each move sequence. This combined use of protocol analysis and the computer chess program allowed us to discriminate between biased and non-biased hypothesis testing in a precise way.

These novel experimental approaches and applications of methodology allowed us to test two main theories of hypothesis testing. The mathematical relationship theory (Klayman & Ha, 1987) predicts not only that people have a tendency to generate hypothesis tests which are consistent with their hypothesis (positive tests), but that falsification of an untrue hypothesis can be predicted more readily by the mathematical relationship between the hypothesis and the available evidence than by the active attempts of the hypothesis tester. The uniformity theory (Poletiek, 2001) predicts not only that people find falsification extremely difficult, but that confirmation and falsification are different outcomes of the same cognitive process. The results from these experiments question the tenets of both theories. The results highlight how expert domain knowledge, the consideration of alternative hypotheses, and competition make hypothesis falsification possible, regardless of whether people know what the relationship between the hypothesis and the evidence is. The results also indicate that confirmation is distinctly different from falsification in several ways, specifically when it is intentionally biased, that is, when participants expect falsifying triples to confirm.

In this chapter, I will first summarise the findings and consider how they inform our understanding of how people search for evidence to falsify untrue hypotheses. Second I will address how the results bring current theories of hypothesis testing into question and the consequences they have for current theories of cognitive expertise and expert thinking. Third, I discuss the broader implications the results have for human cognition, namely, social hypothesis testing and reasoning.

In short, the results suggest that a new theory of hypothesis testing may need to be developed to specifically address the role of domain knowledge, the consideration of alternative hypotheses, and competition in hypothesis testing.

In what follows I identify two new theoretical problems that must be tackled in future studies, one which is important in order to start to build a new theoretical framework of hypothesis testing. That is, whether a logical or Bayesian framework is more appropriate for the development of a new theory of hypothesis testing.

## **Summary of findings**

### ***Alternative hypotheses and falsification***

In Chapter 2 I described the Imaginary Participant 2-4-6 task. Previous research had found that participants understood the falsifying implication of negative tests when they were presented with them (Kareev & Halberstadt, 1993); and this thesis wished to examine whether participants could generate their own negative tests to falsify a hypothesis. The central idea in the introduction of an Imaginary Participant was to create a laboratory hypothesis testing situation akin to an everyday hypothesis testing situation, for example, when a teacher must correct a student's incorrect hypothesis by providing a counterexample to it (Cowley & Byrne, 2005; 2015).

I suggested that a major purpose of falsification in human reasoning was to identify untrue hypotheses and if people experience falsification as possible, then perhaps it was the case that experimental psychologists had previously overlooked facilitative factors in hypothesis falsification (e.g., Poletiek, 1996). To this end how participants tested untrue (low-quality) and true (high-quality) hypotheses in Experiment 1 were compared. Participants were told that a researcher had a rule in mind that the number sequence 2-4-6 conforms to. The experimenter's rule was the typical 'any ascending numbers' rule. They were told that an imaginary participant called Peter hypothesized that the experimenter's rule was either the untrue 'even numbers ascending in twos' or the true 'any ascending numbers' rule. Participants tested if Peter's rule was the experimenter's rule by generating their own number triples in such a way as to show Peter whether his hypothesis was correct.

Experiment 1 investigated whether falsification was consistently possible in the 2-4-6 task. The results showed that high-quality hypotheses were confirmed more often than low-quality hypotheses, and low-quality hypotheses were falsified more often than high-quality hypotheses. Does this imply that hypothesis quality alone affects hypothesis testing (Poletiek, 1996)? Participants falsified reliably more often when they *knew* they were testing a low-quality



hypothesis than when they did not know. Half of the participants in the low and high-quality conditions were given additional information ‘you know that the experimenter’s rule is in fact any ascending numbers’. From this additional sentence they knew whether they were testing a high or low-quality hypothesis. The critical condition was when participants tested Peter’s low-quality ‘even numbers ascending in twos’ hypothesis and were also told that ‘you know that the experimenter’s rule is in fact any ascending numbers’. The only way participants could test Peter’s low-quality hypothesis in order to show him that it was untrue was by generating a negative falsifying test. Consider, the triple 5-10-15, to which they receive a ‘yes’, and Peter could then interpret that the rule does not pertain to evenness or ascending in twos. In addition they must show that they *intend* this negative test to falsify by indicating that they expect this ‘yes’ response from the researcher (See Table 1.1). More negative falsifying tests (90%) were generated in this condition than any other. For example, when participants tested Peter’s same low-quality hypothesis and they did not know the experimenter’s rule, they generated reliably fewer falsifying tests (30%). Participants did not appear to be able to falsify regardless of whether they knew they tested a high-quality hypothesis (0%) or not (10%). Perhaps participants realise that when a hypothesis is very high-quality, it may accurately represent the truth and there is scant falsifying evidence. The results from this experiment demonstrated that people can consistently engage in falsification, that is, they can generate negative tests that genuinely falsify a hypothesis. They can falsify in the most difficult hypothesis testing situations, that is, when the hypothesis is low-quality and it is embedded within the true rule. They falsified with the intention that their chosen tests would falsify; they expected falsification. They appear to be aware of the implications of their test choice because they announced that the imaginary participant would realize from the test results that a hypothesis was low-quality (Poletiek, 1996).

What factors facilitated this high rate of negative falsifying tests? Hypotheses have to be untrue in order for participants to falsify in a useful way. The extra sentence ‘you know that the experimenter’s rule is in fact any ascending numbers’ may facilitate this falsification in several ways. The sentence presents an alternative hypothesis to the participants; they not only test Peter’s hypothesis ‘even numbers ascending in twos’ but are aware that the experimenter’s rule is ‘any ascending numbers’. The alternative ‘any ascending numbers’ is a higher quality hypothesis than the hypothesis under test; it makes

known to the participant that it is higher quality; it is an explicitly stated hypothesis; and it presents the participant with two hypotheses to represent from the outset, both Peter's and the alternative. In the experiments that followed I examined which of these factors affected falsification.

Experiment 2 investigated whether the quality of the alternate hypothesis affected falsification of an untrue hypothesis. The results showed that participants falsified Peter's untrue hypothesis 'even numbers ascending in twos' with negative falsifying tests more often when they knew the alternative 'any ascending numbers' was the experimenter's rule (61%) than when they did not know it was the experimenter's rule (51%) but this result was not significant. Participants falsified 'even numbers ascending in twos' more when the alternative was the high-quality 'any ascending numbers' (51%), than when it was medium quality 'numbers ascending in twos' (48%) or when it was low-quality 'even numbers ascending in twos ending in the digits 2,4,6' (33%). These results tentatively suggest the possibility that knowledge that a hypothesis is an accurate representation of the truth is not enough; the alternative must be higher quality than the hypothesis under test to falsify it. It is possible that the consideration of an alternative hypothesis presents the hypothesis tester with a set of possibilities from which to generate falsifying triples, perhaps to generate a higher quality alternative. Yet falsification led participants to subsequently announce the correct experimenter's rule reliably more often when the alternative was high-quality (50%), and medium quality (44%), but not lower quality than the hypothesis under test (12%). The implication is that not all falsifications or alternative hypotheses are helpful. The alternative need not necessarily be an accurate representation of the truth, as a medium quality alternative hypothesis led to rule discovery as often as the alternative that was the experimenter's rule. The alternative hypothesis need not represent the truth, but it must lead the participant towards the truth rather than away from it to be helpful in rule discovery.

Experiment 3 investigated whether the explicitness of the alternative hypothesis affected falsification of an untrue hypothesis. Participants tested the imaginary participant Peter's hypothesis 'even numbers ascending in twos' and were given either an explicit, non-explicit, or no alternative hypothesis to consider. Participants did not falsify reliably more often when the alternative was non-explicit (53%) and when there was no alternative (54%), than when the alternative was explicit (43%). Yet falsification led to rule discovery more often

when there was an alternative that was explicit (56%) than non-explicit (33%), or when there was no alternative at all (19%). The implication is that falsification could not be predicted by how explicit the alternative hypothesis was, but the consideration of an explicit alternative may help people to use their falsification in order to discover the truth.

The results of the three experiments in Chapter 2 imply that falsification *and* alternative hypotheses may go hand-in-hand in discovering the truth in hypothesis testing in most circumstances (e.g., Wason & Johnson-Laird, 1972). While falsification by itself may only lead to the discovery that the hypothesis is untrue, the generation or consideration of an explicit higher quality alternative hypothesis may help towards the eventual discovery of the truth.

### ***Competition and falsification***

Even without the consideration of an alternative hypothesis, the rates of falsification in our experiments were higher than usual in the literature (see Poletiek, 2001 for a review). One possible explanation is that people had a tendency to falsify a hypothesis belonging to someone else (the imaginary participant Peter) more than their own hypotheses.

Experiment 4 investigated whether a competitive factor such as hypothesis ownership affected hypothesis falsification. Participants were presented with the imaginary participant ‘Peter’s hypothesis: even numbers ascending in twos’ in one condition, and presented with ‘Your hypothesis: even numbers ascending in twos’ in another condition. They were instructed to test if the hypothesis is the experimenter’s rule. The hypotheses were equally untrue, and they were not presented with any other conditions to help them to falsify such as an explicit alternative hypothesis (Klayman & Ha, 1989). Participants did not generate their own hypothesis; ‘even numbers ascending in twos’ was simply given to them as their hypothesis and they were instructed to test it, ruling out a personal investment explanation. The results showed that participants generated reliably more negative tests (46% vs 24%) and fewer positive tests (54% vs 76%) of the imaginary participant’s hypothesis than their own equally untrue hypothesis. Participants confirmed their own untrue hypothesis reliably more (84%) than the imaginary participants (60%), and falsified the imaginary participant’s untrue hypothesis somewhat more (40%) than their own (16%). Falsification was used to abandon the untrue hypothesis reliably more often when it belonged to the imaginary participant (62%) than when it belonged to themselves (25%).

Hypothesis ownership not only affected hypothesis falsification, but whether this falsification was used to abandon an untrue hypothesis.

Experiment 5 investigated whether a second competitive factor affected hypothesis falsification, namely, contending with an opponent hypothesis tester. I examined whether participants falsified their own hypothesis more under conditions in which they were told an opponent hypothesis tester was also testing their hypothesis, than when they were not told anything about an opponent. I wanted to see whether the awareness of an opponent testing their hypothesis would encourage or discourage falsification.

Unexpectedly, participants generated a similar amount of confirming tests whether there was an opponent hypothesis tester (91%) or not (88%), but the types of confirming tests did differ. The results showed that participants generated reliably more negative test triples for their own hypothesis when an opponent was present (54%) than when an opponent was not present (13%). But participants in the opponent condition expected their negative triples not to be consistent with the experimenter's rule either, thereby confirming their hypothesis. A triple that is considered by a participant not to be consistent with the experimenter's rule, or consistent with the hypothesis they test can only confirm the hypothesis further, albeit using a negative test (see Table 1.1). The results suggest that the consideration of an opponent hypothesis tester may prompt participants to try harder to falsify their hypotheses, and they generate negative tests of these hypotheses, but they may hope that the opponent does not see how to falsify their hypotheses (much like the chess novices in Chapter 4).

Reliably more participants abandoned the low-quality hypothesis when there was an opponent (56%) than when there was no opponent (38%) suggesting that participants are more likely to use falsifying evidence to admit that their hypothesis is untrue when an opponent hypothesis tester is present. The results show that the consideration of an opponent hypothesis tester affects hypothesis testing. Participants not only generated more negative tests when there was an opponent than when there was no opponent, but they were more likely to abandon an untrue hypothesis.

### ***Theoretical implications of results from the Imaginary Participant 2-4-6 task***

The results from our Imaginary Participant 2-4-6 experiments challenge current theories of hypothesis testing in several ways. The first tenet of the uniformity theory (Poletiek, 2001) predicted that falsification was not consistently possible

for people because they do not know where to find information enabling them to generate a falsifying test (see Table 1.3). Experiments 1 to 5 showed that participants could generate negative tests, and intend these negative tests to falsify when an alternative hypothesis was presented to them. In some cases the alternative hypothesis was the experimenter's rule, or a higher quality hypothesis, or simply a non-explicit hypothesis such as 'something else', but participants could falsify using negative falsifying tests.

The second tenet of the uniformity theory proposed that confirming and falsifying were one and the same process; they represent the process of carrying out a test. The experiments showed that confirming and falsifying differed; hypothesis testers chose different types of tests (negative or positive) depending on whether they themselves or someone else owned a hypothesis, or whether an opponent hypothesis tester was present.

The third tenet of the uniformity theory proposed that the result of a hypothesis test may be as much a consequence of the quality of the hypothesis under test, than of any specific strategy employed by the hypothesis tester. The experiments showed that while this tenet was true under some circumstances, particularly when testing high-quality hypotheses, participants can play an active role in their test choice for low-quality (untrue) hypotheses. When testing an untrue hypothesis participants were able to choose negative tests with the intention to falsify when alternative hypotheses were presented to them, and they were able to choose negative tests more often when the hypothesis belonged to someone else than themselves regardless of whether an alternative hypothesis was present or not.

Similarly the results have generated several questions for the mathematical relationship theory (Klayman & Ha, 1987). The first tenet of the mathematical relationship theory predicts that people have a tendency to use a positive test strategy in any hypothesis testing situation they encounter, and the second tenet of the mathematical relationship theory predicted that people do not know when a positive test strategy is accurate and when it is not. (See Table 1.4). Our results showed that participants could readily engage in a negative test strategy when it was appropriate to do so, whether they were presented with an alternative hypothesis that was the actual experimenter's rule, or a higher quality hypothesis, or simply a non-explicit hypothesis such as 'something else'.

The third tenet of the mathematical relationship theory proposed that (a) the mathematical relationship between the hypothesis test and the experimenter's

rule affects the effectiveness of positive and negative test strategies, and (b) that the only way people can overcome their tendency to use a positive test strategy in the embedded relationship typical of the standard 2-4-6 task was to consider an alternative hypothesis. The first part of this tenet is true, because only a negative test can lead to a falsifying result when a hypothesis is embedded within the truth as Figure 1.1 in Chapter 1 shows. The second part may not be true. The results showed that participants could choose negative tests more often than positive tests when they were aware of an opponent hypothesis tester. They could generate negative tests of a hypothesis without being presented with an alternative hypothesis that could indicate to them what the relationship was. While the mathematical relationship can constrain the effectiveness of positive or negative tests in terms of whether they can objectively lead to falsification in the 2-4-6 task, the mathematical relationship cannot constrain the intention of the hypothesis tester. For example, participants who chose negative tests when they were aware of an opponent hypothesis tester actually intended them to confirm. Participants also generated negative tests readily when the hypothesis belonged to someone else and when they considered an opponent hypothesis tester. Participants showed they could determine when a negative test strategy was accurate, even though they were not presented with an alternative that would make the relationship between the hypothesis and the truth explicit to them.

In sum, the results from our Imaginary Participant experiments have shown that future theories of hypothesis testing should give a prominent place to the consideration of alternative hypotheses and how they are mentally represented (Tweney *et al.*, 1980; Gale & Ball, 2003; 2005; Oaksford & Chater, 1994; Wason & Johnson-Laird, 1972), and hypothesis testers should be seen as playing a more active role in their strategy selection (Poletiek, 1996; 2005; Klayman & Ha, 1987).

The thesis has shown then that people are capable of falsifying in a useful and rational way, and the experiments described examine how participants can falsify hypotheses belonging to somebody else. The final experimental analysis addressed whether people are able to falsify their own hypotheses once they have domain expertise.

### ***Expert thinking in chess and falsification***

The results of the experimental analysis on chess players' hypothesis testing corroborate the prediction that masters try to falsify their hypotheses. Chapter 4 presented a detailed protocol analysis of chess players' hypothesis testing. The think-aloud protocols of five masters and five novices who thought about choosing a move for play in normal and randomized chess positions were selected. I created problem behaviour graphs of their verbalized move sequences, and conceptualized their hypothesis testing by examining the moves they hypothesised to be good ones, which they subsequently tested by evaluation of alternative possible moves that their opponents could play. The subjective evaluations of the move sequences they thought about to the objective evaluations produced by the chess program *Fritz 8* were compared. Chess masters thought about how their opponent might refute their plan in move sequences that objectively led to error in normal chess. Falsification in this expert domain appears to be possible, useful and rational (see Popper, 1959; 1963; Kuhn, 1993; and Poletiek, 1996). The result also indicates that when domain knowledge is eliminated in random chess, masters do not falsify more successfully than novices; domain expertise is necessary to falsify one's own hypotheses.

Several properties in the problem behavior graphs of masters and novices were analysed, which were indicative of the type of search process the players used to generate hypothesis tests. Chess masters searched to the same depth as novices for normal positions, but to a somewhat greater depth for random positions. They also searched to the same breadth as novices for normal positions, but to a somewhat greater breadth for random positions. The results are suggestive that masters may have superior search skills than novices for dealing with novel problems in their domain.

Masters and novices' did not differ systematically from each other in their ability to articulate their move sequences. Half of all masters and novices' think-aloud protocols contained completely articulated sequences, and the remainder contained skipped moves, that is, moves that were not articulated, either during the sequence (skipped move sequences) or at the outset (base skip sequence). Masters and novices both tended to articulate their own moves more than their opponent's moves, but the articulation of opponent moves was essential to generating an objective falsification.

Based on the implications of these results presented in Chapter 4, I presented hypothesis testing as a new component of expert chess problem solving (Cowley & Byrne, 2004; 2016). By examining the search process in chess masters' hypothesis testing we were able to map not only the mental representations that are required in order to falsify one's own hypotheses, but how expert knowledge may facilitate falsification in hypothesis testing. I revealed that falsifying one's own hypotheses involves searching for the ways in which hypotheses may not lead to what is expected to be true (e.g., Popper, 1959), and involves making the alternative ways of testing our hypothesis as explicit as possible in order to identify evidence counter to the hypothesis currently represented (e.g., Byrne & Tasso, 1994). The result implies that future theories of hypothesis testing may usefully pay particular attention to the mental representation of hypotheses and how people search for falsifying counterexamples. The emphasis on mental representation in hypothesis testing may improve an understanding of how hypothesis testing is related to similar processes involved in other types of human reasoning such as deduction (e.g., Johnson-Laird & Byrne, 1991).

By examining the search process in chess masters' hypothesis testing, theories of cognitive expertise may be able to move beyond theoretical frameworks grounded in systematic pattern recognition (e.g., Chase & Simon, 1973a; 1973b), which relies on the assumption that cognitive expertise is largely reproductive (e.g., Gobet; 1998; Ericsson *et al.*, 1993). Our results suggest that there is a more adaptive view of the interaction between the expert and the problem at hand, for example, how novel an expert judges a problem to be. Little is known about how experts adapt their knowledge to novel problems in order to acquire new knowledge structures (Tei Leine & Saariluoma, 2001), detect errors (e.g., Green & Gilhooly, 1992), or learn from experiencing mistakes (Frensch & Sternberg 1993).

The hypothesis testing component of expert thinking aims to understand how experts detect errors, recover from them and adapt to novel problems in order to create new knowledge. Many disciplines require experts to test hypothetical predictions or solutions to problems whether they are scientists (e.g., Kuhn, 1993; Fugelsang *et al.*, 2004), legal experts (e.g., Britton, 1997), medics (e.g., Koriat *et al.*, 1980) or even creative individuals (e.g., Eysenck, 1995).

In our conceptualization of chess players' hypothesis testing it became apparent that not only search, but the evaluation of search results was involved in biased hypothesis testing. When a planned move objectively led to a negative



outcome, novices tended only to evaluate how their move could be confirmed. Previous research has not distinguished between the search and evaluation process in hypothesis testing. Some theorists see search and evaluation of the result as one and the same process (e.g., Poletiek, 2005; Howson & Urbach, 1993). But in our chess experiments we have shown that while a search can lead to an objectively falsifying result the interpretation can be biased. In the next section the implications of these findings in a broader context for theories of cognition are addressed, and in the final section I address the problem of test evaluation in the role of hypothesis revision.

### **General Implications for Human Learning and Cognition**

In this section I consider the implications the results have for our understanding of hypothesis testing in general in human cognition.

#### ***Implications for reasoning***

The results have two main implications for reasoning. First, I will outline why the results on falsification are important to the consideration of negative information in reasoning, and second I will outline what the results on the consideration of alternative hypotheses suggest about the consideration of alternatives in reasoning in general.

Hypothesis falsification implies that people can think about negative instances under certain conditions (e.g., Khemlani, Orenes, & Johnson-Laird, 2014; Klayman & Ha, 1987; Kareev & Halberstadt, 1993; Vallee-Tourangeau *et al.*, 1995). Chess masters were capable of falsifying the moves they hypothesized as leading to good outcomes but which in fact allowed opponent counter moves that would lead to bad outcomes (Cowley & Byrne, 2004). The consideration of bad outcomes to a hypothesized plan of action implies that chess masters were thinking about instances which were negative, and chess players' evaluations of move sequences are verbalized as falling into negative, positive or neutral categories (e.g., DeGroot, 1965; Newell & Simon, 1972; Holding & Reynolds, 1982).

Consider the example of someone who is given the following statement to think about: 'If Sharon is in Spain, then Justina is in Holland', and they encounter a piece of information that is inconsistent with this statement such as 'Justina is not in Holland', they can deduce that 'Sharon is not in Spain'. This type of inference is called *Modus Tollens* and it has been investigated

extensively in the literature on deductive reasoning (e.g., Byrne, 1989; Byrne & Tasso, 1994), and is logically equivalent to hypothesis falsification (e.g., Popper, 1959; Klayman & Ha, 1987).

To consider the inference people may construct a *counterexample* that is a possibility which is inconsistent with the possibility currently under consideration. A counterexample may be similar to an opponent *counter-move* in chess (Hartston & Wason, 1982), or a refutation of a theory in science (Kuhn, 1993). Little is known about how people search for counterexamples when they reason (e.g., Byrne, Espino, & Santamaria, 1999), and our research on chess masters hypothesis testing suggests that accessing domain knowledge aids the discovery of falsifying instances, and this implies that accessing relevant domain knowledge may have an important role in the retrieval of counterexamples when reasoning in everyday life in some circumstances. In addition, our research on chess players' hypothesis testing suggests that the explicit representation of opponent moves in the move sequence was partly responsible for the discovery of falsifying instances. Only the move sequences in which all opponent moves were articulated could objectively lead to falsification. Where players skipped moves for their opponents by verbalizing their own moves in a sequence and omitting their opponents' chances for countermoves, they had no opportunity to represent how their plan might not work. This suggests that it may be important to explicitly mentally represent what is being reasoned about as fully as possible in order to discover counterexamples, otherwise deductive errors could be made, such as concluding 'nothing follows' when you are told 'Justina is not in Holland' (e.g., Johnson-Laird & Byrne; Johnson-Laird & Byrne, 2002). Chess masters may actively seek counterexamples in their reasoning about the possibilities for play. Indeed, the ability to falsify, and consider what may be false or negative may be part of what makes an expert, and helps them to avoid making mistakes in their reasoning regardless of the domain (e.g., Hale, 2014).

Another implication is for the consideration of alternatives. The experimental analysis in the imaginary participant 2-4-6 task showed that falsification by itself could not be predicted by how explicit the alternative hypothesis was. Yet the representation of an alternative hypothesis as explicit such as 'ascending numbers' as opposed to non-explicit such as 'something else' was critical in the use of falsification to abandon an untrue hypothesis. This condition may parallel scientific reasoning that tends not to abandon a falsified theory unless a viable

alternative theory presents a better explanation (e.g., Kuhn, 1993), or labels falsification as an anomaly until a better alternative theory is generated (e.g., Koslowski, 1996).

How do chess players reason about falsification and alternative hypotheses?. Chess players may generate their own counter move to an anticipated counter move from an opponent (e.g., DeGroot, 1965; Camerer, 2004; Hedden & Zhang, 2002; Zhang & Hedden, 2003; 2003). The result of the experimental analysis of chess masters' hypothesis testing suggests that it may be necessary to switch attention in the search space from one move sequence to another, in order to evaluate alternative possible plans (Cowley & Byrne, 2004; 2016). As chess masters have better problem representations due to template structured knowledge, it may be easier for them to switch from one alternative to another and to maintain the results of previous tests (e.g., Ericsson & Kintsch, 1995; Perfect & Lindsay, 2014; Packiam-Alloway & Alloway, 2013). Novices on the other hand may find it difficult to disengage from a current line of investigation and may find switching to another line of investigation difficult (Cowley & Byrne, 2004; 2016). It may also be difficult for novices to retrieve the results from earlier tests; they may not remember the results and tend to reinvestigate moves already examined more than experts (Newell & Simon, 1972; Saariluoma, 1995). A similar phenomenon may occur when people reason in general, for instance when they tend to perform identical tests numerous times (Wason & Johnson-Laird, 1972), and they do not appear to be able to maintain as much information as experts in working memory (e.g., Baddeley, 1999; Ericsson & Kintsch, 1995; Tukey, 1986).

When people do not have expertise in a domain they may find it difficult to generate and represent many needed alternative possibilities, because they do not have the knowledge to generate an alternative (e.g., Cowley & Byrne, 2005; 2015), or because they do not have the practice built up that allows people to manipulate and mentally represent more information in working memory when they have domain expertise (e.g., Baddeley, 1999; Ericsson & Kintsch, 1995).

### ***Implications for planning***

The ability to falsify may allow people to anticipate how a hypothesized plan may go wrong, for example, by an opponent responding to a plan in a way that would cause their plan to fail (Hedden & Zhang, 2002; Zhang & Hedden, 2003; 2003). The results showed that chess masters can use their knowledge to

anticipate how an opponent could falsify their plan, and that novices tend to only see how their opponent's moves could confirm their plan. On the one hand this result suggests that masters used their domain knowledge to facilitate better planning, and the ways that a plan may be falsified could be retrieved from a store of domain relevant knowledge (e.g., Ericsson & Kintsch, 1995).

On the other hand the result suggests that masters also differed in the way they search for the alternative ways a plan could be falsified by limiting their search not to every conceivable possible counter move their opponent could make, but to the strongest countermoves their opponent could make in response to their plan. In other words chess masters tend not to examine irrelevant countermoves and search more efficiently than novices by testing their hypothesized plans as severely as possible (Popper, 1959; 1963; Poletiek, 2001; 2005). The implication is that in real life it may be helpful to consider the possible alternative ways a plan may be falsified than confirmed whether the situation is a political debate, diplomatic endeavour, military strategy or even a game of tic-tac-toe or poker (e.g., Koslowski, 1996; Mallie, 2001; Camerer, 2004).

When falsification of a plan results in error it may provoke the generation of an alternative plan addressing how things could have worked out better (e.g., Walsh, 2001). In this way the experience of a planning falsification may help people learn from their past mistakes, for example, by thinking of a way in which a past action might have been avoided (e.g., Mandel & Lehman, 1996; Byrne & McEleney 2000), or what a better solution to a problem might have been (e.g., Anzai & Simon, 1979). As a result the experience of planning falsification may help people plan better for the future and avoid past mistakes (e.g., Roese & Olson, 1995; Chevallier, 2016; Crowley & Zentall, 2013).

### ***Implications for social hypothesis testing***

Theories of social hypothesis testing predict that the smaller the number of alternative hypotheses a person considers the more confidence they may have in their initial hypotheses (Kahneman, Slovic, & Tversky, 1982). When we consider prejudiced beliefs as starting out as tentative hypotheses pertaining to traits we attribute to other persons or groups (e.g., Kruglanski & Webster, 2000), we may discern how important the consideration of alternative hypotheses might be. For example, if a person is only beginning to engage in prejudiced thinking about Jews as outlined by Anne Frank, then the presentation

of propaganda material by the media to persuade people that this prejudice is true may be effective; the truthful alternatives are suppressed (Aronson, 1999; Kruglanski & Webster, 2000). Without the presentation of alternative possibilities a piece of falsifying evidence such as the case of Anne Frank may be classed as an anomaly (e.g., Koslowski, 1996); or misleading evidence (see Gardenfors, 1988). For example, people with the prejudiced belief about Jews have concluded that Anne Frank's diary is not authentic (see the forward to Anne Frank's diary by Otto Frank & Mirjam Pressler, 2001). Social psychologists have termed such thinking '*motivated closing of the mind*' (Higgins & Kruglanski, 2000). Social knowledge constructions draw heavily from examples held in background knowledge, such as negative personal experiences with members of an ethnic minority (Ajzen & Fishbein, 2000), and such experiences are difficult to incorporate into a coherent knowledge base where the negative examples of a group of people are the exception rather than the rule (e.g., Harman, 1986; Gardenfors, 1988).

When we are making social inferences about someone's personality (e.g., Snyder & Swann, 1978), or about a group of people (Kruglanski & Webster, 2000), or even about what the social consequences of some action might be (Roese & Olson, 1994), it would be beneficial to search for examples which might not support our social hypotheses. Further it would be helpful to consider the alternative hypotheses which offer better quality explanations and can explain any counterexamples we do find (e.g., Cowley & Byrne, 2005; 2015). It has been said that when we are 'given a thimbleful of facts we rush to make generalizations as large as a tub' (Allport, 1979; p.8), but by challenging such generalizations by searching for falsification and alternative explanations we may avoid mistaken and even prejudiced thinking (Wason, 1960; Popper, 1959).

### **Future Questions**

How can hypothesis testing and the consideration of falsifying evidence in particular facilitate the acquisition of new knowledge? Second, how can hypothesis testing research contribute to the theoretical debate between logical and probabilistic theories as explanatory frameworks for understanding human cognition? I will now contemplate two questions that need to be addressed if the processes and applications of hypothesis testing are to be adequately integrated into future research.

### ***How does hypothesis testing lead to the acquisition of new knowledge?***

Falsification as the result of a negative test which exists outside of a current hypothesis may give the hypothesis tester information about where the current hypothesis is wrong (Klayman & Ha, 1989; Koslowski, 1996). In other words negative tests that lead to falsification can give a hypothesis tester information about what should be included in a future hypothesis, for example when a hypothesis tester discovers that odd numbers are consistent with the experimenter's rule they should try to include this new information in their hypothesis.

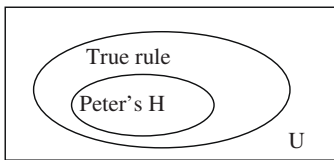
Reasoning with information outside of a current knowledge representation is sometimes referred to reasoning with inconsistency (e.g., Elio & Pelletier, 1997; Walsh & Johnson-Laird, 2005). Reasoning with inconsistency is important in the accumulation of new knowledge in human and artificial information processing (e.g., Harman, 1986; Franklin, 1997; Tough, 2013), but little is known about how inconsistencies such as falsifying results are accommodated by an existing state of knowledge (e.g., Elio & Pelletier, 1997; Gardenfors, 1988).

One possible way reasoning with falsification may lead to new knowledge is by the generation of plausible alternative hypotheses which accommodate the falsifying result (e.g., Kuhn, 1972; Harman, 1986). When entertaining a prejudiced belief as is the case in the Anne Frank example it is essential to search outside of the initial knowledge base for evidence to show that this assertion is untrue. The search for negative tests of a hypothesis may be required in order to acquire new knowledge especially when the hypothesis is untrue, as it tends to involve the situation where the untrue hypothesis is embedded within the truth (Wason, 1960; Klayman & Ha, 1989).

The embedded situation where an untrue hypothesis is embedded within the truth may be present in many situations. For example, I present a similar analogical comparison of an embedded low-quality hypothesis in the 2-4-6 task with one that occurs in another cognitive domain: chess playing. It is instructive to focus on the chess domain because falsification is possible for chess masters but novices find it difficult. The diagrams were designed according to prescriptions based on the principles of set theory. Klayman and Ha's work was adapted to classify a hypothesis testing situation in chess where U represents the

universe or the total number of possibilities in terms of hypothesis tests as shown in figure 5.1 below. In the 2-4-6 task U represents the total number of possible number triples a participant can generate (Klayman & Ha, 1987). In chess U represents the total number of possible moves that can be chosen from in a given chessboard position. An embedded low-quality hypothesis in the 2-4-6 task is labeled Peter's hypothesis which is 'even numbers ascending in twos'. Even numbers ascending in twos is contained in a smaller circle within the true rule (experimenter's rule) 'any ascending numbers', therefore embedded.

Embedded false hypothesis in the 2-4-6 Task:



Embedded false hypothesis in the domain of chess:

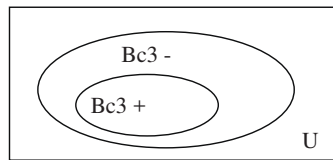


Figure 5.1: Embedded low-quality hypotheses in the 2-4-6 task and chess.

An embedded low-quality hypothesis in chess is a move hypothesized to be a good one as the plus sign shows for the move Bc3+ (Bc3 refers to where a player thought about moving their bishop piece to the square on the chess board with a coordinate called c3. A player must search for evidence that the move is a good one by mentally generating move sequences that consider opponent responses 'if I play this, you might play that' etc. The move Bc3 has some possible opponent moves that lead to a good outcome for the player thereby there is confirming evidence that Bc3 is a good move. But if we consider the set of all possible opponent moves that exists as tests of Bc3, there may be some that lead to a bad outcome. The circle Bc3- labels the set of possible hypothesis tests that falsify the hypothesis that Bc3 is a good move.

The subtle point is that even if there is only one possible response by the opponent that falsifies the hypothesis that Bc3 is a good move, then the move Bc3 in total terms is a bad move. In line with a Popperian analysis whether the hypothesis tester is testing an embedded low-quality hypothesis in chess or in the 2-4-6 task or even testing a prejudiced belief, they do well to test their hypothesis as severely as possible by searching for inconsistent evidence in terms of a negative test.

Whether an untrue hypothesis concerns a numerical rule in a laboratory task, a chosen move in a game of chess, a scientific theory or even a prejudiced belief it is necessary to search for evidence which is inconsistent with that hypothesis (e.g., Wason, 1960; Cowley & Byrne, 2004; Kuhn, 1993; Kruglanski & Webster, 2000). The search for evidence inconsistent with a belief or hypothesis is sometimes equated with the search for new knowledge in both human reasoning (e.g., Popper, 1963) and machine learning (e.g., Franklin, 1998; Cowley & Macdorman, 2006). But little is known about how people or machines accommodate this type of inconsistency once it is encountered (Harman, 1986; Gardenfors, 1988). In the next section I consider two frameworks which could address knowledge acquisition through reasoning with inconsistency.

### ***When is one refutation enough? Logical and Bayesian frameworks***

I showed that extensive research has been carried out to examine what types of information people seek out in order to test their hypotheses. But once people encounter evidence which is inconsistent with their hypotheses, we know little about how they accommodate that information. That is, how do they revise a hypothesis once it has been falsified? Theories of belief revision may provide some hints about what theoretical framework may be appropriate for the development of a new theory of hypothesis testing. For example, both the fields of belief revision and hypothesis testing deal with how people reason with evidence counter to that which they expect, that is, evidence that is inconsistent with a hypothesis (e.g., Klayman & Ha, 1989; Klayman, 1995; Cowley & Byrne, 2005; 2015) or a belief (Elio & Pelletier, 1997; Walsh & Johnson-Laird, 2005). This evidence is termed *refutation* and theories of hypothesis testing tell us very little about how people use refutations to evaluate and revise their hypotheses. Previous research has not been successful in discovering situations in which hypothesis falsification is facilitated (see Poletiek, 2001 for a review).

I have shown that people do not always abandon untrue hypotheses that have been refuted, whether these are hypotheses presented in the psychological laboratory, or real theories of scientists in a laboratory (Dunbar, 2000). Why do people sometimes abandon a hypothesis once they encounter inconsistent evidence, and why do they sometimes not? On the one hand theories of belief revision show that when people come across a fact that conflicts with their beliefs the revision to their beliefs is minimal (Harman, 1986). For example,



people abandon specific beliefs, such as facts, that require a small change in belief, rather than giving up a general belief, which requires a large change in belief, such as giving up an entire theory (Elio & Pelletier, 1997; Kuhn, 1993). On the other hand theories of belief revision have found that people who were asked to create explanations to resolve inconsistencies, tended to refute the general rather than the specific belief (Walsh & Johnson-Laird, 2005). These findings from belief revision and hypothesis testing give rise to the discrepancy that refutations sometimes lead to the revision of hypotheses and beliefs and sometimes they do not.

There are currently two classes of reasoning theories which make predictions about how people reason with refutations of a belief, logical and Bayesian theories, and an important future question for hypothesis testing is which one of these theories can provide the best explanation of how people reason with refutations (falsifications) of their hypothesis. One view of hypothesis revision has been inspired by the work of Popper (1959; 1963). He proposed that no amount of confirming evidence is sufficient to prove a theory, and that it should be abandoned in light of any refuting evidence. Popper's theory presents a logical framework to hypothesis revision. In our example from Anne Frank, the hypothesis that 'Jewish people are lesser beings' should be rejected. But as we have seen people sometimes have difficulty abandoning hypotheses about plans of action that have been proved not to work, and they tend to abandon a refuted hypothesis when it belongs to somebody else more than when it is their own.

In contrast, Bayesian theorists propose that every piece of evidence leads to a small increase or decrease in belief in a hypothesis. Bayesian theories present a probabilistic framework to hypothesis revision. From this perspective the weaker the initial belief or evidence in favour of a hypothesis, and the stronger the evidence against it, the more likely people will revise it. In our example, people may simply decrease the degree of belief they have that 'all Jewish people are lesser beings' from 100% certainty to, say, 90% when they encounter Anne Frank (Howson & Urbach, 1993).

Future theories of hypothesis testing will need to address what logical and Bayesian frameworks can offer to an explanation of hypothesis revision (e.g., Taleb, 2007). For example, by examining if the strength of a refutation, the amount of refutation, or the amount of initial investment in a hypothesis prior to testing, we may be able to discern whether people consider one refutation as

enough to logically abandon a hypothesis or whether people tend to reason in a probabilistic way.

### **Conclusion**

In this thesis I have described a study of the factors which make hypothesis falsification possible in human reasoning. The results of the experiments suggest not only that people find falsification consistently possible, but that they have an active role in hypothesis testing by showing insight into the implications of their test choices counter to the main theories of hypothesis testing (i.e., Poletiek, 2001; Klayman & Ha, 1987). The results of our experiments suggest that the consideration of an alternative hypothesis, competing with an opponent hypothesis tester, and accessing expert knowledge facilitate hypothesis falsification. A new hypothesis testing theory may need to be developed if we are to address how these factors contribute to an explanation of how people test their hypotheses, and future experiments would need to address whether a logical or Bayesian framework is more appropriate for a new hypothesis testing theory to accommodate these results. Finally, these findings have important implications, not only because they reveal ways for people to overcome untrue hypotheses or false beliefs, but because they suggest that people may be more rational than previously thought.



## References

- Allport, G. W. (1979). *The nature of prejudice* (2<sup>nd</sup> Ed.). London: Addison-Wesley.
- Anderson, J. R. (1983). *The architecture of cognition*. Harvard: Harvard University Press.
- Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, 86, 124-180.
- Aronson, E. (1999). *The Social Animal*. New York: Worth/ W. H. Freeman.
- Baddeley, A. D. (1999). *Essentials of Human Memory*. Sussex, UK: Psychology Press.
- Beevor, A. (1998). *Stalingrad*. London: Viking.
- Britton, P. (1997). *The Jigsaw Man*. UK: Bantam Press.
- Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley and Sons.
- Burns, B. D. (2004). The effects of speed on skilled chess performance. *Psychological Science*, 15, 442-447.
- Byrne, R. M. J. (2005). *The Rational Imagination*. Cambridge, MA: MIT Press.
- Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, 40, 347-373.
- Byrne, R. M. J., & Mc Eleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1318-1331.
- Byrne, R. M. J., & Walsh, C. A. (2005). Resolving contradictions. In V. Girotto & P. N. Johnson-Laird (Eds.), *The Shape of Reason: Essays in honour of Paolo Legrenzi*, 91-105. Sussex, UK: Psychology Press.
- Byrne, R. M. J., & Tasso, A. (1994). Counterfactual reasoning: Inferences from hypothetical conditionals. In A. Ram & K. Eiselt, (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 124-120. Hillsdale, NJ: Erlbaum.
- Camerer, C. F. (2004). *Behavioral Game Theory: Experiments in strategic interaction*. Princeton NJ: Princeton University Press.
- Carnap, R. (1950). *Logical foundations of probability*. Chicago, IL: University of Chicago Press.
- Carroll, L. (1994). *Alice's adventures in wonderland*. London: Puffin books.
- Chabris, C. F., & Hearst, E. S. (2003). Visualisation, pattern recognition, and forward search: Effects of playing speed and sight of the position on

- grandmaster chess errors. *Cognitive Science*, 27, 637-648.
- Charness, N. (1991). Expertise in chess: The balance between knowledge and search. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits*, 39-63. Cambridge, England: Cambridge University Press.
- Chase, W. G., & Simon, H. A. (1973a). Perception in Chess. *Cognitive Psychology*, 4, 55-81.
- Chase, W. G., & Simon, H. A. (1973b). The mind's eye in chess. In W. G. Chase (Ed.), *Visual Information Processing*. New York: Academic Press.
- Cherubini, P., Castelvechio, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2-4-6 Task: An information theory approach. *The Quarterly Journal of Experimental Psychology*, forthcoming.
- Chevallier, A. (2016). Strategic thinking in complex problem solving. Oxford: Oxford University Press.
- Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in physics problem solving. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence (Vol. 1)*, 1-75. Hillsdale, NJ: Lawrence Erlbaum.
- Christensen-Szalansky, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 928-935.
- Cowley, M. (2016). Chess masters' hypothesis testing in games of dynamic equilibrium. SSRN eJournal Series Including: *Cognition in Mathematics, Science, & Technology eJournal*, vol. 8, Issue 2: Jan 12, 2016.
- Cowley, M. (2015). Hypothesis falsification in the 2-4-6 numbers test: Introducing imaginary counterparts. SSRN eJournal Series Including: *Cognition in Mathematics, Science, & Technology eJournal*, vol. 7, Issue 42: December 3, 2015.
- Cowley, M. (2006). The relevance of intent to human-android strategic interaction and artificial consciousness. *Proceedings, 15th International Conference on Robot-Human Interaction, IEEE*, 580-585. University of Hertfordshire, UK. SSRN e-library Classification Catalogue Topic 'Consciousness', 'Innovation & Cognitive Science', and 'Cognitive Neuroscience'.
- Cowley, M., & Byrne, R. M. J. (2005). When falsification is the only path to truth. In B. G. Bara. L. Barsalou, & M. Bucciarelli (Eds.). *Proceedings of*

- the Twenty-Seventh Annual Conference of the Cognitive Science Society*, 512-517. Mahwah, NJ: Erlbaum. Stresa, Italy.
- Cowley, M., & Byrne, R. M. J. (2004). Chess Masters' Hypothesis Testing. In K. D. Forbus, D. Gentner, & T. Rogers (Eds.). *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*. pp. 250- 255. Mahwah, NJ: Erlbaum. Chicago, USA. (Cognition & Neuroscience Seminar Series TCIN)
- Cowley, M. (2002). Confirmation bias as a default heuristic in novice chess players' problem solving. Undergraduate thesis: Trinity College Dublin.
- Cowley, S. J., & Macdorman, K. F. (2006) What Baboons, babies and Tetris players tell us about interaction: A biosocial view of norm-based social learning. *Connection Science*, 18(4), 363-378.
- Craik, K. (1943). *The nature of explanation*. Cambridge, England: Cambridge University Press.
- Crowley, P. H., & Zentall, T. R. (2013). Comparative decision making. Oxford: Oxford University Press.
- De Groot, A. D. (1965). *Thought and Choice in Chess*. The Hague: Mouton.
- De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioural sciences*. The Hague: Mouton.
- Dunbar, K. (2000). What scientific thinking reveals about the nature of cognition. In Crowley, K., Schunn, C. D., & Okada, T. (Eds.), *Designing for Science: Implications from Everyday, Classroom, and Professional Settings*. LEA. Hillsdale: NJ.
- Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85, 386-416.
- Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, 21, 419-460.
- Elo, A. (1978). The rating of chess players, past and present. New York: Arco.
- Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working Memory. *Psychological Review*, 102, 211-245.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Ericsson, K. A., Krampe, R. Th., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.
- Evans, J. St. B. T. (1989). *Bias in Reasoning*. Hove, UK: Erlbaum.

- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove, UK: Lawrence Erlbaum.
- Eysenck, H. J. (1995). *Genius*. Cambridge, UK: Cambridge University Press.
- Eysenck, M. W., & Keane, M. T. (2000). *Cognitive Psychology: A student's handbook*. Hove, UK: Psychology Press.
- Farris, H., & Revlin, R. (1989). The discovery process: A counterfactual strategy. *Social Studies of Science*, 19, 497-513.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Frank, A. (2001). *The diary of a young girl*. London: Penguin books.
- Franklin, S. P. (1998). *Artificial Minds*. Cambridge, MA: MIT Press.
- Frensch, P. A., & Sternberg, R. J. (1991). Skill related differences in game playing. In R. J. Sternberg (Ed.), *Complex Problem Solving: Principles and mechanisms*. USA: Lawrence Erlbaum Associates.
- Fugelsang, J., Stein, C., Green, A., & Dunbar, K. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, 58, 1392-1411.
- Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, MA: MIT Press.
- Gale, M., & Ball, L. J. (2003). Facilitation of rule discovery in Wason's 2-4-6 task: The role of negative triples. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, 438-443. Boston, MA: Cognitive Science Society.
- Gale, M., & Ball, L. J. (2005). Dual-goal facilitation in Wason's 2-4-6 task: What mediates successful rule discovery. *The Quarterly Journal of Experimental Psychology*, 59, 1-13.
- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic books.
- Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (1998). *Cognitive neuroscience: The biology of the mind*. London: W. W. Norton.
- Giroto, V., Legrenzi, P., Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78, 111-133.
- Gobet, F. (1998). Expert memory: A comparison of four theories. *Cognition*, 66, 115-152.
- Gobet, F., & Campitelli, G. (2002). *Education and chess: A critical review*. Forthcoming.

- Gobet, F., & Simon, H. A. (1996a). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology*, 31, 1-40.
- Gobet, F., & Simon, H. A. (1996b). The roles of recognition processes and look ahead search in time-constrained expert problem solving. Evidence from grandmaster chess. *Psychological Science*, 7, 52-53.
- Gobet, F., & Simon, H. A. (1996c). Recall of rapidly presented chess board positions is a function of skill. *Psychonomic Bulletin and Review*, 3, 159-163.
- Gobet, F., de Voogt, A., & Retschitzki, J. (2004). *Moves in mind: The psychology of board games*. Hove, UK: Psychology Press.
- Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.
- Gooch, S. (1981). *The Secret Life of Humans*. London, UK: J. M. Dent & Sons Ltd.
- Gorman, M. E. (1995a). Confirmation, disconfirmation and invention: The case of Alexander Graham Bell and the telephone. *Thinking and Reasoning*, 1, 31-53.
- Gorman, M. E. (1995b). Hypothesis Testing. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason*. Hove, UK: Laurence Erlbaum Associates Ltd.
- Gorman, M. E. & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory and a control strategy on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 36A, 629-648.
- Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory, and combined strategies affect group problem solving. *British Journal of Psychology*, 75, 65-79.
- Green, A. J. K., & Gilhooly, K. J. (1992). Empirical advances in expertise research. In M. T. Keane & K. J. Gilhooly (Eds.), *Lines of Thinking: Volume 2*. Chichester: Wiley.
- Gross, R. (1999). *Key studies in psychology*. London: Hoddon & Stoughton.
- Hale, B. (2008). *Philosophy looks at chess*. Chicago: Open Court.
- Harman, G. (1986). *Change in View*. Cambridge, MA: MIT Press.
- Hartston, W., & Wason, P. C. (1983). *The Psychology of Chess*. London: Batsford.
- Hedden, T., & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games. *Cognition*, 85, 1-36.



- Holding, D. H., & Reynolds, R. I. (1982). Recall or evaluation of chess positions as determinants of chess skill. *Memory and Cognition*, 10(3), 237-242.
- Holding, D. H. (1985). *The Psychology of Chess Skill*. New Jersey: Erlbaum, Hillsdale.
- Hollander, M., & Wolfe, D.A. (1999). *Nonparametric statistical methods*. New York: John Wiley & Sons.
- Howson, C., & Urbach, P. (1993). *Scientific reasoning* (2nd Ed.). Chicago: Open Court.
- Hsu, F., Campbell, M. S., & Hoane, A. J. (1995). Deep Blue system overview. *In Proceedings of The Ninth International Conference on Supercomputing, 240-244*. USA: ACM Press.
- Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678.
- Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from Inconsistency to Consistency. *Psychological Review*, 111, 3, 1-23.
- Kareev, Y., & Halberstadt, N. (1993). Evaluating negative tests and refutations in a rule discovery task. *Quarterly Journal of Experimental Psychology*, 46A, 715-727.
- Kerlinger, F. N. (2000). *Foundations of behavioural research* (3<sup>rd</sup> Ed.). London: Harcourt College Publishers.
- King, D. (1997). Kasparov v Deeper Blue: The ultimate man v machine challenge. London: Batsford.
- Klahr, D., & Dunbar, K. (1988). Dual search during scientific reasoning. *Cognitive Science*, 12, 1-55.
- Klayman, J., & Ha, Y-W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review*, 94, 2, 211-228.
- Klayman, J., & Ha, Y-W. (1989). Hypothesis Testing in Rule Discovery: Strategy, Structure and Content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15 (4), 596-604.
- Klein, S. W., Wolf, S., Militello, L., & Zsombok, C. (1995). Characteristics of skilled option generation in chess. *Organisational Behaviour and Human*

- Decision Processes*, 62, 63-69.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107-118.
- Koslowski, B. (1996). *Theory and Evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.
- Kotov, A. (1971). *Think like a grandmaster*. London: Batsford.
- Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard: Evidence from the Tower of Hanoi. *Cognitive Psychology*, 22, 143-183.
- Kruglanski, A. W., & Webster, D. M. (2000). Motivated Closing of the Mind: "Seizing" and "Freezing". In E. T. Higgins & A. W. Kruglanski (Eds.), *Motivational Science: Social and personality perspectives*, 354-375. USA: Taylor & Francis.
- Kuhn, T. S. (1993). *The Structure of Scientific Revolutions*. (3<sup>rd</sup> Ed.). Chicago: Chicago University Press.
- Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology*, 53, 636-647.
- Kunda, Z. (2000). The Case for Motivated Reasoning. In E. T. Higgins & A. W. Kruglanski (Eds.), *Motivational Science: Social and personality perspectives*, 313-335. USA: Taylor & Francis.
- Laing, R. D. (1999). *The Politics of the Family*. London: Routledge
- Lakatos, I. (1970). Falsification and methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of scientific knowledge*, 91-196. New York: Cambridge University Press.
- Larkin, J. H., Mc Dermott, J., Simon, D., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, 208, 1335-1342.
- Lesgold, A. M., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise is a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Luria, A. R. (1987). *The Mind of a Mnemonist* (2<sup>nd</sup> Ed.). Cambridge, MA: Harvard University Press.
- Mallie, E. (2001) *Endgame in Ireland*. London: Hoddon & Stoughton.
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and

- ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 70, 450-463.
- Manktelow, K. I. (1999). *Reasoning and Thinking*. Hove, UK: Psychology Press.
- McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, C. (1981). Knowledge organisation and skill differences in computer programmers. *Cognitive Psychology*, 13, 307-325.
- Milgram, S. (1963/1974). *Obedience to Authority*. New York: Harper Torchbooks.
- Miller, G. E.(1956).The magical number 7 plus or minus 2: Some limits on our capacity for processing. *Psychological Review*, 63, 81-97.
- Mitroff, I, (1974). *The subjective side of science*. Amsterdam: Elsevier.
- Molloy, A. M., & Scott, J. M. (2001). Folates and prevention of disease. *Public Health Nutrition (Review)*, 4(2b), 601-609.
- Mynatt, C. R., Doherty, M. E., & Dragon, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology*, 46A, 759-778.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29, 85-95.
- Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, 30, 395-406.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs NJ: Prentice-Hall.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Nixon, M. (2002). *The Little Oxford Thesaurus*. Oxford: Oxford University Press.
- Nunn, J. (1999). *Nunn's Chess Openings*. London: Everyman.
- Oaksford, M., & Chater, N. (1994). Another look at eliminative behaviour in a conceptual task. *European Journal of Cognitive Psychology*, 6, 149-169.
- Packiam-Alloway, T., & Alloway, R. G. (2013). *Working memory:The connected intelligence*. New York: Psychology Press.

- Peirce, C. S. (1992). *The essential Peirce, vol. 1*. In N. Houser, C. Kloesel, & the Peirce Edition Project. Bloomington, IN: Indiana University Press.
- Perfect, T. J., & Lindsay, D. S. (2014). *The Sage handbook of applied memory*. London: Sage.
- Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.
- Poletiek, F. H. (1996). Paradoxes of Falsification. *The Quarterly Journal of Experimental Psychology*, *49*, 447-462.
- Poletiek, F. H. (2001). *Hypothesis Testing Behaviour*. UK: Psychology Press.
- Poletiek, F. (2005). The proof of the pudding is in the eating: Translating Popper's philosophy into a model for testing behaviour. In K. I. Manktelow (Ed.). *Historical and Theoretical Perspectives on Reasoning*, forthcoming.
- Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
- Popper, K. R. (1963/1978). *Conjectures and Refutations* (4<sup>th</sup> ed.). London: Routledge and Kegan Paul.
- Popper, K. R. (1992). *Unended quest: An intellectual autobiography*. London: Routledge.
- Roese, N. J., & Olson, J. M. (1995). *What might have been: The social psychology of counterfactual thinking*. Hillsdale, NJ: Lawrence Erlbaum.
- Saariluoma, P. (1995). *Chess Players' Thinking*. UK: Psychology Press.
- Saariluoma, P., & Laine, T. (2001). Novice construction of chess memory. *Scandinavian Journal of Psychology*, *42*, 137-146.
- Simon, H. A., & Gilmarin, K. (1973). A simulation of memory for chess positions. *Cognitive Psychology*, *5*, 29-46.
- Simon, H. A., & Hayes, J. R. (1976). The understanding process: Problem isomorphs. *Cognitive Psychology*, *8*, 165-190.
- Sloboda, J. A. (1976). Visual perception of musical notation: Registering pitch symbols in memory. *Quarterly Journal of Experimental Psychology*, *28*, 1-16.
- Snyder, M., & Swann, W. B., Jr. (1978). Hypothesis-testing in social interaction. *Journal of Personality and Social Psychology*, *36*, 1202-1212.
- Sperber, D., & Mercier, H. (2010). Reasoning as a Social Competence. In H. Landmore, & J. Elster (Eds.). *Collective Wisdom*. UK: Cambridge University Press.
- Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. USA: Penguin.
- Tough, P. (2013). *How children succeed: Grit, curiosity, and the hidden power*

- of character. UK: Random House.
- Tukey, D. D. (1986). A philosophical and empirical analysis of subject's modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 38, 5-33.
- Tversky, A., & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, 201-208. New York: Cambridge University Press.
- Tweney, R. D. (1989). Fields of enterprise: On Michael Faraday's thought. In D. Wallace & H. Gruber (Eds.), *Creative people at work: Twelve cognitive case studies*, 91-106. Oxford: Oxford University Press.
- Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (1981). *On Scientific Thinking*. New York: Columbia University Press.
- Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A. & Arkkelin, D. L. (1980). Strategies of rule discovery on an inference task. *Quarterly Journal of Experimental Psychology*, 32, 109-123.
- Vallee-Tourangeau, F., Austin, N. G., & Rankin, S. (1995). Inducing a rule in Wason's 2-4-6 task: A test of the information-quantity and goal complementarity hypotheses. *Quarterly Journal of Experimental Psychology*, 48A, 895-914.
- Van der Henst, J. B., Rossi, S., & Schroyens, W. (2002). When participants are not misled they are not so bad after all: A pragmatic analysis of a rule discovery task. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, 902-907. Mahwah, NJ: Erlbaum.
- Van Someren, M. B., & Sandberg, J. (1994). *The Think-aloud Method*. London: Academic Press.
- Vygotsky, L. S. (1986). *Thought and language* (A. Kozulin, Ed. ). Cambridge, MA: MIT Press.
- Wagenaar, W. A., van Koppen, P. J., & Crombag, H. F. (1993). *Anchored narratives: The psychology of criminal evidence*. London: Harvester Wheatsheaf.
- Walsh, C. A. (2001). *The role of context in counterfactual thinking*. Unpublished PhD Thesis. School of Psychology, Trinity College, University of Dublin.
- Walsh, C. A., & Johnson-Laird, P. N. (2005). Changing your mind. *In submission*.

- Wason, P. C. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129-140.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273-281.
- Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.
- Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.
- Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology*, *14*, 246-249.
- Wharton, C. M., Cheng, P. W., & Wickens, T. D. (1993). Hypothesis-testing strategies: Why two goals are better than one. *Quarterly Journal of Experimental Psychology*, *46A*, 743-758.
- Zhang, J., & Hedden, T. (2003). Two paradigms for depth of strategic reasoning in games: Response to Colman. *Trends in Cognitive Sciences*, *7*, 4-5.

## List of Figures

<b>Figure</b>	<b>Title</b>	<b>Page</b>
1.1	Embedded relationships between a participant's hypothesis (H) and the experimenter's rule (True Rule).	
3.1	Percentages of positive and negative tests generated in Experiment 4.	
3.2	Percentages of positive and negative tests generated by participants for their own hypotheses when an opponent hypothesis tester was absent or present.	
4.1	A representation of a chess board middle game, in which it is white to play.	
4.2	A section of a problem behaviour graph constructed from the chess master's protocol.	
4.3	Grandmaster (participant 4), normal position 1, black.	
4.4	Novice (participant 12), normal position 1, black.	
5.1	Embedded low quality hypotheses in the 2-4-6 task and chess.	





## List of Tables

<b>Table</b>	<b>Title</b>
<i>Page</i>	
1.1	Confirming and falsifying test types in the 2-4-6 task for the hypothesis 'even numbers ascending in twos'.
1.2	The different ways hypothesis testing strategies have been conceptualised in hypothesis testing research over the past forty-five years.
1.3	Tenets of the uniformity theory (Poletiek, 2001)
1.4	Tenets of the mathematical relationship theory (Klayman & Ha, 1987; 1989).
2.1	The number of triples generated in each condition of experiment 1.
2.2	The percentage of confirming and falsifying triples generated for high and low quality hypothesis when quality type was known or unknown.
2.3	Percentages of confirming and falsifying positive and negative test types generated in experiment 1.
2.4	The number of triples generated in for each type of alternative hypothesis quality in experiment 2.
2.5	The percentages of participants who correctly announced that Peter's hypothesis was incorrect and the percentages who subsequently discovered the experimenter's rule.
2.6	The percentage of confirming and falsifying triples when the alternative hypothesis was high quality, medium quality and

- low quality.
- 2.7 The percentages of positive and negative confirming and falsifying triples generated when the alternative hypothesis was high quality, medium quality and low quality.
  - 2.8 The percentages of participants who correctly announced that Peter's hypothesis was incorrect and the percentages who subsequently discovered the experimenter's rule.
  - 2.9 The percentages of confirming and falsifying triples when the alternative hypothesis was explicit, non-explicit and when there was no alternative.
  - 2.10 The percentages of positive and negative confirming and falsifying triples generated when the alternative hypothesis was explicit, non-explicit, and when there was no alternative.
  - 3.1 Percentages of hypothesis test types generated in each condition of experiment 4.
  - 3.2 Percentages of abandoned and endorsed hypotheses in each condition of experiment 4.
  - 3.3 Percentages of hypothesis test types generated in each condition of experiment 5.
  - 3.4 Percentages of abandoned and endorsed hypotheses in each condition of experiment 5.
  - 4.1 The nine possible hypothesis types based on the subjective and objective evaluations of move sequences.
  - 4.2 An example of a segmented expert protocol. This protocol corresponds to the fourth master problem-behaviour graph in the Appendix F.

- 4.3 The mean number of different types of hypothesis tests for the complete move sequences for the normal and random board positions by the masters and novices in Experiment 6.
- 4.4 The mean number of positive, negative and objective tests by the masters and novices for normal and random board positions (with standard errors in parentheses) in Experiment 6.
- 4.5 The mean ply depth and ply breadth of move sequences, and the mean number of individual moves in the generated move sequences in Experiment 6, with standard deviations in brackets.
- 4.6 The mean number of different sorts of move sequences generated by the five masters and five novices in Experiment 6 for the normal and random board positions (standard deviations are in parenthesis).

## **List of Appendices**

Appendix A (i): Materials used in Experiment 1

(Title: Falsification of a low-quality hypothesis)

Appendix A (ii): Recording sheet used in the 2-4-6 experiment

Appendix B: Materials used in Experiment 2

(Title: Quality of alternative hypotheses)

Appendix C: Materials used in Experiment 3

(Title: Explicit and non-explicit alternative hypotheses)

Appendix D: Materials used in Experiment 4

(Title: Hypothesis ownership)

Appendix E: Materials used in Experiment 5

(Title: An opponent hypothesis tester)

Appendix F: Materials used in Experiment 6

(Title: A protocol analysis of hypothesis testing by chess masters)

Appendix G: Set up for chess computer program *Fritz 8*

Appendix H: Segmented protocols of chess players' thinking

Appendix I: Selected problem behaviour graphs of chess players' thinking

Appendix J: The experimenter's think aloud script used in Experiment 6

Appendix K: The experimenter's recording sheet used in Experiment 6

Appendix L: Ethics approval for the experiments

## **Appendix A(i): Materials used in Experiment 1**

### **(Title: Falsifying a low-quality hypothesis)**

#### **Instructions (page 1, common to all the 2-4-6 experiments carried out in the thesis)**

Thank you for participating in this study. This study is interested in the kinds of strategies people use when thinking. It should take about 10 to 15 minutes to complete. A full explanation of the main aims of the study will be provided at the back of your booklet. Please do not look at this until you have finished. You may also ask the researcher to answer any further questions you may have at the end. Thank you so much for your time.

For the purpose of the study please write your gender (male = m, or female = f) age and the date in the spaces provided.

Gender: \_\_\_\_\_

Age: \_\_\_\_\_

Date: \_\_\_\_\_

#### **The scenario instructions used in Experiment 1 (page 2)**

(\*Please note that the term ‘researcher’ was used in place of the term ‘experimenter’ in the participants’ instructions because it was a more neutral term.)

Hypothesis quality

High-quality hypothesis: *‘any ascending numbers’*

Low-quality hypothesis: *‘even numbers ascending in twos’*

Knowledge of hypothesis quality

The additional sentence given at the end of the first paragraph to give participants knowledge about the quality of their hypothesis in the known conditions was:

*The researcher’s rule is in fact ‘any ascending numbers’.*

First paragraph (common to all conditions)

In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the researcher's rule was: '\_\_\_\_\_'. (*The additional sentence – The researcher's rule is in fact 'any ascending numbers' - was placed here for the known conditions*)

Second paragraph (common to all conditions)

Your aim is to go about testing if Peter's rule '\_\_\_\_\_' is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if Peter's rule '\_\_\_\_\_' is the researcher's rule and not to test any ideas of your own that you think the researcher's rule might be.

Last paragraph (unknown conditions)

You should try to go about testing if Peter's rule '\_\_\_\_\_' is the rule the researcher has in mind by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if Peter's rule is the researcher's rule, *and not before*, you are to write down 'Peter's rule is the researcher's rule' or 'Peter's rule is not the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

Last paragraph (known conditions)

The word *not* was placed where indicated in the last paragraph in the condition when the participants knew that their hypothesis was low-quality and not the researcher's rule.

You should try to go about testing Peter's rule '\_\_\_\_\_' in a way that would help him discover that his rule is (*not*) the researcher's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is (*not*) the researcher's rule, *and not before*, you are to write down 'Peter now knows his rule is (*not*) the researcher's rule'. You are to write this under your most recent

number sequence. The experimenter will then write whether or not you are correct beside your announcement.

**Check questions (page 3, common to all 2-4-6 experiments carried out in this thesis)**

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

**or**

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.

**Debriefing paragraph (page 3)**

Thank you once again for taking part. This study intended to examine the way people test their hypotheses and ideas about incidences and relationships in the world around them. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false. This study was interested in the extent to which people followed confirmation or falsifying strategies when thinking about how to test another person's hypothesis. If you have further questions feel free to ask the researcher. If you would like to request the overall result of the study please use details on the contact information sheet.

**Appendix A (ii): Recording sheets**

				<b>*Feedback from experimenter</b>
Number sequence	Reasons for choice	Do you expect it to conform to Peter's rule	Do you expect it to conform to the researcher's rule	Does your number sequence conform to the researcher's rule
2,4,6	...	yes	yes	y

---

--	--	--	--	--

Recording sheet (18 lines per participant, common to all conditions in which participants tested Peter's hypothesis).

				<b>*Feedback from experimenter</b>
Number sequence	Reasons for choice	Do you expect it to conform to your rule	Do you expect it to conform to the researcher's rule	Does your number sequence conform to the researcher's rule
2,4,6	...	yes	yes	y

---

--	--	--	--	--

Recording sheet (18 lines per participant, common to all conditions in which participants tested their own hypothesis).



## Appendix B: Materials used in Experiment 2

### (Title: Quality of alternative hypotheses)

#### The scenario instructions used in Experiment 2 (page 2)

Hypothesis under test (common to all conditions)

*Peter's hypothesis:* 'even numbers ascending in twos'

Hypothesis quality of alternative hypotheses (experimental conditions)

*High-quality hypothesis:* 'any ascending numbers'

*Medium quality hypothesis:* 'numbers ascending in twos'

*Low-quality hypothesis:* 'even numbers ascending in twos that end in the digits 2,4,6'

First paragraph (experimental conditions)

In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence **2,4,6** conforms to. Peter hypothesised that the researcher's rule was: 'even numbers ascending in twos'. You know that another participant called James hypothesised that the researcher's rule was '\_\_\_\_\_'.

Second paragraph (experimental conditions)

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

Last paragraph (experimental conditions)

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the researcher's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the researcher's rule, *and not before*, you are to write down 'Peter now knows his rule is the researcher's rule' or 'Peter now knows his rule is not

the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

Control condition (the rule is known: a replication of the high-quality known condition in Experiment 1)

Knowledge of hypothesis quality

The additional sentence given at the end of the first paragraph to give participants knowledge about the quality of their hypothesis in the known conditions was:

*The researcher's rule is in fact 'any ascending numbers'.*

First paragraph (control condition)

In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the researcher's rule was: 'even numbers ascending in twos'. (*The additional sentence for the known conditions was placed here*).

Second paragraph (control condition)

Your aim is to go about testing if Peter's rule 'even numbers ascending in twos' is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if Peter's rule 'even numbers ascending in twos' is the researcher's rule and not to test any ideas of your own that you think the researcher's rule might be.

Last paragraph (control condition)

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is not the researcher's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is not the

researcher's rule, and not before, you are to write down 'Peter now knows his rule is not the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

**Check questions (page 3)**

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

**or**

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.

**Debriefing paragraph (page 3)**

The 2-4-6 study aims to examine the way people test their hypotheses about incidences and relationships in the world around them. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false. This study was interested in the extent to which people followed confirming or falsifying strategies when thinking about how to test another person's hypothesis. If you have any further questions, the experimenter will be happy to discuss them with you.

## **Appendix C: Materials used in Experiment 3**

### **(Title: Explicit and non-explicit alternative hypotheses)**

#### **The scenario instructions used in Experiment 3 (page 2)**

Explicitness of alternative hypothesis

*Explicit alternative hypothesis:* You know that another participant called James hypothesised that the researchers' rule was 'any ascending numbers'

*Non-explicit alternative hypothesis:* You know that another participant called James hypothesised that the researcher's rule was 'something else'.

*No alternative hypothesis:* no alternative hypothesis was given

First paragraph (common to all conditions)

In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the researcher's rule was: 'even numbers ascending in twos'. (*The relevant alternative hypothesis was placed here for each condition*).

Second paragraph (common to all conditions)

Your aim is to go about testing Peter's rule 'even numbers ascending in twos' in a way you think would help him to discover if his rule is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if they conform or do not conform to the rule the researcher has in mind.

Final paragraph (common to all conditions)

You should try to go about testing Peter's rule 'even numbers ascending in twos' in a way that would help him discover that his rule is or is not the researcher's rule by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have helped Peter discover that his rule is or is not the researcher's rule, *and not before*, you are to write down 'Peter now knows his rule is the researcher's rule' or 'Peter now knows his rule is not the researcher's rule'. You are to write this under your most recent number

sequence and raise your hand. The experimenter will then write whether or not you are correct beside your announcement.

**Check questions (page 3)**

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

**or**

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.

**Debriefing paragraph (page 3)**

The 2-4-6 study aims to examine the way people test their hypotheses about incidences and relationships in the world around them. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false. This study was interested in the extent to which people followed confirming or falsifying strategies when thinking about an alternative hypothesis. If you have any further questions please feel free to ask the experimenter.

**Appendix D: Materials used in Experiment 4**  
**(Title: Hypothesis ownership)**

**The scenario instructions used in Experiment 4 (page 2)**

*Hypothesis owned by the imaginary participant: 'even numbers ascending in twos'*

*Hypothesis owned by the participant themselves: 'even numbers ascending in twos'*

*Instructions: Hypothesis owned by the imaginary participant*

First paragraph

In a previous study investigating human thinking a participant called Peter was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. Peter hypothesised that the researcher's rule was: 'even numbers ascending in twos'.

Second paragraph

Your aim is to go about testing if Peter's rule 'even numbers ascending in twos' is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if Peter's original rule 'even numbers ascending in twos' is the researcher's rule, and not to test any new ideas of your own that you think the researcher's rule might be.

Final paragraph

You should try to go about testing if Peter's rule 'even numbers ascending in twos' is the rule the researcher has in mind by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if Peter's rule is the researcher's rule, *and not before*, you are to write down 'Peter's rule is the researcher's rule' or 'Peter's rule is not the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

*Instructions: Hypothesis owned by the participant themselves*

First paragraph

In a previous study investigating human thinking you were a participant who was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. You hypothesised that the researcher's rule was: 'even numbers ascending in twos'.

Second paragraph

Your aim is to go about testing if your rule 'even numbers ascending in twos' is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if your original rule 'even numbers ascending in twos' is the researcher's rule, and not to test any new ideas of your own that you think the researcher's rule might be.

Final paragraph

You should try to go about testing if your rule 'even numbers ascending in twos' is the rule the researcher has in mind by citing as few number sequences as you can. Please note that you have three pages on which to test your number sequences if you need to. When you feel highly confident that you have discovered if your rule is the researcher's rule, *and not before*, you are to write down 'My rule is the researcher's rule' or 'My rule is not the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

**Check questions (page 3)**

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

**or**

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.

**Debriefing paragraph (page 3)**

The 2-4-6 study aims to examine the way people test their hypotheses about incidences and relationships in the world around them. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false. This study was interested in the extent to which people followed confirming or falsifying strategies when thinking about how to test another person's hypothesis. If you have any further questions please feel free to contact the experimenter.



## **Appendix E: Materials used in Experiment 5**

### **(Title: An opponent hypothesis tester)**

#### **The scenario instructions used in Experiment 5 (page 2)**

##### *Instructions: participants do not consider an opponent hypothesis tester*

###### First paragraph

In a previous study investigating human thinking you were a participant who was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. You hypothesised that the researcher's rule was: 'even numbers ascending in twos'.

###### Second paragraph

Your aim is to go about testing if your rule 'even numbers ascending in twos' is the researcher's rule. You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if your original rule 'even numbers ascending in twos' is the researcher's rule, and not to test any new ideas of your own that you think the researcher's rule might be.

###### Final paragraph

When you feel highly confident that you have discovered if your rule is the researcher's rule, *and not before*, you are to write down 'My rule is the researcher's rule' or 'My rule is not the researcher's rule'. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct beside your announcement.

##### *Instructions: participants consider an opponent hypothesis tester*

###### First paragraph

In a previous study investigating human thinking you were a participant who was asked to discover a rule a researcher had in mind that the number sequence 2,4,6 conforms to. You hypothesised that the researcher's rule was: 'even numbers ascending in twos'.

### Second paragraph

Your aim is to go about testing if your rule ‘even numbers ascending in twos’ is the researcher’s rule. However, an opponent participant called Peter is also testing if your rule ‘even numbers ascending in twos’ is the researcher’s rule. You must discover if your rule ‘even numbers ascending in twos’ is the researcher’s rule before the opponent participant Peter does.

### Final paragraph

You are to do this by writing down other number sequences with sets of three numbers. You will then be informed if these number sequences conform or do not conform to the rule the researcher has in mind. Please remember your aim is specifically to test if your original rule ‘even numbers ascending in twos’ is the researcher’s rule, and not to test any new ideas of your own that you think the researcher’s rule might be. When you feel highly confident that you have discovered if your rule is the researcher’s rule, *and not before*, you are to write down ‘My rule is the researcher’s rule’ or ‘My rule is not the researcher’s rule’. You are to write this under your most recent number sequence. The experimenter will then write whether or not you are correct and whether or not you have discovered if your rule is the researcher’s rule before the opponent participant Peter does.

**Check questions (page 3)**

For the purpose of the study please circle yes or no where applicable below if you have ever done a problem like this before Yes / No,

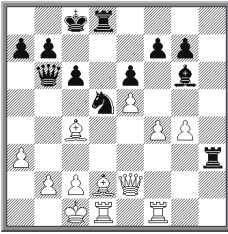
**or**

if you have taken courses dealing with the concepts of confirmation and falsification in the past Yes / No.

**Debriefing paragraph (page 3)**

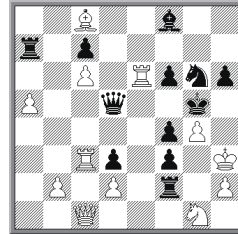
Thank you once again for taking part. This study intended to examine the way people test how accurate their ideas about the world around them are. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false, even if their ideas are incorrect. This study was interested in the extent to which people followed confirming or falsifying strategies when competing about how to test their own idea as opposed to another person's idea. If you have further questions please ask the researcher.

**Appendix F: The three normal board positions and three random board positions used in the experiment.**



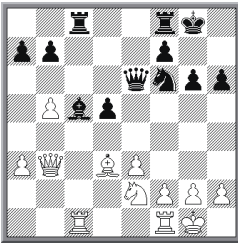
**Position 1**  
(normal)

**White to play** (Rook on h3 is on h8 when black to play)



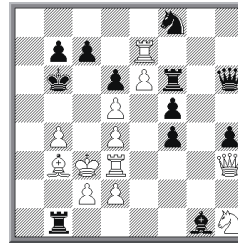
**Position 4**  
(random)

**White to play** (black pieces were transposed on white piece coordinates when black to play)



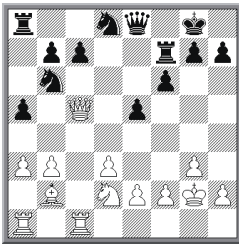
**Position 2**  
(Normal)

**White to play** (Pawn on h6 is on h7 when black to play)



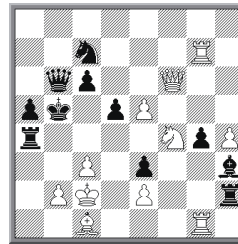
**Position 5**  
(Random)

**White to play** (black pieces were transposed on white piece coordinates when black to play)



**Position 3**  
(Normal)

**White to play** (Knight on d8 is on c6 when black to play)



**Position 6**  
(Random)

**White to play** (black pieces were transposed on white piece coordinates when black to play)

## Appendix G: Set up for chess program *Fritz 8*

The *Chessbase* analysis engine *Fritz 8* was used to estimate objective evaluations of chosen move sequences. Each node in the sequence was evaluated. These evaluations were essential in comparing chess players' expected outcomes with realistic outcome expectancies. The program has the playing strength of at least a world champion candidate grandmaster. The following procedures were followed for the computerised evaluation:

### a) Maximising speed of evaluation

- 1) The hash table size was increased to a power of two, and was set at 256 MB to speed up the processor
- 2) All multimedia functions such as 'talking to you while you play' were switched off so as not to slow down the processor thus maximising accuracy of evaluation and reducing horizon effects
- 1) The openings book database was set at optimum strength
- 2) The number of lines of play considered in parallel was increased from two to three lines
- 3) Analysis was a function of ensuring at least a ply depth of 11 for each examined line using the engine's '*infinite analysis*' mode. This analysis was used to evaluate in place of a ply depth setting alone so as to counter any horizon effects likely to occur after 11ply (This analysis was undertaken by recommendation of a correspondence with professional chess grandmaster, Elo 2578. Although Chabris & Hearst, 2003 set their analysis to 10ply, and this is still very precise). The infinite analysis mode reduces horizon effects as it explores what it identifies as critical lines to a greater ply depth than others. Once the numerical evaluation stops fluctuating during the infinite analysis it is a sign that a precise estimate of the true evaluation has been reached. Estimated evaluations are expressed in 1/100ths of a pawn. The evaluation had to have at least stopped fluctuating for 15secs when the engine was processing positions at a speed of at least 750 k/Ns.

## Appendix H: Segmented protocol transcripts for complete set of think-aloud data

Masters: S1, S3, S4, S8, S17

Novices: S2, S11, S12, S13, S20

### *Segmented protocol transcripts for normal positions*

S3 Position 1: W

1. Em, the position looks so much better for white,
2. em maybe because of his adv on the k-side,
3. em aggressive moves like f5
4. are looking quite promising.
5. Eh apart from f5 I don't seem to have that many ideas,
6. eh, I can't see anything really against f5 at the moment.
7. No, f5 seems to be critical,
8. em exf5
9. gxf5
10. Bh5
11. em Qg2,
12. em Bxd1
13. Qxh3,
14. I've got a lot of attractive options.
15. His B has massive trouble getting out,
16. em in fact I don't think that I can,
17. all the squares seem to be covered.
18. So after f5
19. exf5
20. gxf5
21. Qg2
22. more critical for him is a move like h4
23. attacking Bc4
24. and keeping his b attacked on d1.
25. Bg5 would skewer his two rooks
26. but em to take on c4

27.doesn't seem to be too bad.  
28.Em, ok I'm looking at Rh1 now  
29.because I want to,  
30.if I can get in f5 without any of these tricks  
31.then I should definitely be clearly better in this position.  
32.Em I think Rh5  
33.Rgh8  
34.ah doesn't seem to be a massive thing for me.  
35.I could probably play Qg2  
36.Rh1  
37.Rdh8  
38.Qg8  
39.em when I get the h-file  
40.possibilities of penetrating  
41.doesn't seem to be too much he can do about that  
42.if he Rh1  
43.Rdh8  
44.Qg2  
45.Rxh1  
46.Qxh1  
47.then Qd8  
48.to prevent Qa8.  
49.It's not quite clear what I've accomplished,  
50.I can't play f5 in for the moment.  
51.So, are these the right ideas?  
52.Ne3 looks like one of his ideas if I don't do anything,  
53.em forcing me to take on e3 em,  
54.that position doesn't look too clear,  
55.doesn't look too good for me anyway.  
56.Em so I need to...

S17 Position 1: W

1. Ok, em, this is a position I would have regularly played as black,
2. em and I don't really know why I used to.
3. Em, basically white has the advantage of extra space

4. and I would regard the two bishops as being a big advantage.
5. Em, and white's only problem really is that the pawn structure is a wee bit stuck,
6. and that ... em black has well controlled place on the h-file.
7. The ideas I'd be looking for... at play for white here.
8. Actually there's a few initial candidate moves.
9. I would consider playing Qe1
- 10.with the idea of blocking Bh5.
- 11.I'd ... look at trying to make the tactic of playing f5...
- 12.doesn't work right now
- 13.as after exf5
- 14.gxf5
- 15.Bh5.
- 16.Oh...that actually might work with the ...
- 17.probably play Qg2.
- 18.If white achieves that then white has again fluid pawn movement,
19. which would be good for the two bishops.
- 20.Eh... and an-another kind of normal idea in this position is just to eh,
- 21.playing in c4
- 22.which removes that N
- 23.which is well placed.
- 24.Em, and that would have the advantage of opening the d-file.
- 25.And black would be a bit vulnerable on the d-file
- 26.as his rook is already committed to playing on the open h-file...
- 27.Now the one idea that looks as though it would cut his bishop off is f5...
- 28.So f5
- 29.exf5
- 30.gxf5
- 31.Bh5
- 32.Qg2
- 33.R...
- 34.N black can play,
- 35.hmm, that's interesting...



S8 Position 1: W

1. Em, ok let's have a look.
2. Looks like Caro-Kann position.
3. Em yea, let's check the material first,
4. yeah ok equal,
5. black has got the h-file.
6. We might be able to break through with f5
7. although there's a tactic...
8. aha not really
9. f5
- 10.exf5
- 11.gxf5
- 12.Bh5
- 13.Qg2
- 14.so Bxd1
- 15.Qxh3
- 16.and that looks very good.
- 17.Any other ideas,
- 18.Qg2 first, ....
- 19.Anything here for black
- 20.Rxa3,
- 21.no not really a threat,
- 22.c2 is covered,
- 23.Nc3
- 24.no I don't think I believe in all this kind of stuff.
- 25.Ok Qg2
- 26.or f5 seem to be the moves.
- 27.Hmm, could also think of taking on d5
- 28.Bxd5
- 29.cx
- 30.no that's counter play.
- 31.Em yea, f5
- 32.is there any threat at all?
- 33.I can't see one there
- 34.ok Ne3

35.Qx  
36.Qx  
37.f5...

S4 Position 1: B

1. So, I'm black in this position.
2. It's some kind of Alekhine Defense,
3. Caro-Kann something in that line,
4. black to play
5. and white has immediate threat f5
6. at the same time all of my pieces seem to be ok,
7. except maybe for the Bg6.
8. So I now need to find a way to stop f5,
9. I have semi-open file h,
10. which could be leading somewhere,
11. em I can stop f5 by playing Ne7
12. but that could lead to Bb4
13. threatening to,
14. well obviously wanting to get the N in
15. or planting the B on d6.,
16. probably need to play c5,
17. which I can do if I have to if...
18. or next...
19. now I think this is ...
20. I don't see any other active moves at the moment...
21. nothing which comes to mind really.
22. f5 is a big threat,
23. it would be very unpleasant
24. and get the bishop to move away.
25. I guess my ... Nd5...
26. ah ok, so Ne7,
27. what comes next?
28. So after Ne7
29. f5
30. I don't have to worry about it...

31. ideally I would like to see the bishop moving to e4
- 32....d5
33. maybe I get the queen before that.
34. now, there is a problem with Ne3
35. which is Be3
36. probably because then I need now move out and get some...
37. rid of p on a-file.
38. c5 will then be much risky
39. so maybe b4 is possible
40. Q... oh right...
41. anything else here,
42. em not really
43. ok f5 for the moment is not a big threat
44. because I can take on f5
45. px f5
46. then Bh5
47. win exchange
48. so not a big deal.
49. Now let me think,
50. got one move when I can do something.
51. What I want to do is not really clear to me...

#### S1 Position 1: W

1. First the position is kind of a caro-kann em type of position...
2. I would eh, like to ,
3. the knight in the former caro-Kann Bf5 variation...
4. leaving the B on g6
5. em which is obvious place ...
6. but well outside of the pawn chain
7. which is supporting condensing the black pawn configuration of pawns on the light squares at e6, c6.
8. It is important to discern...that the black position ...em... eh is based around two ideas.
9. They may involve Nd5
10. as the strength of it as a piece

11. which has very large em...control of the board
12. controlling indeed the whole centre...
13. and eh the black bishop on g6
14. which eh is looking towards the white k on c2
15. em it seems to be a more or less evenly balanced position...
16. it's difficult to see how black is actually going to make any progress here
17. when white probably has more chances
18. due to the pair of bishops
19. em...the centrally placed rooks
20. and the threat of f5
21. which could prove annoying.
22. If black is forced to go Bh7 there ...
23. then eh the bishop is very much out of play...
24. and eh then white possibly move his own Bd3
25. and perhaps with preparation play c4
26. kicking the N back
27. always gaining space...
28. em and then once the N is kicked back
29. the Be4
30. pc4
31. somehow manage to play his Bb4...
32. em attacking the dark squares
33. or weak squares...
34. d6...
35. and if black supports the play of b5
36. then eh c4
37. might leave the rook more exposed
38. eh... firstly white eh... think white has to be a bit careful about playing f5 too early
39. as particular f5
40. px
41. px
42. Bh5 em
43. which pins the eh... which pins the eh the queen or the em queen and rook
44. although it might be even possible now straight off is to play px

45.px  
 46.Bh5  
 47.Qg2...  
 48.and ...blacks rook has become somewhat embarrassed  
 49.with QxR the black pawn...  
 50.or the black bishop is takes on d1  
 51.has nowhere to go  
 52.and white also has the idea of b6  
 53.or f6+ f6 discovered check  
 54.so win in the house.  
 55.f5 seems to be the critical move...  
 56.objectively white has possibly a set of chances here.  
 57.Black has got a superficially good position.  
 58.An active Rh3  
 59.but I don't see how he can make any progress against the white king  
 60.due to the fact that the queenside is very well consolidated by the two  
 bishops...  
 61.it's eh more than likely.  
 62.The course of play...

### S3 Position 2: W

1. Ok em, isolated d pawn position,
2. equal material,
3. opposite coloured bishops,
4. em basically black's got reasonably active pieces
5. but white has definitely got the better position due to the structural weakness,
6. Nf4 attacks Queen
7. and pawn on d5,
8. em seems to lead to quite a favourable position for white.
9. Nf4
- 10.Qd6
- 11.Rfd1 seems to be logical
- 12.oh but the pa3 hanging in that line,
13. em so I don't really want to analyse that
- 14.because letting him get my a3 pawn for d5

15. which is a big let off for him.
16. Em, so do I have any other moves apart from Nf4 in this position?
17. Em Nd4 is possible
18. but not very good.
19. He can take it for one Bxd4
20. after which the structural defect is nullified
21. because I have an isolated p too.
22. But that position might be slightly in my favour as well
23. because my B is strong
24. but he doesn't have to take it.
25. So eh I'm looking at a4,
26. a4 is prophylactic but,
27. it doesn't do much else otherwise.
28. Em, maybe maybe if I played Rfd1 immediately,
29. then if Qd6
30. just a4 there...
31. Ba3
32. eh so is a little bit annoying
33. but Rxc8
34. Rxc8,
35. yeah getting in Q as well.
36. I definitely feel like I should play a move
37. just to cement my advantage here.
38. Em it's not quite clear what it is.
39. His pieces are all defended em,
40. Rfd1
41. Qd6 hmm,
42. maybe Rc3
43. in that position black is better,
44. clearly mobilised his Ne4 there.
45. Rfd1
46. Qd6,
47. yea my king,
48. I'm beginning to like my position a bit less than I did at the outset.
49. Let's think
50. It's quite difficult to cement an advantage here...

## S17 Position 2: B

1. [Coughs]. Ok, well we have an isolated queen's pawn position (IQP)...
2. which normally means that black should be looking to try em...
3. black has the isolated queen pawn
4. and should be trying to get some activity
5. and normally in these positions you don't like to have swapped off too many pieces,
6. ...and there are a couple of minor pieces missing
7. which means that I'm not happy.
8. Which means that em I'm looking perhaps at trying to equalise in this position.
9. Em, the advantages for trying to equalise includes the fact that the bishops are of opposite colours
10. and I think that em white has probably erred
11. em...by pushing the queenside pawns too far.
12. Under those kind of circumstances...em,
13. the moves that I would like to play would be...
14. initially candidate moves would be Bb6,
15. Qe5,
16. and perhaps just Qd7.
17. The ideas I would like to pursue in this position would be to gain control over the d4 square
18. so I could...pushing the pawn.
19. The other idea
20. which Qe5 kind of helps...
21. is that it allows Ne4
22. which is a good place for the N to go to.
23. It's not not really in white's favour to exchange
24. so Ne4
25. Bxe4
26. pxe4 in that position
27. 'cause then em...I've got rid of my IQP
28. and em...I, I've got rid of minor pieces

29.and got a bit more space...  
 30.so actually Qe5  
 31.looks like a fairly decent move...  
 32.what kind of things can white do against that?  
 33.White would probably , would like to play maybe Qc3...  
 34.and so I would...try bring in Bb6 first  
 35....white...  
 36.oh yea the a3 pawn is under attack  
 37.with Qd6 as well.  
 38.Qd6  
 39.forces a4  
 40.and then I can just play Na4  
 41.or ...Qd6  
 42.a4  
 43.I can play d4...d4.  
 44.there is a lot to think about here.  
 45.Qd6  
 46.a4  
 47.Ng4  
 48.is there compensation developing there?  
 49.There's actually some very decent play developing...  
 50.ah now I'm beginning to see some nice ideas for...for black  
 51....Nd4 just of its self,  
 52.em threatening taking on e3.  
 53.Threatening Qe5...  
 54.Ok...Qe5  
 55.Qc3  
 56.then that's the end of the line really,  
 57.of no actually I can't...

S8 Position 2: W

1. Gosh, again a middle game position.
2. I wish we would have an endgame.
3. Ok let's have a look here
4. 2-4-6, 2-4-6 ok, yea material is equal,



5. hmm b5 looks already very committal,
6. ok so isolated pawn
7. but it doesn't look too bad.
8. Yea there are some tactics in the position.
9. In principle I'd say black looks quite healthy,
10. em unless the tactic works,
11. Nf4 Nf4 hmmm,
12. I don't really like that move
13. I think it's better to kind of probably play something like Bb1
14. with the idea of doubling on the d-file
15. and putting Nd4
16. and then later on try to attack pd5.
17. what I don't like about white's position is that b5 has already been made
18. and Bc5 looks very strong.
19. Does d4 work?
20. No it doesn't,
21. ok that's still pinned.
22. e4 doesn't work.
23. Nf4
24. just attacks the Q
25. doesn't do anything.
26. Yea I think Bb1 looks like the move I would play in this position.
27. Rd1 first possible
28. but might be more,
29. might safe to protect pe3 with the Q first.
30. Hmmm after h4
31. no that's too slow.
32. Any attacks on g6?
33. No I don't see any
34. Bb1 is my move.

#### S1 Position 2: W

1. Eh here we have a position which clearly appears to be better for white at first glance.
2. Opposite coloured bishops

3. eh, white's only real weakness is the pe3
4. which is also in the centre...
5. the strength is its connected to the rest of the pawn chain.
6. Black e...black's isolated pd5 seems quite weak,
7. the Qe6 would not be a good defender of the pawn
8. as its exposed itself.
9. So, an immediate Nf4
- 10.would expose the eh...drawback of the position.
- 11.I suppose black does have the...
- 12.two ideas in this position are central.
- 13.First of all the idea would be to play d4 at some point
- 14.and exchange off the weak d pawn
- 15.at which point the position would be eh...would be equal.
- 16.For black anyway,
- 17.get rid of the central...
- 18.even though quite weak on g6
- 19.and h6 on the kingside
- 20.threatening things like...
- 21.there's a tactic tricks have which eh...
- 22.the idea of Bxe3
- 23.px
- 24.Qx+
- 25.Kh1
- 26.and eh Ng4 at that point
- 27.threatening ideas like eh N...
- 28.Rxc1
- 29.and Nf2+
- 30.could be quite embarrassing
- 31.although, white could always throw the B in the way.
- 32.Nf4
- 33.take g pawn
- 34.Nh3
- 35.and a more or less finishing in a perpetual in some particular lines
- 36.on the N
- 37.if the N is guided certainly to g1...
- 38.As for what white should actually do in this position...

39.well bishops influence on b1 is not so...  
40.I think if N could actually become quite a reasonable... (unable to translate for 2 seconds)...  
41.not on e2 as it stands.  
42.I'd tend to almost to... cover the threat of eh...to cover the threat of e3.  
43.I'm reluctant to possibly play Nf4 immediately  
44.as the Q goes back e5  
45.which is quite a good placement.  
46.The ideas Bd6  
47.and e5  
48.would give black some activity.  
49.But the weakness of the d pawn is not so clear then.  
50.Essentially it's a structural weakness  
51.black offers eh some kind of dynamic activity...  
52.Perhaps doubling on the c file  
53.would eh eh...a way to try to...adjust the situation...  
54.something tells me that if white could arrange the position to have eh h3  
55.in with d4  
56.could be answered with e4  
57.and f4  
58.and exchange of queens possibly  
59.and also after d4  
60.px  
61.or B sac on g6  
62.would no longer work then  
63.as Ng4 would no longer follow.  
64.So h3 might be a useful prophylactic move.  
65.White actually should aim to strengthen the position gradually  
66.as opposed to doing anything drastic.  
67.As it is a kind of established position  
68.given the play Bb1  
69.em with the idea to play Rfd1  
70.and Nfxd4...d4 pawn.  
71.It seems a bit passive however...  
72.em B would actually,  
73.I think, I think h3 would be a useful move...

74.so decide...the idea...

S4 Position 2: B

1. Ok, I'm black here.
2. I've got the isolated d pawn
3. with most of my pieces rather active...
4. b pawn has gone a bit too far up to b5
5. which gives me a nice place on c5 for the B.
6. Also it ... (unable to translate for 2 seconds)...
7. I have, no I have sufficient...
8. but the position now might be at least equal.
9. Normally I would move the pawn on the kingside
- 10.trying to create an attack there
- 11.and the fact that his Q on b3 should help me
- 12.because the Q has gone a bit too far.
- 13.Now, also I don't have too many pieces to play with
- 14.so I can't try to do something on the kingside for that reason
- 15.I might consider moves like Qe5
- 16.then Ng4.
- 17.Em, another thing I might play is Ng4 straight way
- 18.because that eyes both f2 and e3.
- 19.That somewhere and move my pieces.
- 20.I might take the pawn on e3 in some variations
- 21.now I can consider immediately taking Bxe3
- 22.fxe3
- 23.Qxe3+.
- 24.Whether that is a good move...it's hard to tell
- 25.but it does look sort of promising
- 26.Bxe3 is the most forcing of them
- 27.and white would have to take on e3 really
- 28.either immediately or after taking on c8.
- 29.he has this intermediate move.
- 30.Ok let's start looking at the rook take first
- 31.so Bxe3
- 32.Rxc8

33.nothing there with the bishop yet  
 34.so I have to take back with the R Rxc8.  
 35.He has moves with the B such as Bxg6  
 36.but they are not really good  
 37.because I can always take back on f2  
 38.or p back on g6 I think...  
 39.So that's ok.  
 40.Bxe3  
 41.Rxc8  
 42.Rxc8  
 43.fxe3  
 44. Qxe3+,  
 45.Now I've got 2 pawns for the B  
 46.Rf2 is not good  
 47.because Ng4.  
 48.So he has to play Kh1  
 49.then I can but my Ng4,  
 50.and that leads to a very unpleasant threat of Nxf2,  
 51.and if in this position I can win his R for N I'll be a R and 2 pawns up.  
 52.And even if I drop one of the pawns I'm most likely to be better in the  
 position  
 53.because he...Bxe3.

#### S4 Position 3: W

1. So this is from the English opening
2. and the position looks about equal.
3. I would say it looks about equal
4. it's a real even
5. so not much seems to be happening.
6. I have a real problem with the Bb2
7. which seems to be blocked
8. so ideally I would like to play d4 at some point.
9. Simply to improve the B.
- 10.It may not be possible right now
- 11.but is definitely a plan.

12.Semi-open file  
13.but black is defended that and reasonably well...  
14.about playing, I can Nc4  
15.whether that is a big improvement is not clear  
16.but then I might take back with a piece  
17.and then I might take back with a pawn b pawn...  
18.slightly but that's not really an improvement  
19.because the Nb6 is not doing much as such  
20.so, black's position is pretty solid  
21.so maybe I don't have any advantage (cough)  
22.so, to have to have advantage is not to be sure in this position.  
23.Black on the other hand doesn't really have a threat  
24.Ne6 but that's not a threat as such,  
25. em perhaps one way of playing would be Nf3  
26.and then at some point if I can manage I might go for e4.  
27.I have to watch out for the d4 pawn...  
28.something has to be done with this B (Bh2).  
29.It's just sitting there doing really nothing.  
30.Now the other idea would be to Nc5 maybe via e4,  
31.so I can probably start by playing Ne4  
32.then moving the queen out  
33.if back to c3  
34.and then the Nc5...  
35.not a big plan but it's something.  
36.So Ne4 (cough)  
37.now e4 allows him to play Ne6  
38.and after taking everything on b4  
39.it's nothing no I don't think it does.  
40.The e5 pawn is hardly a weakness  
41.and I don't think I can pressurise this  
42.although Bc3...  
43.(inaudible) a4  
44.would be good enough for black.  
45.Probably Ne4 I would play  
46.with the idea of bringing Q out  
47.then maybe Nc5.

48. Another option is to play Nf3  
49. and watch out for eventual,  
50. it could be equal.

S1 Position 3: W

1. Ah ok, we have a typical position from the English opening
2. white's fianchettoed.
3. A reversed Sicilian,
4. eh black has well a good,
5. actually white em should superficially seems to be quite well on the queenside
6. with pressure on the c file
7. with black a backward c pawn
8. fianchettoed B
9. em...and eh it seems that he's doing ok.
10. These positions are quite unclear
11. as is this because eh eh black has a very simple idea which is Ne6
12. followed by Qe3
13. followed by Qb5 / Q
14. Re8
15. Ne6
16. will help have a very strong influence across the board
17. discourage d4
18. em maybe in conjunction with a move like Qe7
19. give the possibility of Ng5
20. pressurising h3 pawn or h3 square
21. allowing the black Q in
22. some tactics eh...
23. it's difficult to say.
24. White should probably try to break up the position,
25. liberate the B on B2.
26. there's two ways of doing this...
27. firstly try playing b4.
28. b4
29. is is eh is not such a hot idea

30.because of eh Na4 immediately  
31.em nasty little trick probably Nx8  
32.Qx  
33.Qb5  
34.maybe pressurising b4  
35.and eh Black is doing quite well.  
36.He could play a4  
37.but that would reduce really re...tension in this position,  
38.and eh perhaps after Ne6  
39.and an eventual a5  
40. Nb5-b4  
41.he may have trouble eh ...bringing Q in ...  
42.ah that being said I wouldn't be hasty about playing that move of  
43....moves pawns  
44.with the knight into e4  
45.which could pressurise the d6 square  
46.which could be quite vulnerable...  
47.I don't like a4 though  
48.because it's very committal  
49.and eh its losing any chance of taking in the future.  
50.I'd problem with e3 is that after Ne6  
51. eh a move like Qc2  
52.Qd5  
53.followed by Rd8  
54.Rd7  
55.then e3 becomes weak.  
56.So, what has actually black  
57.or eh white's best course of action  
58.is simply tricky to assess...  
59.and indeed I'd go back to the move I didn't like earlier which is a4  
60. is not so stupid as it looks  
61.followed by Ne6  
62.Qb5  
63.after an exchange of queens  
64.white has gained...  
65. eh I mean white will have doubled pawns



- 66.but the a file will be opened
- 67.Nc4 attack
- 68.a5
- 69.and if Nxc4
- 70.take with the b pawn
- 71.or if white so wanted
- 72.although I more likely to take with (the) a pawn...
- 73.the idea of stinging the pawn.

### S8 Position 3: B

- 1. Again a complex position.
- 2. Open basically.
- 3. Now let's have a look.
- 4. Material level again.
- 5. The king is fianchettoed.
- 6. Ah hmmm, let's have a look,
- 7. what do we see in this position?...
- 8. I think black has rather a nice position.
- 9. R on seventh.
- 10.Bb2 doesn't do anything.
- 11.Em ok b4-b5 might be an idea for white.
- 12.Or black can double on the d file
- 13.but then how do you proceed?
- 14.Ok move Q
- 15.with Rd5
- 16.and then at some stage, yea maybe, ah that's an idea yea, ok double on d file
- 17.play Nc8
- 18.Nd8
- 19.then e4 later on.
- 20.Do you have to take care of anything?
- 21.No I don't think so,
- 22.em alternatively well-developed black,
- 23.f5
- 24. e4 is also possible

- 25.but that would make Bb2 alive again.
- 26.Rad8
- 27.and Rd7
- 28.so Rd7
- 29.I might prefer Ne4-c5
- 30.could be a problem
- 31.so maybe Rd8 after all
- 32.with idea of playing Nd4
- 33.attacking e2
- 34.might weaken e3
- 35.might be weak
- 36.oh that is better.
- 37.c7 is covered,
- 38.a5 little problem,
- 39.easily just sac the pawn.
- 40.So, Rad8
- 41.looks like the logical move now...
- 42.Rad8 is my move.

S17 Position 3: B

1. Ok, em...hah, I know what opening and everything this came from.
2. It's an English...
3. em...yea I hate these positions as black
4. because em...they're just equal probably
5. em except they're far easier for white to play...
6. this position...solid pawn structures
7. white would have the edge
8. because white has an extra central pawn.
9. White has a Bishop...
- 10.em black wants to use the open d file...
- 11.and play things like Nd4
- 12.to stunt the bishop
- 13.but it's very...lots of organising of
- 14.the c pawn would be often un pres in those lines.
- 15.Rook can't move

16.as protecting (the) a pawn.  
17.Candidate moves...  
18.I would be looking at what white might play  
19.b4 in that position...  
20.no white can't play b4  
21.as I reply Na4.  
22.Em...then white might play moves like Nc4  
23.to get rid of that half decent Nb6.  
24.Em...because the Nd2 would be a much worse piece  
25.so he might just exchange it.  
26.So Nc4  
27.move my Nd7  
28.Qb5  
29.but my Nb6 and pawn b7 vulnerable  
30.so perhaps just exchange...  
31.again play Qd6 here  
32.because the Q looks like its half decent on that square.  
33.Or play Rd7 here  
34.with the idea of playing Nd5  
35.then d4.  
36.Actually that looks like a more sensible suggestion.  
37.Rd7...  
38.coaxes the idea of playing Rd5  
39.em...it attacks the a pawn  
40.and after that bring the rook over.  
41.Eh, allowing me just to play Nd4.  
42.There always a Q +,  
43.and actually Rd7 allows white to play it now.  
44.Na4  
45.will no longer work because of Qc4+.  
46.So Rd2-d4  
47.is not the biggest problem in the world  
48.but ...just reply to it by...well reply a few ways maybe  
49.a4  
50.would be the best way eh,  
51.but he can also play Rd4

52.try to stop...the Nd4.

53.Ah...oh yea the other idea is that the white Kg2...

### S3 Position 3: W

1. Definitely looks better for white.
2. My pieces are considerably more active than his,
3. my,
4. for the moment he's holding his weakness on c7,
5. em the real question is how I can bring a little more pressure down on his position.
6. Em pa5 is a little bit protruding.
7. Bc3 looks reasonable tempting.
8. He can play a4 he might,
9. d4,
10. a4 immediately
- 11.could be the... the move,
- 12.em because Bc3,
- 13.I don't seem to be able to do because Nd7 is possible
- 14.Qxc7,
- 15.so he would try a4...
- 16.then c6
- 17.Bc3
- 18.Nd7 in that position
- 19.which would force my Q from e5,
- 20.so other than that it's difficult to get anything going in the centre
- 21.or kingside
- 22.because f4
- 23.would leave my e2 pawn undefended
- 24.and my pieces aren't posted there anyway.
- 25.I would like to make more use of my B,
- 26.em it doesn't seem that prosperous at the moment.
- 27.Ne6 is coming up next move as well
- 28.so it will be quite difficult to keep Qc5,
- 29.eh ...so if N,
- 30.a4

31.Ne6  
32.Qb5 is interesting  
33. 'cause if he takes px  
34.he can't get at my pawn weaknesses I don't think.  
35.I've got a lot of activity in that position.  
36.a4 is definitely looking like the move, em.  
37.Other than that I'm not sure.  
38.I'd like to make some use my Ra1.  
39.I really don't see how that possible for the moment.  
40.Em, so in that case a4  
41.I think definitely strongest move here.  
42.Ne4 but that's  
43.Ne4  
44.Ne6  
45.Qe3  
46.Nd5  
47.I don't think I'd like that.  
48.Yea definitely a4...  
49.Ne6  
50.then Qb5  
51.can't play c6,  
52.N undefended.  
53.a4 would be good I think...

S13 Position 1: W

1. Em, I'm thinking about playing Bxd5,
2. then cxd5 em,
3. maybe Rf3
4. to way off black's r
5. and threatening to check on the c file.
6. Em, eh I think an immediate f5,
7. seems to lose the exchange after exf5
8. gxf5
9. Bh5...em,
- 10.do do I'm thinking about Bxd5

11.cxd5,  
12.em Rf3  
13.Rxf3  
14.Qxf3,  
15.maybe Bd4,  
16.em Qc3+  
17.Kb8 something like that,  
18.em I'm thinking about playing Qg2...  
19.and to move to chase the rook.  
20.I don't see any threats for the N  
21.or the B.  
22.I don't see any way he can defend the Rh3.  
23.Em, I don't see any good square along that rank for it to move to,  
24. so perhaps defend with other R.  
25.Qg2  
26.Rd8-h8  
27.Bxd5  
28.cxd5...  
29.maybe Rh1  
30.looks quite even.  
31.I'm thinking about a4  
32.to be playing a5  
33.and attacking Qb6  
34.or maybe Bxd5  
35.cxd5  
36.and then a4  
37.or maybe and then Bb4.  
38.So, Bxd5,  
39.cxd5  
40.Bb4  
41.view to going to d6  
42.where the Bishop is very well placed,  
43.em I don't see a good defense for black to that at all,  
44.ok so Bxd5  
45.cxd5  
46.Bb4

47.maybe he plays Be4...

S20 Position 1: W

1. I'm white again and...
2. I do think generally speaking white would appear to have the best of the play on the board...
3. he has more scope for his pieces.
4. The one hole in this and I think that is Bg6 on that diagonal
5. is a threat in front of the king
6. and eh,... I imagine we'll have to try to do something with this.
7. I don't see that there is any immediate threat.
8. I could of course play pawn to f5
9. which would lock off that Bg6
- 10.and px
- 11.px back
- 12.threatening the Bishop
- 13.and eh forcing it back,
- 14.and if those two advanced pawns might be good
- 15.but they're up against three
- 16.so I would be a little concerned about doing away with that,
- 17.I think the most attacking move would be with the Qg7 sorry Qg2
- 18.attacking the R.
- 19.Now has he any counter play with that?
- 20.The Nd5 could be threatening
- 21.Qe1
- 22.Qb5 has to be watched,
- 23.so but I don't think there's any threat of it.
- 24.The N even if it moves exposing the B to attack by Rd8.
- 25.That Bishop is still defended by the R on d...
- 26.So I think probably the Qg2 might be best.
- 27.I have to guard here too against a possible sacrifice of the Rh3,
28. Rxa3
- 29.and opening up a gap on the queenside there where the N,
- 30.or where his Qb6 has an attack,
- 31.so I wouldn't.

32. But I think it can be sustained,
33. so I think, all things considered I would play Qg2.

#### S12 Position 1: B

1. Ok, I'll have a quick look at the position here.
2. Eh 3-6 pawns for me
3. 6 pawns for white,
4. N and B versus B and B,
5. so material is level
6. em Rh8,
7. I've a nice open file here.
8. B attacking c2 eh,
9. none of my major pieces are under attack,
10. 3 pawns,
11. white has 3 pawns in the centre,
12. or on the queenside rather than those coming through
13. looking good,
14. eh N is being attacked on the open file,
15. potential for an open file, on d file,
16. so Rd8
17. looks like a good move,
18. em white's white B lined up against my king's corner
19. BxN
20. pxB,
21. Bh5
22. or Ba5 I should say
23. is a potential move.
24. If I don't actually move Rd8 em,
25. need to be aware of that,
26. no danger there at the moment
27. 'cause queen will just take,
28. eh strong pawns coming through on q side.
29. Attacking moves going forward,
30. N move.
31. Nxp



- 32. well protected by two pieces.
- 33. Eh N...
- 34. g4 pawn is over protected by the queen.
- 35. Eh strong Nd5,
- 36. I think my move would be Rd8. Yea...

#### S11 Position 1: B

- 1. Well, hits me as rather a complicated position.
- 2. I know that I am equal on material,
- 3. black (white) has two bishops,
- 4. immediate way to make progress.
- 5. Actually I have an open h file,
- 6. black (white) has a chance possibly to eh to advance f pawn ...,
- 7. increase pressure that way.
- 8. My Ne5 (d5) well supported
- 9. but don't see any options for it,
- 10. em I think I might actually, possibly bring Ne7
- 11. and eh perhaps get some pressure,
- 12. if I bring it right back Ne7
- 13. also helps to help to prevent f5
- 14. hold that up.
- 15. Em, I think eh that might be the best move to make.
- 16. I'm looking at my B isn't doing very much where it is (Bg6).
- 17. I don't see how I can improve it
- 18. eh I could possibly I think try and double rooks on h file
- 19. maybe Rh7
- 20. then Rd8-h8,
- 21. white could try and swap off those rooks
- 22. but I don't think I would mind that too much,
- 23. I would still control the h file,
- 24. of course, em...em I think that is probably the best way to go about it.
- 25. I don't think my king is in any danger either.
- 26. And I don't see any way of getting at the em white king.
- 27. I'm particularly sure my choice would be between...

## S2 Position 1: W

1. Ok, ...first impression I get from this position is that its pretty even.
2. I'm looking at the eh, rook being uncovered by the N.
3. If black moves his N
4. the R is bearing down on the position.
5. But, for my move I think...I think I would probably try to get my b down in front of my pawn blocking the eh...R,
6. unfortunately the knight is covering the square which would be nice to get my B on...
7. b4 em,
8. however, if he takes that I have a check
9. with my RxR
10. and then moving my own R+
- 11.if he takes with either the king or queen.
12. Eh, so I think that would be the, the lines I would be conceiving along,
- 13.Bb4
- 14.if he moves the pawn up against the Bg6 (f5)
15. e x pawn.
- 16.The R on the site doesn't worry me too much unduly,
- 17.black's rook,
- 18.so I think my move would be Bb4
- 19.opening up my rook threatening something to happen,
- 20.he can decide whether to swap the N for B,
- 21.and eh give me an open file there.
- 22.I can retake
- 23.but lose a pawn,
- 24.but I think in a game I would do that
- 25.because time is just as valuable as the piece.
- 26.Is that ok?

## S2 Position 2: W

1. Right, now first examination
2. again it looks like I don't have any particular advantages.

3. Em the material is pretty equal.  
4. A queen two rooks and 2 pieces,  
5. bishop and knight,  
6. black has eh a pawn in the centre.  
7. It looks good  
8. but it's also isolated  
9. and there's little threats on here that we may not fully comprehend  
10.like eh at some stage the bishop taking the rook  
11.taking on Bc5 bringing one of the rooks to the back rank.  
12.So, the only piece that's really unprotected so far, the one pawn on the  
right hand side of the board on h6 for black  
13.but I don't think it's very weak as such  
14.em...but if I move my Nf4  
15.threatening his queen  
16.threatening the centre pawn  
17.and threatening the pawn as well on g6,  
18.you know, it's going to be dangerous,  
19.especially if I worked out the sacrifice on taking the g6 pawn  
20.might be good,  
21.because my queen is on the same diagonal as his king  
22.and em it looks like pressure.  
23.So the knight at the moment is doing nothing  
24.except protecting d4 should he advance the pawn  
25.offering to swap queens  
26.and if I move Nd4  
27.he could Bxd4  
28.and after RxC8  
29.Rxc8  
30. his rook would be on c8  
31.controlling that file  
32. so my move at the moment I'm thinking of is immediately Nf4  
33.which apart from other considerations threatens his queen  
34.and he has to move it.  
35.So, and where he moves it to I would wait to see  
36.but eh obviously can't move it away from the pawn  
37.or it falls by the queen

38.so I'd be happy enough with that move if I were to play this position as white.

39.That 3 minutes? Yep.

S20 Position 2: W

1. Ok, right. Now first of all suppose it's as well to see how we're balanced on material...

2. we're pretty well equal,

3. so starting from scratch here,

4. and eh, he has , black has some threats here.

5. Again it's a question of whether it's worth sacrificing

6. if Bc4 (Bc5) was to take the pe3,

7. pawn takes back

8. fxe3

9. then Qxe3+

10.and is checking

11.and at the same time the file where the Bc was I the c file is opened up to black's Rc8.

12.Now the alternative is that white Bxe3

13.RxR,

14.but the other rook,

15.that Rf8 would take Rf8xc8

16.and hold that open file

17.and then the Qxe3

18. could be very serious.

19.I think from white's point of view that's the most immediate threat,

20. so how to answer that is the question for white.

21.How for white to try and advance himself?

22. What has he got? And they say attack is often the best form of defense.

23.Would be Ne2-f4,

24.that threatens the Qe6,

25.it doesn't really threaten the pawn black pawn on d5

26.because it has the queen and knight covering it.

27.But, the queen has to move.

28.And the question is where does the queen have to move to?

29.Em, ... the queen might consider moving to e5.  
 30.now there's an inherent threat in this on Qe5  
 31.if black plays Ng5.  
 32.there's a threat of mate on h2  
 33.but it can be defended by g3,  
 34.so I don't think I'd worry too much about it.  
 35.I'd see, where would I move that queen,  
 36.or what would be my follow on to there with that threat.  
 37.Eh...took...  
 38.I think white eh eh...is eh...his pawns are pretty strong.  
 39.Black has the isolated pawn d5.

#### S11 Position 2: W

1. Well, the eh, first thing I notice about the position is that ah material once again appears to be level.  
 2. However, in this situation em first thing that comes to mind is the fact that black has a weak, potentially weak pd5.  
 3. it can't really advance I think without being lost  
 4. and possibly eh I might have some opportunity of ganging up on it,  
 5. em maybe by playing Rf1-d1  
 6. and pressurising it that way,  
 7. em also I notice that my own king seems to be pretty safe.  
 8. There are no threats against it at the moment.  
 9. And eh there are, some of blacks pieces are not in very favourable positions.  
 10.The queen for example can be attacked by Nf4,  
 11.which also helps put pressure on the pd5.  
 12.so maybe some move in conjunction with Rfd1  
 13. followed by Nf4.  
 14.don't know which one of them I'd pick first.  
 15.And also em, there's another weakness in black's position.  
 16. I don't see any weaknesses in my own position.  
 17.Em he could try and swap off the pd5 by advancing it maybe  
 18.if I exchanged queens  
 19.but then I can just recapture

20. and he's swapping one weakness for another
21. because immediately if he swaps off queens (after d4)
22. he captures with pawn
23. then the pg6 is weak as well.
24. And the other factor of the position is ph6 is weak
25. and eh, so I might have some possibilities there,
26. but Nf4 as well as attacking the queen
27. would also, em look at pg6 as well
28. and maybe some pressure around it there
29. Qc2 and possibly attack,
30. attack there so that seems to be a weakness.
31. Em the only underdeveloped piece that I see that I have is Rf1
32. so Rd1...

#### S12 Position 2: B

1. Again lots of pieces on the board
2. eh pawns 6 versus 6.
3. Two bishops,
4. two knights on the board,
5. two rooks,
6. two queens
7. so material is level
8. white has advanced b5 pawn
9. supported by queen and bishop.
10. White's rook is attacking my knight or my bishop.
11. Bc5 em by the Nf6 protected by the queen.
12. White doesn't have a black square bishop to attack it
13. so, that's ok, by queen
14. or knight.
15. Eh, immediate danger to me
16. queen is supporting a pawn
17. which is protected twice.
18. N move f4 (Nf4)
19. hit my queen make that move.
20. Rd1 from f file for white

21. looks like potential good move.
22. Em, that would put more pressure on my d pawn
23. which is isolated.
24. Probably reinforce the defense of that by potential Rfd8.
25. White pb5 could that move forward to b6
26. but do I want it to go to b6?
27. so by playing b6
28. that would stop that advancing a bit more.
29. Em put my pawns on black squares
30. potentially good as white has a white square bishop
31. so b6 looks good
32. that also adds more defence to Bc5,
33. eh ...so two crucial things here
34. pd5
35. and Bc5
36. protect them I think
37. I would play Rfd8. Yea...

#### S13 Position 2: W

1. Ok, I'm noticing Bc5,
2. em I'm checking for material equality,
3. it seems that material is equal.
4. I see that black has an isolated d pawn,
5. em I'm considering sacrificing on g6
6. although that's very premature at this stage.
7. I'm looking at Nf4
8. which attacks the queen
9. and the isolated queen's pawn at the same time.
10. Nf4
11. Qd7 possible,
12. of so Qe7
13. bad because Bxg6
14. pxg6
15. Nxc6
16. with fork on queen and rook

- 17.so ok, Nf4
- 18.Qd7
- 19.and perhaps Rfd1
- 20.in that case to put more indirect pressure on the pawn (IQP).
- 21.Or perhaps the man oeuvre Be2-f3
- 22.or perhaps the idea to blockade the d4 square in that case.
- 23.So I'm looking at Nd4
- 24.Qd7
- 25.and Qb2
- 26.then attacking knight,
- 27.perhaps if black can just play Kg7
- 28.defending knight
- 29.and then maybe I can play Rcd1
- 30.to double,
- 31.indirect attack from the rook and the pawn.
- 32.Looking at, I don't see any threats for black
- 33.and my king is em.
- 34.I think f4,
- 35.that's bad because it loses pe3.
36. Bb1
- 37.and then Rfd1
- 38.with good control over the d4 square
- 39.but that might allow black to immediately play d4.
- 40.Bb1
- 41.d4
- 42.can't take 'cause Ne2 is undefended.
- 43.So that allows him to liquidate his isolated pawn.
- 44.As perhaps I play Rfd1 immediately...

### S13 Position 3: W

1. Em, I've noticed that my bishop (Bb2) is blocked by the pawn chain e5,
2. I'm thinking about playing Ne4
3. and with the possibility of maybe sacrificing on f6.
4. I'm thinking about playing Qe3
5. em with the idea of transferring my queen to kingside.



6. Although all my pieces are on the queen side  
7. so perhaps I should be playing over there.  
8. That's what well against my instincts.  
9. I'm thinking about Nc4  
10.Nxc4  
11.dxc4  
12.and then double rooks on d file  
13.or perhaps Rc2  
14.followed by Rac1  
15.and try and gang up on thee c7 pawn  
16.which looks weak.  
17.I'd say that's the main weakness in black's position.  
18.And eh I see his Nd8 is passively placed.  
19.I'm thinking about Nc4  
20.Nxc4  
21.Qxc4  
22.pinning rook,  
23.I then playing Qh4  
24.and maybe transferring some pieces to the kingside.  
25.I'm thinking about playing em Rh1  
26.with the idea of h4-h5 with the kingside  
27.and maybe attacking there,  
28.em should move e4  
29.but that creates a long-term weakness on e3,  
30.do something about offering controlling d5.  
31.I see that black is ready to play Ne6  
32.winning a tempo on my queen,  
33.so I don't think Nc4  
34.would be such a good idea.  
35.Em, so I'm inclined to move my queen Qe3 I think,  
36.possibly Qe3  
37.Ne6  
38.a4  
39.with the the idea of playing Ba3 where it stands well,  
40.small weakness of d3  
41.em, I'm trying to see f5

- 42. looks to be a good square for my knight
- 43. so perhaps I would play Nf3
- 44. Nh4
- 45. Nf5
- 46. a4
- 47. in conjunction with a3 Ba3,
- 48. attacking dark squares...

S12 Position 3: W

- 1. Ok, lots more pieces this time,
- 2. 3-7 pawns,
- 3. knight and bishop,
- 4. black has 7 pawns
- 5. two knights versus bishop and knight.
- 6. My queen is in very open play.
- 7. Looking a bit tied up at the back
- 8. although my Rc1 is in a good position.
- 9. My king is on a white square
- 10. and no white square bishop for black.
- 11. That's ok.
- 12. I'd be worried if he had a light squared bishop.
- 13. Eh I'm attacking the pc7.
- 14. Eh, with my queen,
- 15. and I'm attacking it twice with queen and rook.
- 16. Looking over towards the centre eh squares my knight can go to my knight has control of e4/c4
- 17. look like nice squares for it to go to.
- 18. On the defensive side black isn't particularly aggressive at the moment.
- 19. Nb6 can come in to d4 d5 I should say
- 20. to not great advantage
- 21. be lost if it went to there,
- 22. em a5 pawn not doing very much,
- 23. attacking it with the queen.
- 24. I could hit it with the bishop by going Bc3.
- 25. Eh, play then pawn up

- 26.no real in that.
- 27.My knight is protecting c4,
- 28.can the queen be attacked if I play,
- 29.black could play Ne6 and hit my queen
- 30.and protect the c7 pawn.
- 31.Queen back to e3 eh
- 32.stay on the c file? ...
- 33.just looking again queen move back along c file
- 34.or diagonally across?
- 35.Might play Ne4
- 36.and that is looking like the move I would make
- 37.Ne4.
- 38.Eh I can get , gets hit by f5,
- 39.Qxe5
- 40.I can do that
- 41.or Bxe5
- 42.pxe5
- 43.so I could put pressure on the e5 pawn...by playing Nf3
- 44.and I don't think that does anything positive for...
- 45.I think in this position I would play Nc4...

### S11 Position 3: B

1. Well, this is a normal position, thanks be to God!
2. ...em I eh, see that my king is quite safe.
3. Em, the knights look possibly em threatening at some stage to go to good squares.
4. I eh black's white's eh some of white's pieces don't seem to be very well placed
5. eh his rook two rooks, his queen and rook lined up on the c file
6. eh don't seem to be very threatening
7. and his Ra1 is not doing very much.
8. Em I can't see any immediate threats,
9. but the knights and possibly...
- 10.I think maybe move my own rook
- 11.and probably play, either of my rooks

12. even both over to d file,
13. em yes and em I don't see any eh, I don't see much scope for my knights at the moment
14. eh, eh my queen is on e8,
15. possibly queen,
16. no rook Rf7-d7
17. with the prospect of maybe doubling them
18. and getting queen out somewhere
19. and get a threat on the king.
20. Position wise I don't see any serious weaknesses for either team, either side.
21. Eh looking at pawns so if
22. em, as I say his bishop is basically, on the long diagonal,
23. is em ...not actually threatening very much the em
24. and also the queen and rook lined up,
25. I think eh I could possibly play Rd7
26. with the possibility of doubling up. Rd7-d5,
27. or even the Nb6-d4,
28. and maybe some pressure going em...

### S20 Position 3: B

1. Well, I'm ok I'm ready to go.
2. I'm looking at black first of all this time
3. and eh, so far as the minor pieces are concerning the rest is equal
4. but its two knights against a knight and bishop.
5. Now oddly enough I'm the sort of guy who likes to play with knights
6. so, I think I have possibilities here.
7. There's nothing to be won immediately
8. but as well as that his Qc5 is pretty well exposed out there.
9. And I think the three pawns on the queenside should bring somewhere
10. but it's a very very equal position.
11. I think as a first off goal what I would be looking at would be probably Nb6-d7 (I corrected the notation here, he had said Nb3-d2 by mistake)
12. attacking the queen
13. that forces a move of some sort from the queen.

14. And I think that would be a starting point
15. because then you could get the pawns that are behind the knights to move.
16. Now queen can't take c7
17. because it's covered by the pawn
18. and would have to move pretty well back away from the influence of the two knights
19. which cover almost up as far as the fifth rank,
20. the way they're placed.
21. So, the queen back so (Qe3 or Qc3?)
22. and probably, neither of them appear to be great.
23. He can go to eh...he can go to e6 (Qe3)
24. probably the best move.
25. On the other hand, if he could move back to c6 or c7 (c3 or c2)
26. the queen is pretty locked in there.
27. Back to Qc3.
28. there's a move with the knight,
29. not yet!
30. But, eventually...eh there would be Nd4
31. threatening the queen again there.
32. Eh black, as I see it would have to start the moving those knights to get value out of them.
33. So my choice move would be Nb3-d2 (Nb6-d7).

## S2 Position 3: W

1. The thing open is to be to try and force a move in the centre,
2. so I would probably be playing something like Nb4
3. or d4
4. because the bishop obviously has to get more freedom along that line...
5. eh the other thing would be going on to f4
6. but unfortunately I think that would be giving black too much.
7. If he took the pawn on f4 with his e5 pawn
8. his queen is then ranging down on e2 for an easy check...
9. em eh the rook is in front of the king (on f7)
10. and unfortunately I don't have a white bishop to pin that somewhere
11. and the one obvious move of getting down on c7 (Qxc7)

- 12.but its covered by that rook.
- 13.Even though I could take but I'd lose material.
- 14.So...I'm thinking at the moment that my best move is to play d4
- 15.even allowing black to retake
- 16.or to move ahead with it...
- 17.eh I just think it gives me a little bit more freedom
18. and if I could at one stage shift the... the rook off by a knight move or something
19. I would looking dangerously down to c7
- 20.with queen protected by my rook
- 21.and there's always the possibility of some sort of dramatic take by the bishop
- 22.eh again as its lined up against the king
- 23.although I can't see it working at the moment
- 24.but its always the... the type of thing that worries,
- 25.would worry me
- 26.eh like the bishop continually focusing down on g7 in front of the king.
- 27.So as to what move I would actually make it would be in the centre
- 28.d3-d4 pawn
- 29.or the knight form d2-e4
30. or even Nc4
- 31.and even the back pawn e4
- 32.forcing the pawn to stay where it is until I can attack it on the other side...

*Segmented protocol transcripts for random positions*

S3: Position 4: W

1. It's a mess, em ...
2. yea eh well just look Rc5 briefly.
3. The position is quite difficult to make sense of.
4. Em ok basically both kings are in massive danger.
5. Well maybe black's king move
6. because black doesn't have a check.
7. If I could deflect his Qd5

8. and Rf2
9. from defended f3 pawn,
- 10.then Nxf3++.
- 11.Em so the question is... how do I manage to do that?
- 12.Em I've a feel like I should be doing something else,
- 13.but maybe mate I is the only thing to play for in this position.
- 14.Material it doesn't matter.
- 15.But, I don't really see any other way.
- 16.Right eh ok, so let's see Rc5
- 17.isn't possible because Bxc5 em,
- 18.although I could take on c5 with queen
- 19.then queen takes queen possible
- 20.deflect queen from f3,
- 21.deflecting rook as well more difficult.
- 22.Eh so, yea... Rc5
- 23.how can I do that.
- 24.My king can't move.
- 25.My knight for the moment can't move.
- 26.Qe1 idea of threatening R.
- 27.Qe1 is probably one of the more promising ones.
- 28.Em just trying to get rid of his Rf2
- 29.em give me option of playing Qd3
- 30.which in some ways might be useful.
- 31.Like after Qe1
- 32.Rg2
- 33.Re8
- 34.Bf5
- 35.some combination of Qh5.
- 36.Yea if I can get his queen off that square.
- 37.Em, ok eh Qe1... eh Qe1 I think does,
- 38.I might just look Rg2,
- 39.I'd imagine that's quite weak.
- 40.Em let's see if I have any ideas.
- 41.Ok once again if I can deflect the queen em...
- 42.How do I deflect the queen though?
- 43.Qe4

- 44.Qx
- 45.Rx
- 46.then Rf2
- 47.can't ,
- 48.Ne5.
- 49.I definitively have to proceed with forcing moves.
- 50.Once he gets Ne5
- 51.he's better
- 52.although I don't know how anyone could be better in this position.

S17 Position 4: B

- 1. Ok we have a complicated position.
- 2. Em, where it looks as though ...it wouldn't surprise me as black, eh, if I was to be mated in the next few moves.
- 3. So, em I just better count the material.
- 4. 1-2-3-4-5-6, ok 1-2-3-4-5-6 (counting pawns).
- 5. Ok we're even on material.
- 6. Em, white has a whole load of pieces around my king
- 7. and em I don't like that.
- 8. Em, (laughing) and really in this position I would be looking at eh like I suppose I do have some decent moves.
- 9. Em eh things in my position.
- 10.No I don't! eh (laughing).
- 11.Well actually I don't really have anything significant.
- 12.I'm playing for a draw here.
- 13.Eh eh, I have to try and find a way eh of white's threats are
- 14.and countering them.
- 15.Now ...white would like to play...oh white would like to get his bishop
- 16.and queen involved
- 17.Qe5 would be a decent move for white,
- 18.em because it's just threatening Rxa7+
- 19.eh Rxa7+
- 20.Kxa7
- 21.Qc7+ mating him
- 22.and actually that looks extremely difficult to avoid for white.



23. So, do I play a move like Rf8  
24. to try just to stop that  
25. em perhaps I have to  
26. em do I have any other nice ideas if I play,  
27. I suppose I might play Qe8  
28. after Qb5 I might be able to get away with  
29. Nc6,  
30. so harassing.  
31. Eh oh hold on does Rf4,  
32. no there isn't Rf4,  
33. em because the Bc1.  
34. Is there any way I can make that tactic work?  
35. No there isn't,  
36. eh my king doesn't have a move,  
37. my knight doesn't have a move.  
38. And so let's just look at the pieces I can actually move in this position.  
39. Em, Rf4  
40. just the idea of Nd4,  
41. looks really strong for em.  
42. Oh perhaps it doesn't  
43. Nd4  
44. RxR,  
45. Nx...  
46. interesting.  
47. Let's just see.  
48. Eh, can I play Qd8.  
49. Qd8!  
50. Qd8 attacks the rook em.  
51. Qe5 no longer has the same threats  
52. because Queen covers the square that's going to  
53. ok Qd8 looks like a decent thing,  
54. and on the following move I would play something like Rd1  
55. so that my bishop on f1 in protecting that pawn on b5  
56. and I can try and get that bishop back into the game.  
57. Oh hang on maybe I should just play Rd1 immediately?  
58. Because that threatens Rxc1

- 59.Nxc1
- 60.Rf4
- 61.and if white plays Qe5
- 62.I have my defense against that of
- 63.Qd8.
- 64.not that I'm entirely happy with that.

S4 Position 4: B

- 1. Well, here some kids have been messing again,
- 2. and there have another random position on the board.
- 3. At least this one is good for me.
- 4. My king looks safer than the enemy.
- 5. And or does it?
- 6. Yes it does.
- 7. Material seems to be...equal.
- 8. Right material is equal
- 9. and I hope to bring my bishop.
- 10.The white king is not a very powerful creature.
- 11.My king although more in the centre is the safer one
- 12.Rc5 is not a threat
- 13.because I control that square.
- 14.Nxf3 is not a threat at all.
- 15.I would be happy to see that (Qxf3++).
- 16.Now can get at his king
- 17.that would be pretty nice thing to do.
- 18.So how about h5
- 19.so I
- 20.and well checkmate and everything.
- 21.So he cannot take
- 22.because Qf5 ++ as well.
- 23.So that, he would have to move the rook somewhere to stop that.
- 24.Which sure enough gives me an idea really
- 25.maybe I should start with f5
- 26.because that doesn't give him the same chance
- 27.but f5 exposes my king a little bit,

28.he ok  
 29.so if I whatever reason would be better,  
 30.also if I can now get my other rook on a7 to h-file  
 31.now that would be checkmate immediately.  
 32.h5  
 33.he moves the rook somewhere  
 34.I might move Rh7  
 35.if the bishop goes  
 36.I take on g4.  
 37.Bxg4  
 38.my bishop comes somewhere say Be6  
 39.and rook-h  
 40.and checkmate is coming.  
 41.Oh I would say that is the move,  
 42.what else what else have I no no yea h5  
 43.looks the best  
 44.and I don't see other,  
 45.ok if I can get my Qh4  
 46.that is checkmate too.  
 47.Some one idea is to play Kh6  
 48.where the cute idea of Qg5  
 49.Qh4++.  
 50.Now how he is going to defend that I don't know.  
 51.Maybe, how about this, how would have to take off Rxf6  
 52.with rook  
 53.and then if I play Qg5...

#### S8 Position 4: B

1. Hmm, this is Alice in Wonderland chess (laughs) oh gosh, how did this position arise? Eh ok this could only be a position out of a psychological test!
2. Em yea takes time to understand this.
3. Wonder if this position is possible?
4. Ok so black to move.
5. Both kings are in danger hmm.
6. Ok so, what are the features of the position.

7. Right probably can only be mate,
8. em better look at the material 2-4-6, 1-4-6 (counting pawns),
9. equal material
- 10.but that's not very important in this position.
- 11.God and I'm expected to come up with a move.
- 12.Oh, ok so how do you get at the king?
- 13.Hmm, Rf4
- 14.doesn't work
- 15.Bc1 can take that.
- 16.Nb3 covers a5.
- 17.It's a tricky position.
- 18.So ok the Rc7 not protected.
- 19.What can we do there?
- 20.Em I'm supposed to talk all the time.
- 21.I prefer to think in this position (laughs).
- 22.Ok can't find a sensible move.
- 23.Glad I'm not in time trouble.
- 24.Ok so it's, if I look just logically here...
- 25.Rd1 any sensible move,
- 26.trying to get rid of that Bc1
- 27.then Rf4.
- 28.But after Rd1
- 29.Be3
- 30.hasn't really achieved anything.
- 31.On the other hand, ok is black really in danger of being in mate here?
- 32.Could be, hmm.
- 33.Ok if I come up with a move it's because I have to not because I really like this position.
- 34.Em these are not my type of positions.
- 35.Ok Nc6
- 36.that would also be something
- 37.so Qd8
- 38.no
- 39.then Rd7
- 40.sacking the exchange.
- 41.Qa5 is covered

- 42.hmm very difficult em you will probably tell me that my time is up soon.
- 43.Ok, Rf5 is also a possibility,
- 44.don't ask me why,
- 45.hmm probably Rd1 looks most sensible
- 46.Be3 than Re1
- 47.but it doesn't really do anything.
- 48.At least it covers b5 with the bf4.

S1 Position 4: W

- 1. Well, first this position strikes me as being eh, composition straight off
- 2. or eh a random position
- 3. due to the placement of all the pawns everywhere.
- 4. The f pawns the 'Irish pawn centre'.
- 5. Also the eh...the way eh the pieces the pawns are all just the white pawns are all just atypically spread.
- 6. The actual position to assess, phew em quite tricky
- 7. because the placement of the pieces suggest that both teams are in danger,
- 8. they actually look quite vulnerable
- 9. it seems to me that the white king is in more danger...
- 10.eh than the black so to speak.
- 11.Em obviously eh obviously eh white would, white would like to play Rc5 at some point
- 12.eh taking advantage of the queen and king unfortunate alignment of the queen.
- 13.The black pieces seem more well placed.
- 14.The material here in the position is quite irrelevant actually
- 15.from eh looking at it its actually...equal.
- 16.Eh the knight of course would be more useful in this kind of position
- 17.bishop should,
- 18.tends to have a more limited role.
- 19.So the bishop when you scale it sown essentially worth nothing
- 20.protected the Re6,
- 21.now , plans in this position might be eh...moves like f5.
- 22.it is white to play.
- 23.I need to support, to assess the kind of threats black has

- 24.as I said I don't see any immediate eh chances of there being eh an advantage for white.
- 25.That being said the Ng1 could spring into action at some point,
- 26.the Rf2 seems to gain control,
- 27.covering it as regards seeing what eh white should do in this position its quite tricky to say the least ...(keep on talking)
- 28.essentially it's going to be difficult for white by
- 29.I don't see eh the winning path for black either
- 30.but eh all the activity suggests there maybe moves like h5
- 31.gxh5
- 32.then gxh6
- 33.which immediately might just win the house straight out. That eh...

#### S2 Position 4: W

1. Well this looks like what I would call a hairy position.
2. The king is out on the edge of the board (Kh3).
3. Risky position indeed.
4. And the black pawns are at very advanced position.
5. So, I , I'm looking in consequence at the black king
6. and I see that he is also in a very dangerous position
7. because he can't actually move the king anywhere (Kg5)
8. and he will be checked because on every square he could move to he would be walking into check.
9. So, if I could force the issue like swinging the Rc5
- 10.planting it in front of the queen.
- 11.Unfortunately it's covered by the bishop (Bf8).
- 12.So it would be Bxc5.
- 13.the other alternative would be to play Nxf3+,
- 14.getting the queen out of the way first,
- 15.but unfortunately that would give black the game
- 16.because Qxf3 mate.
- 17.So, very delicately poised...
- 18.em another possibility would be to try bringing the rook over.
- 19.That doesn't seem to work.

20. So what would I move, I ... the key lies somewhere in shifting the queen off the file...
21. off the rank in front of the king
22. so I get the check in with the rook.
23. So, can I take the bishop first? (E: keep on talking)
24. ok, it's all hinging around there so maybe if I could ... no move.
25. So, supposing I play my rook in front of queen...

#### S20 Position 4: B

1. This is certainly a most unusual position for both white and black with both kings. The position there in their absolutely the game is wild.
2. I couldn't describe it as anything else.
3. Eh: Black has queen and rook on a file (rank)
4. and a very far advanced pawn
5. which I don't think will be tenable for some time.
6. The black king and the white king are in most disastrous looking positions.
7. So, what would I do as black.
8. Well I think obviously.
9. One is going to be attempting to force the position onto the white king.
10. Now the difficulty, again even the Rd3 is defended by the Bf1.
11. So white or black with the move at the moment could I think immediately play Qd8. now that looks like the most immediate response.
12. It threatens the rook from c7.
13. the white rook which has really eh, no-where to go.
14. The white pawns of course can be advancing .
15. But I think this is the move, now play that and eh.
16. The Rb7 in which,
17. ah there's a knight up there too
18. which makes life rather difficult
19. because there was a lovely move with the queen,
20. but however, Qa5
21. would have been very nice
22. except there's a knight there so you can't do that.
23. It's very hard to see how one makes progress in this position.

24. But I think maybe on second thoughts the Qe8 might be better than Qd8
25. now this threatens the pawn
26. if there is a way of shifting something there
27. but where it's going to shift to I just do not know.
28. Is there anything else in this at all... a most confusing position.

S11 Position 4: W

1. Well the first thing that occurs to me when I look at this is that it is a mirror image of the position,
2. eh, take me a moment or two to realise what colour I am.
3. I think I'm white (laughs) (E: yes that's right).
4. Em ok, I em, well first thing is my king is in a very exposed position
5. but em maybe eh black has some threats against that,
6. em my Re3 is defended by bishop at moment.
7. Maybe black could generate threats against that
8. eh by playing Rh1
9. and maybe getting the rook over,
10. well Rh1
11. threatens RxB
12. the QxR.
13. Also maybe trying something, his rook over to the em h-file
14. his king is his king is in his way but I don't see the ,
15. any immediate yes, does it? There's a threat there.
16. I'd better do something about it. I could, I'd like to be able to play Rf4 (c5)
17. that would win the queen
18. but I can't at the moment because of Bf8...
19. em, I don't see any maybe if I played em, if I play Re1 (Re8)
20. the double threat of swiping off the Bf8
21. and then playing Rf4,
22. that's a major threat,
23. so what happens if I play Re1 (Re8),
24. eh well, if bishop moves to g7...



S12 Position 4: B

1. Ok, I'm black in this position,
2. piece count 2-5-6 pawns.
3. White has six pawns two bishops, two knights (for black),
4. so material level.
5. White has rook on seventh (Rc7).
6. Two passed pawns.
7. White's kings exposed on b4,
8. can't really move so have the potential in there for checkmate playing something like Nxc6+
9. but that loses to Qxc6++ eh,
10. I can pin the queen by playing Rf4
11. that wins the queen as
12. QxR
13. QxQ+
14. and must lead to mate.
15. Pawn up Qxp,
16. Nd4
17. to block it
18. and then the something
19. there must be a potential mate in there.
20. Em just spotted the Bc1 there on
21. and that screws up the Rf4 move.
22. Em to way of getting that in there, play something like Rxc3
23. KxR,
24. doesn't really get me very far.
25. Bishop back into play.
26. My pawn on h4 is attacked,
27. doubly attacked by queen and rook
28. so I might advance my queen pawn to g5.
29. what can white play in terms of getting me in trouble
30. with Rxa7
31. Kxa7
32. pc7
33. Na6,

34. should be paying probably more attention to my k-side defence rather than my q-side defence.
35. Em very double edged position here.
36. Looking at bishop moves.
37. My Bf8 has to be a way of taking advantage of the fact that white has no squares for his king.
38. Eh, rook move
39. Rd3
40. can I hit the queen can I move the bishop?

S13 Position 4: W

1. Em, I looked at Nxf3 immediately
2. but that's not good because that leads to mate (Qxf3++).
3. I see black's king is in a spot of bother
4. and if I can play h4
5. with h5 protected
6. its mate.
7. I'm looking at a way to em, move my king from the square
8. because he's precariously placed at the minute as well.
9. I'm thinking about Qe1
10. but I think Bc5
11. defends rook adequately,
12. I think black has the possibility of creating a passed d-pawn
13. 'cause em, the pawn is quite advanced.
14. Apply more pressure
15. and I see a threat is Bb4
16. with a skewer with the rook through d-pawn.
17. So, something I'd have to be aware of.
18. Em, so now I'm looking for a counter to that move.
19. But perhaps I play Kb1
20. if Bb4
21. Rxd3
22. eh I think Rxd3...
23. perhaps em, I think I well ok 'cause the Qxa5
24. and I think my d2 pawn's a gonner.

25. So I don't think that leading to any particular advantage.  
 26. Em consider playing Bd7  
 27. just to defend my c-pawn  
 28. and the bishop  
 29. and consequently the rook in the long-term  
 30. but that's not give me any threats  
 31. and it's also easy to play Bb4.  
 32. I'd look briefly at Ne2  
 33. but that just gets taken.  
 34. I can't see a good way of removing the rook..  
 35. especially as its on dark square  
 36. as I don't have a dark squared bishop.  
 37. I think the Rf2  
 38. my pawn on as is attacked twice at the minute.  
 39. Em I'm thinking about sacking my rook on f6.  
 40. I'm, also bit doubtful how that works.  
 41. Also possible to play Rc5  
 42. but its only briefly  
 43. because its covered by the bishop.  
 44. Em, I don't see any checks.  
 45. I like the idea of playing Qb1  
 46. Bd4  
 47. Rxb1  
 48. Qxa5 '  
 49. cause now I realise I can play Rd5  
 50. no I can't.  
 51. trying to get my queen onto the light squares  
 52. maybe thinking about checkmate eventually  
 53. on f5 with my queen there.  
 54. Now I'm trying to think of a way to do that now. Em, perhaps I could...

S17 Position 5: B

1. (laughing). Complicated.
2. Count pieces 1-2-3-4-5-6, 1-2-3-4-5-6.
3. Actually kind of initially like certain aspects of my position here.

4. My king is very safe.
5. Em white has a lot of weak pawns eh on the queenside
6. whereas my pawns are safe.
7. I'd be wondering what happened earlier on in the game to get tripled e-pawns.
8. But eh and actually I'm not too happy with my Na7.
9. ok so ways for me to improve my position here.
10. Em well I'd like to get my Bg6 into the game,
11. my Na7 into the game.
12. Getting the Na7 sorry Na8 into the game,
13. the only way to do that is via c7
14. and there are particular problems with that
15. because if I play Qb7
16. which is the way to do it
17. eh then eh white plays the move that white wants to play which is Rb3.
18. So I'd like to discourage Rb3
19. or how would I do it
20. so not eh Bh7
21. is a potential move
22. because that rook hasn't actually
23. Bh7 yea.
24. I kind of like that idea
25. because the white king is vulnerable
26. although white pieces around it n' stuff
27. so Bh7
28. when followed by kg7
29. followed by Rh6
30. followed by Qg6
31. followed by Qg5
32. looks extremely strong.
33. Actually that could be just winning.
34. Bh7 rook somewhere along the back rank.
35. Is there a particular problem with that?
36. There's certainly a problem.
37. Ah there's a problem with the whole idea
38. because eh pawn e5 will be un pres when I carry out that man oeuvre.

39.I'd have to look at precisely where the rook actually goes after Bh7  
40.em it'd go to d8  
41.because all the other squares whether allow me to,  
42.are covered by  
43.I gain tempos em now  
44.I'd probably just pursue.  
45.Oh the king can go to f5.  
46.Perhaps a nicer way of doing it.  
47.Potentially nicer anyway,  
48.and black ok em wow there in this position.  
49.I think Bh7 would definitely by my move.  
50.100% of the time  
51.and after Rd8  
52.I would definitely play either Kf5  
53.or Kf(g)7.  
54.Kf5  
55.might have the ...to allow Rd5  
56.and which could give some kind of counter play against the d5 pawn.

S4 Position 5: W

1. Gosh, what is this?
2. Let's something...its completely random.
3. Phew, I'm white so I guess I should make a move.
4. My, all my pieces are weird.
5. The only good thing I can see is my e6 pawn.
6. I think I can queen that pawn.
7. I don't think there are particularly dangerous threats which black has, now  
...
8. one idea is to, a central idea is Rxf7
9. K x followed by
- 10.Kxc7
- 11.get Queen up,
- 12.eh probably wouldn't work at the moment,
- 13.now the other thing I might consider is Rf7
- 14.then with the same idea of playing e7.

- 15.Em, Rf7
- 16.Rx and
- 17.pxf7 my pawn is blocked,
- 18.but the good thing I might take the pawn on f5
- 19.(Qxf5).
- 20.Still my Re7 is probably a good one so I may not do it.
- 21.Now if I can move my pieces somewhere closer that should help
- 22.but I don't think I can move them,
- 23.ok so I should really find something...now so.
- 24.The material is level that's a good thing.
- 25.Oh well all my pieces should be probably better,
- 26.I have equally bad pieces
- 27.but I have at least a passed pe6.
- 28.em, centre Re8 and then
- 29.e7 perhaps,
- 30.now that's that's it. Am I finished? I am finished now phew.

#### S8 Position 5: B

1. Hmm, really wonder what this test is about? (laughs).
2. Ok an unusual position
3. and, more sensible than the last one,
4. black to move
5. although black looks like he's got a good pawn.
6. King is in a cage
7. but doesn't seem to be in danger immediately.
8. Pawn d3 looks very promising.
9. How do you get at it that is the question.
- 10.Ok, eh what's the plan?
- 11.Ok white, hmm ok Qb7
- 12.could be an idea try to get in on the second rank
- 13.and once the rook has moved
- 14.just play d2
- 15.followed by d1!
- 16.Does white have any threats?
- 17.Not really.

18.Pawns are weak, not going anywhere.  
 19.Ok the Ne/a8 not doing anything either.  
 20.Qb7 might be sensible  
 21.any way to get at the king,  
 22.h-file yea but that takes a long time.  
 23.Get rook out,  
 24.bishop out Kf5  
 25.looks quite safe.  
 26.Could get the king maybe,  
 27.the queen to h-file  
 28.and then try to mate with Qh4  
 29.but there are too many pieces in the way.  
 30.A more prophylactic approach is probably to play Qb7  
 31.Qb1/b2  
 32.then Rd1  
 33.just to get rid of knight on c1  
 34.and then queen the pawn.  
 35.Alternatively Kf5  
 36.followed by Qb1  
 37.d1  
 38.Qg4  
 39.could also be a good idea.  
 40.em yeah, yea Qb7 looks like the move that I would play in this position.

S1 Position 5: W

1. Ah, ok here we have another extremely unclear position
2. which is actually possible to occur....
3. Well likely... well anyway the actual position as I see it.
4. Once again the kings are in very compromised positions.
5. Black is more active than white,
6. white seems to have more immediate position plusses.
7. The Rd7-e7 em
8. what white needs to do there is to somehow... untangle his forces on the kingside seems to be maybe to do...
9. the actual opportunity where for instance where eh... moves such as Kc4

- 10.followed by c3
- 11.is not immediately observable
- 12.should actually... perhaps with the idea of b5
- 13.followed by Kb4
- 14.em would be interesting
- 15.as em the R +
- 16.the queen then would come into the play on the third rank.
- 17.And while the king is on b4
- 18.c4 would be possible
- 19.followed by c5+
- 20.em king forced a7
- 21.and after a move like Rd8...
- 22.em mate
- 23.or check on the a-file
- 24.it would be mate...
- 25.even b6 +
- 26.px
- 27.px
- 28.Kx eh
- 29.Rc3
- 30.followed by Rc6+
- 31.sacrificing rook
- 32.bring the queen in in some lines,
- 33.but that being said... instance Kb5
- 34.c4
- 35.c5+
- 36.Rxc7
- 37.right were black as black do anything to counteract this?
- 38.Ideally, if playing b5
- 39.black could...
- 40.why I like the idea of playing b5
- 41.is black had looked to take advantage of the weakness on the dark squares...
- 42.by just opened up such as eh eh by playing Qa8 to a5+
- 43.with the queen misplaced on the other side of the board,
- 44.and it doesn't look like it's going anywhere fast...



45.but there is the threat of the h-pawn  
46.em but perhaps key to blacks defense of this position is actually after b5  
47.to play a move like Ka5  
48.to stopping the idea of Ka4  
49.it looks quite absurd...  
50.but maybe it is not as stupid it looks...  
51.unless if white was to take on b5...  
52.then Rxc7  
53.would surely give a big advantage to white  
54.followed by Rc4...  
55.Rb4+  
56.and just eh...  
57.Bc4+ clearly be,  
58.the main problem that white has is in this position the Rb1  
59.while shuts off the king on b3  
60.and divides the board in two.  
61.As the solid pawn structure is defended all of the action is taking place on  
the a: B and g files.

## S2 Position 5: W

1. Right, ... well again I see a position where white has an advanced passed pawn
2. and this time it's important.
3. I'd probably would be thinking if it doesn't win the game what black will have to give up to stop it queening,
4. so, ... the first move I'm thinking of is Rf7
5. or Rd7.
6. Rd7 looks better
7. because if I go to Rf7
8. RxR
9. pxR
- 10.forces knight to go away,
- 11.but at the same time the queening square is covered by the queen.
- 12.Rd7
- 13.is a little more threatening

14. especially if I can get... something else behind it
15. like the other rook
16. which at the moment I can't
17. because of the eh p f4.
18. So, that would be the lines I'm thinking of ,
19. em what threats has black got against me?
20. Eh, seem to be pretty well caged in by all my own pieces
21. and look safe
22. unless I'm misunders' estimating.
23. One move he has maybe is Bxp+
24. drawing me out
25. but I don't see where it leads him.
26. So , I would be thinking of Rd7
27. protecting a further advance of the pawn (e7)
28. dividing the game
29. I would try to angle myself back somewhere to Qe2
30. to protect it further,
31. but I know I couldn't stay there long
32. because he could just push the pawn up against it
33. and eh force me to move
34. but I would be looking along those lines.
35. Maybe not getting in to get a queen
36. but of forcing black to give up a piece
37. soon maybe eventually end of the game win on material...
38. that's em, is that 3 minutes?
39. (E: no, you can keep going if you want).
40. The black rook looks dangerous
41. but I think it's just looks,
42. em because even if I have to move my queen at any stage
43. my bishop is protected by my rook and queen.
44. And I can see the knight protecting the pawn.

S20 Position 5: B

1. Ok, I take it I'm black again here,
2. and again very wild open position

3. with the kings out of the way.
4. Now, I have to say for starters.
5. I'd see black with three pawns on are eh file
6. so, you can look at that and say is that going to go anywhere,
7. whereas, actually has two passed pawns.
8. Now they've a bit to go yet
9. and I'd say that this is going to be difficult maybe to stop,
- 10.eh also the Rg8 is certainly pinning the king in there
- 11.if there was to be any sort of.
- 12.Where white will go is difficult to say,
- 13.there is his, eh Bb8
- 14.the white bishop is probably not that well placed there
- 15.and I think the... the eh... what would he play,
- 16.his Na8 is trapped.
- 17.So really he is going to have to try to extricate that Na8,
- 18.the queen is defending pa5...
- 19.oh no sorry that pawn is defended by the queen
- 20.if he moves back there's the possibility of moving.
- 21.Eh, I think... again the prospects look like it has to be also
- 22.the black Rd2 has no support
- 23.and no defense
- 24.and I think the queen will have to become more active
- 25.and still it's going to let that pawn advance
- 26.so what can we do to stop that...
- 27.the thing is the rook the black rook really has nowhere to go.
- 28.Move onto the back rank.
- 29.That might be the best thing,
- 30.all things considered
- 31.eh with the possibility on the h-file
- 32.and still is not threatening or doing anything.

#### S11 Position 5: B

1. Ok, the first thing that strikes me looking at this position is that it certainly didn't occur in a real game.
2. It's a composed position.

3. My first impression as I look at the board is that from the wrong side?
4. So it's very difficult.
5. Looking very complicated.
6. Very hard to think about the position.
7. I see that em, that my king is safe,
8. maybe not as safe as white's.
9. Ah actually my king may not be safe
10. because its if white can somehow generate threats against it somehow,
11. it cannot go back to g7,
12. em also eh white has eh maybe has some threats of advancing his pf4
13. and eh queening eh hmm.
14. I em, don't see any particular, any way of getting eh at white's king.
15. I don't see any way for white to get at mine,
16. eh I I my knight h1 (a8) stuck in corner.
17. I can't see how I can move that.
18. I could possibly try to re-man oeuvre my queen
19. by playing Qb7...
20. eh now with possibility of coming down to g2 (b2).
21. Well that would probably force off an exchange of queens,
22. em but why do I have to be in so such a hurry,
23. I could play Qg7 (Qb7)
24. with threat of going to b2
25. but not immediately.
26. I can then re-deploy my Re6 to em h6...

S12 Position 5: W

1. Ok, very complicated position here.
2. Lots of pawns 1-2-3-4-5-6, white six pawns.
3. White Nh1 Qh3.
4. Rook stuck up there on e7.
5. Eh black has Rb1.
6. King exposed although maybe not that bad.
7. This position here does not look good.
8. Black has pawns on f-file.
9. Ok they're doubled

- 10.but they're eh fairly clean run through to promotion.
- 11.Same could be said about my pe6
- 12.eh what how can I advance that.
- 13.Rook supporting it.
- 14.Immediate danger is from my perspective bishop moves
- 15.Bxd4+,
- 16.... Rxe6
- 17.pxN,
- 18.black isn't going to play that,
- 19.bishop into play
- 20.Rook Bishop out to get his bishop active
- 21.drop that...
- 22.potential moves for white
- 23.Rf7
- 24.RxR
- 25.pxR
- 26.no and queen just wins that back
- 27.em, Rf3 attack pawn
- 28.what against it can be done.
- 29.Rf3
- 30.Nf2
- 31.knight active,
- 32.Ra1
- 33.Bxd4
- 34.RxB
- 35.back pawn up pf3
- 36.white's Nf8 is fairly inactive...
- 37.Ng6 to hit my rook.
- 38.Where does my rook got then, Rxc7...

S13 Position 5: W

1. Em, ok I see I've got tripled pawns on the d-file.
2. I see I've got a protected e-pawn on e6.
3. I would consider playing Rxf2 (c7)
4. I'm also considering playing Ba4

5. at some stage.
6. Ba4 em,
7. ... just think that's a good square for the bishop.
8. I'm thinking about playing Ba2 em,
9. perhaps that's a little pointless.
10. I think about playing Kc4
11. with the idea of Rc3
12. and then Kd3.
13. I'm also thinking about trying to vacate the d3 square for the queen.
14. So maybe I'll play Rf3
15. em, with the idea of Qf1
16. attacking-
17. ok I can't play that.
18. Rf3
19. with the idea of Rf1
20. Rxf1
21. Qxf1
22. Bh2
23. Qc4
24. maybe something like that
25. with a strong attack on c2, c7 pawn.
26. Em, I', thinking of ways to support advancing the e-pawn,
27. em I see my knight is dominated by the bishop
28. but the black bishop also relatively out of play.
29. Em my king position is a little restricted.
30. I'm thinking about transferring a rook somehow to the a-file
31. keeping blacks king locked in,
32. so I'm thinking about playing Kc4
33. followed by Ba2
34. rook moves
35. and then Ra3
36. with the idea of following all that up with Qd3
37. and aim for mate
38. or something like that.
39. The black king is very restricted.
40. I don't see any threats for black immediately

41.and looking for them,  
42.eh I see he's double pawns on c-file  
43.but they are passed pawns.  
44.I control the queening square  
45.and the foremost pawn is well supported.  
46.Em I'm thinking about playing Qg2  
47.attacking the bishop.

S17 Position 6: B

1. Again I'll count pieces 1-2-3-4-5, 1-2-3-4-5 (counting pawns).
2. We're even on material
3. em and it's my move
4. and I've to work out whether I can take that R on h6.
5. Basically because if I take that rook its game over
6. (N.B. the rook was half on h6/h5, I then adjusted it and chess clock).
7. Em, ah (E: sorry about that, S: laughing).
8. Ok, eh ok we have an interesting position then,
9. my kingside is a bit vulnerable
- 10.and white's king does seem to be safer.
- 11.My Bf8 isn't a good piece
- 12.because, can't really isn't getting into the game at the moment
- 13.em but I do have positives in so far as Ba6.
- 14.White's Ba6 isn't great
- 15.and em I've a nice passed pd4.
- 16.So what kind of things would I want.
- 17.Well I think an endgame would suit me in this position
- 18.because em, the pd6 would then become a weakness
- 19.and my king would be less vulnerable.
- 20.So do I want to look at playing moves like Qe1
- 21.which is position ally will swap queens off
- 22.because the knight has to move.
- 23.Em the knight doesn't have many good squares.
- 24.Oh then the d pawn falls
- 25.so that might force white into doing some radical action.
- 26.Ok Qe1

27.em what kind of radical action does ,  
28.oh sorry my knight would be un pres from my queen.  
29.Ok, how do I sort that one out?  
30.Em d6 looks just equal.  
31.He just,  
32.I can take the eh Ba6 with the knight.  
33.It gets rid of my knight.  
34.That's one way of want to play.  
35.I could play,  
36.play Rb6 (8?)  
37.attacking the d-pawn again.  
38.The interesting problem with that is I would be worried about a few  
moves for white in reply to Rb6.  
39.e5 which protects the pawn  
40.and threatens to open up my kingside  
41.and gives a good square for the Nf2.  
42.So ok Nxa6 looks like it could be a move then.  
43.Rxa6  
44.Rxa6  
45.Qe1,  
46.oh see that allows Nf2 going to a difficult square...  
47.d3  
48.em there's actually go to be some tactical difficulties  
49.because here like Rc7  
50.is a threat for white.  
51.Maybe I should just move the Ne6  
52.stops opening up my king,  
53.em, problems.  
54.Ne6  
55.Na6 wins the pawn  
56.allowing me to build up slowly on the d6 pawn,  
57.push a pawn.  
58.c7 is a problem there.  
59.Ne6  
60.white can play e5 in that position.  
61.I think that's opening the white king.



62.I'd be better with the knight over protecting my king,  
63.better, ok don't need the extra time  
64.Ne6 is the move I would play.

#### S8 Position 6: B

1. Eh, ok again quite messy.
2. Why aren't I getting any positions that I like?
3. Controlled positions with clear simple plans.
4. Ok so what does this position look like?
5. Ok d7 is covered.
6. Material equal,
7. d-pawn looks strong,
8. b-pawn not so,
9. em the end position looks,
- 10.yea quite good for black,
- 11.em the only problem is d7
- 12.we have to take care of that eh otherwise,
- 13.the e5 square would be very nice for the knight
- 14.but he can't get there.
- 15.Em: Black's king looks quite safe.
- 16.Ok the obvious move would be d3
- 17.and say, now does that have any disadvantage?
- 18.Ok, eh could xc5
- 19.xd7,
- 20.that doesn't really do anything.
- 21.He can always hide there on g8,
- 22.hmm ok so still have a-pawn.
- 23.d3 looks like obvious move.
- 24.I wonder what Michelle is thinking there? (laughs).
- 25.Ok tactical solution here
- 26.xf2,
- 27.Qe1 is also possible.
- 28.Qe1 getting rid of that Q first
- 29.could make the whole thing safer even?
- 30.Yea how about Qe1,

31. what happens then?
32. Qe1
33. where does the knight go.
34. Knight doesn't go anywhere,
35. go to h3.
36. It's probably better even yea
37. maybe play Qf1/h1
38. followed by Rg2.
39. Kf5
40. is that really dangerous?
41. Qg6+
42. oh perpetual check is coming up.
43. Yea and I see so we do have to be careful.
44. Ok Qe1
45. Kf5 ok
46. if we take Qg6+,
47. now that looks too dangerous,
48. at least perpetual.
49. Ok the, what do we do now, have to hurry up.
50. Have to find something against Kf5
51. so maybe you need a move like Qc4,
52. yep I probably Qc4,
53. Rd5 blocks that!
54. Hmm not so simple at all!
55. Qg5
56. Qg6+,
57. quite tricky.

S1 Position 6: W

1. Now eh once again its very unclear position.
2. White's queen and rook are quite a way far down the board...
3. g6 is threatened queen
4. gaining an essential distilled advantage.
5. There seems to be some danger though for white with this position

6. at the that eh black is fairly well placed to launch some kind of eh desperate attack. The Rh2 on the seventh em looking at e2  
7. which is the only...the only only pawn separating the rook form the king.  
8. A move like Kc4  
9. for black with threaten  
10.Qb3 for black  
11.then Kb1,  
12.Kb1  
13.followed by eh Nb5  
14.which immediate mate  
15.while the Na3+  
16.Ke1  
17.Na3+  
18.Ra1++  
19.a nice little combination.  
20.I'm not sure what white can actually do against eh Kc4 threat  
21.but I'm thinking about b3  
22.to attack the rook  
23.and in Ra3+  
24.Bb2  
25.and perhaps a move like a4 anyway...  
26.just probing for more reasons  
27.a4  
28.followed by a3  
29.and Rxb2+  
30.and pxa4  
31.might be followed by Kc4 again.  
32.Em this time with the Qxb3+  
33.coming probably resulting in mate.  
34.So white probably has eh a couple of problems here  
35.although it looks like say well...  
36.the weakness of the light squares is is the key in this position.  
37.Another thing I'm quite tempted to sacrifice, ach to consider sacrificing eh a rook for knight  
38.Rxc7  
39.Qx

- 40.followed by Ne6
- 41.Nd4+
- 42.eh of course it won't work
- 43.because Rxe2+.
- 44.So probably the best idea in this position is for white to try and take control of the light squares
- 45.and he might be... be able to play moves such as Qf5
- 46.Qb3+
- 47.at which point there will be possibility of of eh playing em playing the knight back into play.
- 48.The only drawback with that eh is that the Ra4 is hitting the Nf4
- 49.white now has to be very careful as to how he does it...
- 50.course the other thing that would be possible Qf5
- 51.d3
- 52.followed by NxB
- 53.and if Rx Ba4
- 54.give white eh clear advantage
- 55.as the e3 pawn is quite weak
- 56.em although his bishop is poorly placed on c1...
- 57.the bishop is actually doing a good job of protecting...

## S2 Position 6: W

1. Ok, now this position... I see as very even again.
2. Eh, I'm looking at the business of the knight attacking the pawn in the centre
3. and the other pawn being pinned in a way against the queen protected by a king.
4. Also protected by the knight.
5. So the lines I'm looking at at the moment is to
6. RxN...
7. it's a dramatic type of move
8. but that doesn't give me much against the king.
9. Now we need to have something focused on... the queen...
- 10.so that would give me, if I prepare to shift the queen off the square
- 11.Bxp on e3

12.moves Qe3  
13.if he takes  
14.and I'm not sure whether that gives me what I'm looking for...  
15.eh black is dangerously close here  
16.with the pawns and also, ...  
17.eh has a rook bearing right down on my king position...  
18.so to make a move... I would ... leave the pe3  
19.and I would try to advance the e-pawn myself,  
20.because I have lots of squares covered all the way down to the queening square  
21.bar the last one  
22.and if I was to kind of take off the knight at an opportune stage  
23.it would be looking very good for me.  
24.So, I probably should analyse the black position a bit more in terms of the threats that...

#### S20 Position 6: B

1. Once again I'm black and the assessment is that black has 2-5 pawns two minors (minor pieces) two rooks and queen.  
2. White has queen, two rooks, two minors 2-4-5 pawns.  
3. So material looks pretty well equal.  
4. And black effectively has two passed pawns  
5. if he can do anything with them.  
6. So I think that is a possibility now  
7. so far white's attacks are concerned he hasn't anything very much.  
8. He has the queen  
9. which isn't going a whole lot of places...  
10.eh the knight black Nc5 has prospects  
11.and might be worth bringing it into play.  
12.e3, e6 sorry.  
13.Eh it doesn't affect anything immediately  
14.whereas at the moment if the pawn advances it is supported  
15.but it's already well supported with the eh, ... yea...  
16.oh the trouble is if that pawn advances on d3  
17.white might well then play eh his pf3 to f4

18. which means the pawn would be pinned against the queen
19. and eh, I don't eh yea.
20. White can't very well move the queen
21. because the knight is threatened then by the rook
22. and let me think.
23. All things considered
24. I'd be inclined to play d3, with the pawn.

#### S11 Position 6: B

1. Well on first glance it looks like my king is in some trouble.
2. My position black has some threats
3. or white has some threats along the h-file
4. and also along the seventh rank for example
5. he would like to play, I think eh
6. Rxe7
7. only for my knight is protecting that.
8. So em on the other hand I have some pressure against em against white position, white's king is not in a favourable position.
9. He seems to be leading the attack with the king
10. so, that might be of some benefit to me.
11. Em I think eh, a move I would look at, em I would actually like to play
12. Ne6.
13. Once again I have this problem with the eh e-pawn.
14. Em, but em, I think I could probably play this move em, (E; keep on talking),
15. eh I think I could probably play this move.
16. I could actually, there is another move I could play
17. Nxa6
18. and recapture the b-pawn
19. which might be an idea.
20. Actually I couldn't capture the b-pawn
21. because Ra5 cover it.
22. Em, possibly a good move might be to play Qd2
23. attacking knight
24. threatening to win the Nf2.

25.It is defended by the queen  
 26.and I don't see any other way for em, for white to hold on to Nf2.  
 27.He can play, he can play  
 28.Nh1,  
 29.eh I would also like possibly,  
 30.to bring my bishop into the attack if I could,  
 31.eh by playing Bxd6  
 32.except for that the queen is covering that,  
 33.so I think, I think a good move might be Qd2,  
 34.em I don't see any threats  
 35.any immediate  
 36.threats for white is Rh5,  
 37.so probably go after that.  
 38.I think, I think that would probably be best.  
 39.I think the position seems to be safe enough.  
 40.Qxd2 em, em bishop...

#### S12 Position 6: W

1. Ok, I have , I'm white I have 1- 5 pawns, against 1-5 pawns.
2. Eh I have bishop and knight vs bishop and knight two rooks queen.
3. Opposite colour bishop.
4. Queen quite advanced
5. Rg7 quite advanced
6. but black has Rh2
7. also well advanced on the seventh rank.
8. Black could play Rxe2+
9. But that's protected by the knight.
- 10.Can I attack in perspective, could play eh,
- 11.Nxd5
- 12.NxN
- 13.not good for me.
- 14.Pushing my pawn on my e5 pawn come racing through
- 15.so if I, can white sack the exchange
- 16.so white push pain.
- 17.RxN

- 18.QxN
- 19.Rxe2+
- 20.takes away the knight from my protected e2 pawn.
- 21.Nothing in that for black.
- 22.Ok, Kb1 may be reasonably well protected
- 23.Q attacks on b2,
- 24.so there's a danger of that happening
- 25.em, yet I need to be aware of that
- 26.so Rh2 attacking e2.
- 27.R+
- 28.give me big big problem.
- 29.But the Bc1 protects d2
- 30.so swing knight into action.
- 31.e6 looks like a reasonably good move.
- 32.Eh Re4
- 33.Queen back push on again,
- 34.still can't take,
- 35.push again Ne8
- 36.do something about that.
- 37.Re4 looks like a reasonable move.
- 38.In this position I think I would play e6. Yeah ...

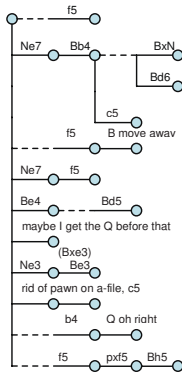
#### S13 Position 6: W

1. ...Em ok so
2. NxB
3. px
4. RxB
5. and I see NxB
6. Ne8
7. a strong intermediate move threatening the queen,
8. Qc5
9. ok queen ok
- 10.so I'm now considering Qg7
- 11.defending the eighth square
- 12.stopping the fork

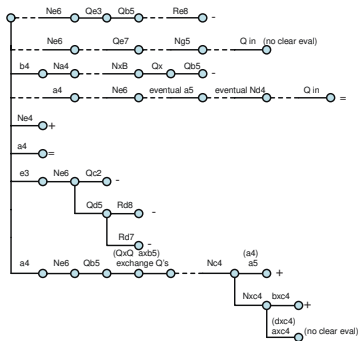


- 13.and threatening Qd3+
- 14.em, I'm also looking at the possibility after Qd6/f5 of
- 15.playing Kg3
- 16.so I'm attacking the rook
- 17.and if I play Qd3+
- 18.the king moves or approaches the rook.
- 19.So Qc5
- 20.Kh6
- 21.Qd3+
- 22.Rc4 (E: is that Qb3+ or)
- 23.No Qd3
- 24.Qf5..
- 25.R...
- 26.ok I'm looking at anyway that black can destroy get rid of my Nf4
- 27.which is defending my vital pawn e2.
- 28.so perhaps he can play...

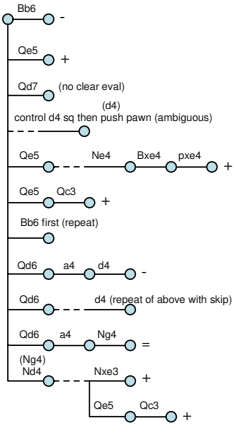
**Appendix I: A sample of twelve problem behavior graphs (three masters and three novices for the three normal positions and the three random positions).**



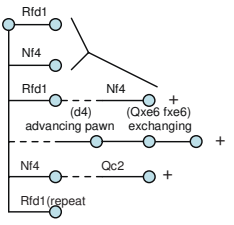
Example 1: Grandmaster (participant 4), normal position 1, black



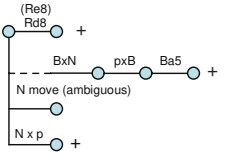
Example 2: International Master (participant 1), normal position 1, white



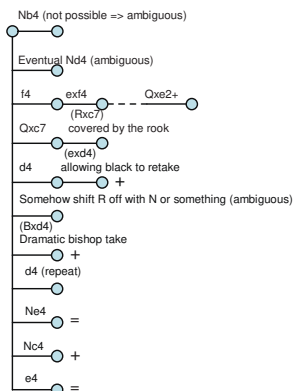
Example 3: Fide Master (participant 17), normal position 2, black



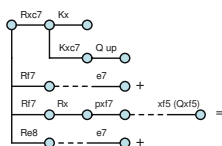
Example 4: Novice (participant 11), normal position 2, white



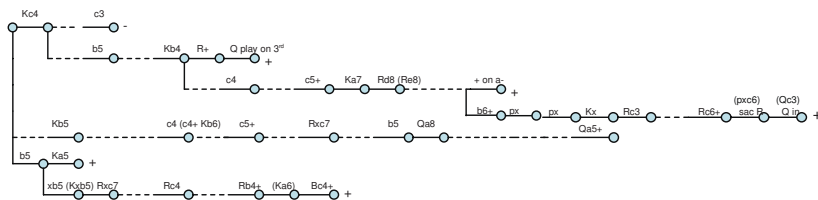
Example 5: Novice (participant 12), normal position 1, black



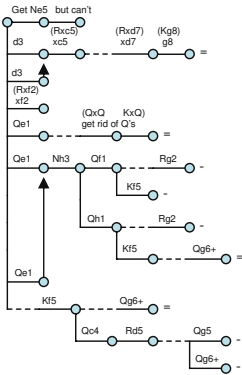
Example 6: Novice (participant 2), normal position 3, white



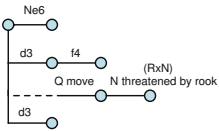
Example 7: Grandmaster (participant 4), random position 5, white



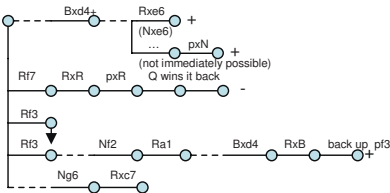
Example 8: International master (participant 1), random position 5, white



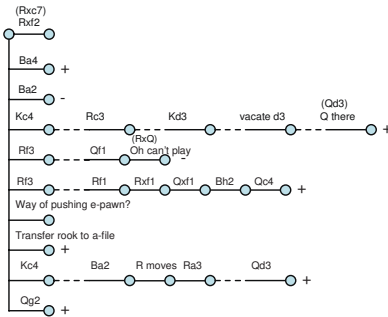
Example 9: International master (participant 8), random position 6, black



Example 10: Novice (participant 20), random position 6, black



Example 11: Novice (participant 12), random position 5, white



Example 12: Novice (participant 13), random position 5, white

## Appendix J: The experimenter’s think-aloud script used in Experiment 6

### *Instructions*

#### *Experimenter:*

“Thank you for participating in this study. I would like to remind you that your participation is anonymous and the data will be handled strictly confidentially. Should you wish to withdraw at any stage you are free to do so. This study is interested in the way chess players think about chess positions, and not in unconscious emotions or hidden thoughts.

You will be shown 7 chess positions including one practice position to ease you into the procedure. You will have 3 minutes to think out loud about each chess position. If you are finished thinking about the chess position before the time is up please feel free to say so. I will be recording you with a dictaphone and taking notes while you are thinking aloud. When your time is up I will ask you some questions about what you were thinking in the time.

In a moment you will be shown a (practice) chess position. You are asked to choose a move you would play in the way you are used to going about choosing a move in a real game. When time is up you will be asked to declare your chosen move. It is important that you say aloud everything that you think while choosing a move. If you stop talking at any stage I will prompt you with the words ‘keep on talking’. Also there is no need to explain why you are thinking about something, it is more important that your protocol is natural than comprehensible to me. Would you like to ask me any questions at this stage?”

*Response:*

*Experimenter:*

“Are you ready to start?”

***Stoppage:***

After a 3 second stoppage the first one to two seconds could be “ok, what do I do now?” on the third second the participant is likely to start doing something or looking for something so the words “keep on talking” was uttered quietly as outlined in van Someren *et al.* (1994).

***Retrospective Evaluation Script***

“Now I will ask you some questions about what you were thinking in the time you were analysing this chess position. I will remind you of some of the moves/move [perhaps a subject particularly non-expert will just analyse variations for one chosen move] you considered. I'll ask you to tell me whether or not a particular move led to a positive outcome for your position, in other words, an improvement of your position from its present state, or to a negative outcome for your position, in other words, to a worsening of your position from its present state, or whether it neither positively nor negatively affected your position. Is there anything you would now like to ask me about that?”

*Response:*

*Experimenter:*

“Okay, your first move was: (x)  
“Moves that followed on from (x) were (y) and then (z)  
(r) and then (s)  
...”

“Did the line {x, y, z, ...} lead to a positive or negative outcome, or neither?”  
(For example did the line Ne4, Bf5, Nd2 lead to a positive or negative outcome or neither?)

*Response:*

*Experimenter:*

“Did the line {**x**, r, s} lead to a positive or negative outcome, or neither?”  
(i.e. **Ne4**, Bd6, Qd5).

*Response:*

Single moves were also indicated.

*Experimenter:*

“So, did the move (**x**) lead to a positive or negative outcome, or neither?”

(e.g. “So, did the move Ne4 lead to a positive or negative outcome, or neither?”)

*Response:*

***Initial instructions for the remainder six positions were then modified as follows:***

*Experimenter:*

“Once again I will ask you to think-aloud while choosing a move in a chess position. Again it is very important that you say everything you think while thinking aloud. If you are finished thinking about the position before the three minutes are up, feel free to let me know. Ok are you ready to start?”

*Response:*

*Experimenter:*

“Again I will be reminding you of moves you looked at while choosing a move in this position and asking you to tell me whether they led to a positive negative or neither positive or negative outcome for your position. Ok are you ready to start?”

*Response:*



### *Verbal Debriefing*

“Thank you once again for participating in this study. Now I will tell you about what the study aimed to examine. In everyday thinking people generate hypotheses or ideas about incidences and relationships in the world around them. Psychologists have found that people tend to look for evidence to confirm their own ideas rather than look for evidence to prove their ideas false. This study was interested in how different levels of expertise affected confirmation or falsification of ideas in a specific domain — that is, the domain of chess. Is there anything else you would now like to ask me?”

*Response:*

**Appendix K: The experimenter's recording sheet used in Experiment 6**

Chess Position Number: \_\_\_\_\_

White/Black side: \_\_\_\_\_

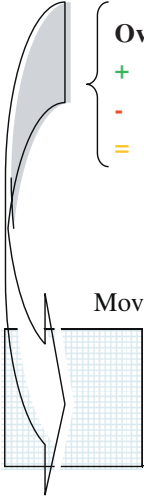
Move chosen overall: \_\_\_\_\_

**Overall Chosen Move Evaluation box**

+ : positive outcome for participant's position

- : negative outcome for participant's position

= : neither positive nor negative outcome for participant's position



Move/s chosen	Variation/s examined	Variation Evaluation	Chosen Move
	↓	↓	↓



**More  
Books!** 



**yes**  
**I want morebooks!**

Buy your books fast and straightforward online - at one of the world's fastest growing online book stores! Environmentally sound due to Print-on-Demand technologies.

Buy your books online at  
**[www.get-morebooks.com](http://www.get-morebooks.com)**

Kaufen Sie Ihre Bücher schnell und unkompliziert online – auf einer der am schnellsten wachsenden Buchhandelsplattformen weltweit!  
Dank Print-On-Demand umwelt- und ressourcenschonend produziert.

Bücher schneller online kaufen  
**[www.morebooks.de](http://www.morebooks.de)**

OmniScriptum Marketing DEU GmbH  
Bahnhofstr. 28  
D - 66111 Saarbrücken  
Telefax: +49 681 93 81 567-9

[info@omniscrptum.com](mailto:info@omniscrptum.com)  
[www.omniscrptum.com](http://www.omniscrptum.com)

OMNIScriptum 

