Assessing the Difficulty of Chess Tactical Problems

Dayana Hristova*, Matej Guid, and Ivan Bratko

Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
* On leave of absence from University of Vienna, Vienna, Austria

***Abstract*—We investigate experts' ability to assess the difficulty of a mental task for a human. The final aim is to find formalized measures of difficulty that could be used in automated assessment of the difficulty of a task. In experiments with tactical chess problems, the experts' estimations of difficulty are compared to the statistic-based difficulty ratings on the Chess Tempo website. In an eye tracking experiment, the subjects' solutions to chess problems and the moves that they considered are analyzed. Performance data (time and accuracy) are used as indicators of subjectively perceived difficulty. We also aim to identify the attributes of tactical positions that affect the difficulty of the problem. Understanding the connection between players' estimation of difficulty and the properties of the search trees of variations considered is essential, but not sufficient, for modeling the difficulty of tactical problems. Our findings include that (a) assessing difficulty is also very difficult for human experts, and (b) algorithms designed to estimate difficulty should interpret the complexity of a game tree in the light of knowledge-based patterns that human players are able to detect in a chess problem.**

Keywords–Task Difficulty; Assessing Problem Difficulty; Eye Tracking; Problem Solving; Chess Tactical Problems; Chess

I. INTRODUCTION

In this article, we investigate the ability of experts to assess the difficulty of a mental task for a human, and study the possibilities for designing an algorithmic approach to predicting how difficult the problem will be to solve by humans [1], [2]. Modeling the difficulty of problems is a topic becoming increasingly salient in the context of the development of intelligent tutoring systems [3], neuroscience research on perceptual learning [4], and dynamic difficulty adjustment (DDA) for gaming [5], [6]. However, as-of-yet there is no developed methodology to reliably predict the difficulty for a person of solving a problem. This work therefore seeks to explore different ways of assessing difficulty, including human experts, and statistical analysis of performance data.

In our study, we use chess as an experimental domain. In our case, a problem is always defined as: given a chess position that is won by one of the two sides (White or Black), find the winning move, or *a* winning move in cases when several moves lead to victory. A chess problem is said to be *tactical* if the solution is reached mainly by calculating possible variations in the given position, rather than by long term positional judgement with little calculation of concrete variations. The starting point of our investigation is scrutinizing the relationship between a player's chess expertise and their ability to assess the difficulty of a tactical problem.

The term 'difficulty' requires further explanation. We are primarily concerned with *task difficulty*, which mediates between "subjective experience of difficulty" (that cannot be objectified) and "task complexity" (an inherent quality of a task; e.g., the properties of its state space). We define the difficulty of a problem as the probability of a person failing to solve the problem. Solving a problem is associated with uncertainty. Even in the case that a person solving a problem has complete knowledge relevant to the problem, she may occasionally miss the solution. In chess, there are well known cases of blunders when a chess grandmaster failed to see an obvious winning move. Accordingly, the difficulty depends on both the problem and the person.

The more experienced the person is in the area of the problem, the easier the problem is for that particular person. For a group of people of similar expertise and problem-solving skills, the problem's difficulty will be similar for all of them. In such cases, when talking about difficulty, we may leave out the reference to any particular individual within the group. We thus make the following assumption regarding the ranking of problems according to difficulty. For two people with different experience in the problem area, the ordering of two problems according to difficulty is the same for both people. That is, if problem 1 is easier than problem 2 for person A, then problem 1 is also easier than problem 2 for person B. Of course, this assumption may be debated, but we believe it is true in large majority of cases.

The aim of our investigation is to find underlying principles of difficulty perception and estimation for a defined group. This will allow us to omit the reference to individual persons and to focus on regularities that are required for modeling the difficulty of particular tasks.

In the case of chess tactical problems, human players will encounter difficulty when the problem exceeds the limitations of their cognitive abilities, i.e., their ability to detect relevant motifs and to calculate variations in [7]. The perception of difficulty can also be influenced by psychological factors, and from the way a particular problem is presented [8]. De Groot [9] and Jongman's [10] are among the first contributions to the academic research on thinking processes in chess. Both authors focus on the ability of players of different expertise to memorize chess positions. Research on expertise in chess has been mostly focused on the perceptual advantages of experts over novices [11], [12], [13], [14], [15].

Our study aims to explore the connection between task difficulty and expertise, as well as the variability among individuals. Although relatively little research has been devoted to the issue of problem difficulty, it has been addressed within the context of several domains, including Tower of Hanoi [16], Chinese rings [17], 15-puzzle [18], Traveling Salesperson Problem [19], Sokoban puzzle [20], Sudoku [21], and also

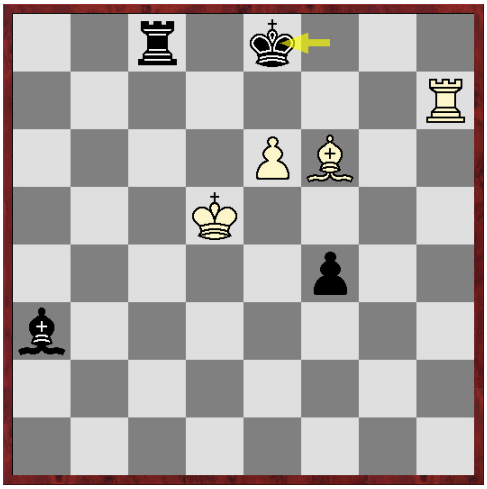

Figure 1. Chess Tempo: White to move wins. Black's last move: Kf8-e8.

chess [2]. To the best of our knowledge, no related work has been focused either on domain experts' abilities to estimate the difficulty of a mental task for a human, or on modeling the difficulty of chess tactical problems.

To approach task difficulty we are using performance measures (accuracy of solution, time, variations considered, ranking positions), psychophysiological measures (eye tracking), and qualitative retrospective reports (on perceived difficulty and on variations considered). The paper is organized as follows. In Section II, we introduce the difficulty ratings, state our hypothesis and explain why modeling the difficulty of chess tactical positions is problematic. Section III describes our methodology. We present our results of experimental data analysis in Section IV, which is followed by a thorough discussion of illustrative examples from the eye-tracking experiment. The final section of the paper is reserved for concluding remarks and directions for future work.

II. Toward Modeling Difficulty

A. Difficulty Ratings

We have adopted the difficulty ratings of Chess Tempo – an online chess platform available at www.chesstempo.com – as a reference. The Chess Tempo rating system for chess tactical problems is based on the Glicko Rating System [22]. Problems and users are both given ratings, and the user and problem rating are updated in a manner similar to the updates made after two chess players have played a game against each other, as in the Elo rating system [23]. If the user solves a problem correctly, the problem rating goes down, and the user's rating goes up, and vice versa: the problem's rating goes up in the case of incorrect solution. The Chess Tempo ratings of chess problems provides a base from which to analyze the ability of human experts to estimate the difficulty of a problem, and in our case to predict the statistically calculated measure of difficulty.

Fig. 1 shows an example of a Chess Tempo tactical problem. Superficially it may seem that the low number of pieces implies that the problem should be easy (at least for most chess players). However, this is one of the top rated Chess Tempo problems, ranked as the 52^{nd} out of 48,710 problems at the time of this writing, with the rating of 2450 rating points (other Chess Tempo statistics of this problem include: 211 users attempted to solve it, spending 602 seconds on average and with success rate of 31.75%).

What makes a particular chess tactical problem difficult? In order to understand it, we must first get acquainted with the solution. The solution of the problem in Fig. 1, shown in standard chess notation, is 1.Rh7-h8+ Ba3-f8 2.Bf6-g7! (2.e6-e7? Ke8-f7!=) Ke8-e7 3. Bg7-h6!! and Black loses in all variations, e.g.: 3... Rc8-a8 4.Rh8-h7+! Ke7-f6 5.Rh7-f7+ and the black bishop is lost. White's 3^{rd} move (3.Bg7-h6!!), virtually giving an extra move to the opponent, is particularly difficult to see in advance. Note that 3.Bg7xf8? Rc8xf8 4.Rh8-h7+ achieves nothing after 4... Ke7-f6!, with a draw. In the present case, it was not only the case that white was required to make the highly unexpected and counterintuitive move 3.Bg7-h6!!, there were also some seemingly promising alternatives that actually fail to win.

B. Hypothesis

Our hypothesis is that one's ability to estimate the difficulty of a problem is positively correlated with his or her expertise and skills in the particular problem domain. In chess, for example, such expertise and skills are usually measured by the World Chess Federation (FIDE) Elo rating. However, we conceive of chess strength as only one among multiple factors influencing the ability to make good predictions. For example, in the case of teaching, one should develop skills related to estimating difficulty in order to select appropriate tasks for one's students. Exhibiting greater expertise in a domain (e.g., being a stronger chess player) should (in principle) increase the chances of making better predictions – due to increased awareness of various possibilities and their potential consequences. However, for a group of people of similar expertise, the problem's difficulty may vary due to their specific knowledge and individual style. Moreover, it is important to note that FIDE Elo rating does not solely reflect chess players' tactical skills, but also their strategic knowledge etc. Hence, we do not necessarily expect a high linear correlation between player's FIDE Elo rating and their success in ranking the positions.

C. Modeling the difficulty of tactical positions

Guid and Bratko [2] proposed an algorithm for estimating the difficulty of chess positions in ordinary chess games. However, we found that this algorithm does not perform well when faced with chess tactical problems. The reason for this is that computer chess programs tend to solve tactical chess problems very quickly, usually already at the shallowest depths of search. The above mentioned algorithm takes into account the differences in computer evaluations when changes in decisions take place with increasing search depth, thus the computer simply recognizes most of the chess tactical problems to be rather easy, and does not distinguish well between positions of different difficulties (as perceived by humans). Estimating difficulty of chess tactical problems therefore requires a different approach, and different algorithms. It is therefore necessary to investigate the way the players of different strength solve tactical problems and estimate their difficulty, and to better understand what may be the properties of such difficulty estimation algorithms. Hence, we have used physiological measures that gauge performance in chess players' ability to assess the difficulty of tactical problems, in

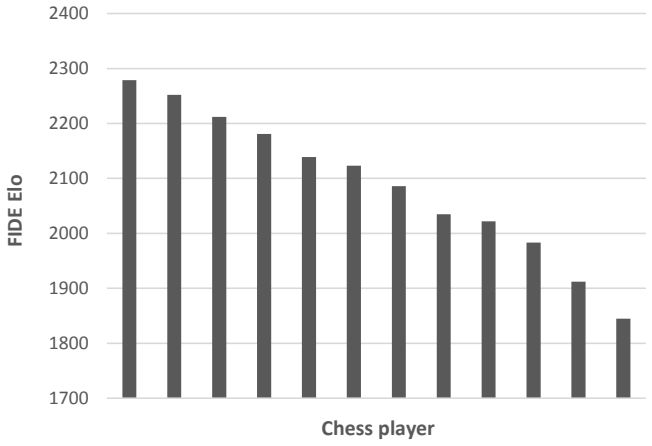

Figure 2. FIDE Elo ratings of the participants.

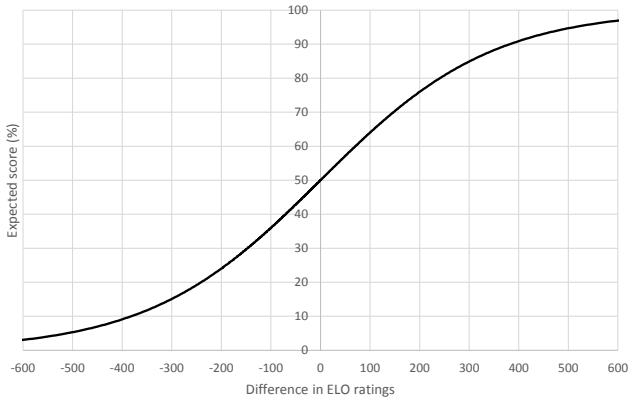

Figure 3. The Elo curve and expected scores.

addition to qualitative reports on perceived difficulty and on variations considered during problem solving.

III. METHODOLOGY

In the experiment, 12 chess experts solved and then ranked a selection of Chess Tempo problems according to their estimated difficulty. Only problems with established difficulty ratings (each attempted by at least 575 Chess Tempo users) were used. The participants consisted of 10 male and 2 female chess players (average age: 48 years). Their FIDE Elo ratings vary between 1845 and 2279 (average: 2089) and are given in Fig. 2. The Elo rating system [23] is adopted by FIDE (World Chess Federation) to estimate the strength of chess players.

Fig. 3 shows the Elo curve, i.e., a plot of the expected score at particular rating differences between two players. It is shown here in order to give the reader an approximate idea about the relative strength of the participants. Assume, for example, that two players are rated r_1 = 2200 and r_2 = 2000. The difference between r_1 and r_2 is 200 rating points in this case. According to the Elo rating system, the expected success rate of the higher rated player playing against the lower rated player is 76% and the expected success rate of the lower rated player is 24%. The expected scores do not depend on the actual ratings r_1 and r_2, but only on their difference. The expected score between two players would also be 76:24 according to the Elo curve if their ratings were, say, r_1 = 2050 and r_2 = 1850, because the rating difference in this case is also 200 points.

Eye tracking was used in order to gather perceptual data about performance and difficulty. One of the main advantages of eye tracking is that there is no appreciable lag between what is fixated and what is processed [24]. The aim was to have a grip on what is happening when the players were solving the problems, in order to understand better why a particular player missed the correct solution, what happened when a particular problem was underestimated, what piece movements did the player focused upon etc. In the experiments, the chess problems were displayed as ChessBase 9.0 generated images, 70 cm from the players' eyes. Participants' head was stabilized by a chin rest. Fig. 4 shows the experimental setting in the eye-tracking room. The players' eye movements were recorded by an *EyeLink 1000* (SR Research) eye tracking device, sampling at 500 Hz. Nine-point calibration was carried out before each part of the experiment session.

Participants were presented with 12 positions – chess tactical problems – randomly selected from Chess Tempo according to their difficulty ratings. Based on their Chess Tempo ratings, the problems can be divided into three classes of difficulty: "easy" (2 problems; their average Chess Tempo rating was 1493.9), "medium" (4; 1878.8), and "hard" (6; 2243.5). While the problems within the same difficulty class have very similar difficulty rating, each of the three classes is separated from the other by at least 350 Chess Tempo rating points. Some problems may have more than one single correct solution. Table I displays the statistics for the 12 tactical chess problems: Chess Tempo rating, success rate and the number of attempts by Chess Tempo users, average problem solving times, the number of correct solutions, and our difficulty class.

The 12 positions were presented in 3 blocks of four positions, randomized within the blocks and between blocks to avoid a sequence effect. There were short breaks to prevent the accumulation of fatigue. The experiment with each player lasted between 20 and 45 minutes. The subjects were instructed to input their solution (their suggested best move) as soon as they have found a winning solution. They were not allowed to exceed the time limit of three minutes for each position.

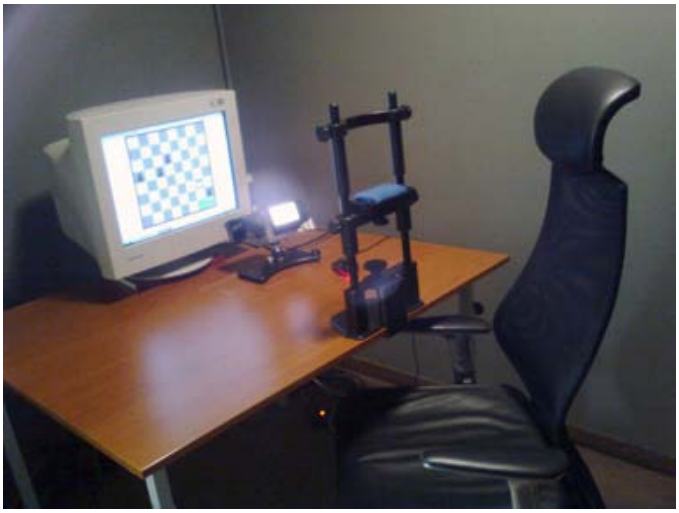

Figure 4. The experimental setting in the eye-tracking room.

TABLE I. CHESS TEMPO STATISTICS OF THE PROBLEM SET.

#	Rating	Success	Attempts	Average time	Solutions	Difficulty
1	1492.5	61%	789	3:50	2	easy
2	1495.3	62%	712	2:12	2	easy
3	1875.2	49%	669	4:08	3	medium
4	1878.1	51%	626	3:31	1	medium
5	1878.6	52%	774	3:16	1	medium
6	1883.3	53%	694	6:39	2	medium
7	2230.9	37%	809	6:53	1	difficult
8	2233.1	36%	815	6:13	1	difficult
9	2237.5	34%	575	7:01	1	difficult
10	2238.5	38%	751	5:20	1	difficult
11	2243.4	40%	572	8:49	1	difficult
12	2274.9	38%	580	9:41	1	difficult

TABLE II. THE PROBLEM-SOLVING STATISTICS.

#	Rating	Success	First moves	Pieces	Avg. time (sec)
1	1492.5	83%	4	3	71.5
2	1495.3	100%	2	2	65.5
3	1875.2	100%	2	2	67.4
4	1878.1	33%	5	3	105.0
5	1878.6	42%	4	3	101.3
6	1883.3	100%	1	1	91.6
7	2230.9	25%	2	2	78.5
8	2233.1	42%	5	3	95.0
9	2237.5	67%	3	2	113.5
10	2238.5	75%	3	2	96.3
11	2243.4	33%	3	1	120.0
12	2274.9	33%	3	1	123.5

Retrospective reports were obtained after the completion of the experiment. These reports serve as a key to understanding the way experts approached the presented position, and to the variations they considered. Chess experts are able to remember variations and are capable of reconstructing even full chess games. Hence, the retrospective reports obtained should have high validity. After the experiment, participants were asked to rate the problems (from 1 to 12) in ascending order of their difficulty. They were *not* told that the problems were divided into three difficulty classes, in order to avoid the bias introduced by this information.

The data types of primary importance to our investigation were: success rate in solving and in ranking the positions, and the type of solutions that players considered (also the incorrect ones). Success rate is an objective parameter, associated with the difficulty of the problem. It shows whether the person was able to solve the problem correctly. In combination with the retrospective reports, it provides an additional framework for understanding participants' estimation of the difficulty of particular problems. On the other hand, the measure of success rate does not account for the way that people went about solving the problem. We analyzed the success rate of the participants in ranking the positions while using Chess Tempo's (well established) difficulty ratings as a frame of reference, in order to observe how good chess players were at estimating the difficulty of problems. We found that in the cases when players did not solve the problem correctly, they tended to make a gross error in their estimate of the difficulty of the position.

The program *DataViewer* was used to generate reports about the participants' eye-gaze activity: saccades, fixations, interest areas, and trial reports. The data analysis will be discussed in the next section.

IV. ANALYSIS OF EXPERIMENTAL RESULTS

A. Statistical Analysis

We computed the correlation between various difficulty rankings for the set of chess positions. The rankings come from individual players that took part in the experiment, and from the Chess Tempo database. The Chess Tempo ranking order was derived from the Chess Tempo difficulty ratings of individual positions (see Table I). The players did not estimate difficulty ratings, but produced their ranking orders directly. That is, they were asked to rank the positions in order: from easiest to most difficult. We used Kendall's tau (τ) rank correlation coefficient which we applied to our data as follows. Given two rankings, Kendall's τ is defined by:

$$\tau = \frac{n_c - n_d}{n * \frac{n-1}{2}} = \frac{n_c - n_d}{n_c + n_d} \quad (1)$$

Here n is the number of all chess positions in the rankings, and n_c and n_d are the numbers of concordant pairs and discordant pairs, respectively. A pair of chess positions is *concordant* if their relative rankings are the same in both ranking orders. That is, if the same position precedes the other one in both rankings. Otherwise the pair is *discordant*. In our data, some of the positions were, according to Chess Tempo, of very similar difficulty. Such positions belong to the same difficulty class. To account for this, the formula above was modified. In the nominator and denominator, we only counted the pairs of positions that belong to different classes.

Table II shows respectively: position numbers and their Chess Tempo ratings (see Table I for more details about the problem positions), the rate of correct solutions by the participants, the number of different first moves tried, the number of different pieces considered for the first move, and the participants' average time spent on the problem.

Fig. 5 shows the relation between Kendall's τ and FIDE Elo ratings for each of the 12 participants. Pearson product-moment correlation coefficient (Pearson's r) was computed in

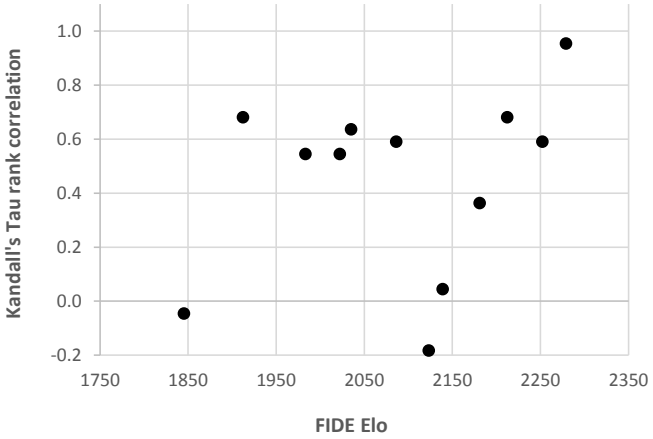

Figure 5. The relation between Kendall's τ and FIDE Elo ratings.

TABLE III. PARTICIPANTS' RESULTS OF PROBLEM SOLVING AND DIFFICULTY ESTIMATION.

Participant #	1	2	3	4	5	6	7	8	9	10	11	12	–
FIDE Elo	*2279*	*2252*	*2212*	*2181*	*2139*	*2123*	*2086*	*2035*	*2022*	*1983*	*1912*	*1845*	*Chess Tempo*
"easiest"	2	3	2	3	3	7	2	2	4	3	8	3	1
	1	2	1	10	2	8	3	3	5	2	2	9	2
	3	6	3	1	12	12	10	4	6	5	11	10	3
	5	1	10	2	5	6	1	7	2	1	7	1	4
	6	10	6	7	6	3	6	1	9	7	1	2	5
	10	4	9	6	4	2	4	10	3	8	6	11	6
	4	7	7	9	1	1	8	5	7	4	9	12	7
	9	12	5	8	10	9	12	9	1	6	5	7	8
	11	9	4	5	7	4	7	6	10	10	12	6	9
	12	8	12	4	9	10	9	12	8	11	10	4	10
	7	5	8	12	8	11	11	8	11	12	3	5	11
"hardest"	8	11	11	11	11	5	5	11	12	9	4	8	12
Discordant pairs	1	9	7	14	21	26	9	8	10	10	7	23	–
Kendall's τ	0.95	0.59	0.68	0.36	0.05	-0.18	0.59	0.64	0.55	0.55	0.68	-0.05	–
Solved correctly	11	8	8	5	7	8	6	8	5	8	9	6	–

order to determine the relationship between Kendall's τ and the chess strength of the participants (reflected by their FIDE Elo rating). There was a moderate positive relationship that is statistically not significant between Kendall's τ and FIDE Elo ratings (r = .30, n = 12, p = 0.34). Clearly, there is no linear correlation between player's Elo rating and their success in ranking the positions.

Table III demonstrates big discrepancies between ChessTempo rating and participants estimation of difficulty. It shows the difficulty rankings each participant gave to the positions they solved. For example, the chess player with FIDE Elo rating of 2279 ranked the positions in the following order: 2 (the easiest one according to the player), 1, 3, 5, 6, 10, 4, 9, 11, 12, 7, 8 (the most difficult one). The "correct" order according to the Chess Tempo ratings is given in the last column of the table. Notice that the numbers of positions refer to the position numbers given in Table I: Positions 1-2 are from the difficulty class *easy*, Positions 3-6 are from the difficulty class *medium*, and Positions 7-12 are from the difficulty class *difficult*.

As it can be seen from the table, on several occasions our participants ranked a position from the class *difficult* to be easier than a position from the class *easy*, and vice versa. Keep in mind that the difficulty classes are clearly separated by more than 350 Chess Tempo rating points. Although Chess Tempo ratings only resemble FIDE ELO ratings (they are not on the same scale), a difference of 350 points – or even 700 points, i.e., the minimal distance between the difficulty classes *easy* and *difficult* – represents a huge difference in difficulty.

We were mainly interested in the number of mistakes made in the comparison of pairs that belong to different difficulty classes, and not the ones within a class. Thus, when computing the value of Kendall's τ, we only counted the pairs of positions that belong to different classes as discordant pairs. The above mentioned player ranked Position no. 2 before Position no. 1, however, this is not a discordant pair, since they both belong to the difficulty class *easy*. The only discordant pair of this player is 10-4, since Position no. 10 is from the difficulty class *difficult* and Position no. 4 is from the difficulty class *medium*. As another example, let us briefly mention discordant pairs by the second-best rated chess player (FIDE Elo 2252): 3-2, 3-1, 6-1, 10-4, 10-5, 7-5, 12-5, 9-5, and 8-5. At the bottom of the table the number of correctly solved problems is displayed for each of the participants.

Chess players obtain their FIDE Elo ratings based on chess tournament games. However, they may not be a reliable predictor of the players' tactical skills. Even the correlation between their FIDE ratings and the performance at solving the experimental problems was surprisingly unclear. In order to verify this, we observed the relation between players' FIDE Elo ratings and the number of correctly solved tactical problems that were the subject of our experiment. The results are demonstrated in Fig. 6. Players' FIDE Elo ratings were rather poor predictors of the players' success in solving the given tactical problems. This is not completely surprising, as chess strength is dependent upon multiple factors in addition to the tactical ability. Nevertheless, this result provides an explanation for why estimating difficulty of chess tactical problems cannot be strongly correlated with players' FIDE Elo ratings. Perhaps Chess Tempo ratings would be a more reliable predictor for this purpose, however, these ratings were unavailable, since several of our participants were not Chess Tempo users.

We then observed the relationship in players' success

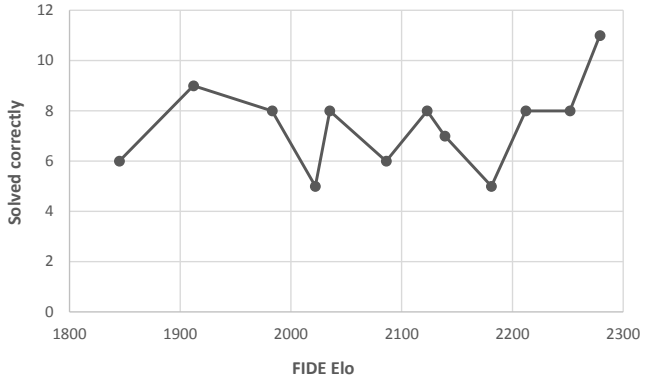

Figure 6. The relation between players' FIDE Elo ratings and their success in solving tactical chess problems.

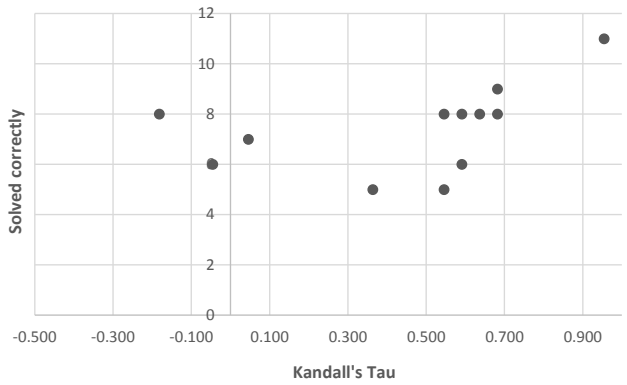

Figure 7. The relation between players' success in estimating difficulty of the problems and their success in solving these problems.

in estimating difficulty of the problems (according to the Kendall's τ rank correlation coefficient) and their success in solving the problems correctly. The results are demonstrated in Fig. 7. There was a moderate positive (statistically not significant) relationship between Kendall's τ and the problem-solving success rate (r = .44, n = 12, p = 0.15). It is interesting to notice that this relationship is slightly stronger than the relationship between Kendall's τ and FIDE Elo ratings (given in Fig. 5), which is in line with the observations stated in the previous paragraph.

Questions remained about the reasons why some rather strong players (according to their FIDE Elo ratings) performed rather poorly at estimating the difficulty of the problems, as well as at solving them correctly (and vice versa). For this purpose, we analyzed the data from the eye-tracking sessions and from the players' retrospective reports. This analysis is the subject of the following section.

B. Eye Tracking

A crucial part of eye tracking data processing is the analysis of fixations and saccades in relation to the squares of the chess-board, defined as interest areas (IAs) [25]. We analyzed what percentage of the fixations fall on a particular interest area: 1) for each individual, 2) for all fixations of all participants. For the purpose of the analysis, the following phases were focused upon: 1) the first 10 seconds after presentation; 2) overall duration of the trial. The first 10 seconds represent the *perceptual phase* according to [26].

De Groot [27] conducted several think-aloud protocols with chess players of different strengths, and discovered that much of what is important to decide on the best move occurs in the player's mind during the first few seconds of exposure to a new position. He noted that *position investigation* always comes before the investigation of possibilities. Furthermore, he divided the initial phase of the thought process into *static*, *dynamic*, and *evaluative investigation*, and found that considering the position from these three points of view typically occurs in this fixed order. Eye movement studies showed that during a few seconds exposure of a chess position, masters and novices differ on several dimensions, such as fixation durations and the number of squares fixated. Retrospective protocols indicated that very little search is conducted during these first few seconds [28].

Fig. 8 demonstrates two *EyeLink* duration-based fixation maps (visualized as "heatmaps") of Position 3. The displayed heatmaps depict the areas upon which two of the participants spent the greatest amount of time looking at. The left-hand diagram depicts the fixations made by Participant 1, and the right-hand diagram the fixations by Participant 4. The FIDE Elo ratings of the two participants are 2279 and 2181, respectively, and the first participant was more successful both in terms of ranking the positions according to their difficulty as well as in solving them correctly (see Table III for details). Position 3 has three possible solutions. The quickest way to win is mate in 4 moves: 1.b3-b4 (avoiding drawing due to stalemate – i.e., when the player to move has no legal move and his king is not in check) a5xb4 2.Kg3-f2 b4-b3 3.Kf2-f1 b3-b2 4.Sh3-f2 checkmate. However, there are two alternative solutions, which begin with the White Knight jumping to squares g5 (1.Nh3-g5) and f2 (1.Nh3-f2+), respectively. In this case, the two motifs (sacrifice a pawn and deliver checkmate vs. merely move the knight to avoid stalemate) are neatly separated on the board so that eye activity can be reliably attributed to each variation.

The heatmaps show that Participant 1 (depicted in the left-side diagram), i.e., the stronger player according to the FIDE Elo ratings, focused upon the quickest path to checkmate, while Participant 2 (see the right-side diagram) looked at the first of the alternative moves. Interestingly, the stronger player correctly assessed this position as the third easiest one, while the other one assessed it as the easiest position of the whole set (see Table III). This may be contributed to a possible message by the two heatmaps: the second player (right-side diagram) most likely did not notice that there exists a quick and effective solution which however demands a sacrifice of a pawn in order to avoid stalemate. It is stalemate in this position that causes some players to go wrong by moving White King to f2 (not noticing that this move results in no legal moves for the opponent), thus contributing to the higher rating of this problem (compared to the lower-rated Positions 1 and Position 2). We briefly note that the stronger player also spent less time on this position (20 seconds vs. 36 seconds).

Fig. 9 shows an alternative type of *EyeLink* fixation map for Position 4 – one of the positions that was regularly estimated by participants to be more difficult than its Chess Tempo rating (1861) indicates. The problem has only one correct solution - attacking Black Queen on b3 with the move 1.Nc2-a1. The retrospective accounts of the variations the players considered indicate the presence of two main motifs that all participants attended to: 1) weakness of Black King on e8; 2) trapping Black Queen on b3. The diagrams from the perceptual phase (see the left-side diagram Fig. 9) and the data from players' retrospective reports confirm that all participants spotted the first motif. The players considered different variations aiming at exploiting this motif (see the solid arrows in the right-side diagram Fig. 9): attacking with Re4xe7 or strengthening their attack through playing Qc1-e3. During the perception phase and for the overall duration of the trial, the e7 square is the most attended IA – accounting for 9.5% of the fixations in perceptual phase, and 9.3% of the fixations in overall duration of the trial, respectively. Another main piece in this motif, Re4, is the third most visited area, accounting for 7.3% of the fixations in the perception phase.

The other salient motif in Position 4 has also been reported

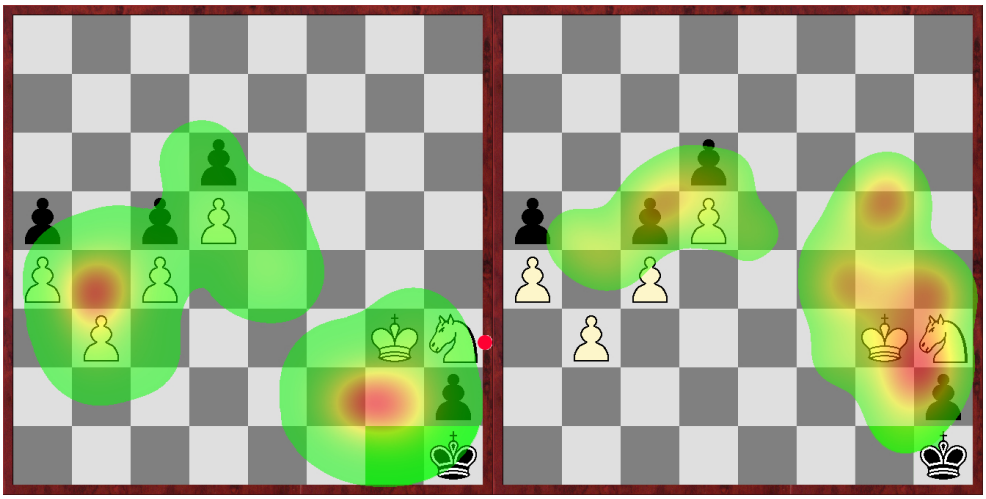

Figure 8. The *EyeLink* fixation maps for Participant 1 (left) and Participant 4 (right), showing the areas that the two players were focused on.

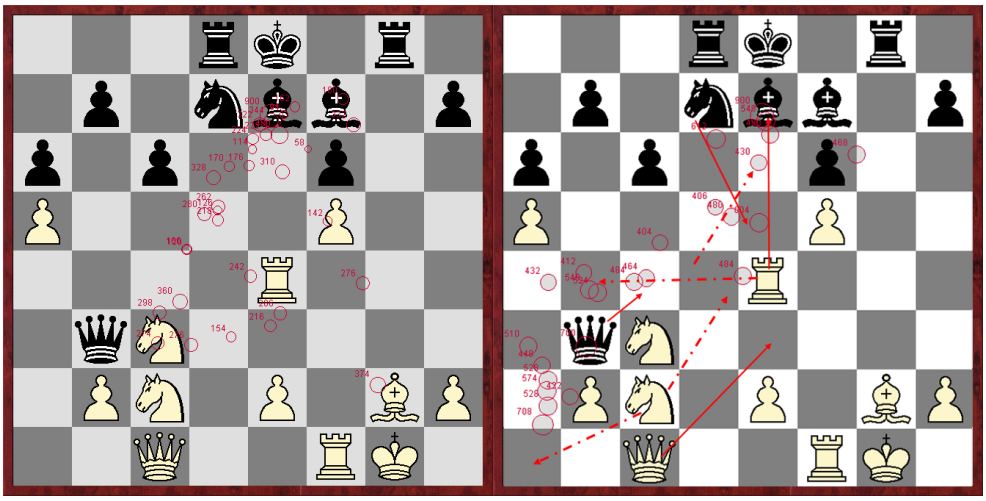

Figure 9. The *EyeLink* fixation maps of a random participant for the first 10 seconds (left) and overall duration of the trial (right), for Position 4.

in the retrospective accounts provided by all participants: trapping Black Queen on b3. As shown on Fig. 9 (right side, see the dashed arrows) three moves were considered by participants: 1.Re4-b4, 1.Nc2-d4 or 1.Nc2-a1. The percentage of fixations recorded on a1 is low – 0.3% of the whole trial. A possible explanation is that once the potentially winning move Nc2-a1 is spotted, the calculations should be focusing on the squares surrounding the Qb3 – to verify whether this move leads to a success in trapping the Queen. Also, the rate of the fixations on a1 may be influenced by the fact that a1 is a corner square. During the perceptual phase the White Knights on c2 (2.9%) and c3 (8.9%) – note that they are both on the squares surrounding the Qb3 – were among the fixations attended to for the longest period of time.

Our data shows that despite their differences in strength, participants' line of thought focused on the above two motifs. This position has only one good solution (1.Nc2-a1), but two salient motifs (two families of branches of the search tree). The first motif triggers variations that do not contain the right solution. It is evident and invites for violent moves in the center of the board and along the e-file. This motif is even more appealing as White has two Knights at her disposal – pieces that are usually strong in the center of the chess board. The candidate moves are: Re4xe7 - direct attack; Qc1-e3 - strengthening White's attack. The second motif's candidate moves appear less intuitive. Choosing to move a Knight to the edge, or even to the corner (a1), is a rather counterintuitive move since Knights are considered to be strongest in the middle of the chessboard. Ultimately, the aforementioned characteristics of the problem create predisposition for increased difficulty even for skilled chess players. Hence, the success rate for this position was 33% only.

The White Knight on c2 was identified as the piece that should be used in the first move of the winning variation in this tactical position by 66% of the participants. However, half of these players were simply unable to see the move 1.Nc2-a1, most likely because all chess players are taught not to move a knight into a corner. Putting the knight on such square reminds chess experts on the well-known expressions like "A knight on the rim is dim" or the French "*Cavalier au bord, cavalier mort*" ("A knight on the edge is dead"). Neiman and Afek [29], who analyzed the reasons why some moves are

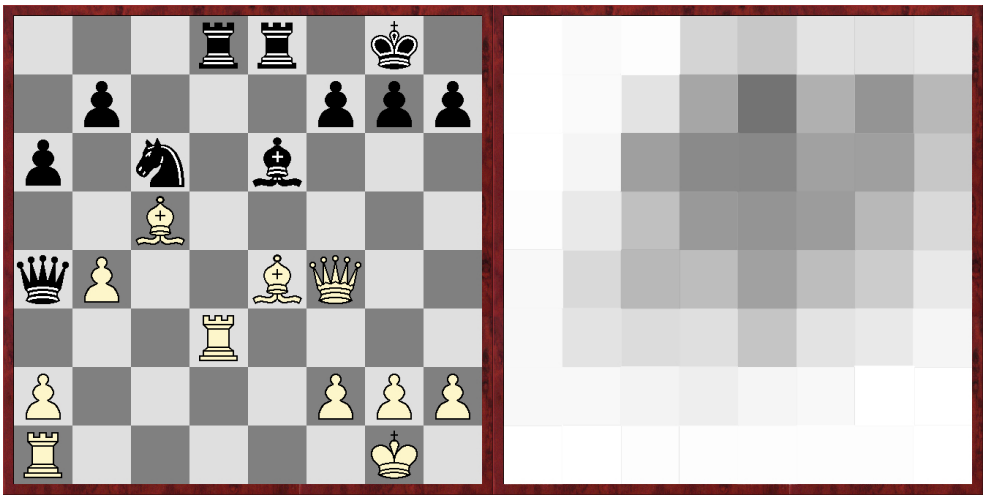

Figure 10. Left: Position 10; right: the *EyeLink* fixation map in this position for overall duration of the trial and averaged across all participants.

often "invisible" to chess players, discovered that amongst all the possible backward moves with the chess pieces, the hardest to spot are those by the knight. Actually, the incorrect alternative 1.Nc2-d4 – putting the night in the center – is so natural that it makes the correct (but backward!) move 1.Nc2-a1 very difficult to find for many players. This is an example of a mistake made due to negative knowledge transfer [30] when the player overlooks the solution of the problem as a result of their training. In other words, seemingly good moves can increase the difficulty of a chess position due to a simple (but misleading) heuristics that people may use in order to solve the problem. A famous example of the negative impact of prior experience is the so-called Einstellung (mental set) effect, which applies to a wide range of problem-solving settings where the presence of a familiar pattern can actually block the discovery of better solutions [31], [32], [33].

Fig. 10 (the left-side diagram) demonstrates Position 10, which was one of the most difficult positions in the experimental set (Table I). However, most of the participants underestimated its difficulty.

The solution is a sequence of moves based on a geometrical motif:

- Step 1: White Queen moves to h4, where it simultaneously attacks both h7 (thus threatening checkmate) and Black Rook on d8.
- Step 2: Black should use the next move to defend against checkmate, thus has no time to protect or move the Rook.
- Step 3: White exchanges the White Bishop for Black Knight (attacking Black Queen at the same time), to remove the crucial defender of Black Rook on d8.
- Step 4: Black should recapture the White Bishop, since Black Queen is under attack.
- Step 5: White wins Black Rook on d8, taking it with White Rook, supported by White Queen.

According to Chess Tempo statistics, about 60% of users failed to solve this problem. In this particular case, good combinatorial vision is required in order to recognize the geometrical pattern. Once the motif is spotted, the solution may seem rather easy. In our experiment 75% of participants solved this problem correctly, which is probably the reason for the underestimation of its difficulty.

On the right side of Fig. 10, the more frequently viewed squares according to the eye tracking data are shaded in darker grey (and vice versa). This information was obtained by averaging the fixation maps of all participants for overall duration over the trial, thus representing the "collective" fixation map for Position 10. It was interesting to observe, also on the basis of individual fixation maps in perceptual phase, that all participants focused on roughly the same part of the board. However, although one would expect that the squares that play the major role in the above presented geometrical motif (such as h4, h7, d8, c6, and e4) would stand out in this diagram, this is not the case. The most viewed square by the participants was e7, which does not play any particular role in the problem solution – except that it is positioned somewhere in the middle of the above mentioned squares. On several occasions – one of them is also the move 1.Nc2-a1! in Position 4, as explained earlier – we spotted that the players found the best move, although they barely looked at the square with the piece that is about to execute it. This reflects some of the limitations of eye tracking research when exploring higher cognitive functions (as in the case of solving chess tactical problems).

One explanation is that eye tracker records the position of the focus of the eye. However, neighboring squares are also visible to the person. In the case of Position 4, the low amount of fixations on a1 may be due to it being a corner square, or just because the player had to calculate the implications of the move Nc2-a1 for the pieces surrounding Black Queen. In both cases, there is no deterministic one-to-one mapping between the physiological data (fixations) and higher cognitive processes. Hence, in our study, the eye-tracking data proved to be most useful when providing physiological evidence of the areas (groups of adjacent squares) on the chess board that people attended to.

Analyzing eye tracking data together with the retrospections provided the basis for the previously described case studies. Eye tracking data enables the verification that a player's retrospection is a genuine account of her thought process when solving the problem, and not a post-hoc justification for her decision. In this way, they can also provide clues about the source of difficulty of a position.

C. Retrospection Reports Analysis

The retrospective reports represent an important source of information for better understanding of how the participants tackled the given problems, and what were the candidate moves and variations they considered. In this section, we briefly analyze what we learned from retrospection analysis of Position 5 (see Fig. 11). This position is an example of a position with many motifs, although they are very unsophisticated. Each motif is actually a direct threat to capture a piece in one move, as shown by the arrows in Fig. 11: both Queens are under attack (Nc6xa5, Rd8xd1, Ne3xd1) and there are many further direct threats to capture pieces (Nc6xd8, Ne3xf1, Ne3xg4, f5xg4). These single-move "motifs" are so straightforward that they hardly deserve to be called motifs due to their conceptual simplicity.

In their retrospections, the players mentioned all or most of the motifs shown in Fig. 11. Even if the motifs themselves are straightforward, the players' typical comment was "a rather complicated position." Only 50% of the players found the only correct solution b7xc6, and the most frequent incorrect solution was Rd8xd1. What makes this position difficult is the large number of simple motifs (threats) which combine in many different ways. This gives rise to relatively complex calculation of possible variations where various subsets of the "motif moves" combine in different orders. In this particular case, this is enough to make a position difficult for a human.

This case very clearly supports the following tentative conclusions indicated by the retrospections concerning other positions as well. First, the retrospections nicely conform to the early model by De Groot [27] of chess players' thinking about best moves in chess. De Groot's model conceptually consists of two stages: (1) positions investigation (in this paper referred to as "identifying motifs"), and (2) investigation of possibilities, or search (here referred to as "calculation of variations"). Strong chess players have to master both of these two tasks. But an interesting question is: which of the two tasks contributes more to the difficulty? The tentative conclusion from our analysis of retrospective reports is that this is task 2, i.e., calculation of variations. At least for players of Elo rating between about 1800 and 2300 (our players' range), the calculation skill seems to be the more important factor. The motifs detected in our positions are almost invariable between the players. The success in solving the positions however varies considerably, which is due to the different strengths at calculation of variations. These differences are not only reflected in the correctness of the solution proposed by the players, but can also be clearly detected in the players' comments that include many mistakes in the calculations.

It was interesting to notice that missing the correct line of reasoning often leads not only to underestimating, but also to overestimating the difficulty of a position. One of the participants, for example, provided the input 1... Bf8-d6?? (incorrect move) as the solution of the tactical problem in Fig. 11. This move not only fails to win, but also loses very quickly to 2.Nc6xa5 Ne3xd1 3.Bg4xf5+ (the move that the player missed, although 3.Bg4xd1 and several other moves also win for White). However, this participant ranked this position as the most difficult of the whole set of 12 positions – although this position is from the difficulty class *medium*, and therefore its Chess Tempo rating is more than 350 points lower than the ratings of 6 positions in the data set. There were actually two

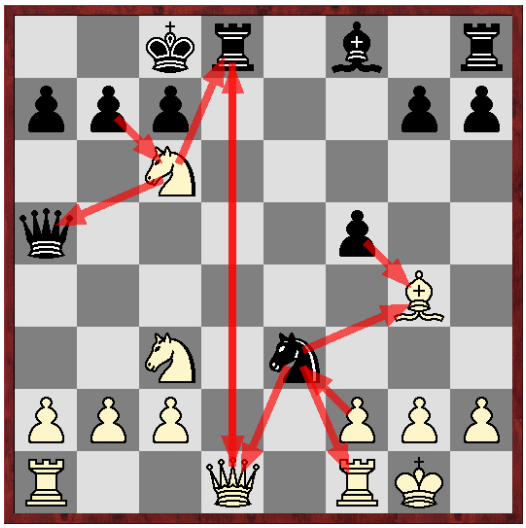

Figure 11. Each arrow indicates a move that corresponds to a separate simple motif, in this case a direct threat to capture an opponent's piece.

participants who labeled this position as the most difficult of all positions in the set (see Table III).

Several participants (5 out of 12) ranked Position 3 (Fig. 8) as the easiest one in the experimental set (refer to Table III). The retrospection analysis revealed that the participants tended to assess this position as a very easy one just because they solved it without much effort after correctly noticing the stalemate motif. However, when assessing the difficulty of such a position, one has to have in mind that not all chess players will notice this motif and that it is likely that many other players may fall into the trap of playing the seemingly logical (but wrong, due to the stalemate) move 1.Kg3-f2, with the idea of putting White King on f1 and then delivering checkmate with White Knight. It is precisely this possibility that caused this problem to score higher, i.e., to obtain a higher Chess Tempo rating. It is interesting to notice that about 50% of Chess Tempo users who attempted to solve this problem failed to solve it correctly.

V. Discussion

As expected, our data indicates that no single measurement directly predicts the difficulty of the task for the players. The best approximation to the difficulty is offered by looking at data such as success rates and solution times.

Difficulty depends on the knowledge of the player and her individual abilities - to spot the most relevant motifs and to calculate concrete variations based on the motifs observed. A tentative conclusion from our retrospection analysis is that the player's strength in calculation of variations is in fact more important than the ability to detect motifs in a position. This seems to be true at least for players in the Elo rating range between 1800 and 2300. This conclusion will be surprising to many since a common view among strong players is that a players' chess strength mainly comes from her deep understanding of chess concepts. The motifs belong to this deep chess knowledge. The calculation of variations is, on the other hand, usually considered as routine activity done without any deep understanding of chess.

Difficulty also depends on the task characteristics, such as the weight of the alternative variations - as this may have an

impact on the degree of uncertainty the player experiences (e.g., the existence of many good or seemingly good solutions may confuse). This is a crucial observation for further attempts to model difficulty.

Regarding the eye tracking data, the analysis of heatmaps and players' retrospections showed that the most attended squares of the heatmap of the player do not necessarily correspond to the squares that the player was thinking about. This is in agreement with general experience in eye tracking research. Instead, a central square of heatmap density should be understood as an indication that the neighboring squares, in addition to the maximal density square, were the specific areas of the players' interest. This is illustrated in Figs. 9 and 10. An interesting future project would be to develop a careful transformation between the heatmaps and the squares on the board that are of genuine interest to the problem solver. Chess knowledge and calculation of variations would certainly be part of such a more subtle algorithm for interpreting eye tracking data.

On the other hand, a potential use of eye tracking data is illustrated by Fig. 8, where the areas on the chess board of the two main motifs were not overlapping. In this and similar cases, the tracking of the player's eye fixations is sufficient to reliably predict what variations are considered.

The players' retrospective reports give important clues on what a mechanized difficulty estimator should look like. It should involve the calculation of chess variations, but not in the way that strong computer chess programs do. The difficulty estimator should carry out a more subtle search guided by the motifs that human players spot in a position. So, only moves relevant to these motifs should be searched, as illustrated in the analysis of the retrospections of Position 4. The complexity of such limited search should eventually produce reliable estimates of difficulty of problems for humans.

VI. Conclusion

The goal of our research is to find a formal measure of difficulty of mental problems for humans. The goal is then to implement such a measure, possibly as an algorithm, which would enable automated difficulty estimates by computers. Obvious applications of this are in intelligent tutoring systems, or in better evaluation of student's exam results, which would take into account the difficulty of exam problems.

In this paper, our study of how to mechanically estimate difficulty was limited to chess problems, more precisely to solving *tactical* chess positions. In solving such problems, humans have to use their knowledge of the domain, including pattern-based perceptual knowledge and the skill of position analysis through calculation of concrete variations of what can happen on the board. Similar kinds of knowledge and skill are required in solving other types of problems, for example in mathematics, everyday planning and decision making, and acting skillfully in unexpected social situations. Therefore, we believe that observations pertaining to difficulty in chess will apply to problem solving in other domains.

Our experiments included observing humans during problem solving (eye tracking, retrospection analysis), and humans themselves estimating the difficulty of problems (ranking of chess positions according to difficulty). One conclusion from this is that estimating difficulty is difficult also for humans, including highly skilled experts. Our experimental results did not confirm statistical significance of the hypothesis that the human's level of expertise correlates strongly with the human's ability to rank problems according to their difficulty. The results in Table III illustrate this point. The players' difficulty rankings of chess problems appear to be almost random!

Also explored was the question of which of the following stages in chess players' thinking about best moves contributes more to the difficulty of chess tactical problem solving: identifying motifs or calculation of variations? The tentative conclusion from our retrospection analysis is that, at least for players of FIDE Elo rating between about 1800 and 2300 (our players' range), the calculation skill seems to be the more important factor in this respect.

In a further analysis of the correlations between the players' rankings and Chess Tempo rankings (considered as the ground truth), and players' Elo chess ratings and the players' success in *solving* the chess problems (not estimating the difficulty), all of these relations turned out not to be statistically significant. The largest correlation coefficient was observed between overall success in difficulty ranking and the overall success in problem solving over all the experimental problems. Although this also turned out not to be statistically significant, it provides an indication that further work in this area may prove to be valuable. Namely, to investigate another hypothesis, i.e., that the success in estimating the difficulty of a particular problem depends on the ability to solve that particular problem.

Acknowledgment

The authors would like to express their gratitude to Kristijan Armeni, Grega Repovš, Anka Slana, and Gregor Geršak for providing support with the preparation of the experiment this study is based on, and to Rob Lee for his comments on an earlier version of the paper.

References

[1] D. Hristova, M. Guid, and I. Bratko, "Toward modeling task difficulty: the case of chess," in COGNITIVE 2014, The Sixth International Conference on Advanced Cognitive Technologies and Applications. IARIA, 2014, pp. 211–214.

[2] M. Guid and I. Bratko, "Search-based estimation of problem difficulty for humans," in Artificial Intelligence in Education, ser. Lecture Notes in Computer Science, H. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Springer, 2013, vol. 7926, pp. 860–863.

[3] B. Woolf, Building Intelligent Interactive Tutors. Morgan Kauffman, New York, 2008.

[4] Y. Wang, Y. Song, and Z. Qu, "Task difficulty modulates electrophysiological correlates of perceptual learning," International Journal of Psychophysiology, vol. 75, 2010, p. 234240.

[5] R. Hunicke, "The case for dynamic difficulty adjustment in games," in Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, ser. ACE '05. New York, NY, USA: ACM, 2005, pp. 429–433.

[6] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, "Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback," International Journal of Human-Computer Interaction, vol. 25, 2009, pp. 506–529.

[7] W. G. Chase and H. A. Simon, "Perception in chess," Cognitive psychology, vol. 4, no. 1, 1973, pp. 55–81.

[8] P. Chandler and J. Sweller, "Cognitive load theory and the format of instruction," in Cognition and Instruction, L. E. Associates, Ed. Taylor & Francis, 1991, pp. 292–332.

[9] A. D. de Groot, "Perception and memory versus thought: Some old ideas and recent findings," Problem solving, 1966, pp. 19–50.

[10] R.W.Jongman, Het oog van de meester [The eye of the master]. Assen: Van Gorcum, 1968.

[11] H. Simon and W. Chase, "Skill in chess," American Scientist, vol. 61, 1973, pp. 393–403.

[12] E. M. Reingold, N. Charness, M. Pomplun, and D. M. Stampe, "Visual span in expert chess players: Evidence from eye movements," Psychological Science, vol. 12, 2001, pp. 48–55.

[13] F. Gobet, J. Retschitzki, and A. de Voogt, Moves in mind: The psychology of board games. Psychology Press, 2004.

[14] E. Reingold and N. Charness, Perception in chess: Evidence from eye movements, G. Underwood, Ed. Oxford university press, 2005.

[15] F. Gobet and N. Charness, "Expertise in chess," 2006.

[16] K. Kotovsky, J. Hayes, and H. Simon, "Why are some problems hard? Evidence from tower of Hanoi," Cognitive Psychology, vol. 17, no. 2, 1985, pp. 248–294.

[17] K. Kotovsky and H. A. Simon, "What makes some problems really hard: Explorations in the problem space of difficulty," Cognitive Psychology, vol. 22, no. 2, 1990, pp. 143–183.

[18] Z. Pizlo and Z. Li, "Solving combinatorial problems: The 15-puzzle," Memory and Cognition, vol. 33, no. 6, 2005, pp. 1069–1084.

[19] M. Dry, M. Lee, D. Vickers, and P. Hughes, "Human performance on visually presented traveling salesperson problems with varying numbers of nodes," Journal of Problem Solving, vol. 1, no. 1, 2006, pp. 20–32.

[20] P. Jarušek and R. Pelánek, "Difficulty rating of sokoban puzzle," in Proc. of the Fifth Starting AI Researchers' Symposium (STAIRS 2010). IOS Press, 2010, pp. 140–150.

[21] R. Pelánek, "Difficulty rating of sudoku puzzles by a computational model," in Proc. of Florida Artificial Intelligence Research Society Conference (FLAIRS 2011). AAAI Press, 2011, pp. 434–439.

[22] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," Applied Statistics, vol. 48, 1999, pp. 377–394.

[23] A. E. Elo, The rating of chessplayers, past and present. New York: Arco Pub., 1978.

[24] M. A. Just and P. A. Carpenter, "A theory of reading: from eye fixations to comprehension." Psychological review, vol. 87, no. 4, 1980, p. 329.

[25] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.

[26] M. Bilalić, P. McLeod, and F. Gobet, "Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect," Cognition, vol. 108, no. 3, 2008, pp. 652–661.

[27] A. D. De Groot, Thought and choice in chess. Walter de Gruyter, 1978, vol. 4.

[28] A. D. De Groot, F. Gobet, and R. W. Jongman, Perception and memory in chess: Studies in the heuristics of the professional eye. Van Gorcum & Co, 1996.

[29] E. Neiman and Y. Afek, Invisible Chess Moves: Discover Your Blind Spots and Stop Overlooking Simple Wins. New in Chess, 2011.

[30] R. J. Sternberg, K. Sternberg, and J. S. Mio, Cognitive psychology. Wadsworth/Cengage Learning, 2012.

[31] A. S. Luchins, "Mechanization in problem solving: the effect of Einstellung," Psychological Monographs, vol. 54, 1942.

[32] F. Vallee-Tourangeau, G. Euden, and V. Hearn, "Einstellung defused: interactivity and mental set," Quarterly Journal of Experimental Psychology, vol. 64, no. 10, October 2011, pp. 1889–1895.

[33] H. Sheridan and E. M. Reingold, "The mechanisms and boundary conditions of the einstellung effect in chess: evidence from eye movements," PloS one, vol. 8, no. 10, 2013, p. e75796.