Chess Masters' Hypothesis Testing

Michelle Cowley (cowleym@tcd.ie)
University of Dublin, Trinity College,
Dublin 2, Ireland

Ruth M. J. Byrne (rmbyrne@tcd.ie)
University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract

Falsification may demarcate science from non-science as the *rational* way to test the truth of hypotheses. But experimental evidence from studies of reasoning shows that people often find falsification difficult. We suggest that domain expertise may facilitate falsification. We consider new experimental data about chess experts' hypothesis testing. The results show that chess masters were readily able to falsify their plans. They generated move sequences that falsified their plans more readily than novice players, who tended to confirm their plans. The finding that experts in a domain are more likely to falsify their hypotheses has important implications for the debate about human rationality.

Hypothesis Testing

People understand everyday and scientific phenomena by generating hypotheses to explain them. They achieve a true understanding only by *testing* hypotheses by searching for proof. There are two main ways people can test the truth of their hypotheses. They can either seek *confirmation*: evidence that is consistent with a hypothesis, or *falsification*: evidence that is inconsistent with a hypothesis. Falsification is generally considered better than confirmation: no matter how much evidence is gathered to confirm a hypothesis, there remains the possibility of refutation later (Popper, 1959). Confirmation could lead to the endorsement of untrue ideas and so if people are rational, they should attempt to falsify their hypotheses. Many cognitive scientists have interpreted experimental findings on hypothesis testing within the framework of falsification (e.g., Wason, 1960). The conclusion has sometimes been reached that when people fail to attempt to falsify, they fail to think rationally.

Early research on hypothesis testing found that people were prone to a *confirmation bias*: they tended to search for confirming evidence and avoid falsifying evidence (e.g., Wason, 1960). Confirmation bias has sometimes been viewed as evidence of human irrationality, for example, it may lead people to form prejudiced beliefs (e.g., Aronson, 1995). But the idea that human hypothesis testing is irrational presents a paradox: How can it be flawed given that it has led to important civil, technological and scientific discoveries? There are two possible answers: one possibility is that testing hypotheses through confirmation is more useful than indicated by a Popperian analysis, and a second possibility is that people are more capable of falsification than experimental evidence has revealed so far. We will first outline the view that confirmation is a useful strategy to test hypotheses and the view that falsification may be conceptually impossible (e.g., Poletiek, 1996). Then we will show that falsification is in fact possible in a domain that has been a trusted test-bed for theories of cognition for almost forty years: chess problem solving. We will consider experimental results that testify to high levels of falsification in the hypothesis testing of chess masters (Cowley & Byrne, 2004).

Confirmation: Vice or Virtue?

Irrational hypothesis testing in the form of confirmation bias was first reported in the 2-4-6 task (Wason, 1960). Participants in this task are required to discover the rule to which the number triple 2-4-6 conforms. The participants are analogous to scientists and the rule is analogous to a law of nature to be discovered. Participants test their hypotheses by generating other number triples and they are told by the experimenter whether each triple conforms to the rule or not. The rule in the 2-4-6 task is the deliberately general rule of 'any ascending numbers'. The salient features of the 2-4-6 triple tend to induce incorrect hypotheses, for example, participants tend to focus on its properties of even numbers and numbers ascending in twos. Participants who generate these hypotheses, 'ascending even numbers' or 'numbers ascending in twos' can discover the real rule 'any ascending numbers' in only one way: by generating triples that falsify their hypothesis. For example, a participant could try to falsify the 'ascending even numbers' hypothesis by generating the triple '3-5-7'. They would discover their hypothesis is false when the experimenter tells them that '3-5-7' is consistent with the real rule. But participants overwhelmingly generated confirming triples such as '10-12-14'. The triple confirms their hypothesis and it is also consistent with the real rule and so the experimenter tells them '10-12-14' is consistent. They announce their incorrect hypothesis as the rule and fail to solve the task correctly.

Confirmation bias has been demonstrated many times in the 2-4-6 task and in other related laboratory tasks, for example, in a task in which participants are required to discover the law governing the motion of particles in an artificial universe displayed on a computer screen (e.g., Mynatt, Doherty, & Tweney, 1978).

But do people confirm their hypotheses only in artificial laboratory tasks? Perhaps they are better able to falsify in real world contexts where they can access their knowledge about the task? In fact, the tendency to confirm has been observed in NASA Apollo mission scientists (Mitroff, 1974), and in the notes of scientists such as Alexander Graham Bell (Gorman, 1995). It is possible that confirmation is useful. For example, the participants who successfully discovered the rule in the complex artificial universe task tended to be those who confirmed their hypotheses in the early stages of their attempted discovery of the rule, and then tried to falsify when they had a well-corroborated hypothesis (Mynatt et al., 1978). Perhaps it is only when a hypothesis worth testing has been established that it is necessary to attempt to falsify it. Confirmation and falsification may be complementary strategies for successful hypothesis testing.

But it is also possible that people do not falsify because they cannot (Poletiek, 1996). According to this view when people generate a hypothesis it is their *best guess* about the truth, and it does not make sense for them to try to show that their best guess is wrong. In a version of the 2-4-6 task, participants were encouraged to generate their best guess about what the rule might be and then they were instructed to perform falsifying tests on it. The instruction to falsify decreased the number of positive triples, such as '10-12-14', which is consistent with the best guess 'even ascending numbers'. The instruction to falsify also increased the generation of negative test triples, such as '3-5-7' which is inconsistent with the best guess 'even ascending numbers'. However, this test is only a falsifying one if the participant expects the experimenter to say that the triple is consistent with the real rule (and then the participant would know that the hypothesis 'even ascending numbers' was wrong because ascending odd numbers are consistent too). But if the participant generates the inconsistent '3-5-7' triple and expects the experimenter to say that it is *not* consistent with the real rule, then they have attempted to confirm their hypothesis (albeit with a negative triple). In fact, participants generated triples that were inconsistent with their hypothesis (negative triples) *but* they expected them to be inconsistent with the experimenter's rule. The participants could not seem to make sense of the instruction to falsify. The instruction to falsify may be impossible to carry out (Poletiek, 1996).

Given the ideas that confirmation is useful and falsification is impossible, does it follow that the normative prescription of falsification is flawed, rather than human rationality? Perhaps, not. Even when a hypothesis is the best guess it is not necessarily an accurate representation of the truth. We turn to the case for falsification next.

Falsification and the Path to Truth

Consider the following example:

> *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 hypothesize 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 SARS epidemic. They concluded that the 'new virus' hypothesis was correct. The case where the 'other virus' was absent falsified the 'other virus' hypothesis and corroborated the 'new virus' hypothesis. The example illustrates how important falsification can be.

There are many situations in which it is helpful to anticipate the ways in which a hypothesis or plan could go wrong. For example, it may be helpful to falsify in interactions with a collaborator or opponent, whether in contexts such as political or social engagement, or in games such as tic-tac-toe or poker. The importance of considering what might go wrong is observed in cases of military strategy, for example, in Northern Ireland (Mallie, 2001). Attempts to falsify hypotheses, particularly plans of action, could help reduce costly mistakes.

The merits of falsification are not lost on experts, as the SARS example illustrates. It may even be the case that the ability to falsify is part of what makes an expert (Cowley & Byrne, 2004). The competitive nature of science may ensure that different groups of scientists attempt to falsify their opponent's theories even if they only attempt to confirm their own. The refutation of a theory is often discovered by someone who did not invent the theory (Kuhn, 1996). Hypothesis testing in scientific discovery may benefit from a strategy of attempting to confirm a hypothesis until there is sufficient corroboration for it to be considered seriously, and then attempting to falsify it, just as in the 'artificial universe' task (Mynatt et al., 1978). Perhaps more importantly, experts may generate high quality hypotheses from the outset. An exceptional scientist such as Alexander Graham Bell may have tended to confirm rather than falsify his hypotheses because his expertise ensured that his hypotheses were exceptionally good (and there is a smaller potential set of falsifying evidence for a good quality hypothesis than for a poor one).

As these observations suggest, a more systematic study of expert hypothesis testing is warranted. We chose the game of chess as our expert domain because it meets the essential criteria: it is possible to identify a large sample of experts whose expertise is objectively defined and categorised relative to each other, and it is a task that draws directly on participants' expert knowledge and experience.

Chess and Hypothesis Testing

Studies of chess have contributed substantially to understanding cognition, including problem solving (Newell & Simon, 1972), chunking in working memory (Chase & Simon, 1973), and expertise (De Groot, 1965). Findings from research on chess have successfully explained expertise in non-game domains such as physics (e.g., Larkin, McDermott, Simon, & Simon, 1980). Chess offers great potential for an investigation of expert hypothesis

testing. Of course, choosing a move in chess may depend on a variety of processes including accessing a large repository of chunked domain knowledge about possible opponent moves (e.g., Chase & Simon, 1973; Gobet, 1998a). Our suggestion is that hypothesis testing may be one of several important processes for selecting a move in chess. We expect that expert master players will be better than novices at falsifying their planned moves by thinking about opponent moves that could ruin their plan (for details see Cowley & Byrne, 2004).

Our key research question is, do experts and novices differ in their ability to find refutations to lines of play in chess? We conceptualize hypothesis falsification in chess as finding opponent moves that refute the moves a player examines for play. The opponent moves could ruin the player's plan and worsen the player's position. We address an important aspect of choosing a move that has never been systematically investigated: the evaluation of move sequences.

Hypothesis Testing in Chess

The overall goal of chess is to checkmate the opponent by attacking the opponent king and eliminating all the possible ways the opponent king can escape the attack. Chess thinking may consist of exploring different alternative paths in a 'problem space' (Newell and 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*, that is, in chess the way chess pieces are allowed to move. For example, a bishop operates diagonally backwards and forwards and captures on the square it lands on for any one move.

At the beginning of a game of chess the two players have equal numbers of pieces and theoretically equal chances of securing a win, loss or draw. To secure the best possible result the players must play moves they hypothesize to be so good that they cannot be refuted (Saariluoma, 1995). Refutation (that is, hypothesis falsification) occurs when the opponent plays a move that results in a worsening of the player's position. For example, a player may play a move that he or she plans to be a good move, but the opponent replies with a move that stops the player's plan. The opponent's play worsens the player's position and reduces the player's chance of a win.

There may be three major processes in the choice of a chess move: exploration, elaboration and proof (DeGroot, 1965). Evidence of hypothesis testing is available in the proof process. A chess player tests how good a move is by mentally generating move sequences following on from that move. For example, a move sequence might be: "If I move my knight to that square, you might move your pawn to attack my knight, and then I will have to retreat, and that is really bad for me…". In this example, the move sequence is evaluated as leading to a negative outcome: a falsification has been found for the knight move. Move sequences can be evaluated as leading to either a positive, negative or neutral outcome for a chess position. We conceptualize move sequences that are judged to lead to a positive outcome as akin to evidence confirming that a particular move is a good move. Move sequences evaluated as leading to a negative outcome are akin to evidence falsifying a move that was thought initially to be a good move. Move sequences evaluated as leading to a neutral outcome are neutral evidence.

Accessing Hypothesis Testing in Chess

We carried out an experiment on hypothesis testing in chess players (see Cowley & Byrne, 2004, for details). The 20 participants (19 men and 1 woman) were registered members of the Irish Chess Union. The participants were classified according to the Elo system, which calculates expected playing strength value on the basis of tournament and league results, and the value varies from approximately 1000 for an absolute novice and over 2800 for the world champion. We tested *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). All international class masters living in Ireland at the time participated in the study (for further information on participant details see Cowley & Byrne, 2004).

We presented the participants with 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. Importantly, they were 'equality outcome' positions, where there were equal chances with best play for both black and white pieces. This constraint ensured that there would be no obvious confirming or falsifying move sequences. The positions were chosen with the assistance of a chess expert (who was not a participant in the study). See figure 1 for an example of a chess position used.

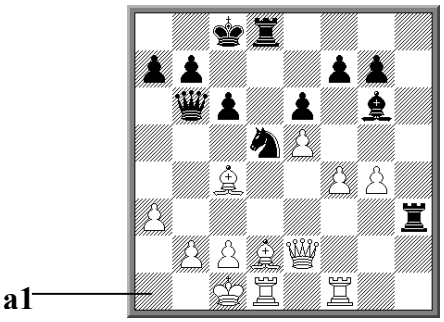

Figure 1: Position 1 with white to play (and the co-ordinate a1 is also illustrated in this diagram).

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. It is instructive to focus on the master level players (for

comparison with masters studied in the chess literature previously) and to this end we selected the think-aloud protocols of five *master level* players (i.e. 1 Grandmaster, 2 International Masters, and 2 Fide International Chess Federation Masters), and compared them to the think-aloud protocols of five novice chess players, chosen at random from the full sample of novices (for other analyses see Cowley & Byrne, 2004).

Moves examined by the player during think-aloud are verbalized using algebraic chess notation, for example a sequence of moves verbalized was: f5 exf5 gxf5 Bh5 Qg2 Rh4. This notation describes each piece and the location of the square it will go to on the chess board. Each square on a board has a location name called an algebraic coordinate. The letters a-h are horizontally along the bottom of the board. The numbers 1-8 are vertically up the board. Each type of piece is given a letter in upper case format. Each coordinate is given a letter in lower case format alongside a number. So for example, the move 'Ra1' refers to a rook piece (R) moving to the a1 square. Or, the rook could go to the b1 square to the right of a1 (Rb1). A sequence of such moves is a move sequence. All of the players were sufficiently fluent with algebraic notation to be able to 'think aloud' using it. 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 we used a combined methodology of think-aloud followed by retrospective evaluation. Verbalized move sequences were recorded not only by dictaphone but also by the experimenter (the first 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 lead to a positive, negative or neutral outcome for their positions.

Scoring confirming and falsifying hypothesis tests

The transcribed think-aloud protocols for the responses to the normal board positions were segmented into episodes, move by move. We constructed 'problem behavior graphs' (using Newell and Simon's guidelines) for the responses to the three normal board positions for each of the ten selected participants (thirty problem behavior graphs in total). These graphs plot each move sequence and its corresponding retrospective evaluation. To illustrate we present in Figure 2 a small fragment of a master's problem behaviour graph for one board position.

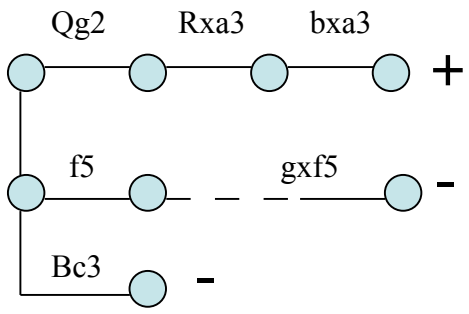

Figure 2: A fragment of a problem behaviour graph constructed from a chess master's protocol.

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, Qg2 means the player thought aloud about the possibility of moving his queen to the g2 square. Next the player thought aloud about a possible reply from his opponent to his Qg2 move, that is, the move Rxa3 where the opponent moves their rook to the a3 square, and the x indicates that the rook captures a piece, in this case a pawn. Next the player thought aloud about his reply to this opponent's move, that is, bxa3 where he would move his pawn on the b file to a3 (pawn moves do not have a P in front of them), and capture the opponent's rook. The plus sign shows that the player evaluated this move sequence as positive for him. The next line sequence begins with the player thinking about f5, that is, a pawn moves to the f5 square. The next utterance the player makes is gxf5, that is, the pawn on the g file of the board captures a pawn on the square f5. This move is only possible for the player and not his opponent. The player has generated a move sequence that mentions only his own moves and does not mention opponent moves. The dashed line captures these *skipped* moves. The minus shows a negative evaluation. Each problem behavior graph incorporates the think-aloud move sequences with the retrospective evaluation (positive, negative, or neutral).

We used *Fritz 8* (one of the most powerful current chess programs) to obtain an objective evaluation of the chess position that occurs at the final move of each sequence (i.e. terminal node). For readers familiar with *Fritz 8*, we used the infinite analysis module, in which each move sequence is evaluated at least from 11ply from the terminal node (see Chabris & Hearst, 2003). The evaluations provided by *Fritz 8* enable us to identify move sequences that would genuinely be positive or negative for a player. We could distinguish between the move sequences that a player indicated as leading to a positive outcome for their position and that the program established would lead to a positive outcome if played, from the move sequences that a player identified as positive, but that the program established would in fact lead to a negative outcome. We conceptualize confirmation bias as a move sequence that a player evaluates as leading to a positive outcome for them, when in

fact it leads to a negative outcome. Likewise, we were able to distinguish the move sequences that a player identified as leading to a negative outcome and that the program established would be negative if played, from move sequences that the player identified as negative, but that the program established were in fact positive for them. We conceptualize falsification as a move sequence that a player evaluates as leading to a negative outcome for them, when in fact it leads to a negative outcome.

Hypotheses Testing in Chess Masters' Thinking

Masters tended to think about 8 move sequences on average for each board position, and experienced novices tended to think about 6 move sequences. A total of 218 move sequences were generated by the 10 players for the three normal board positions (N = 122, M = 8.1 for each board position for the masters, N= 96, M = 6.4 for novices).

Four types of move sequences were identified from the problem behaviour graphs. (1) 50% were *complete move sequences* where every move for the player and his or her opponent was articulated along the move sequence. (2) 25% were *skipped move sequences* where an essential move was not mentioned somewhere in the move sequence. (3) 19% were *base skip sequences* where the first move or 'base move' of the sequence was not mentioned. (4) 6% were *ambiguous move sequences* where the move sequence could not be interpreted. Only the complete move sequences lend themselves to objective evaluation by *Fritz 8*, so we concentrate our analysis here on these hypothesis tests (see Cowley & Byrne, 2004 for further details).

A complete move sequence is scored using the following criteria: (a) whether it is predicted by the player to lead to a positive, negative, or neutral outcome, and (b) whether it is evaluated objectively by *Fritz 8* as leading to a positive, negative or neutral outcome. Thus there are nine possible hypothesis tests for complete move sequences, as Table 1 illustrates. Confirmation bias corresponds to the '+/-' cell in Table 1, and falsification corresponds to the '-/-' cell. These two types of evaluation accounted for 42% of all evaluations.

Table 1: Objective and subjective evaluations of move sequences ('+' refers to a positive evaluation, '-' to a negative one, '=' to a neutral one; '+/-' means the player's evaluation was positive and the program's evaluation was negative).

	Fritz 8's evaluations		
	Positive	negative	neutral
Player's evaluations			
Positive	+/+	+/-	+/=
Negative	-/+	-/-	-/=
Neutral	=/+	=/-	=/=

Falsification In three of the cells of Table 1 (the three on the diagonal from upper left to lower right), the subjective evaluation of the player matches the objective evaluation of the computer program. One of these matching cells is particularly important for our prediction that experts falsify more than novices: genuine falsification occurs in the situation captured by the '-/-' cell in Table 1, in which the player and the program both evaluated the outcome of the move sequence as negative. Chess masters generated more of these falsifying move sequences than novices (M = 3.2 for masters, M = 1.2 for novices) and this difference was reliable (t (8) =2.02, p = .039).

The result indicates that chess masters are capable of falsifying their plans by identifying opponent moves that would worsen the master's position. People are able to falsify (pace Poletiek, 2000). Domain expertise may facilitate this falsification. Moreover, the moves chosen by chess masters for play at the end of each of the three board positions were evaluated by *Fritz 8* as objectively better moves than novices (the quality of moves is measured in terms of 'pawn advantage' or 'pawn disadvantage', and it was +0.309 pawn advantage for masters compared to –1.2 pawn disadvantage for experienced novices). The result is consistent with the idea that the ability to falsify may contribute to making better moves in chess.

Confirmation Bias Confirmation bias occurs when a move sequence is evaluated subjectively by the participant as leading to a positive outcome, but evaluated objectively by the computer program as leading to a negative outcome (the **'+/-'** cell in Table 1). The results show that novices produced somewhat more instances of confirmation bias than masters (M = 2.6 for novices and M = 1.6 for masters). Although the difference was in the predicted direction it was not reliable (t (8) = 1.443, p = .094).

Positive and Negative Testing The nine test types in Table 1 can be categorized into three groups: (1) Objective tests: the player's positive, negative and neutral evaluations matched *Fritz 8*'s evaluations (the three cells on the diagonal from upper left to lower right mentioned earlier), and this category includes the falsification tests. (2) Positive bias tests: the player's evaluation was more positive than *Fritz 8*'s. The three cells in this category include the second and third cells in the first row (**'+/-'**, **'+/='**), and the middle cell in the third row (**'=/-'**), and this category includes the confirmation bias tests. (3) Negative bias tests: the player's evaluation was more negative than *Fritz 8*'s. The three cells in this category include the second and third cells in the first column (**'-/+'**, **'=/+'**), and the middle cell in the third column (**'-/='**).

Chess masters generated reliably more objective tests than novices (M = 6.6 for masters and M = 2.4 for novices). Novices generated somewhat more positive bias tests than the masters (M = 5 for novices and M = 3.4 for masters), but the difference was not reliable. They generated a similar amount of negative bias tests (M = 1.8 for masters, M = 1.2 for novices), as Figure 3 shows.

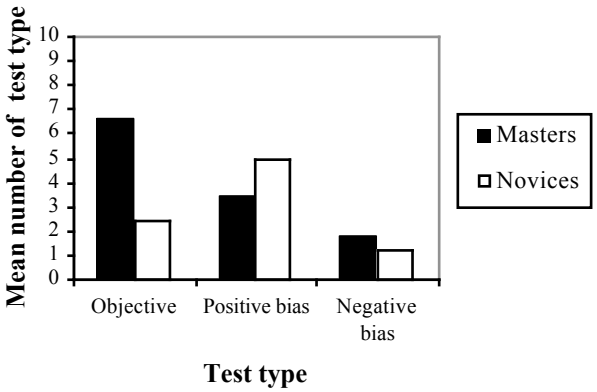

Figure 3: The mean number of objective tests, positive bias tests and negative bias tests generated by masters and novices. (Instances of falsification and confirmation bias are included in these categories).

Conclusions

People are capable of falsifying their hypotheses. Our experimental results show that chess masters falsified their hypotheses: they thought about how their opponent might refute their plan in their move sequences. Chess masters tended to evaluate their moves as good or bad for them more realistically than experienced novices: their judgments matched the objective evaluations of one of the most highly advanced chess computer programs, *Fritz 8*. Experienced novices exhibited something of a confirmation bias: they tended to think about how their opponent would play moves that fit in with their plan, somewhat more than chess masters did. Novices, somewhat more than masters, tended to evaluate their moves as better for them than they were objectively. The evidence that chess masters can falsify suggests that it may be premature to conclude that the normative prescription of falsification is flawed. In this case falsification can be considered a useful and rational strategy.

Hypothesis testing may be influenced by domain expertise. How does domain knowledge affect the ability to falsify by chess experts? We plan to explore this question by examining how masters test their hypotheses for random board positions compared to novices. If falsification relies on domain knowledge, then masters should tend not to falsify their hypotheses about move sequences in the random board positions as often as they do in the normal board positions. Nonetheless, they may attempt to falsify more than experienced novices, if their expertise has helped them to develop a strategy of falsification in this domain.

Acknowledgements

We thank Grandmaster Alexander Baburin, Fintan Costello, Phil Johnson-Laird, Mark Keane and Caren Frosch for helpful comments, and Peter Keating for help with the problem behaviour graphs. Special thanks to Mel O'Cinneide for help with the chess positions. This research was funded by the Irish Research Council for the Humanities and Social Sciences.

References

Aronson, E. (1995). *The Social Animal*. New York: Worth/ W. H. Freeman.

Chabris, C. F., & Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: effects of playing speed and sight of the position on grandmaster chess. *Cognitive Science, 27*, 637-648.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed), *Visual Information Processing*. New York: Academic Press.

Cowley, M., & Byrne, R. M. J. (2004). Hypothesis testing in chess masters' problem solving. *Manuscript in preparation.*

De Groot, A. (1965). *Thought and choice in chess*. The Hague: Mouton.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data.* USA: MIT Press.

Gobet, F. (1998a). Expert memory: A comparison of four theories. *Cognition, 66*, 115-152.

Gorman, M. E. (1995). Confirmation, disconfirmation, and invention: The case of Alexander Graham Bell and the telephone. *Thinking and Reasoning, I (1),* 31-53.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. USA: University of Chicago Press.

Larkin, J. H. Mc Dermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335-1342.

Mallie, E. (2001). *Endgame in Ireland*. London: Hodder & Stoughton.

Mitroff, I. (1974). *The subjective side of science.* Amsterdam: Elsevier.

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., & Simon, H. A. (1972*). Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology, 49A*, 447-462.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Saariluoma, P. (1995). *Chess players' thinking: A cognitive psychological approach.* UK: Routledge.

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