MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„The relation between expertise, performance and difficulty estimation in chess“

verfasst von / submitted by

Aisha Futura Tüchler, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 013
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Degree Middle European interdisciplinary master programme in Cognitive Science UG2002
Betreut von / Supervisor:	Prof. Dr. Ivan Bratko

Abstract

The present thesis aims to investigate the relation between expertise, performance, perceived performance and difficulty estimation in chess. A combination of methods was employed to cast light on the meta-cognitive phenomena of difficulty estimation and assessment of one’s own performance. 20 scholastic chess players participated in an experiment, solving 12 tactical chess problems, as well as assessing their own success in solving the problems and estimating the problems’ difficulty. Our findings suggest that expertise, here reflected by a chess player’s rating, is predictive for the success in solving the problems, but only weakly correlated to success in the difficulty estimation. The overestimation of incorrectly solved problems or the underestimation of successfully solved problems could not be evidenced in our sample. Interestingly, there was an overall correlation between actual and perceived success, albeit with a significant difference between unskilled and skilled participants. The discrepancy between perceived and actual success rate decreased with increased success. The present research thus represents first evidence for the Dunning-Kruger-Effect in the area of chess research. The employed Machine Learning analysis yielded that models for predicting difficulty derived from a previous experiment were useful when tested on the data of this experiment. The surprising results concerning the relation between performance, here success in solving the problem, and difficulty estimation suggest that further exploration of this neglected area of research is indicated.

Die vorliegende Arbeit beleuchtet die Relation zwischen Expertise, Leistung und Schwierigkeitseinschätzung im Schachspiel. Eine Kombination aus Methoden wurde angewandt, um die meta-kognitiven Phänomene Schwierigkeitseinschätzung und Bewertung der eigenen Leistung näher zu beleuchten. 20 Schachschüler nahmen an diesem Experiment teil, welches daraus bestand 12 taktische Schachprobleme zu lösen und die eigene Leistung sowie die Schwierigkeit der Probleme einzuschätzen. Unsere Ergebnisse deuten darauf hin, dass Expertise, hier das Rating der Schachspieler, bestimmend ist für den Erfolg beim Problemlösen, jedoch nur schwach mit Erfolg bei der Schwierigkeitseinschätzung korreliert ist. Das Überschätzen der Schwierigkeit von Problemen, welche nicht richtig gelöst wurden und Unterschätzen der Probleme, die richtig gelöst wurden, konnte in unserer Stichprobe nicht nachgewiesen werden. Die Diskrepanz zwischen tatsächlichem und geschätztem Erfolg wurde kleiner mit einer höheren Erfolgsquote. Die Arbeit stellt erste Hinweise für den Dunning-Kruger-Effekt im Schachspiel dar. Eine Analyse mit Maschinellem Lernen ergab, dass Modelle, welche aus Daten eines vorherigen Experiments abgeleitet wurden, eine gewisse Voraussagekraft bezüglich der Daten aus diesem Experiment haben. Die Ergebnisse bezüglich der Relation zwischen Leistung, in diesem Kontext Erfolg beim Problemlösen, und Schwierigkeitseinschätzung deuten darauf hin, dass dieses bisher wenig beachtete Forschungsgebiet näher zu beleuchten wäre.

Table of contents

1. Introduction

1.1 Topic of the Master thesis

The present Master thesis is concerned with human problem solving and aims to cast light on the relation between expertise, performance, perceived performance and difficulty estimation. This relation is studied in the domain of chess which represents a high-validity environment and is hence suitable for the study of human problem solving and expertise. Expertise in this case refers to the individual's proficiency in the problem solving domain and is reflected by a chess player's rating (see section 2). Performance is the individual's ability to correctly solve a given task – here a tactical chess problem (see section 2.3) – and perceived performance is the assessment of one's own success.

A problem is considered difficult when the person tackling the problem has a hard times solving it. If a given sample of people try to solve a problem, the same problem might be perceived more difficult by some, and less by the others. There are, however, problems that are considered difficult by more people than other problems, and those are presumably the problems, which were also solved by fewer people. It is evident, that when comparing two problems – one solved by half of the sample, and another one merely solved by 15% – the latter problem is probably the more difficult one. It can hence be said that problems with a lower solution frequency and a higher probability of failure are more difficult. But are solution frequency, reflecting objective difficulty, and perceived difficulty really two sides of the same coin?

The objective of this thesis is to detect influential variables of a player's difficulty estimation as well as to investigate the relation between performance and perceived performance in chess.

The investigation of the metacognitive phenomenon of difficulty estimation and perceived performance is of interest to many disciplines and is primarily investigated by cognitive psychology (see section 1.2), educational sciences and neuroscience (Fleming & Dolan, 2012). For the present investigation Machine

Learning methods are employed to unravel hidden patterns concerning the relation of difficulty, performance and expertise.

1.2 Review of Literature

Borg et al. (1971a) showed that perceived difficulty of a test group, which had to solve 9 items of an intelligence test, was inversely related to solution frequency of a second control group with higher sample size. Problems with low percentage of correct solutions in the control group were hence also rated more difficult by the test group with a correlation of r=0.9. This overall strong positive correlation was replicated by other studies involving a visual search task (Borg et al., 1971b) and a different intelligence test, assessing reasoning and spatial ability as well as verbal comprehension (Borg et al., 1971a). These findings suggest an overall straightforward relation between objective and subjective difficulty. Nonetheless, this relation is questioned by one finding of the same study: It could be shown that subjects, who were not able to solve given tasks correctly, estimated those more difficult than subjects who did solve them correctly. This, however, was not true for all the involved sub-test, but only for the reasoning and spatial ability part. In another study using the Raven Progressive Matrices, Borg et al. (1971a) could show that perceived difficulty increased as a monotonically increasing function with order of the items, suggesting that also motivation and fatigue played a role in the perception of difficulty.

In accordance with Borg et al. (1971a+b), in their experiment Touroutoglou & Efklides (2010) show that perceived difficulty, assessed by a 7-point Likert-scale, were higher in tasks with low performance. When comparing perceived difficulty before and after completion of a given task, it increased for the low performance task, whereas for the other task it stayed the same or even decreased. This then suggests that there is a relationship between performance and perceived difficulty, termed Feeling-Of-Difficulty by Touroutoglou & Efklides (2010).

Desender et al. (2017) also demonstrated that perceived difficulty, here termed subjective experience of difficulty, is associated with variables that are indicative of performance. In their study involving a masked priming experiment where the participants had to determine in which direction a shown arrow was pointing, and later report about the perceived difficulty when solving the task (a binary variable), they could show that cues like reaction time and response repetition, amongst others, contribute to perceived difficulty. Furthermore, they showed that training influences the relation of the stated variables to perceived, meaning cues to better estimate difficulty can be learned.

Despite findings suggesting a straightforward relationship between objective and subjective difficulty, it can be concluded that there are nevertheless variables that influence perceived difficulty such as training and learning or motivation, as well as cues that are indicative of performance, suggesting that the relation to objective difficulty is not as straightforward as reported by Borg et al. (1971a+b).

Furthermore, the role of expertise and its relation to actual and perceived performance shall be discussed. Playing chess can be seen as situated in a high-validity environment where moves have determined outcomes, the number of possible moves are finite and the task environment is fully observable, and hence is considered to be a framework where expertise is likely to develop and be effective (Chase & Simon, 1973). In their 2014 paper, Hristova, Guid & Bratko nevertheless found that the chess players' expertise, reflected by their rating, was not significantly correlated to success in solving the given problems. This research will be discussed in more detail in section 3. However, Park & Santos-Pinto (2010) showed that expertise does play a role when trying to predict outcomes of a chess tournament. They reported that when a sample of players were asked to estimate what percentage of the tournament's participants will be eliminated before them, their forecasts, as opposed to poker players' forecasts, were not random guesses. This is presumably due to the fact that the rating system in chess (Elo rating system, see section 3) provides information about their competitors' skills and competences. Nevertheless, Park & Santos-Pinto (2010) could show that the chess player had an overestimation bias concerning their own performance; the forecast error, however, was lower for the participants with a higher rating.

1.3 Hypotheses

In the next section, the reasoning process behind the formulation of the hypothesis is presented and it is highlighted why the relation between difficulty estimation and performance was deemed to be associated.

As stated in the introduction, Borg et al. (1971a) as well as Touroutoglou & Efklides (2010) show that success in a given task and perceived difficulty are related in a way that suggests that success in a given task has an influence on the perceived difficulty. The main hypotheses that will be tested in this thesis are that people tend to over- and underestimate the difficulty of given problems depending on whether they were able to solve these problems correctly or not.

H1a – People tend to overestimate the difficulty of problems they did not solve correctly.

H1b – People tend to underestimate the difficulty of problems which they solved correctly.

Additionally, it will be tested whether their confidence in the correctness of their solution and their judgment about difficulty are associated.

H2 – People tend to underestimate the difficulty of problems which they think they solved correctly.

Park & Santos-Pinto (2010) showed that stronger players with a higher rating are better predictors of their own success in chess tournaments. Thus, it can be suspected that stronger players will also be more accurate in assessing their own success post-hoc.

H3 – Higher skilled individuals will be better in assessing their own success in solving the presented problems.

2. Chess Ratings

As mentioned in the introduction, the present work aims to investigate the relation between performance, perceived performance, expertise and difficulty estimation.

In particular, the discrepancy between difficulty estimation by humans compared to statistical methods is of interest. In chess, these statistics are provided by two main ratings systems, which will be described in the following pages: The Elo rating is used by the international chess federation FIDE (Fédération Internationale d'Échecs) and gives account of a player's strength, and the Glicko rating, a modification of the Elo rating, which here is used to indicate a chess problem's difficulty (Elo, 2008; Glickman, 1998). Contrary to other competitive sports, there is no absolute measure of performance in chess; the players' ratings hence need to reflect their relative strength in comparison to their competitors.

Ratings allow to monitor the process of acquisition of skill and expertise and are useful for the design of tournaments as it becomes possible to let allegedly similarly strong players compete with each other which increases suspense and contentment among the contestants. In the present experiment we use both of these two ratings. Therefore, in this section, the two systems are explained, commonalities and differences are discussed and their predictive power is assessed.

2.1 Elo rating system

The Elo-rating was developed in the late 1950ies by the Hungarian-American physicist Arpad Elo. The United States Chess Federation assigned him to assess the flaws and weaknesses of the currently used rating system whereupon Elo proposed a formula with sound probabilistic underpinning, the Elo rating formula.

As for other characteristics of a human like intelligence or height, also performance in sports is suspected to be normally distributed. This is deemed to be true also for chess. In fact, studies done by Elo (1946), as well as by McClintock (as reviewed in Elo 2008, p. 19) provided evidence for the validity of the claim that chess performances are normally distributed. The Elo rating formula is hence derived from the normal probability function (see Figure 2.1), also called the Gauss error curve or standard sigmoid. The curve displays the probability of a player winning against an opponent in a chess game in as a function of their difference in rating.

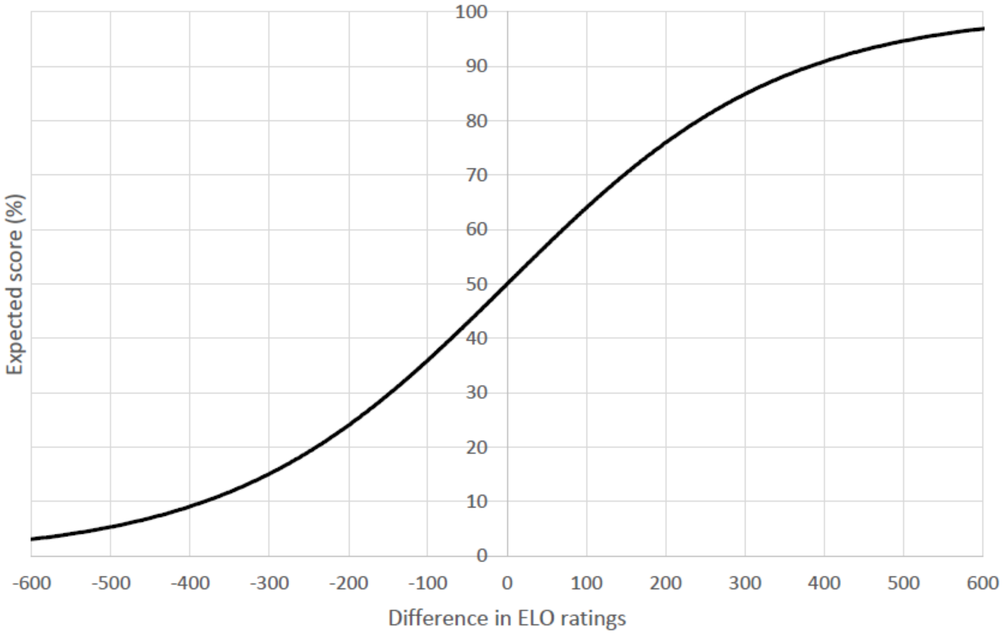


Figure 2.1 The Elo curve with expected scores for the played game. Retrieved from Hristova, Guid & Bratko (2014)

The formula for updating a player's Elo rating after each game played is the following

$$\mathrm{R_n = R_o + k(W - W_e)}$$

R_n is the player's new rating after the game, R_0 the old rating and W the game score (1 for win, 0 for lose and $\frac{1}{2}$ for a draw), W_e the expected game score based on the two players' old ratings.

The factor k is a relative weight, which is chosen for the pre-match rating and the performance rating of the match. A high k gives more importance to the recent performance, whereas a low k values the earlier performances.

The Elo rating being a four-digit number grew out of tradition as the chess federations determined this arbitrarily and it was retained in order to be accepted by players at the time. Also the 200 points range between classes and the 2000 points as a reference point discriminating between amateurs and experts were already common practice before Elo's intervention in the rating system (Elo 2008, p.19). Despite being already developed in the 1950ies, the Elo system was only adopted in 1970 after the FIDE (Fédération International d'Echecs) Congress in Siegen, former West Germany (Elo 2008, p.3).

The Elo rating system proved to be predictive and is also used to validate other approaches for modelling a chess player's strength (Van der Maas & Wagenmakers, 2005), as well as in predicting the outcome of football games (Hvattum & Antzen, 2010). Furthermore, it is also applied for modeling social dominance in primates (Newton-Fisher, 2017) and is used in adaptive educational systems (Pelánek, 2016).

2.2 Glicko rating system

The Glicko rating system builds on the system developed by Arpad Elo but introduces some improvements to the before mentioned. Within the Elo system, it is not relevant, how dated a player's rating is when he is competing in a tournament, the points won or lost only depend on the rating difference between him and the opponent and on the k factor chosen for this occasion. According to Glickman (1998), datedness makes ratings unreliable, or at least less reliable than ratings of players which play with regularity. The Glicko system seeks to incorporate time dependency into the calculation of the rating to account for that.

The Glicko rating system introduces a variable named Rating Deviation (RD). A player's RD increases with time passing and decreases with shorter intervals in playing games (Glickman, 1998). The underlying statistical concept of the Rating Deviation is the same as the Standard Deviation, which means that with a probability of 95% a player's actual – but not known – rating is within the range of the current rating and +/-2 RD.

The incorporation of the RD does not only provide a more realistic rating for a player but also influences the degree to which a player's rating changes after a won or lost game, respectively. Due to this, the Glicko system, unlike the Elo system, is not a symmetrical rating system as two players with different RD might compete with each other; the points won by the one might not be equal to the points lost by his opponent.

2.3 ChessTempo and ChessTempo rating

ChessTempo is an online chess platform with 746444 registered users (as by 20th of September 2018) which provides more than hundred thousand tactical problems to be solved by its users (ChessTempo, n.d.). A problem is denominated tactical “if the solution is reached mainly by calculating possible variations in the given position, rather than by long term positional judgment with little calculation of concrete variations” (Hristova, Guid & Bratko, 2014, p.728). The 12 tactical problems of our problem set were retrieved from the website on January 2nd 2018 and the problem specific statistics presented in Table 3.1 (page 10) represent the information provided by that date. For the ratings on the ChessTempo website, the Glicko rating system is used with an adaptation concerning the consideration of repeated attempts by the players.

ChessTempo ratings are calculated as follows: Both the user and the presented problem receive ratings, starting with a default rating at the beginning, which is then updated according to the Glicko rating system. If a user solves a problem correctly, the rating of this problem decreases whereas the user’s rating increases. Consequently, if a problem is not solved, its rating increases. It foresees penalties for re-attempts in the range of loss of the full credit of 30% for one repeat, and up to 85% for six or more repeats of a problem. Even though this only applies to 1 to 3% of the players on this online platform (Rating system, n.d.), it is nevertheless useful as it prevents the most successful users to be those who have the best memory of already seen problems. This mechanism is termed Duplicate Reward Reduction (Rating System, n.d.). When a user solves a problem at the first repeat, he only gets 70% of the full credit, 55% after the second repeat and only 15% if he has seen it six or more times. If a problem was not presented to a distinct player for 6 months or a year, it will receive 45% and 75% of the full credit, respectively, independent of how many times it was solved or attempted by that player in the past.

ChessTempo hence provides a statistically sound measure of difficulty, which will be taken as basis for the difficulty classification in the present thesis. For the experiment, only problems with more than 6500 attempts are included to ensure reliability of the difficulty ratings.

3. The experiment

The present experiment represents a replication of the experiment done by Hristova, Guid & Bratko (2014) and was slightly modified in terms of sample size (12 subjects in the 2014 experiment, opposed to 20 in the present) and composition and the chess problems used in the experiment. The players in the 2014 experiment had ratings between 1845 and 2279 (2089 on average, +/- SD 134.65), indicating they were much stronger chess players than the players in our sample (see below). The participants solved 12 tactical chess problems of three difficulty classes ‘Easy’, ‘Medium’ and ‘Difficult’. There were 2 problems of the class ‘Easy’ with ChessTempo ratings between 1492.5 and 1495.3, 4 problems of the class ‘Medium’ with ratings between 1875.2 and 1883.3, and 6 problems of the class ‘Difficult’ with ratings between 2230.9 and 2274.9 (1493.9 on average for easy problems, 1878.8 on average for medium problems and 2243.05 on average for difficult problems). After solving all the 12 problems, the participants had to rank them from ‘easiest’ with the rank 1 and ‘most difficult’ with rank 12.

For our experiment, 20 subjects, 18 males and 2 females, between 7 and 15 years old (average age 11.3 years, SD +/-2.30), were recruited from different elementary and high schools as well as from the Chess Association of Slovenia (Šahovska zveza Slovenije) in Ljubljana.

In our sample of participants, the ratings ranged from 1500 to 2000, with an average rating of 1644.5 (+/- SD 222.94). Players with a rating of 1000 to 1199 are classified as novices, this being the fifth lowest class. Players with a rating between 1200 and 1399 classify as class D amateurs, and those with a rating between 1400 and 1599 as class C amateurs (30% of our sample). With a rating between 1600 and 1799 a player classifies as class B amateurs (25% of our sample). 8 players (40%) had a rating between 1800 and 1999, hereby classifying as class A player, i.e. a strong amateur or club player. Beginning with a rating of 2000 and above, a player is classified as expert, this being true for one player in our sample (Elo 2008, p.18. 9 players of our sample had an international Elo-rating in addition to the domestic, Slovenian one. The international rating was always substantially lower than the Slovenian rating, this assumedly due to the fact that these players do not play tournaments with international players often. Keeping in mind that the Elo rating reflects relative strength of a player and his competitors, another reason for the low Elo rating of the participants might be that they competed only to other weak players hence not being able to earn a lot of points. Another thing to mention about scholastic players is that their ratings are mostly generated within a group of scholastic chess players who rarely compete with adults (Glickman 1995). The risk of increasing ratings if once competing outside of their distinct group, which do not reflect a player's strength is hence given.

Analogous to the 2014 experiment, the players had to solve 12 tactical chess problems presented to them on a screen using the chess software ChessBase14. These problems were of three difficulty levels, namely 'Easy', 'Medium' and 'Difficult'. Problems with a ChessTempo rating of 1000 +/-15 are considered easy problems, problems with a rating of 1420 +/-15 as medium and problems with a rating of 1750 +/-15 as difficult. The experimental problem-set contained four problems of each category 'Easy', 'Medium' and 'Difficult' (see Table 3-1).

Table 3.1 ChessTempo statistics of the problem set and classification by the experimenters

#	ChessTempo ID	Rating	Attempts	Average time	Success rate	Difficulty
1	664	999	9325	1:08	69%	Easy
2	61725	1004	8037	0:45	69%	Easy
3	5358	1005	8820	0:39	69%	Easy
4	44557	1005	9315	1:07	69%	Easy
5	26533	1317	11657	2:35	63%	Medium
6	6908	1329	11635	1:04	63%	Medium
7	47751	1318	11431	1:28	63%	Medium
8	50864	1318	11598	1:46	62%	Medium
9	15914	1758	8182	2:04	53%	Difficult
10	73069	1754	6928	2:54	54%	Difficult
11	12973	1765	7448	4:39	53%	Difficult
12	52221	1762	7728	5:09	53%	Difficult

The column ‘Attempts’ refers to the number of users of the website ChessTempo who attempted to solve the given problem and the column ‘Success rate’ gives account of which percentage did so successfully. The time needed to solve the problem, averaged over all the attempts, is listed in the column ‘Average time’.

The participants were presented all the 12 problems of the problem set one by one in a randomized order on a computer screen (see Figure 3.1). After the participants chose their move, the experimenter moved the piece on the screen and, if applicable, moved for the opponent’s side. If the subject’s first move was incorrect, the experimenter would terminate this game and move on to the next problem without communicating if the subject solved it correctly or not. In nine of the problems, it was white to move, in three problems it was black to move (as in Figure 3.1). The number of moves to be made to solve the presented problem ranged from 2 to 6 moves. For the problem in Figure 3.1, the subject would choose a move and communicate this to the experimenter who would execute the move.

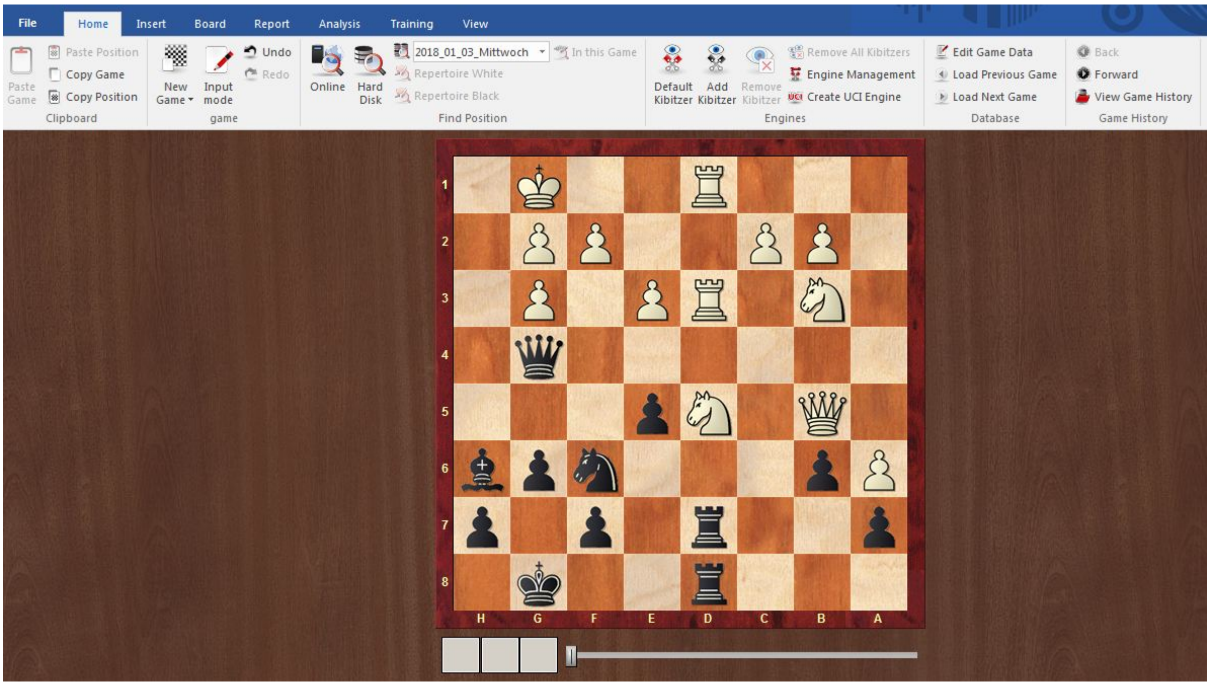


Figure 3.1 Presentation of problem on the computer screen

In this case, the correct move is to move the black rook from d7 to d5 and capture the white knight. Then the experimenter would move for the opponent’s side, in this case move the white rook from d3 to d5, hereby capturing the black rook and wait for the subject to choose a second move and so on. After a wrong move or up to – in this case – maximum 6 moves chosen by the participant, the experimenter would terminate the game and move on to the next problem. All depictions of the problems, including solutions, can be found in the Appendix.

After solving the given problems, either correctly or not, the players were subject to a short semi-structured interview in which they were asked about the characteristics of the given problem (e.g. spotted motifs) and other possible solutions, perceived success, as well as their assessment of the difficulty of each problem. As guiding question to explaining the approach to solving the problem the participants were asked why they chose this distinct move. In addition, they were inquired if they see others ways of solving the problem, if they deemed it easy or difficult and if they think the solved it correctly.

The questions asked after each tackled problem were the following, always in the presented order: 'Do you think you solved the problem?', 'Was this problem easy or difficult for you?', 'Why did you choose this move?' and 'Are there other ways of solving the problem?'. The first questions could only be answered by 'yes' or 'no', if a subject opted for another answer, the experimenter made him or her choose between the mentioned two. The answer for the second questions could also just be one of the possibilities 'easy' or 'hard'. Again, if the subject opted for another response, the experimenter forced a decision between the mentioned. In the question about the choice of move, the subjects had to explain the reason for their respective move. An arbitrarily chosen answer would be the one of Subject 4, who in the problem presented in Figure 3.1, moved the rook to d5 and stated, "It is forced mate. I open the line through which I will be able to take [the rook at] d1 with the queen."

The subjects' accuracy in estimating their own performance was assessed by the Mean Square Contingency Coefficient or phi-coefficient, a measure of association of two binary variables. In our case these binary variables were success in solving the problem (yes or no) and the answer to the question 'Do you think you solved the problem?' (yes or no).

The formula for the phi-coefficient is

$$\varphi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

with a and d being the numbers of occasions a participant's perceived success was congruent with his actual success, and b and c being the numbers of occasions when they were incongruent (see Table 3.2).

		Variable 2 **Perceived success**	
		1	0
Variable 1 **Success**	1	a	b
	0	c	d

Table 3.2 Contingency table with frequency distribution for the variables Success and Perceived success

The phi-coefficient compares the product of the diagonal cells from up left to down right with the product of the cells from up right to down left. The denominator of the formula ensures that the result of the equation is between +1 and -1.

For assessing the participants' difficulty estimates, a pairwise comparison of selected problems of the problem set was done by the player. The participants were presented with two positions at a time (see Figure 3.2) on a computer screen and had to state which of the two they consider more difficult or that the two positions were equally difficult. In contrast to the experiment by Hristova, Guid & Bratko (2014), a pairwise comparison was favored over a full ranking of the problems. This is due to the fact that the sorting that needs to be performed for a full ranking of 12 problems is difficult to carry out for the participants. It might result in them putting more importance to the extremes of the scale, neglecting the ranks in between, hence leading to inaccuracies. It was also favored over the assessment of difficulty via a Likert scale, as these scales pose the problem of calibration.

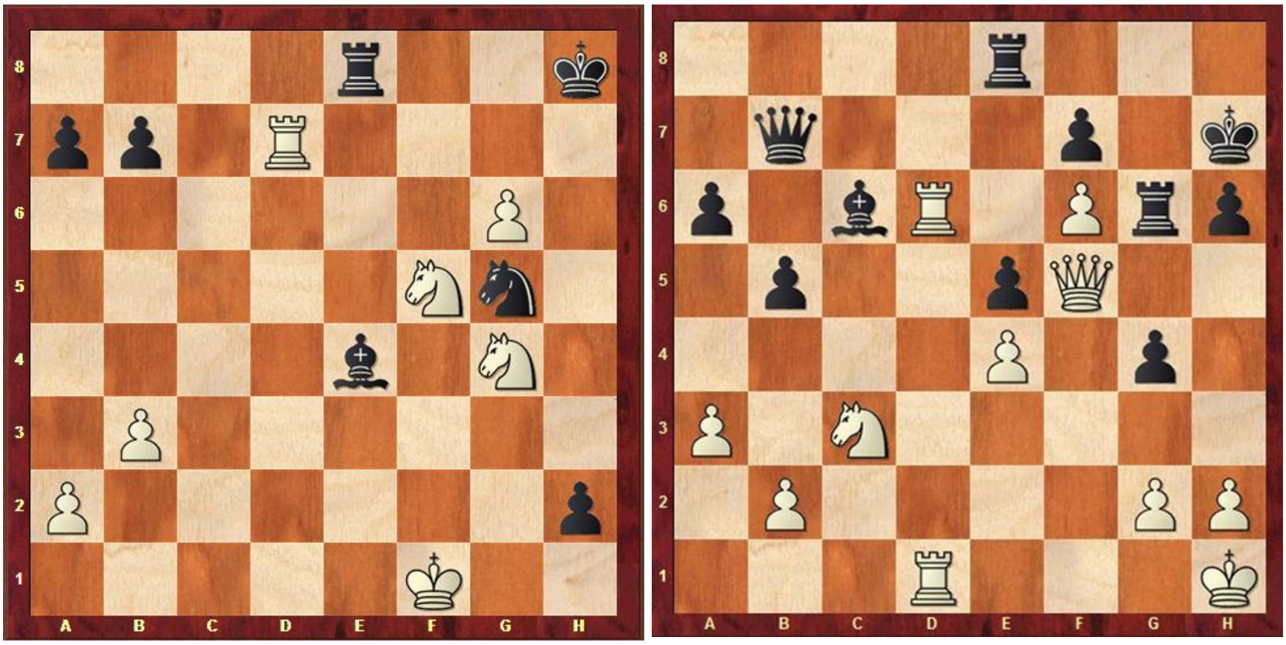


Figure 3.2 Comparison of Problem 9 and Problem 5 of the problem set

After comparing 10 problem pairs, the number of correctly rated pairs was counted, the maximum possible number of correct pairs being 10, the minimum 0. Additionally, an error score was calculated according to the scheme presented in Table 3.3.

		Participant's answer		
		A<B	A=B	A>B
Correct answer	A<B	0	1	2
	A=B	1	0	1
	A>B	2	1	0

Table 3.3 Penalties for incorrect answer in the pairwise comparison

In Figure 3.2, the problem on the left hand's side is a problem of the class 'Difficult', whereas the problem on the right hand's side is of the class 'Medium'. The correct relation between the two is hence '>', as Problem 9 on the left is more difficult than Problem 5 on the right, according to the ChessTempo rating. According to Table 3.3, a participant who would rate the problems' difficulty as equal would get 1 penalty-point added to his error score. If the participant rated the left problem as less difficult, he would receive 2 penalty-points added to the error score. This then reflects the degree to which the difficulty was misjudged.

Subject 20 Pairwise comparison	Subject's answer	Correct answer	Penalty
Prob9-Prob5	<	>	2
Prob6-Prob10	<	<	0
Prob2-Prob3	<	=	1
Prob7-Prob3	<	>	2
Prob10-Prob11	=	=	0
Prob7-Prob8	=	=	0
Prob1-Prob5	>	<	2
Prob4-Prob8	>	<	2
Prob5-Prob6	>	=	1
Prob12-Prob8	=	>	1
Sum of correct pairwise comparisons	3	Error score	11

Table 3.4 Pairwise comparisons and calculation of the error score

In the following the terms underestimation and overestimation will be used if a given problem was under- or overestimated in relation to a distinct pair in the pairwise comparison. In fact, it is more accurately a relative overestimation or underestimation.

To assess whether a participant over- or underestimated the difficulty of a given problem, an estimation score is calculated (see Table 3.4). For every vote in the pairwise comparison, it is determined if the participant over- or underestimated a given problem, or if he estimated the difficulty of the problems correctly. In the example illustrated in Figure 3.2, a participant who would rate the two presented problems as equally difficult (‘=’) or would rate Problem 9 as being easier (‘<’), would underestimate the difficulty of Problem 9 and overestimate the difficulty of Problem 5. This means, that here the terms underestimation and overestimation are used if a given problem was under- or overestimated when estimating the difficulty of distinct pair in the pairwise comparison. In fact, more accurately, it is a relative overestimation or underestimation.

Problem	Over	Under	Correct	Average sum	Estimation
1			0	0	Correct
2	+1			+1	Over
3		-1	0	-1/2	Under
4	+1			+1	Over
5	+1		0 + 0	+1/3	Correct
6	+1		0	+1/2	Over
7			0 + 0	0	Correct
8		-1	0 + 0	-1/3	Correct
9		-1		-1	Under
10		(-1) + (-1)		-2	Under
11	-1			+1	Over
12			0	0	Correct

Table 3.5 Calculation of the estimation score

For calculation of the estimation score, the scheme in Table 3.5 is used. If in a pairwise comparison a problem like Problem 3, is underestimated on one occasion and correctly estimated on another occasion, -1 and 0 are added together. As the problem is voted on twice, the sum of +1 and 0 is divided by the number of votes, resulting in $-\frac{1}{2}$.

For every value between $-\frac{1}{3}$ ≤ x ≤ $+\frac{1}{3}$ the problem is considered to be correctly judged. For values higher than $+\frac{1}{3}$ the problem is considered overestimated, for values lower than $-\frac{1}{3}$, it is considered underestimated. These thresholds were chosen because they well reflect the participant's estimation in ambiguous cases, as can be derived from the participant's judgment of Problem 5 (see Table 3.5): The estimation score results being $+\frac{1}{3}$ when having relatively overestimated the problem in 1 paired comparison, but correctly estimated in 2 other comparisons. As the number of correct estimations is higher than the number of overestimations, Problem 5 is hence considered correctly estimated. The threshold $+/-\frac{1}{3}$ is hence chosen to discriminate between correct estimation and over- or underestimation.

4. Analysis of Data

4.1 Descriptive statistics

20 subjects, 18 males and 2 females, between 7 and 15 years old (average age 11.3 years, SD +/-2.30), from different elementary schools as well as from the chess school of the Chess Association of Slovenia (Šahovska zveza Slovenije) in Ljubljana participated in the experiment. The players' ratings ranged from 1500 to 2000, the average rating being 1728.25 (+/-SD 168.79). The average player in our sample hence qualifies as a class B amateur. On average, the participants solved 6.35 (+/- SD 3.22) problems correctly, this being slightly more than half of the problems presented. The minimum of solved problems per participant was 0 problems solved, the maximum 11. In Table 4.1, the problem specific statistics are presented. The table displays the problems' IDs as found on the ChessTempo website, their respective ratings on January 2nd, their rating on September 20th, and their average success rate on these two dates.

Table 4.1 Table of ChessTempo statistics of the problem set January 2nd 2018 compared to September 20th 2018

#	ChessTempo ID	Attemps 01/2018	Attempts 09/2018	Rating 01/2018	Rating 09/2018	Success rate 01/2018	Success rate 09/2018
1	664	9325	9874	999.0	989.9	68.58%	68.61%
2	61725	8037	8649	1004.4	1012.0	68.65%	68.38%
3	5358	8820	9240	1005.0	999.9	68.93%	68.92%
4	44557	9315	9907	1004.8	983.0	68.38%	68.55%
5	26533	11657	12521	1317.2	1326.1	62.51%	62.34%
6	6908	11635	12197	1329.3	1331.0	63.03%	62.99%
7	47751	11431	12276	1317.6	1305.5	62.77%	62.82%
8	50864	11598	12359	1317.7	1312.6	62.42%	62.51%
9	15914	8182	8433	1757.7	1750.0	53.40%	53.46%
10	73069	6928	7311	1753.7	1755.8	54.11%	54.12%
11	12973	7448	7822	1764.5	1765.4	52.85%	52.84%
12	52221	7728	8116	1761.6	1751.6	51.73%	53.78%

As depicted in Table 4.1, the four problems of each difficulty class are comparable in terms of rating within a class, but differ in this parameter when compared to problems of other classes. Looking at the success rates, the percentage of attempts which lead to solving of the problem, in the different classes, the difference between the classes is much less pronounced, especially not between problems of the class ‘Easy’ and the class ‘Difficult’. According to these statistics, the ratings are generally reliable predictors of success, as problems with higher ratings are solved by fewer users of the website, and vice versa. This is especially true when comparing the class ‘Difficult’ to the classes ‘Medium’ or ‘Easy’. When comparing the class ‘Medium’ to the class ‘Easy’ this is not so apparent.

It can also be seen in Table 4.1 that neither the ratings of the problems and even less so their success rating changed drastically between the date of the selection of the problems and today. This indicates reliability of the ratings provided by ChessTempo.

When looking at the statistics of our sample in Table 4.2, it becomes evident that the class ‘Difficult’ differs significantly from the classes ‘Easy’ and ‘Medium’, and, as in the ChessTempo statistics, not so much between the classes ‘Easy’ and ‘Medium’. The difficulty classes are defined as described in the section where the experiment is explained. Problem 1 in Table 4.1 being the exact same problem as Problem 1 in Table 4.2. There is, however, a difference in time needed to solve the problems, this not so much for the classes ‘Easy’ and ‘Medium’, but very pronounced for the class ‘Difficult’.

Table 4.2 Table of the participants’ statistics in solving the problem

#	Rating	Average time	Success	Perceived success	Difficulty
1	999	0:38	60%	85%	Easy
2	1004	0:17	80%	90%	Easy
3	1005	0:34	65%	85%	Easy
4	1005	0:47	75%	90%	Easy
5	1317	0:53	60%	80%	Medium
6	1329	0:41	70%	80%	Medium
7	1318	0:55	55%	75%	Medium
8	1318	0:53	75%	90%	Medium
9	1758	0:42	15%	75%	Difficult
10	1753	0:40	30%	70%	Difficult
11	1764	1:21	10%	65%	Difficult
12	1762	1:19	20%	65%	Difficult

Compared to the ChessTempo statistics, the difference between the class ‘Difficult’ and the classes ‘Medium’ and ‘Easy’ is nevertheless more pronounced. Remarkably, the perceived success, the percentage of problems the participants thought they solved, is higher than the actual success, the discrepancy being bigger in the problems of the ‘Difficult’ class. This finding will be discussed in more detail in section 6.2.

4.2 Correlation between difficulty estimation and success in solving the problem

In Table 4.3 we can see the subjects’ performance in the difficulty estimation reflected by the number of pairs which they rated correctly in the pairwise comparison and the above described error score (see Table 3.4 and 4.3). On average, the participants rated 3.9 (+/- SD 1.12) pairs correctly, with a minimum of 2 pairs rated correctly, and a maximum of 6. The mean error score was 8 (+/- SD 1.72), with a maximum of 11 and a minimum of 5.

When correlating success in solving the problem including all the 240 instances – 12 problems solved by 20 subjects – and the participants’ error score, a very weak, negative and non-significant (p=.565) correlation of r= -.037 resulted. From this is can be concluded, that success in solving the problem is negatively related to a high error score, even though the relation is very weak and not significant.

The correlation with the participants’ ratings, however, yielded that there is negative relation between rating and error score, meaning the higher the rating of a player is, the lower the error score. Despite being only a weak correlation r=-.144 with p=.026, the participants’ ratings seem to be a better predictor for correct difficulty estimation than it is the case for the success in solving the problem.

Table 4.3 The participants' success, their perceived success, their success in the pairwise comparison and the respective error score

Participant	Rating	Problems solved	Perceived solved	Correctly rated pairs	Error score
1	1559	1	11	4	8
2	1500	3	9	4	8
3	1500	3	4	4	7
4	1850	10	11	5	6
5	1896	8	11	3	8
6	1719	6	10	5	7
7	1919	8	7	6	6
8	1566	4	7	3	8
9	1600	2	11	5	5
10	1500	0	12	3	10
11	1516	6	11	3	11
12	1950	10	11	5	6
13	1700	8	9	5	7
14	1823	6	12	2	10
15	2000	9	11	2	10
16	1650	6	11	4	7
17	1758	8	12	4	8
18	1867	11	11	3	8
19	1812	9	10	3	11
20	1880	9	10	5	6

Additionally, there was an age effect for both the success in solving the problems and the correct assessment of one's performance. When correlating age and number of problems solved there was a significant positive correlation (r=.795, p=.000), and when correlating age and the phi-coefficients, reflecting correct assessment of one's performance, there was as well a significant positive correlation of r=.619 (p=.004)

4.3 Correlation between over- and underestimation and success in solving the problem

When correlating success in solving the problem including all the 240 instances – 12 problems solved by 20 subjects – and the difficulty estimation of the presented problems by means of the Mean Square Contingency Coefficient or phi-coefficient, no significant association between success in solving the problem and bias in the difficulty estimation could be found.

Problems which were not solved by the participants were not significantly (p=.687) overestimated in the pairwise comparison. The correlation being r=.026, it can be stated that there is no or a negligible relationship between these two variables.

The same is true for the inverse, problems which were solved correctly were not significantly (p=.882) underestimated by the participants. The degree of relation being r=.01, so negligible.

4.4 Correlation between difficulty estimation and perceived success

There is a very weak negative and non-significant (p=.239) correlation (r=-.76) between problems the participants thought they had solved correctly and their bias in estimating their difficulty. It can be concluded that there is no or a negligible correlation between perceived success and underestimation of the given problems.

4.5 Correlation between subjects' rating and their success in solving the problems

In contrast to findings of Hristova, Guid & Bratko (2014), this research showed a positive, statistically highly significant (r=.854; p=0.000) relation between expertise, reflected by the subjects' rating, and success in solving the problem. This correlation was computed by including the number of solved problems by the participants and correlating them to the players' ratings. The 2014 findings hence could not be replicated in this experiment, as the correlation of rating and success in solving the problem was merely r=.062. According to the 2014 results, there was no significant correlation between expertise, reflected by the rating, and success in solving the problems. In Figure 4.1, the relation between the participants' ratings and their success in solving the problems in displayed.

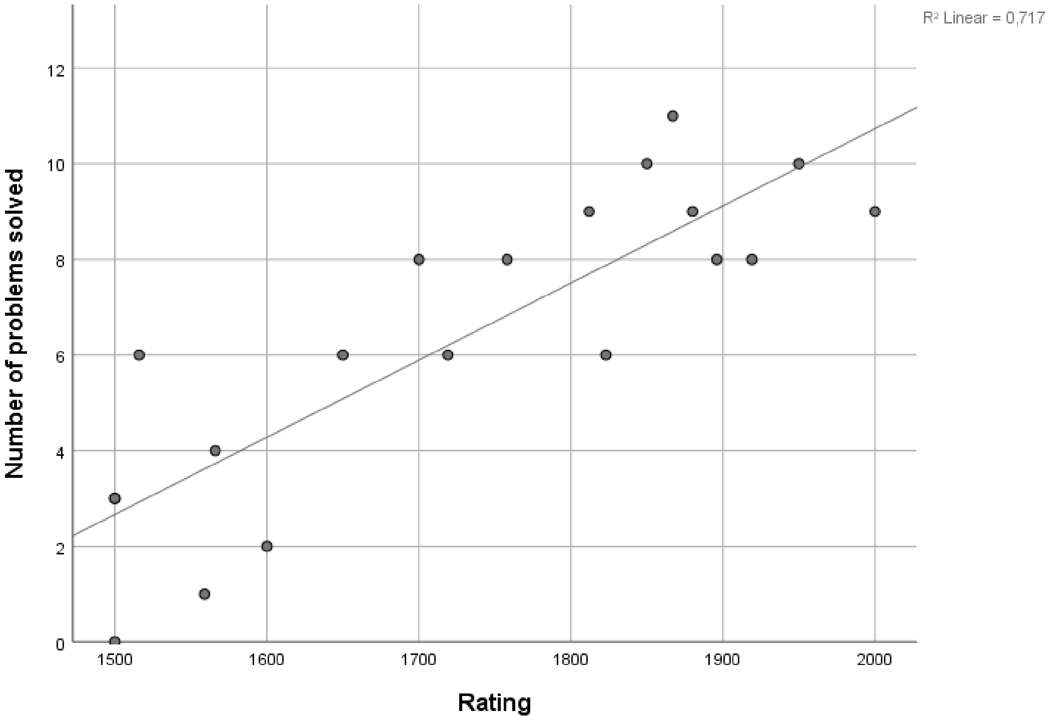


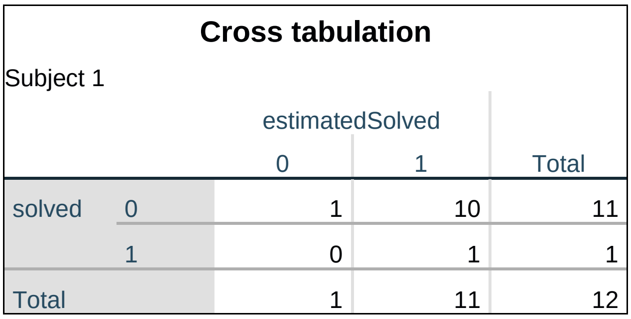
Figure 4.1 Relation between rating and the number of problems solved

4.6 Correlation between success rate and perceived success rate

The data show that the overall correlation of perceived and actual success was moderately high (r=.422) and highly significant (p=.00). However, the participants showed overconfidence in their competencies. While estimating to have solved 10.5 (+/-SD2.01) problems correctly on average, the mean of problems actually solved was in fact 6.35 (+/-SD 3.21). A paired T-Test yielded that the participants overestimated their performance significantly (p=.000), on average thinking they solved 3.7 (+/-SD 3.51) problems more than they actually did.

For participants in the bottom quartile according to their success rate, perceived and actual success only correlated to a low degree (r=.108) and not significantly (p=.413), whereas the upper quartile showed a correlation of r=.767 with a p-value of p=.000. This indicates that worse performers differ from better performers in their judgements about success in solving the given problems making them worse predictors of their own success.

In fact, the data also show that the difference between perceived success rate and actual success rate decreased with increased performance, meaning that subjects who were able to solve more problems correctly were also more correct in their judgments about their success. Perceived success was assessed by a question after each presented problem, enquiring if they think they were able to solve it or not. This overconfidence was more prevalent amongst subjects with little or no success in solving the given problems. The most remarkable example being a subject, who did not manage to solve any of the twelve presented problems, but estimated having solved all of them. In Table 4.4 and 4.5 we can see two cross tabulations of two subjects and their actual and perceived number of solved problems. The subjects were randomly chosen to illustrate the relation of the phi-coefficient and the presented cross tabulations. If most of the data falls along the diagonal cells from up left to down right, like in Table 4.5, the two variables are considered associated

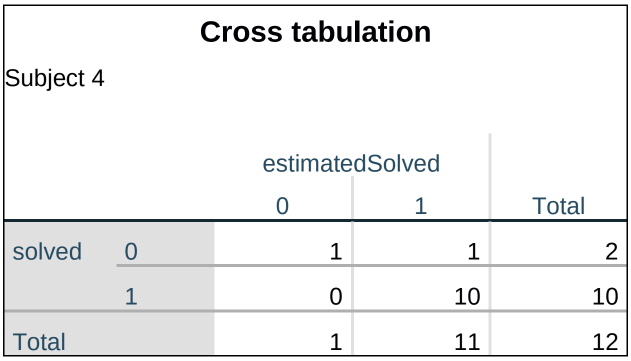


Cross tabulation

Subject 1

		estimatedSolved		
		0	1	Total
solved	0	1	10	11
	1	0	1	1
Total		1	11	12

Table 4.4



Cross tabulation

Subject 4

		estimatedSolved		
		0	1	Total
solved	0	1	1	2
	1	0	10	10
Total		1	11	12

Table 4.5

The relation between actual performance and perceived performance, represented by a high, positive phi-coefficient, as explained in section 3, is displayed in more detail in Figure 4.2. It shows how the correlation between actual and perceived success, the phi-coefficient, is higher in better performing subjects.

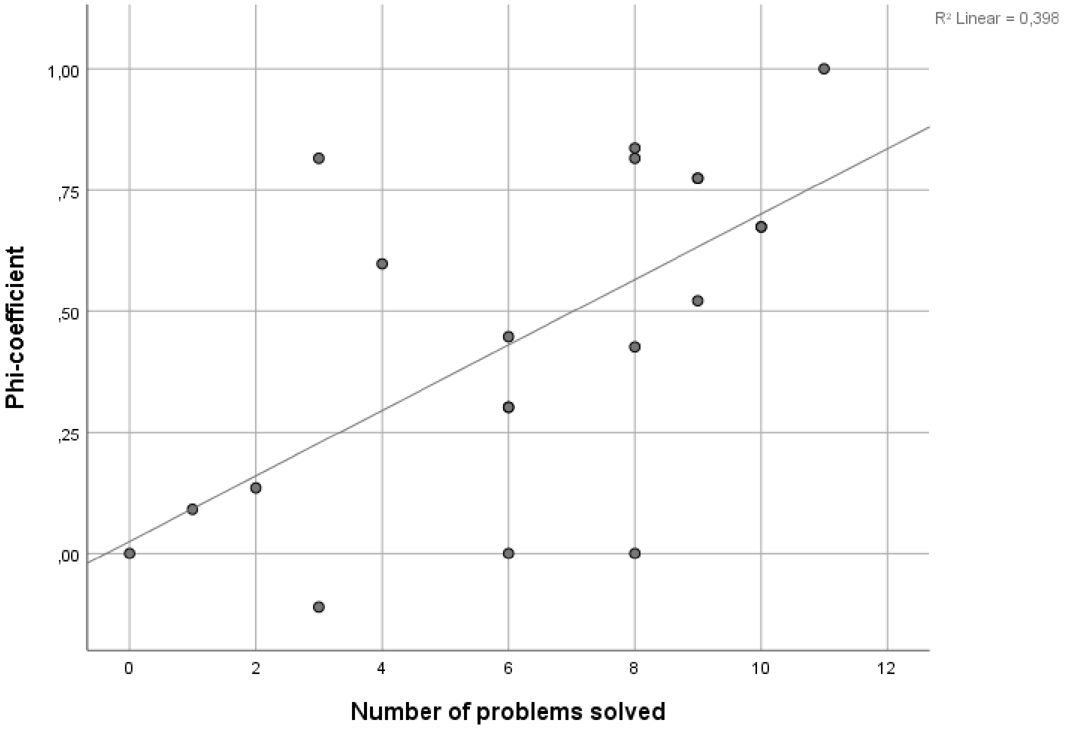


Figure 4.2 Relation between the number of solved problems and the computed phi-coefficient

Computing the correlation between the phi-coefficients of the participants and the number of problems solved, it can be stated that phi-value and number of solved problems are strongly associated with a correlation coefficient of r=0.631. The correlation significant at the 0.01 level with a p-value of p=.003.

This finding goes in line with previous findings of overestimation of someone’s skills and competency, called the Dunning-Kruger effect (Dunning & Kruger, 1999), which will be explained in more detail in the section 6.2.

4.7 Common errors in the pairwise comparison

In this section common errors in the pairwise comparison will be discussed. Overall, the participants estimated the difficulty correctly in 78 of the 200 cases – 10 comparisons done by 20 participants. With 39% of the comparisons correctly judged the participants' correct difficulty estimation is just slightly above chance (as in every comparison there were three possible answers '<', '>' and '='). Of the 10 comparisons performed by every participant, 3 comparisons were detected which show a clear misjudgment of difficulty by the participants. These comparisons involve 5 of the 12 presented problems and will be discussed in more detail below. The order of the presentation follows the order of display in the experiment. The first of the pairwise comparisons is termed Pair 1, the second Pair 2, the last Pair 10 and so on.

Pair 1: Comparison of Problem 9 to Problem 5

In Figure 4.3 we see two problems of different difficulty classes: Problem 9 with a ChessTempo rating of 1758 classifies as 'Difficult' whereas Problem 5 is of the class 'Medium' with a rating of 1317. In the pairwise comparison, though, almost two thirds or 13 of 20 subjects rated Problem 5 to be more difficult than Problem 9, this despite only 15% of the participants solving the latter correctly (opposed to

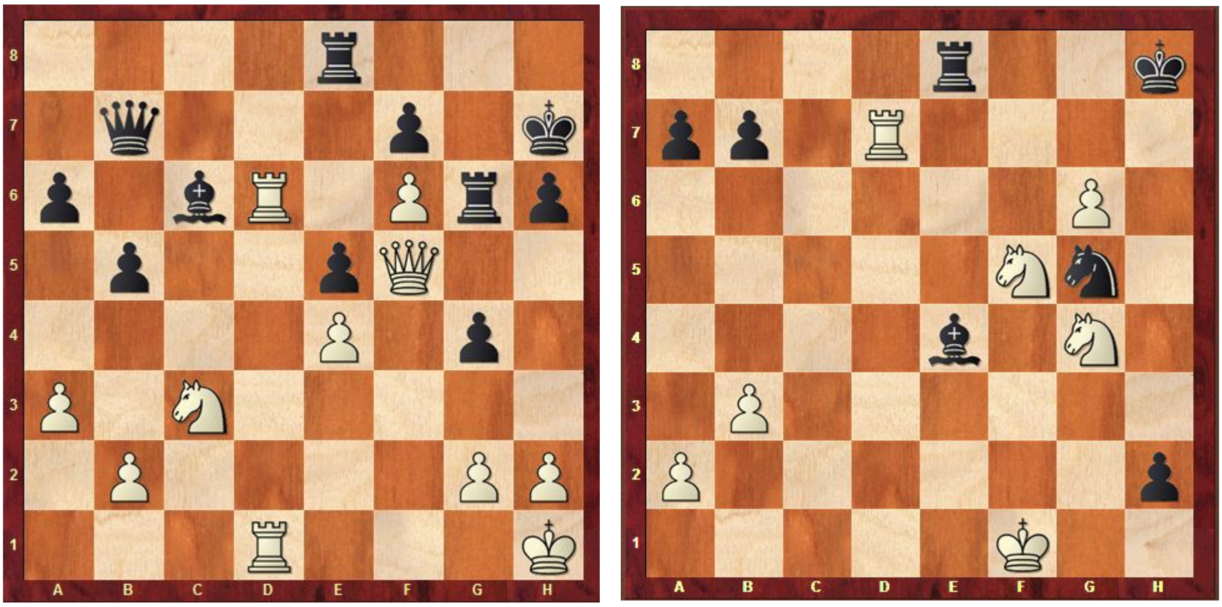


Figure 4.3 Pairwise comparison of Problem 5 (left) and Problem 9 (right)

60% for Problem 5). One participant deemed it equally difficult. In terms of perceived success – how many participants thought that they solved these two problems correctly – the values were comparable: For Problem 5, 16 participants or 80% estimated that they solved it correctly, whereas for Problem 9, the value was 15 participants or 75%. Nevertheless, the majority of participants rated Problem 5 to be more difficult when compared to Problem 9 and hereby relatively overestimating Problem 5. When looking at their respective ratings, the difference between the two problems is significantly more pronounced, namely 441 rating points.

A possible explanation for this misjudgment of difficulty is that in Problem 9 there is an obvious move for White which appears to be the winning move. To establish that White is in reality not winning after that move requires some calculation. Seemingly, many participants were not able to calculate this defensive chance for Black and hence had no motivation to consider the much less obvious rook sacrifice move 1 Rh7+, which indeed leads to a forced mate after 1 ... Nxh7 2. g7+ Kg8 3 Nh6 mate.

Pair 5: Comparison of Problem 10 to Problem 11

The next two problems both belong to the difficulty class ‘Difficult’, hence the correct difficulty estimate when comparing the two should be that they are equally difficult (‘=’).

The problems presented in Figure 4.4 only differ in 11 ChessTempo rating points (1753 for Problem 10 and 1764 for Problem 11), with 54% and 53% as their respective success rates according to ChessTempo. However, the majority of participants of our experiment (namely 55%) considered Problem 10 to be less difficult when compared to Problem 11. Only 5 participants, so 25% rated this Problem 11 to be equally difficult as Problem 11. This goes in line with the finding that Problem 10 had the by far largest success rate, namely 30%, of the difficulty class ‘Difficult’. Especially when opposed to Problem 11, which with only 2 people solving it correctly, had the lowest solving rate (10%) of all the 12 presented problems.

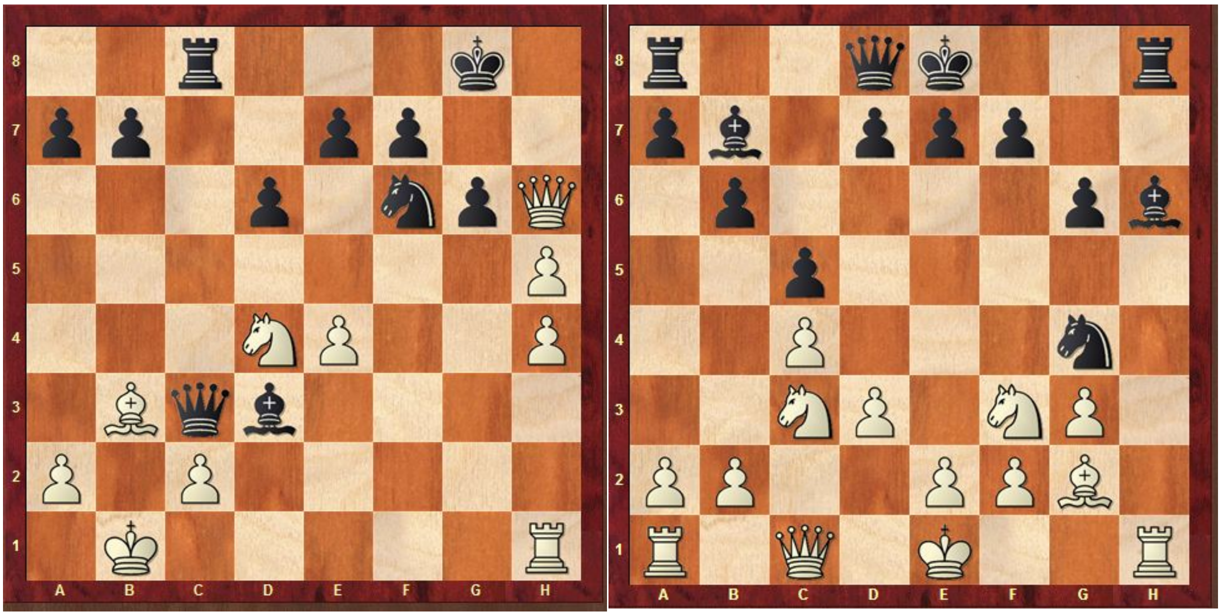


Figure 4.4 Pairwise comparison of Problem 10 (left) and Problem 11 (right)

In Problem 10, there is an obvious and very promising move (1 Qxg6+) which appears to be winning convincingly and might have made the problem appear easy. Therefore there was no motivation for White to look for alternative moves. On the contrary, the winning move in Problem 11 is a rather unnatural move – the White Queen moving so as to be easily captured by Black Bishop – which players might be reluctant to consider.

In this comparison, there seems to be a relation between performance and difficulty estimation, namely the relative underestimation of the problem which was solved by more people (6 participants) and relative overestimation of the problem which was solved by only a few, namely 2 participants.

Pair 9: Comparison of Problem 5 to Problem 6

In Pair 9 (see Figure 4.5), again two problems of the same difficulty class, namely the class 'Medium', are compared. Problem 5, as mentioned, has a ChessTempo rating of 1317 and Problem 6 one of 1329, the difference being 12 rating points. The two problems have comparable solving rates (70% and 60%, respectively) and for both of the problems 80% of participants believed to have solved them correctly. Nevertheless, 12 of 20 participants (or 60%) deemed Problem 5 to be more difficult than Problem 6, and 4 participants stated that the two problems are equally difficult.

It becomes apparent that Problem 5 is not only relatively overestimated when compared to a problem of the same class like in this case, but also when compared to a more difficult class like in Pair 1.

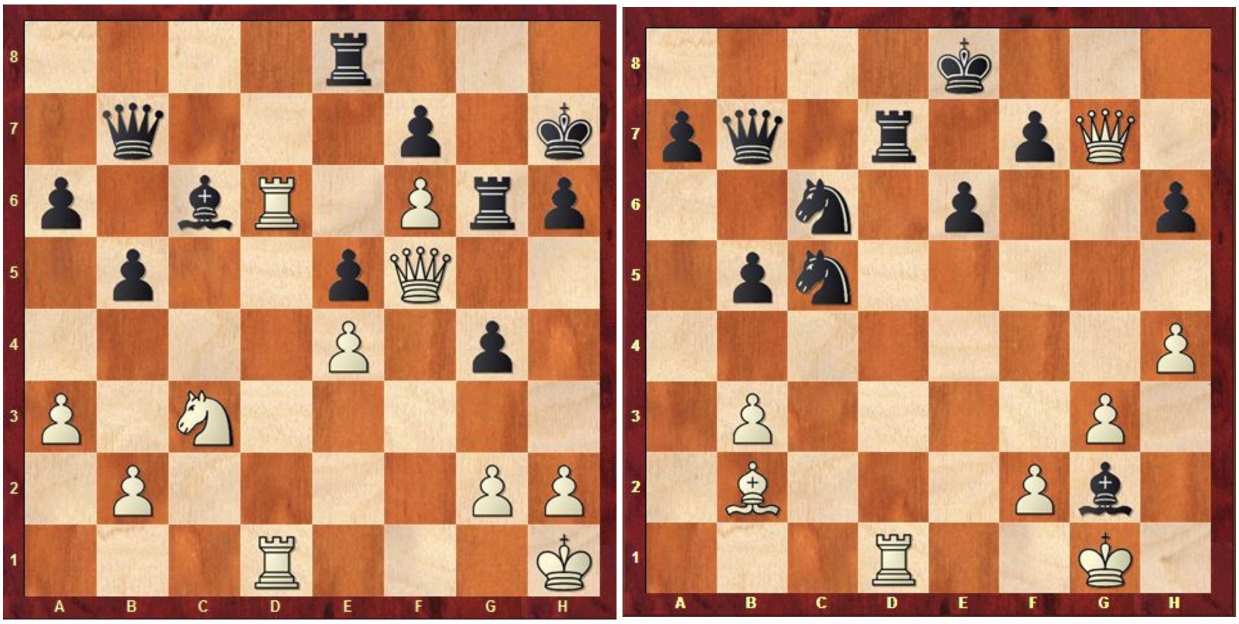


Figure 4.5 Pairwise comparison of Problem 5 (left) and Problem 6 (right)

The misjudgment of difficulty might be explained by the length of the variation to be calculated in the two presented problems. In Problem 5, 4 moves need to be executed whereas in Problem 6 there are just 2 moves needed to mate. A shorter variation is in principle easier to be detected.

5. Analysis of Data with Machine Learning methods

The rationale for applying Machine Learning methods is to characterize problem specific features and their influence on the prediction of variables such as success in solving the problem, as well as difficulty estimation by the participants. By employing these methods on the present data, hidden patterns and relations could be revealed which are not easily accessible to the human's judgment.

In the following sections, predictive variables of difficulty will be analyzed with machine learning methods. There are several problem specific features, which give account of the complexity of a presented tactical problem, like the number of pieces and the presence of a queen on the board, as well as the number of knights, material value, material imbalance and the number of non-pawns. In the game of chess, pawns have limited ability to move on the board and hence are easier to handle than other pieces. A high number of non-pawns therefore increases complexity and, presumably, difficulty. In contrast to pawns, the presence of queens on the board increases complexity as queens have many possibilities to move and can also move to remote areas of the chess board. The knight's ability to perform its characteristic move, which involves jumping over pieces, can lead to oversight and hence increases the probability of error in the calculation of variations, hereby increasing difficulty. In chess, pieces can be assigned a value, which reflects its strategical value in a chess game. The sum of these values of the pieces, termed material value, hence reflects the strength of the pieces on the board. A high material value suggests high strategic value of the pieces on the board, which may increase difficulty. Material imbalance refers to the difference between the material values of the two players' pieces on the board. Unbalanced material values are sometimes indicative of a strategic disadvantage for the player and may hence increase difficulty. In other cases, however, material imbalance decreases difficulty if it is for the benefit of a participant.

In the following, Machine Learning software will be used to compute decision trees and derive rules from data of the present experiment, as well as from the experiment by Hristova, Guid & Bratko (2014). Additionally, several models including decision trees and induced rules which are derived from the experimental data of the 2014 experiment will be tested on the data of the present experiment and evaluated in terms of their predictive power. Predictive variables for difficulty and success in solving the problem will be presented in the following sections.

5.1 Orange Machine Learning software

For the present analysis, Orange, an open source Machine Learning, data mining and visualization software developed by the University of Ljubljana, is used. Orange is a visual programming package based on components, the so-called widgets. These widgets implement algorithms for data analysis and visualization for evaluation of learning algorithms as well as for predictive modeling (Demšar et al., 2013). The version used for the present thesis is version 3.4.5.

5.2 Feature ranking

Feature ranking criteria seek to find the attributes most relevant for the modelling problem, selecting attributes with great prediction relevance. In the present thesis, the feature ranking criteria Gain Ratio and Relief-F were used.

Information gain of an attribute is the degree to which information entropy, the amount of information contained in a data set, decreases by splitting the given data according to the attribute. Information Gain Ratio takes into account both the information gain and the attribute values' entropy and hereby reduces bias towards features with many values. This characteristic of information gain when compared to Gain Ratio is depicted in Table 5.1, where variables with higher values like material value or number of non-pawns are ranked higher when employing Information Gain. As a second feature ranking method, the Relief-F algorithm is employed. In contrast to Gain Ratio, the latter takes into account possible dependencies between the attributes to be ranked (Kononenko et al., 1996). The algorithm takes feature value differences between nearest neighbour instance pairs and assigns this feature a high score if these instances are of a different class. If there is a feature value difference between two nearest neighbour instances of the same class, the assigned score decreases. Apart from ranking the most discriminative features, the scores assigned by ReliefF can also be applied as weights for a model. ReliefF is furthermore proposed for feature ranking for learning algorithms (Kononenko et al., 1996), such as the rule induction algorithm employed in this thesis.

Table 5.1 Feature ranking of the 2014 model's features with 'Difficulty' as Target variable

Features	Inf. gain	Gain Ratio
material value	0.770	0.385
numberNonpawns	0.770	0.385
numberKnights	0.667	0.340
sumPieces	0.563	0.287
material imbalance	0.459	0.500
Queen	0.143	0.347

5.3 Cross validation

In order to test whether a model is descriptive of the input data and consequently predictive for chosen variables – not under- or overfitting the train data – it has to be validated. To give account of how well the model generalizes from the learning data to new, unknown data (so called *unseen* data), an estimation of the model's accuracy needs to be performed. In the present case Leave-one-Out, a special case of *k*-Fold Cross Validation was used due to the scarcity of data. When performing *k*-Fold-Cross Validation, the data is split into *k* subsets, also called folds. The learning algorithm is then trained with all but one subset, the latter being used to test the model and output its prediction error (Witten et al., 2016). This is performed *k* times. The average of all the returned errors of the *k* folds is then taken to be the true error. In our case (leave-one-out), the number of subsets (*k*) was equal to the number of instances (*m*), always testing one example on the model trained with the remaining *m-1* instances. Despite its computational complexity, this cross validation method is often used when there is not abundant amount of data, as in our case. Our data of the 2014 and the 2018 experiment consists of 12 instances, the 12 tactical problems, each with 6 problem specific features.

Parameters of validity of the models used in this thesis are the Area under the Receiver Operating Curve (AUC), Classification Accuracy (CA), Precision and Recall.

Classification Accuracy (CA) is the percentage of correctly predicted among all the predicted instances of a distinct class. If the Classification Accuracy is close to 1, meaning close to 100% of the predicted instances are correctly classified, the model performs well. Precision is the proportion of true positives of all predicted positives, meaning the correctly predicted instances of a given class of all instances predicted to be of this same class. Recall is the proportion of predicted positives of all true positives of the data set, meaning the ratio between instances correctly predicted to be of one class and all the instances actually being of this class. The Area under the Curve (AUC) is a measure of validity of both Precision and Recall. The Fall-out-Rate is plotted against the sensitivity, meaning the true positive rate. The Area under the Curve is a good measure of performance of a model.

In addition to these parameters, the validity of a model can also be illustrated with a confusion matrix (Figure 5.1), which gives more detailed account of shortcomings of a model. The confusion matrix displays the number or percentage of correctly classified instances, and additionally informs about the nature of the incorrect classification. If a model is predictive for the fed in data, the boxes from top left to bottom right should contain the most instances or highest percentage respectively. As depicted in Figure 5.1, for the present experiment, the result is a 3x3 confusion matrix, as in the experiment there are 3 difficulty classes. If the cells from top left to bottom right contain the percentage 100%, the model accurately predicted every instance of the training set, which could indicate a case of overfitting to the train data and might perform poorly when confronted with new, unseen data of a test set.

5.4 Rule induction algorithms

Rule induction algorithms seek to generate a model in form of a set of if-then rules. A rule consists of a premise and a conclusion. The premise is a conjunction of attribute-value expressions, the conclusion is the class that is predicted by this rule. The attributes in the premise can also comprise conjunctions (‘AND’) and/or disjunctions (‘OR’), if applicable. An example of a possible rule for the present experiment is the following:

IF numberKnights≥3.0 **THEN** class=difficult

The premise in this case consists of an attribute, a relational operator, and a value. The rule hence relates the premise of the number of knights on the chessboard being greater or equal to 3 to the conclusion that this then is a problem of the difficulty class ‘Difficult’.

5.4.1 CN2 algorithm

The CN2 algorithm was developed by Peter Clarke and Tim Niblett of the Turing Institute in Glasgow (Clarke & Niblett, 1989). It belongs to the group of Separate-and-Conquer Algorithms which seek to generate rules that accurately predict one class of the target attribute (Conquer) and then exclude the covered instances from the training set (Seperate). The algorithm reiterates this process until all the training instances are excluded. As it does so, refinements to the already generated rules are made so that a measure of accuracy is maximized. The algorithm then returns a decision list, an ordered set of rules.

The CN2 algorithm generates rules in propositional logic, which are intelligible to humans. It is used in knowledge discovery and knowledge acquisition, inter alia for expert systems. It is a modification of the AQ and the ID3 algorithm, taking the advantages and strong features of both. Similar to the ID3 algorithm’s pruning strategy, it employs a top-down search until no further specialization is justified by statistical significance (Clarke & Niblett, 1989). Like the AQ algorithm, also the CN2 algorithm employs logical expressions and propositional calculus. In contrast to the

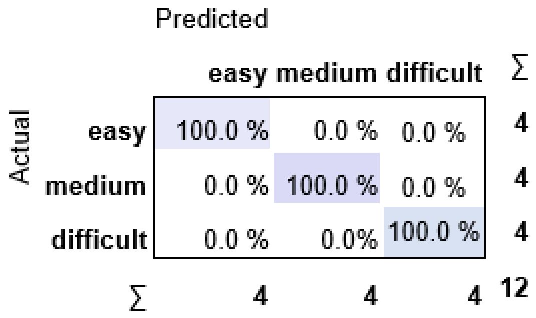
AQ algorithm, it is not dependent on specific examples and hereby increases the search space for possible rules. The CN2 algorithm does not require impeccability of the generated rules by also including those, which do not perform perfectly on the training data (Ruppert, 2006). It serves well to elicit knowledge in the form of IF-THEN rules, which give account of the underlying organization and relation of the features of the given data set.

In this analysis, the CN2 algorithm is employed to generate rules, which give account of the decisive characteristics of the tactical chess problems, which constitute their difficulty.

5.5 Decision Tree algorithms

Decision Tree learning algorithms compute decision trees that give an overview of the dependences in the data. Tree algorithms belong to the group of Divide-and-Conquer algorithms, which divide problems down to sub-problems, which are then easier to solve. A Decision Tree learning algorithm iterates to search for an attribute to split that best separates the different classes. In contrast to rule induction algorithms, the goal of a tree algorithm is to make this split in a way that the purity of its branches are maximized by considering all the classes, whereas rule induction algorithms try to find a rule that is predictive of one class.

In the present analysis, a classification tree is computed to illustrate the relation of the decisive attributes and the difficulty of the tactical chess problems.



Predicted

Actual	easy	medium	difficult	Σ
easy	100.0 %	0.0 %	0.0 %	4
medium	0.0 %	100.0 %	0.0 %	4
difficult	0.0 %	0.0%	100.0 %	4
Σ	4	4	4	12

Figure 5.1 Confusion matrix of a fictitious model

5.6 Modelling difficulty with Machine Learning methods

In the following section, Machine Learning will be employed to analyze predictive variables and generate predictive models of difficulty. The problem specific features, which give account of the complexity of the presented tactical problems, are analyzed in terms of their influence on the difficulty of the problems. This is done for both the experiment performed by Hristova, Guid & Bratko (2014) and the present experiment.

5.6.1 The 2014 experiment

In the following, the results of the analysis of the data of the experiment performed by Hristova, Guid & Bratko (2014) will be presented.

5.6.1.1 Feature Ranking

Table 5.2 Feature ranking with 'Difficulty' as target variable

Features	GainRatio	ReliefF
material imbalance	0.500	0.085
material value	0.385	0.052
numberNonpawns	0.385	0.062
Queen	0.347	0.024
numberKnights	0.340	0.049
sumPieces	0.287	0.044

As depicted in Table 5.2, in the 2014 experiment, material imbalance, material value and the number of non-pawns on the chess board appear to be the most influential variables when employing GainRatio as ranking algorithm. When employing the ReliefF algorithm, the number of non-pawns on the board turns out to be more influential than the material value.

5.6.1.2 Induced rules

Following, the induced rules are depicted. In the first column of Table 5.3, the premise – here termed IF-condition – is shown, followed by its conclusion, the THEN-class.

IF conditions		THEN class	Distribution	Probabilities [%]
material imbalance=0.0	→	Difficulty=difficult	[0, 0, 4]	14 : 14 : 71
Queen=0.0	→	Difficulty=medium	[0, 1, 0]	25 : 50: 25
numberKnights≥1.0 AND material value≥65.0	→	Difficulty=medium	[0, 2, 0]	20 : 60 : 20
material value≥53.0	→	Difficulty=difficult	[0, 0, 2]	20 : 20 : 60
material value≥41.0	→	Difficulty=medium	[0, 1, 0]	25 : 50 : 25
TRUE	→	Difficulty=easy	[2, 0, 0]	60 : 20 : 20

Table 5.3 Induced rules with the target variable 'Difficulty'

The CN2 rule induction algorithm generated 6 rules covering all the 12 instances, the 12 tactical chess problems. 4 of the 6 difficult problems could be classified as such due to their material balance. The remaining difficult problems were classified as such since their material value was greater or equal to 53. 2 of the 4 medium problems were classified 'Medium' as their number of knights was ≥1.0 and their material value was over or equal to 65. The remaining medium problems were either classified as such due to the lack of a queen on the board, or by their material value being ≥41.0, but smaller than 53 (rule 4). The two easy problems were covered by the last rule with the else-condition.

The column Distribution shows how many of the training instances fell into the given class by meeting the condition proposed by the rule. The Probability displayed in the fifth column shows how likely it is that a training instance with the condition met would fall into either the class ‘Easy’, ‘Medium’ or ‘Difficult’. This parameter employs the Laplace estimate of probability.

Summarizing, the generated rules show signs of overfitting which limits their generalizability. However, they are informative in terms of classifying 75% of the difficult problems as such, merely due to the fact that there is material balance. This is plausible, because if there is no clear advantage for the player, this might increase complexity and hence difficulty of the presented problems. The generated rules also suggest that the material value being high discriminates well between the problems of the class ‘Easy’ (lower values) and the two other classes, ‘Medium’ and ‘Difficult’.

5.6.1.3 Decision Tree

For the Decision Tree, the minimum number of instances in the leaves was set to 1 and subsets that are smaller than 3 are not split to increase accuracy and avoid overfitting to the train data. The computed decision tree is shown in Figure 5.2.

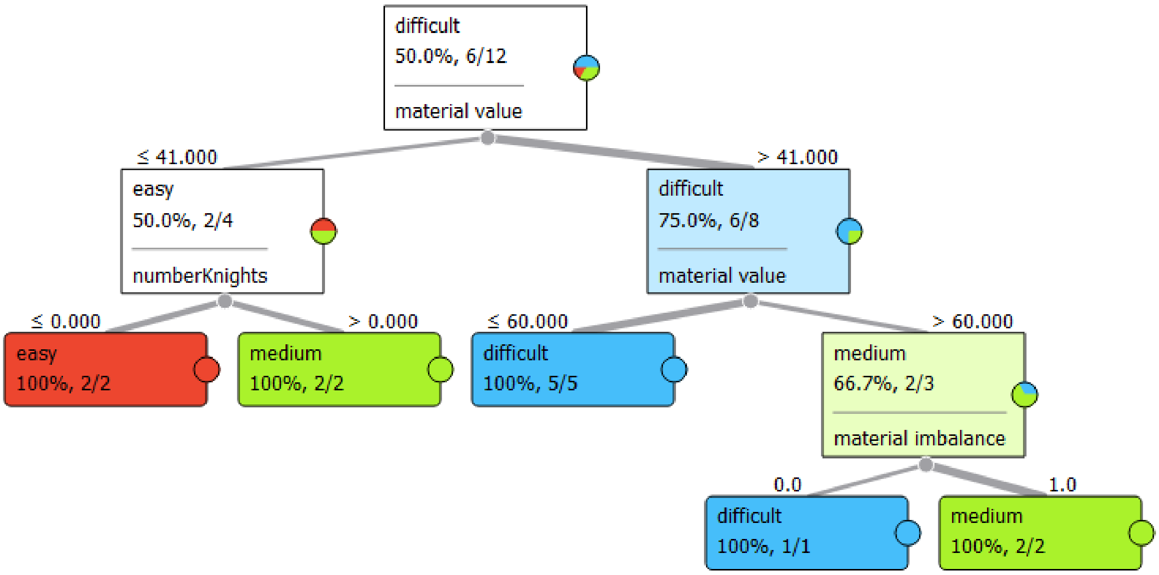


Figure 5.2 Computed decision tree with Difficulty as target variable

As Figure 5.2 shows, the decisive feature for splitting the data in two groups is the material value of the pieces on the board, one group comprising of easy and medium problems (if the material value is ≤ 41), the other of medium and difficult problems (if the material value is > 41). For problems with material values under or equal to 41, the number of knights on the board is decisive for discriminating between the 2 easy problems and 2 of the medium problems. If the material value is over 60 and there is material imbalance, then the problems are classified as medium. Both the rule induction and the decision tree fair models of difficulty for the present training data (see Table 5.4), the tree showing signs of overfitting.

Table 5.4 Results of the cross validation (Leave-one-out) for the 2014 experiment

Method	CA
CN2 Rule Induction	0.667
Tree	0.417
Constant	0.333

The classification accuracy shows that the CN2 model and the Tree model performed reasonably well on the training data. Both models surpassed the Constant method, returning the relative frequency of the classes of the variable to be predicted and hereby informing about the probability of a correct classification by chance.

5.6.2 The 2018 experiment

In the following section, the results of the analysis of the data of the present experiment with 20 chess players of various schools in Ljubljana will be presented.

5.6.2.1 Feature ranking

Table 5.5 Feature ranking with Difficulty as target variable

Features	GainRatio	ReliefF
Queen	0.347	0.018
knights	0.308	0.112

sumPieces	0.234	0.020
nonPawns	0.213	0.031
Material imbalance	0.172	0.054
Material value	0.049	0.019

Table 5.5 shows that that the influential variables for predicting difficulty in the 2018 experiment differ fundamentally from the ones of the 2014 experiment. Material value and material imbalance, the two most influential features of the 2014 experiment, are amongst the three least influential here. When employing Gain Ratio, the presence of a queen on the board was the most influential variable, when employing ReliefF, it was the number of knights.

5.6.2.2 Induced rules

In Table 5.6, the induced rules are depicted. In the first column, the premise – here termed IF-condition – is shown, followed by its conclusion, the THEN-class.

Table 5.6 Induced rules of the 2018 experiment

IF conditions		THEN class	Distribution	Probabilities [%]
sumPieces≥25.0	→	Difficulty=difficult	[0, 0, 2]	20 : 20 : 60
nonpawns≥12.0	→	Difficulty=easy	[2, 0, 0]	60 : 20 : 20
knights≥3.0	→	Difficulty=difficult	[0, 0, 1]	25 : 25 : 50
material imbalance≠0.0 AND nonpawns≥10.0	→	Difficulty=medium	[0, 2, 0]	20 : 60 : 20
knights≥2.0 AND sumPieces≥22.0	→	Difficulty=easy	[1, 0, 0]	50 : 25 : 25
knights≥2.0	→	Difficulty=difficult	[0, 0, 1]	25 : 25 : 50
nonpawns≥10.0	→	Difficulty=medium	[0, 1, 0]	25 : 50 : 25
TRUE	→	Difficulty=easy	[1, 1, 0]	40 : 40 : 20

For the 2018 analysis, the CN2 algorithm induced 2 more rules than for the 2014 experiment, possibly pointing to the prediction of difficulty being more complex in the present case. 3 of the 8 rules have the attribute presence of knights on the board in their premise, which is in line with it being one of the most influential according to the feature ranking.

Again, the fourth column (Distribution) shows how many of the training instances fell into a given class and the Probability displayed in the fifth column gives account of how likely it is that a training instance with the condition met would fall into either the class ‘Easy’, ‘Medium’ or ‘Difficult’. As the rules merely cover one or two instances, the model did not succeed to really generalize from the input data.

The most prominent attribute of the premises, the presence of knights, however, also plays an important role in the computed decision tree in Figure 5.3, discriminating between easy and medium problems on the one hand, and difficult problems on the other hand.

5.6.2.3 Decision tree

To increase accuracy and avoid overfitting of the tree model, the minimum number of instances in the leaves was set to 1 and subsets that are smaller than 4 are not split. In Figure 5.3 the computed decision tree is shown.

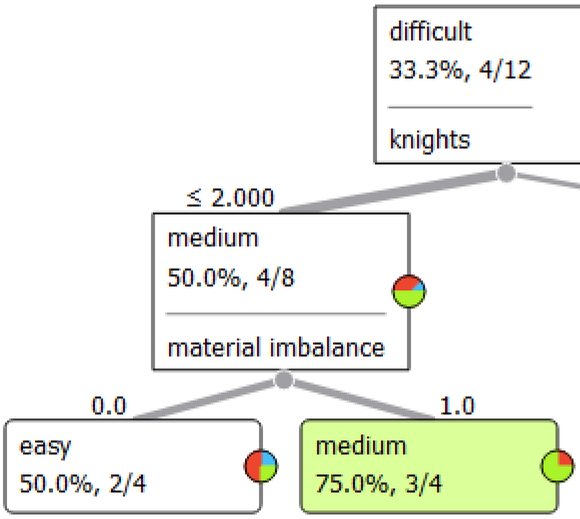


Figure 5.3 Computed decision Tree with 'Difficulty' as target variable

Overall, the decision tree model is not very informative. It only classifies 8 of the 12 problems according to the presence of knights and the material imbalance. When modifying the pruning conditions, the tree model shows signs of overfitting and its informational content does not increase significantly.

Table 5.7 Results of the cross validation (Leave-one-out) for the 2018 experiment

Method	CA
CN2 Rule Induction	0.333
Tree	0.167

Table 5.7 shows that both models performed poorly on the training data. The Tree model even has a classification accuracy below chance, the CN2 Rule Induction performs merely slightly better.

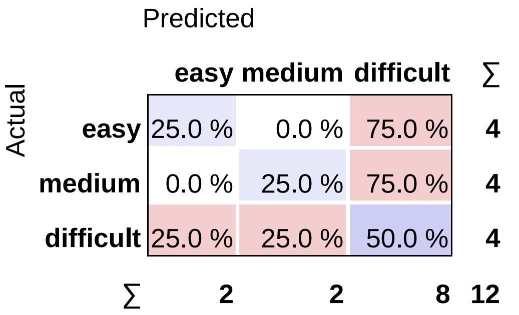
5.7 Testing of the models derived from the 2014 data with the 2018 data

When feeding in the 2018 experiment's data as test data into the models derived from the 2014 experiment, difficulty was sought to be predicted. When looking at the results, it can be stated that the models derived from the 2014 experiment's data are better models for the 2018 data than the models derived from the same data (see Table 5.8).

Table 5.8 Validation parameters for the 2018 data tested on the 2014 model

Method	CA
CN2 Rule Induction	0.500
Tree	0.333

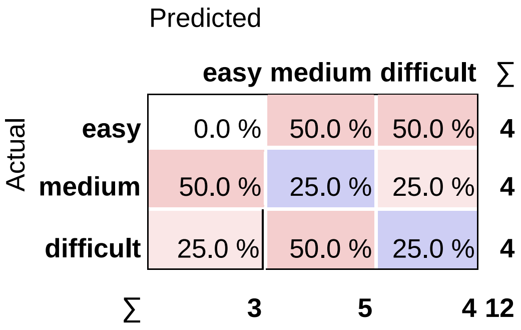
Classification accuracy, the percentage of correctly classified instances, increased for both the CN2 rule induction and the Tree model (see Table 5.8). When looking at the confusion matrix of the Tree model (Table 5.9), the 2 problems of the class ‘Difficult’ were classified as such, and merely 1 problem of each class ‘Easy’ and the class ‘Medium’. The class ‘Difficult’ was also the class that was most often predicted, with 8 instances being classified as such. Overall, the 2014 Tree model was more predictive when tested on the 2018 data as was the Tree model derived from the 2018 data.



Predicted

Actual

	easy	**medium**	**difficult**	∑
easy	25.0 %	0.0 %	75.0 %	**4**
medium	0.0 %	25.0 %	75.0 %	**4**
difficult	25.0 %	25.0 %	50.0 %	**4**
∑	**2**	**2**	**8**	**12**



Predicted

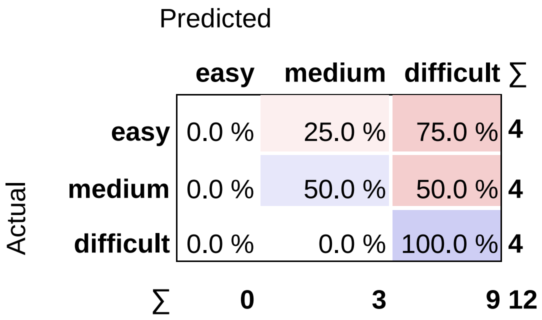
Actual

	easy	**medium**	**difficult**	∑
easy	0.0 %	50.0 %	50.0 %	**4**
medium	50.0 %	25.0 %	25.0 %	**4**
difficult	25.0 %	50.0 %	25.0 %	**4**
∑	**3**	**5**	**4**	**12**

Table 5.9 Confusion matrix for the Tree model for the 2014 model tested on the 2018 data (top), and the results of the cross-validation of the 2018 Tree model (bottom)

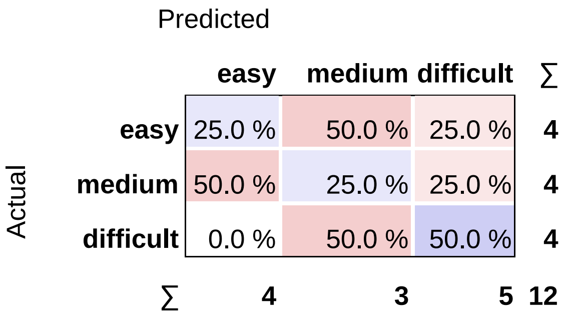
The 2018 Tree model was merely able to classify 2 problems correctly, 1 problem of the class ‘Medium’ and 1 problem of the class ‘Difficult’.

In Table 5.10, the results of the 2014 CN2 rule induction model tested on the 2018 data, as well as the results of the cross-validation (Leave-one-out) for the 2018 CN2 rule induction model are shown. It can be seen that the 2014 model performed better on the 2018 data than did the model derived from the 2018 data. For both the 2014 and the 2018 model problems of the class ‘Difficult’ were classified most successfully, with the 2014 model correctly classifying all the 4 difficult problems, the 2018 model merely 2. One of the ‘Easy’ problems was predicted to be of the class ‘Medium’, 2 of the ‘Medium’ problems and 3 of the ‘Easy’ problems were classified as ‘Difficult’ by the 2014 model.



Predicted

		easy	medium	difficult	∑
Actual	easy	0.0 %	25.0 %	75.0 %	4
	medium	0.0 %	50.0 %	50.0 %	4
	difficult	0.0 %	0.0 %	100.0 %	4
	∑	0	3	9	12



Predicted

		easy	medium	difficult	∑
Actual	easy	25.0 %	50.0 %	25.0 %	4
	medium	50.0 %	25.0 %	25.0 %	4
	difficult	0.0 %	50.0 %	50.0 %	4
	∑	4	3	5	12

Table 5.10 Confusion matrix for the CN2 rule induction for the 2014 model tested on the 2018 data (top), and the results of the cross-validation of the 2018 CN2 model (bottom)

Summarizing, the 2014 models performed better on the 2018 data than did the models derived from the 2018 data. Both the CN2 rule induction model and the Tree model were most successful when classifying problems of the class ‘Difficult’.

6. Discussion and further research

6.1 Summary

In contrast to the 2014 experiment, the present experiment yielded that the players' ratings were positively correlated to their success in solving the problems. The results of Hristova, Guid & Bratko (2014) could not be replicated. The player's ratings were significantly, but weakly correlated to a low error score in the pairwise comparison, indicating that better players are also slightly better in estimating the difficulty of the presented problems. The participants' success in solving the problem was a worse predictor for correct difficulty estimation, as the correlation between success and error score was very weak and non-significant.

However, the reported relation between difficulty estimation and success in solving a problem (Borg et al., 1972; Touroutoglou & Efklides, 2010) could not be replicated in this experiment. Participants who did not solve a problem correctly did not significantly overestimate its difficulty. The hypotheses that people tend to overestimate the the difficulty of problems they did not solve correctly and to underestimate the ones they solved correctly could not be confirmed. Neither the overestimation of incorrectly solved problems, nor the underestimation of successfully solved problems seems to be prevalent in our sample. The coefficients are r=.01 and r=.026, respectively, the correlation is hence very weak to almost nonexistent.

Interestingly, the experiment yielded that even though the overall correlation between actual and perceived success was moderately high, there was a significant difference between unskilled and skilled participants. The hypothesis stating that higher skilled individuals, in this case strong chess players according to their rating, will be better in assessing their own success in solving the problems was confirmed. The difference between perceived and actual success rate actually decreased with increased success. This phenomenon is termed the 'Dunning-Kruger effect' and is described in more detail in the next section.

The present Machine Learning analysis yielded that the models derived from the data of the experiment by Hristova, Guid &Bratko (2014) were more successful in classifying the data of the present experiment. However, both the 2014 and the 2018 models were merely fair models of difficulty with overall low predictive power.

6.2 The Dunning-Kruger effect

The Dunning-Kruger effect describes a cognitive distortion that results in an unskilled individual not recognizing his own deficits in skill and competence and hence overestimating them. This bias of illusionary superiority was first described by the psychologists David Dunning and Justin Kruger from the Cornell University in New York (Dunning & Kruger, 1999). The effect can be observed not only in everyday life situations like examinations in school (Sinkavich, 1995), assessment of ones abilities at the work place, but also when assessing the correctness of one's diagnosis of mental illnesses (Garb, 1989).

In their paper Dunning and Kruger argue that the skills needed for solving a given task are the same as those needed for assessing one's performance in doing so which provides an explanation of this discrepancy between actual and perceived competence or performance. According to their inference, the meta-cognitive process of judging one's own skills and expertise is related to the cognitive task assessing these same skills and expertise (Dunning & Kruger, 1999). More evidence for this assumption is presented in studies that show that the discrepancy between actual and perceived performance is smaller when the skills or prerequisites needed for assessing one's performance are different from the ones that are needed to succeed in solving the task, or performing well in the given task respectively. In archery, the skills needed for hitting the target (controlled limb movement, precision in shooting) and assessing one's performance in doing so (simply watching the arrow hit the target) are properly different from each other. It is hence not surprising that Mabe & West (1982) found correlations of actual and perceived performance to be higher for athletic performances (r=.47), than for managerial skills (r=.04), for assessing one's own interpersonal abilities like empathy and openness towards the other (r=.17), or intellectual tasks (r=.34).

The Dunning-Kruger effect, however, does not suggest that an individual's ignorance or inability and the estimation of competency are inversely related, meaning that the worse one's performance is, the more successful it will be perceived. Actually, in their study, Dunning and Kruger (1999) found that individuals with lower skills and competences rated their skills overall lower than did those with higher expertise and abilities. The crucial point is in fact the discrepancy between the perceived and the actual performance, and not so much the mere overestimation of their performance.

In the present case incomplete knowledge about the moves to make and therefore incomplete knowledge about best strategy, led to the selection of a suboptimal strategy that is nevertheless deemed to be the optimal one. Analysing the participants' responses in the semi-structured interview showed that those who did not perform well in solving the tactical problems were also not able to accurately explain their reasoning behind their decision to make the move. This lack of insight concerning one's deficiencies, the lack of meta-comprehension and not optimal self-monitoring are considered to be the underlying reasons for this discrepancy between actual and perceived success as witnessed in other studies observing the Dunning-Kruger effect.

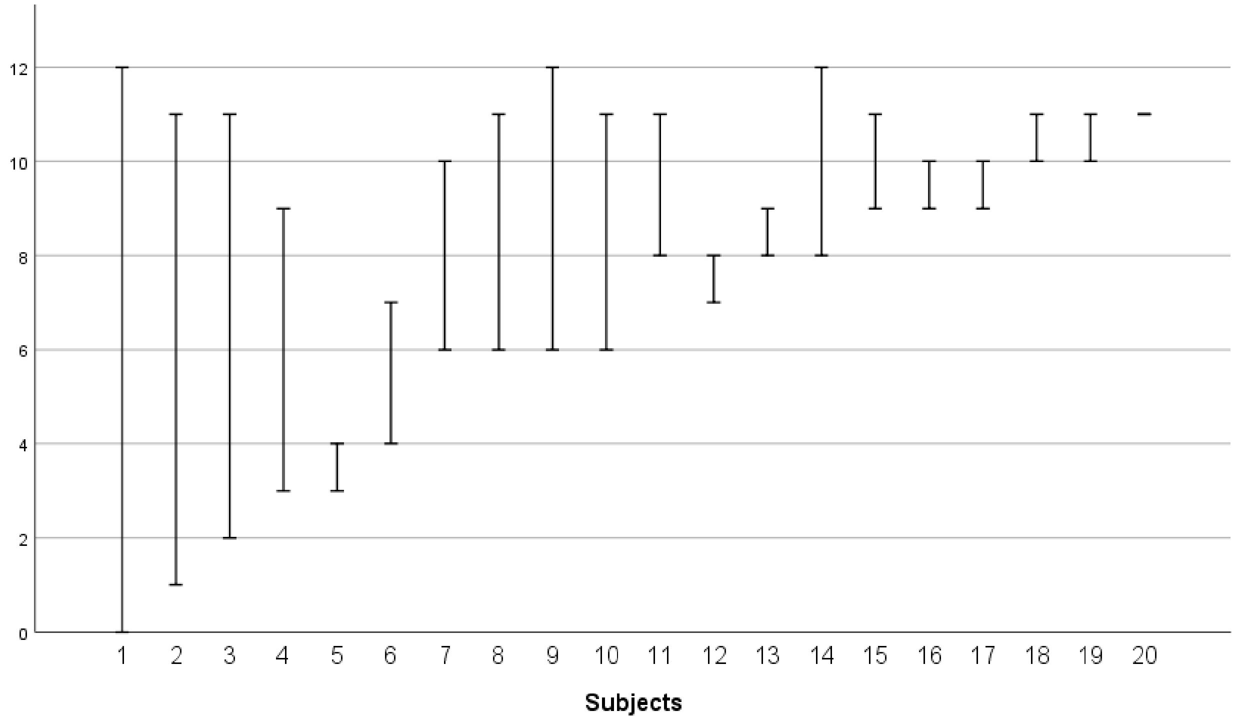


Figure 6.1 Error in the subjects' estimation of performance

Figure 6.1 shows the relation between perceived and actual success in solving the problems. The participants are ordered according to their success in solving the problems, Subject 1 in the graph having solved 0 problems correctly, subject 20 having solved 11 correctly. The vertical line depicts the error of the participants when estimating their performance, the difference between perceived and actual success. In all but in the case of subject 12, the lower end of the line depicts the number of problems solved and the upper end the number of perceived solved problems. Subject 12 believed to have solved only 7 problems correctly, while in reality he solved 8. The longer the vertical line, the higher the error. As we see in Figure 6.1, the first subject did not solve any of the problems correctly, hence the vertical line starts at the x-axis, but estimated to have solved all of them, resulting in the line to end at 12 on the y-axis. For this participant the vertical line, representing the error in estimating his performance, was the longest. In all but one case the participants were optimistic regarding their success.

Similarly to what Dunning and Kruger (1999) found, subjects in our sample who did poorly in solving the given problems did rate their performance as overall worse than the subjects who performed well (see Figure 6.1). The lower quartile of worst performers estimating to having solved 9.4 problems (+/-SD 3.21) on average, opposed to 10.6 (+/-SD 0.55) for the upper quartile.

There is a significant positive relation between the subjects' ratings and their computed phi-values as we can see in Figure 3. This positive relation (r=.54; p=.036) between rating and phi-value, a measure of association of success and perceived success, supports the assumption that participants with higher rating were also able to assess their success more accurately.

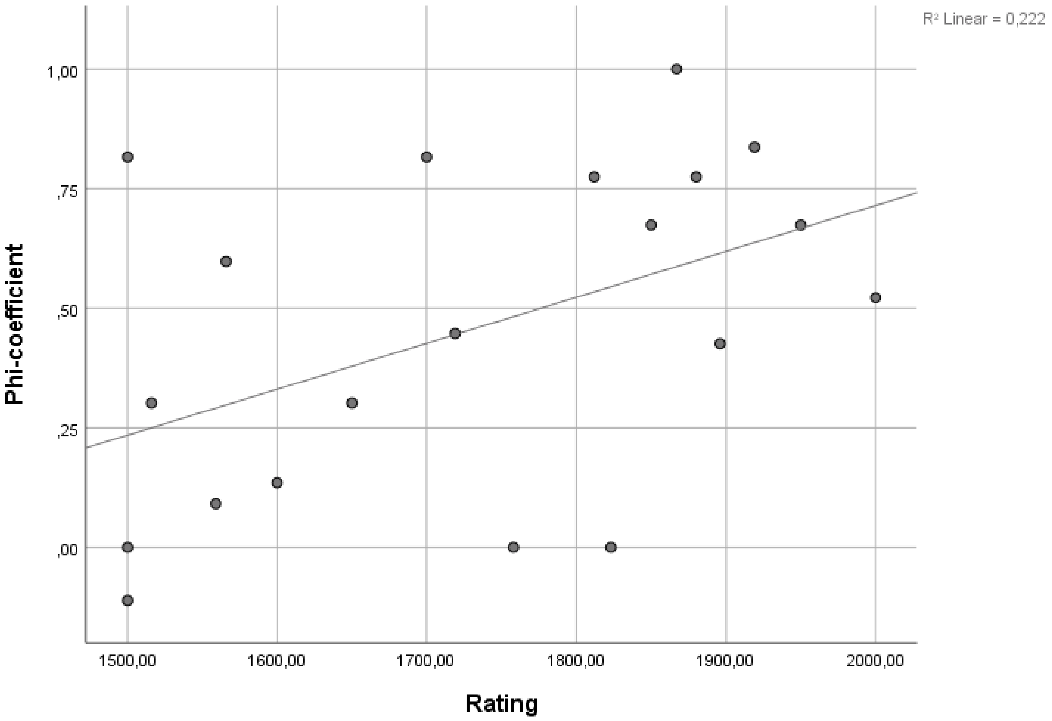


Figure 6.2 Relation between the subjects' rating and the Phi-coefficient

Interestingly, the resulting CN2 and Tree models for classification of the tactical problems of the 2014 experiment was predictive in terms of difficulty of the 2018 problems even though they were easier problems according to ChessTempo ratings. The average rating of the difficult problems was 2243.05 in the 2014 experiment and 1759.25 in the present one. The present problems also required more moves to make than the ones used in the 2014 experiment. However, neither the 2014 nor the 2018 models allow for meaningful generalization in terms of difficulty categorization due to the very small number of learning examples (11 and 12 tactical problems, respectively).

6.3 Critical points

In the present thesis the relative overestimation of a problem had to serve as general overestimation as formulated in the hypotheses H1a and H1b, which represents the main weakness and imprecision of this research. This circumstance is due to the unavailability of a complete ranking of the problems to be obtained from the pairwise comparisons of the participants. Considering the time constraint with which the experimenters had to reckon with given the time window provided for the experiment, the number of pairwise comparisons was limited to 10 comparisons. For the bottom quartile of worse performers this also turned out to be a reasonable decision, as the time scheduled for the experiment was just sufficient. For better players, time turned out not be a limiting factor, which would have allowed for more paired comparisons, hereby increasing the usefulness in terms of analysis and interpretation (see below). In order to be able to rank the subjects' difficulty estimates more accurately, more paired comparisons should be performed. If replicating the experiment, the number of paired comparison should be an acceptable tradeoff between the ideal of comparing each problem to all of the other problems, resulting in 66 comparisons ($\frac{n(n-1)}{2}$ with $n=12$ for the 12 presented problems) and a number that is low enough not to induce fatigue, which might lead to inaccuracies in judgment. This might then yield more accurate insights into the participants' rankings as well as their over- and underestimation of the given problems.

The difficulty with the obtained results as described above then arose in the course of the interpretation of the pairwise comparison, this not only due to the few comparisons but as crucial comparisons were missing as shown in Figure 6.3. The three rows are composed of the four problems of each difficulty class. The arrows refer to the relation of two compared problems, the arrowhead pointing to the problem that was deemed to be less difficult. The lack of an arrowhead signifies that the compared problems were considered equally difficult. The participants' answers allowed for inferences beyond the direct comparisons, as for instance the relation of Problem 1 with Problem 9. As shown in Figure 6.3, the problem set resulted to be divided into two groups, and the relation between two groups of problems could not be identified. This was due to the lack of a comparison of the problems of the left side with the ones on the right side as shown in Figure 6.1. Neither Problem 1, Problem 6 nor Problem 11 were compared to one of the Problems 2, Problem 7 and Problem 12, which would have allowed further inferences, and following perhaps a complete ranking according to the three difficulty classes as proposed by the experimenters.

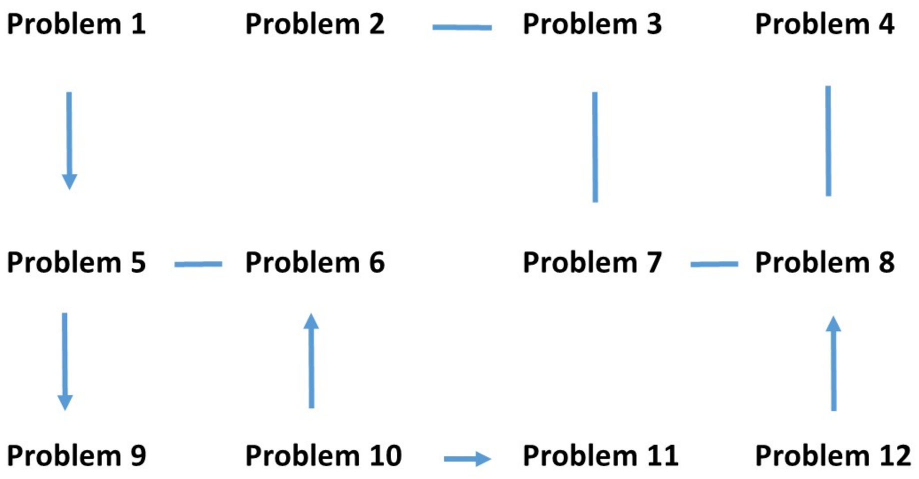


Figure 6.3 Illustration of a pairwise comparison by a randomly chosen participant

Additionally, one of the participants stated in the interview that the experimenter not making another move after the participant’s first move was taken as indication that the given problem was not solved correctly. This presumably due to the fact that previous, already tackled problems involved more than one move. As it was only clearly stated by one of the participants, it is nevertheless suspected to have played a minor role, but surely does represent a weakness of the experimental design. The reliability of this cue for the participant might be restricted by the experimenter clearly stating that to solve a presented problem, one or more moves need to executed, even though this was not the case for the present experiment because 2 to 6 moves needed to be made. Another way of circumventing this circumstance is to also include problems, which can be solved correctly by making only one move.

The discrepancy between actual and perceived success being more pronounced in lower skilled participants might be due to the fact that there was a general overestimation of performance and probability of being wrong just decreased for higher skilled individuals. The experiment design of the present research allowed for identifying the exact problems that were assessed incorrectly. The participants were not asked to estimate their success rate (in percent or in relation to other participants) but were interrogated after each problem individually, if they think they solved it or not. All but one participants overestimated their success in solving the problems, participants from the upper quartile just did so to a minor degree. From the present data it hence cannot definitively be concluded that better players are more objective predictors of theirs own success.

The present models obtained by the Machine Learning analysis do not make meaningful generalization possible. This is due to the very small number of examples from which it was sought to derive predictive models of difficulty.

6.4 Conclusion and further research

The present experiment demonstrates that the relation between expertise, performance and perceived difficulty is still not very clear. Even though the data show a relation between expertise reflected by the participants' ratings and performance, here success in solving the presented tactical problems – conflicting with results found by Hristova, Guid & Bratko (2014) – the relationship between performance and difficulty estimation is not as pronounced.

In this research, the well-studied Dunning-Kruger effect could be evidenced. In contrast to Park & Santos-Pinto (2010), the present research investigated the assessment of one's performance post-hoc, and hence in the spirit of the effect described by Dunning & Kruger (1999). The present research thus represents first evidence for the effect in the area of chess research. The underlying reason for this discrepancy between actual and perceived success suspected by Dunning & Kruger (1999), namely that the skills needed for solving a given task are the same as those needed for assessing one's performance, could also be detected in our research. Participants who played an incorrect move also mostly lacked insight concerning their error. In the interview following every tackled tactical problem, they still argued in favour their wrong move or did not provide a useful explanation at all. This also led them to assess their performance incorrectly, meaning that the participants were confident in their judgement due to the wrong reasons.

Evidence from the present Machine Learning analysis suggests that the models derived from the data of the experiment by Hristova, Guid &Bratko (2014) were more successful in classifying the data of the present experiment. Despite the relative superiority of the 2014 models, they were merely poor models of difficulty with overall low predictive power.

In the present research the relation between success in solving the problem and over- or underestimation, respectively, could not be evidenced. Neither was there a significant underestimation of problems which were solved correctly, nor an overestimation of incorrectly solved problems. This is both in conflict with results from Borg et al (1971a+b) and indications by Touroutoglou & Efklides (2010), as well as with intuitive judgements about this relation. Further research on the relation between difficulty estimation and performance is indicated.

7. Literature

Adamic, P., Kakiashvili, T., Koczkodaj, W. W., Babiy, V., Janicki, R., & Tadeusiewicz, R. (2009). Pairwise comparisons and visual perceptions of equal area polygons. *Perceptual and motor skills*, *108*(1), 37-42.

Borg, G., Bratfisch, O., & Dornic, S. (1971a). On the problems of perceived difficulty. Scandinavian journal of psychology, 12(1), 249-260.

Borg, G., Bratfisch, O., & Dornic (1971b). Perceived Difficulty of a Visual Search Task. Institute of Applied Psychology, Report No.16.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. Visual information processing, 215-281

Rating System. Retrieved from https://chesstempo.com/user-guide/en/tacticRatingSystem.html

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine learning*, *3*(4), 261-283.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14(Aug):2349−2353.

Desender, K., Van Opstal, F., & Van den Bussche, E. (2017). Subjective experience of difficulty depends on multiple cues. *Scientific reports*, *7*, 44222.

Elo, A. E. (1965). Age changes in master chess performance. *Journals of Gerontology*, *20*(3), 289-299.

Elo, A. E. (2008). The rating of chessplayers, past and present. Ishi Press International

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B*, *367*(1594), 1338-1349.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*(3), 387.

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, *3*, 59-102.

Glickman, M. E. (1998). The Glicko system. *Boston University*.

Hristova, D., Guid, M., & Bratko, I. (2014). Assessing the difficulty of chess tactical problems. *International Journal on Advances in Intelligent Systems*, *7*(3), 728-738.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting*, *26*(3), 460-470.

Kononenko, I., Robnik-Sikonja, M., & Pompe, U. (1996). ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Artificial intelligence: methodology, systems, applications*, 31-40.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of applied Psychology*, *67*(3), 280.

Newton-Fisher, N. E. (2017). Modeling social dominance: Elo-ratings, prior history, and the intensity of aggression. *International journal of primatology*, *38*(3), 427-447.

Park, Y. J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: evidence from the field. *Theory and Decision*, *69*(1), 143-166.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169-179.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

Ruppert, M. (2006). Vergleich von AQ, CN2 und CN2 mit Weighted Covering. (Diploma thesis) Retrieved from:
www.ke.tu-darmstadt.de/bibtex/attachments/single/105 - 12.09.18

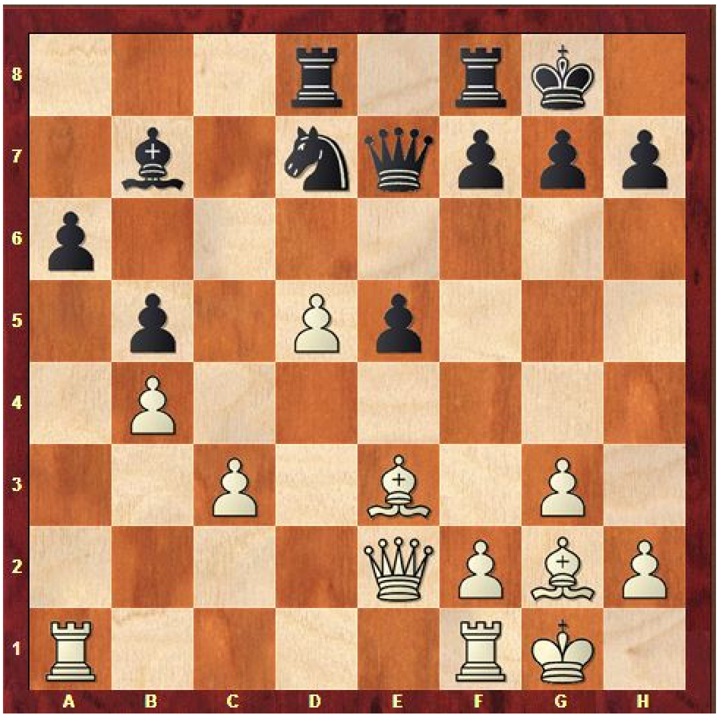
Touroutoglou, A., & Efklides, A. (2010). Cognitive interruption as an object of metacognitive monitoring: Feeling of difficulty and surprise. In Trends and prospects in metacognition research (pp. 171-208). Springer, Boston, MA.

Van Der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American journal of psychology*, 29-60.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

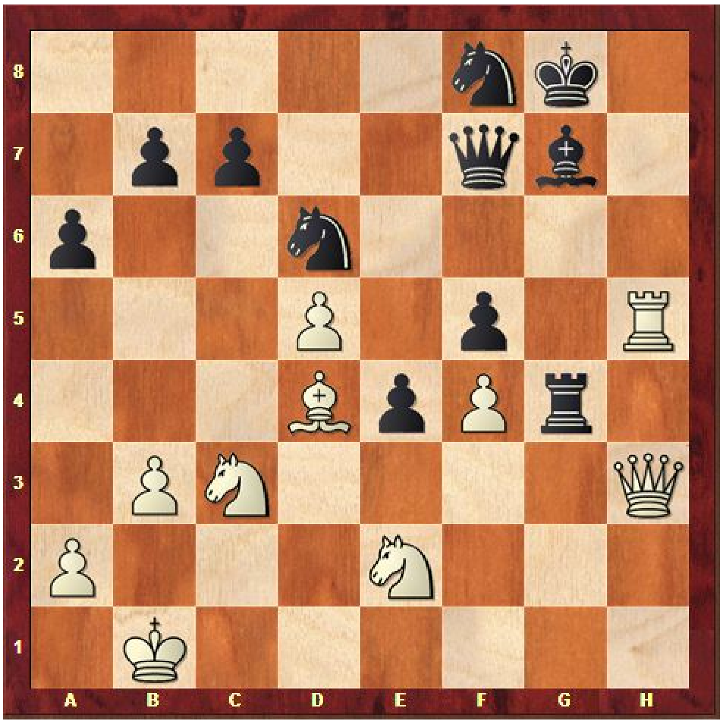
8. Appendix

Problem 1



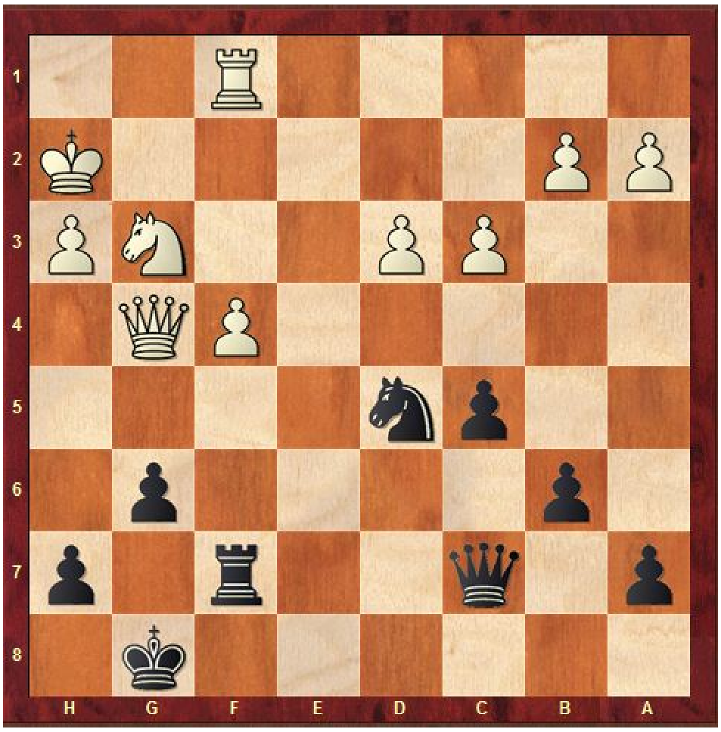
Correct solution: 1 d6, Qxd6 2 Bxb7

Problem 2



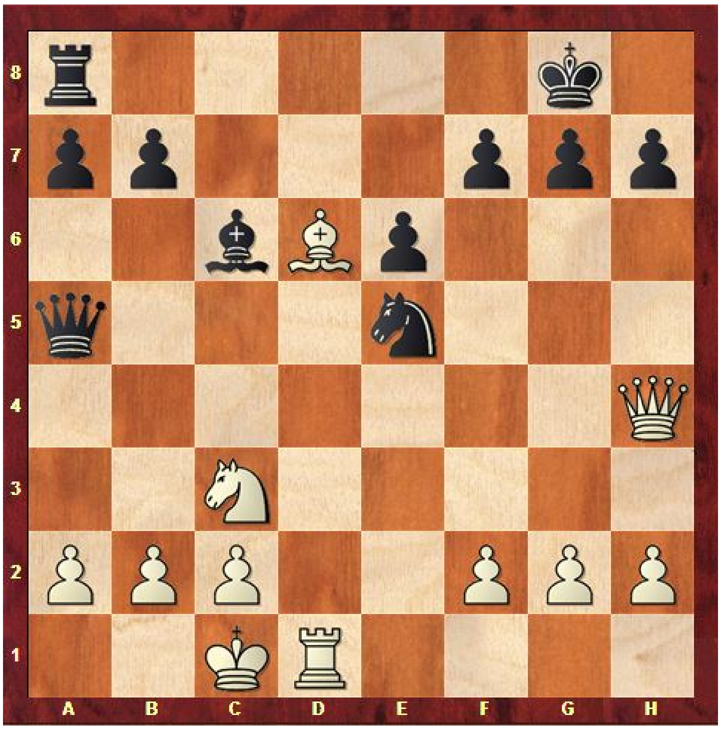
Correct solution: 1 Rh8+ Bxh8 2 Qxh8 mate

Problem 3



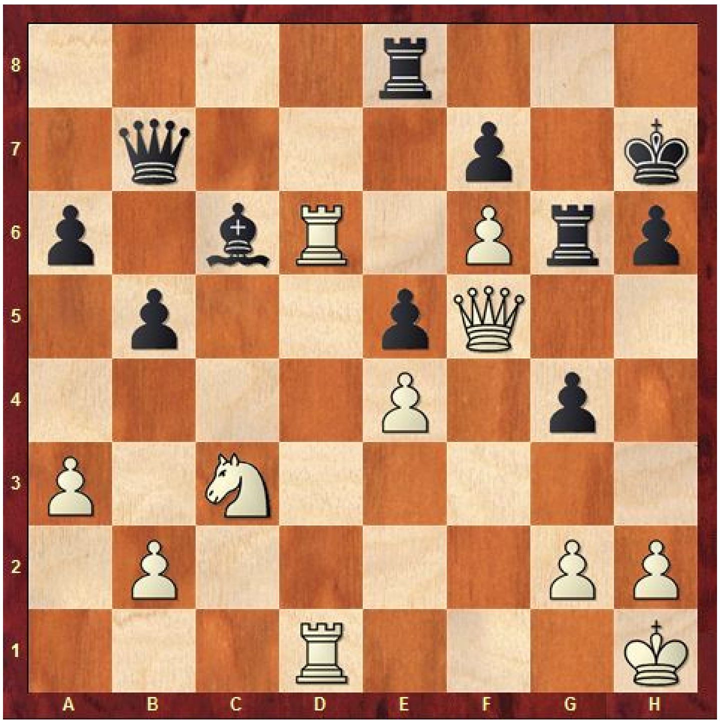
Correct solution : 1 ... Ne3   2 Qf3 Nxf1+

Problem 4



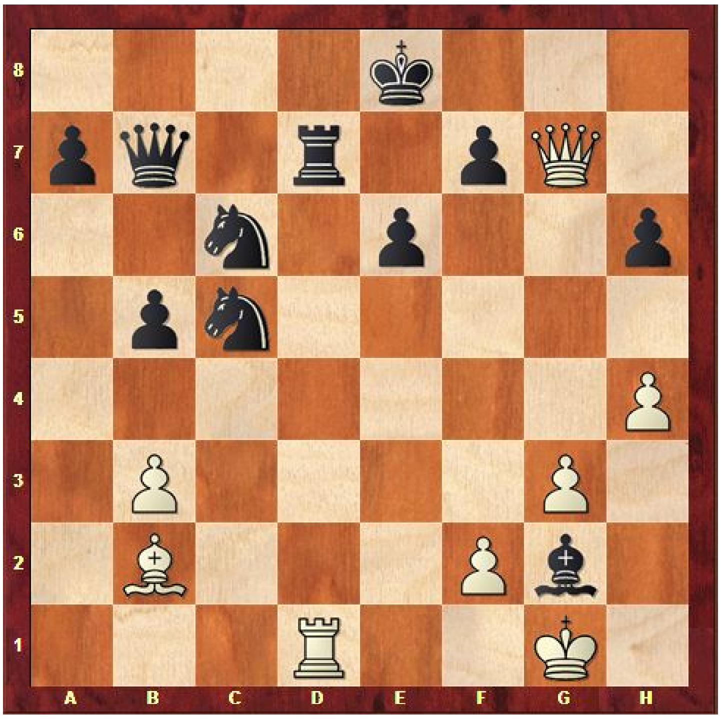
Correct solution : 1 Bxe5 Qxe5 2 Rd8+

Problem 5



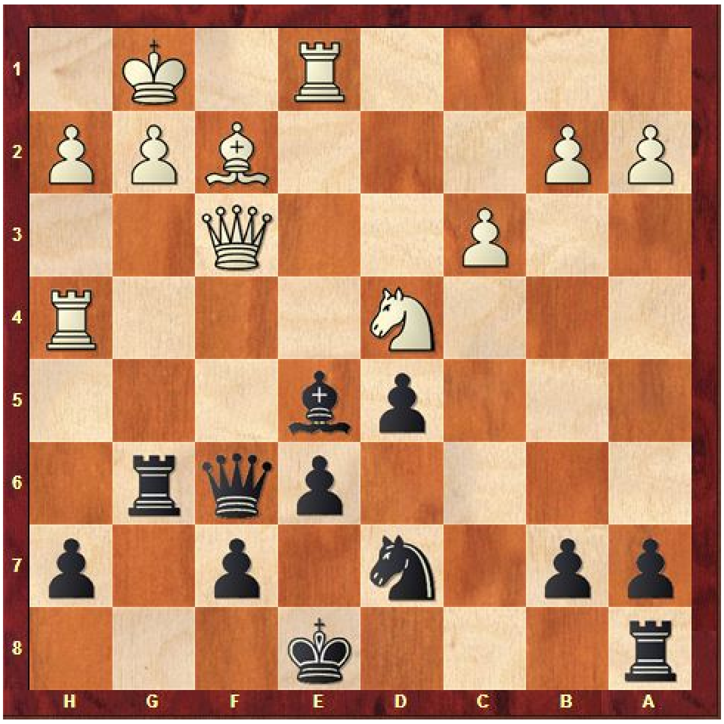
Correct solution: 1 Rd7 Bxd7 2 Rxd7 Qc6 3 Rxf7+ Kh8 4 Qxg6

Problem 6



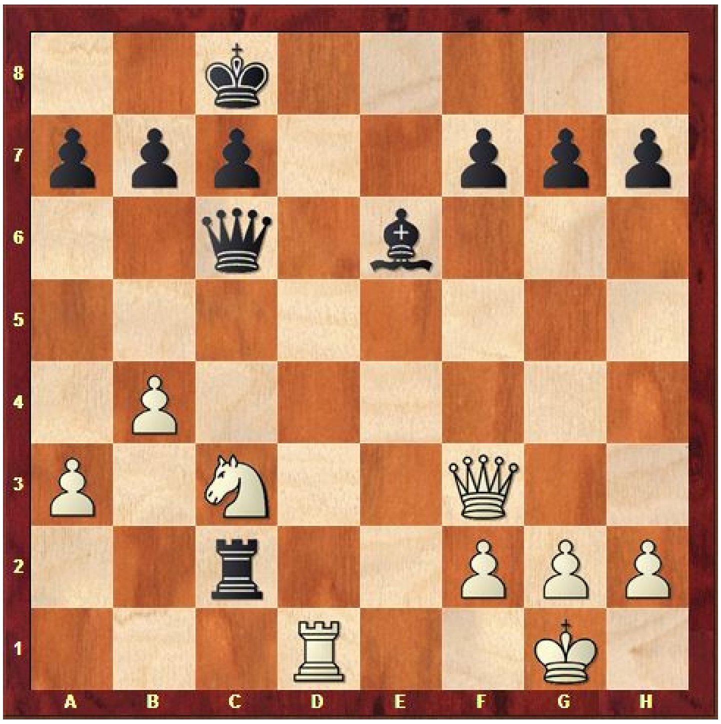
Correct solution: 1 Qh8+ Ke7 2 Bf6 mate

Problem 7



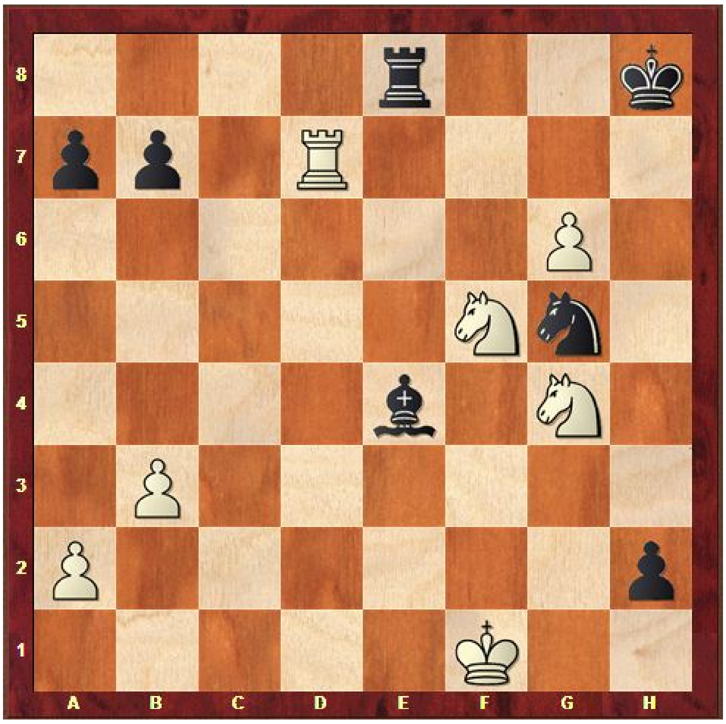
Correct solution: 1 ... Bxd4   2 Qxf6 Bxf6

Problem 8



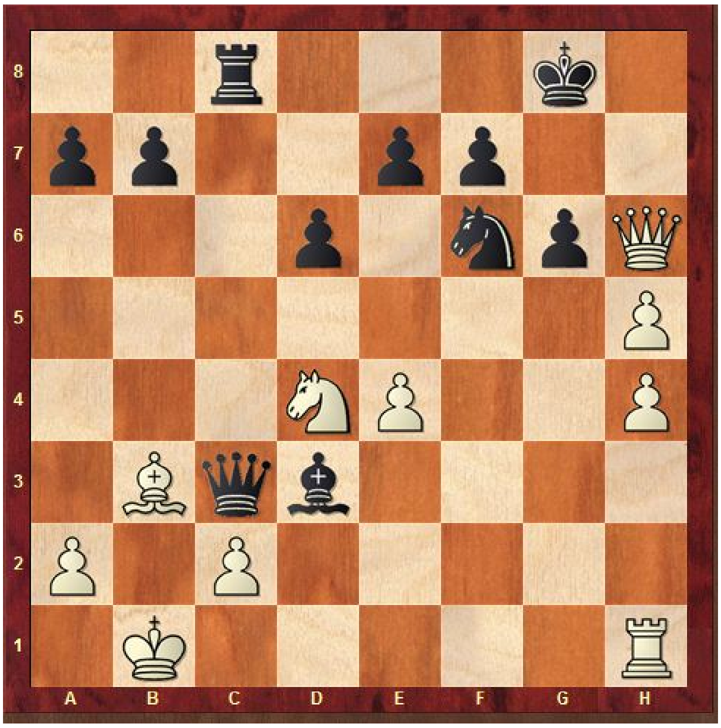
Correct solution: 1 Qd3 Qxc3 2 Qd8

Problem 9



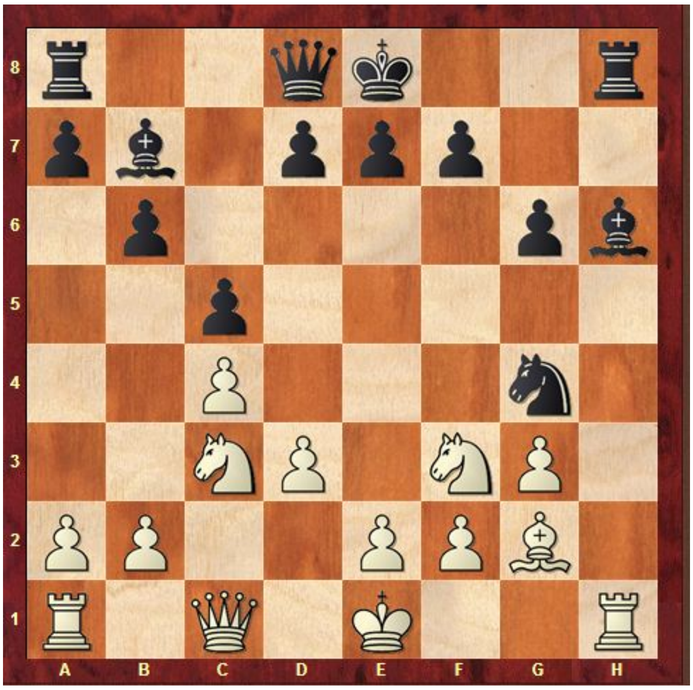
Correct solution: 1 Rh7+ Nxh7 2 g7+ Kg8 3 Ngh6 mate

Problem 10



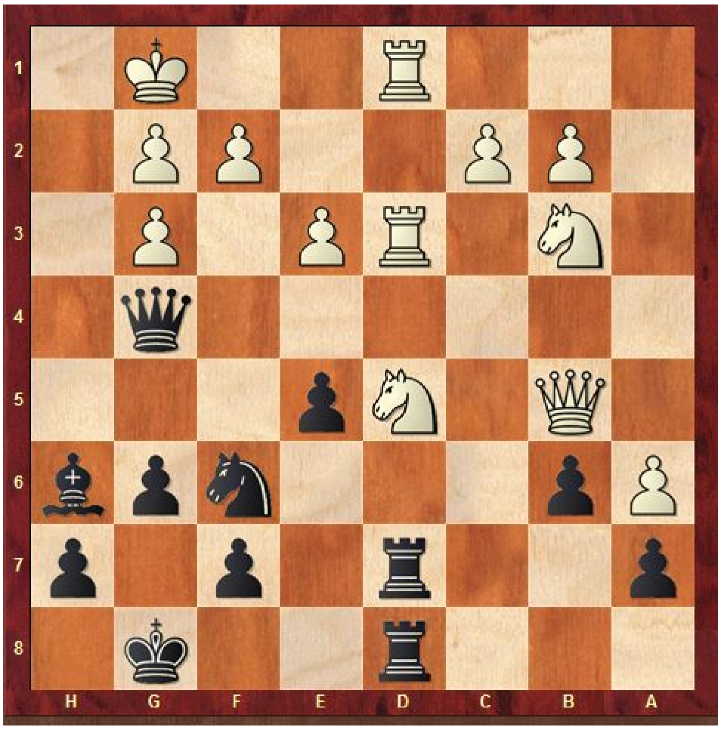
Correct solution: 1 Bxf7+ Kxf7 2 Qxf7+ Kf8 3 Ne6 mate

Problem 11



Correct solution: 1 Qg5 Bxg5 2 Rxh8 mate

Problem 12



Correct solution: 1 ... Rxd5 2 Rxd5 Qxd1+ 3 Rxd1 Rxd1+

4 Kh2 Ng4+ 5 Kh3 Nxf2+ 6 Kh2 Rh1 mate

9. Acknowledgements

I would like to thank my supervisor Prof. Dr. Ivan Bratko for his time and effort accompanying the preparations for and creation of this master thesis.

I also want to kindly thank Marjan Butala and Anton Praznik for reliably organizing the meetings with their chess students.

I would like to express particular gratitude to the chess students participating in our experiment and making it a very interesting and fun experience. Last but not least, I owe very special thanks to my co-experimenter Mr Timotej Volvavšek who was both a rock and an inspiration during this research.