Toward Modeling Task Difficulty: The Case of Chess

Dayana Hristova, Matej Guid, Ivan Bratko
Faculty of Computer and Information Science
University of Ljubljana
Ljubljana, Slovenia
a0902496@unet.univie.ac.at, matej.guid@fri.uni–lj.si, bratko@fri.uni–lj.si

Abstract— **We investigate the question of experts' ability to estimate task difficulty through a case study that asks players to rate tactical chess positions. In an eye tracking experiment, experts' estimations are compared to the statistic–based difficulty ratings of the chesstempo.com website. The subjects' solutions of chess problems and their considered chess variations are analyzed in connection to ChessTempo's solutions. In addition, eye tracking and performance data (time and accuracy) are used as physiological indicators of subjectively perceived difficulty. In the course of our research, we also aim to identify the attributes of tactical positions that induce difficulty. Understanding the connection between players' estimation of difficulty and the properties of the search trees of variations considered is essential for modeling the difficulty of tactical positions.**

Keywords– Task Difficulty, Problem Solving, Search Trees, Chess, Chess Tactical Problems, Eye Tracking, Chesstempo.com

I. INTRODUCTION

Modeling the difficulty of problems is a topic becoming increasingly salient in the context of the development of tutoring systems and dynamic difficulty adjustment (DDA) for gaming [1]. Since, in chess, as in other domains, there is no developed methodology to reliably predict difficulty for each person solving a problem, we are attempting to understand different ways of assessing difficulty. 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.

We are primarily concerned with "task difficulty" that mediates between the "subjective experience of difficulty" (that cannot be objectified) and the "task complexity" – as an inherent quality of a task (e.g. the properties of its space state). In order to approach task difficulty we are using psychophysiological measures (eye tracking), performance measures (accuracy of solution, time, variations considered, ranking positions), as well as qualitative retrospective reports (on perceived difficulty and on variations considered). We define the difficulty of a problem as the probability of a person succeeding in solving the problem. Hence we have adopted the difficulty ratings of chesstempo.com – an online chess platform – as a reference. These ratings are based on two principles: 1) the success rate for the particular position; 2) the ChessTempo score of the user, who has attempted to solve it. These ratings provide a basis to analyse the ability of human experts to estimate the difficulty of a problem, and in our case – to predict the statistically accumulated measure of difficulty.

In the case of chess, the difficulty for humans is induced by exceeding the limitations of player's cognitive abilities: to detect relevant motifs, to think strategically, to calculate a variation, and find a solution. The perception of difficulty is also influenced by psychological aspects, e.g. when the player is not able to calculate a variation all the way through to a checkmate, they have to deal with uncertainty (stemming from the incompleteness of the information set [2]). To our knowledge, in chess no work has been conducted that explicitly focuses on modeling the difficulty of chess tactical problems. Also, current research on expertise in chess has been mostly focused on the perceptual advantages of experts over novices [3]. Our study aims to explore the connection between task difficulty and expertise, as well as the variability among individuals.

The paper is organized as follows. In Section II, we state our hypothesis and explain why modeling the difficulty of chess tactical positions is problematic. Section III describes our methodology. We present our preliminary results of data analysis in Section IV, which is followed by a thorough discussion of an illustrative example 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. Hypothesis

Our hypothesis is that the players' ability to estimate the difficulty of a position is positively correlated with the players' chess strength measured by 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. E.g. in the case of teaching, one should develop skills related to estimating difficulty in order

ISBN: 978-1-61208-340-7

to select appropriate tasks for the students. Being a greater expert in a domain (e.g., being a stronger chess player) should (in principle) increase the chances of making better predictions – due to the better overview over the mass of possibilities. However, for a group of people of similar expertise, the problem's difficulty may vary due to their specific knowledge and individual style. Hence, we do not expect a high linear correlation between player's Elo rating and their success in ranking the positions.

B. Modeling the difficulty of tactical positions

We observed that the algorithm for estimating difficulty of chess positions in ordinary chess games proposed by Guid and Bratko [4] fails to perform well on chess tactical problems for the following reason: the programs tend to solve the problems very quickly, usually at the shallowest depths of search. Since the algorithm takes into account the differences in computer evaluations when changes in decisions take place with increasing search depth, the computer simply recognizes most of the chess tactical problems to be rather easy, and does not distinguish between positions of different difficulties (perceived by humans). Estimating difficulty of chess tactical problems therefore requires a different approach, and different algorithms. We ought to gain an insight into the way the players of different strength solve the tactical problems, and to better understand what may be the properties of such algorithm. Hence, we use physiological measures that gauge performance in chess players' ability to assess the difficulty of tactical problems, and qualitative reports.

III. Methodology

In the experiment, so far conducted with 11 strong chess players, eye tracking is used in order to gather perceptual data about performance and difficulty. In our experiment, chess experts are solving and then ranking according to their difficulty a selection of ChessTempo problems with established difficulty ratings (each solved by minimum 600 people). The participants who have completed the experiment are 10 male and 1 female (avg. age= 48 years) chess experts. Their FIDE Elo ratings vary between 1900 and 2300. The chess problems were displayed as ChessBase 9.0 generated images, 70 cm from the players' eyes. The players' eye movements were recorded by an EyeLink 1000 eye tracking device (SR Research), sampling at 500 Hz. Nine–point calibration was carried out before (each part of) the experiment session.

Participants were presented with 12 positions randomly selected from ChessTempo according to their difficulty ratings: 6 hard; 4 medium; 2 easy. The estimation of the difficulty level is relative to the level of skills of the participants, who, as already mentioned, are strong chess players. Each of the three difficulty classes is separated from the other by 350 points. The problems within each class have very similar difficulty rating. The 12 positions were presented in 3 blocks of four positions: randomized within the blocks and between blocks to avoid a sequence effect. 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 variation that occurs to be winning. For each position, they were not allowed to exceed the time limit of three minutes. 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 have high validity. After the experiment, participants were asked to rate the problems (from 1 to 12) in ascending order of their difficulty.

DataViewer is used to generate reports about the participants' eye activity: saccades, fixations, interest areas, and trial report. The data analysis will be further discussed in the next section.

IV. Data Analysis

A. Kendall's τ rank correlation coefficient

We computed the correlation between various difficulty rankings of our set of chess positions. The rankings come from individual players that took part in the experiment, and from the ChessTempo database. The ChessTempo ranking order was derived from the ChessTempo difficulty ratings of individual positions. The players did not estimate difficulty ratings, but produced their ranking orders directly. We used Kendall's τ rank correlation coefficient which we applied to our data as follows. Given two rankings, Kendall's τ is defined by:

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

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 ChessTempo, 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 position that belong to different classes.

Fig. 1 shows for the 11 players the relation between the player's Kendall rank correlation coefficient with ChessTempo ranking, and the player's FIDE Elo rating.

Pearson product–moment correlation coefficient (Pearson's *r*) was computed in order to determine the relationship between Kendall's τ and the chess strength of the participants (reflected by their rating). There was a medium correlation that is statistically not significant between Kendall's τ and FIDE Elo ratings (r = .436, n = 11, p = 0.174). While a higher number of participants in this experiment is required, these initial results suggest that stronger players indeed tend to produce more correct rankings of chess tactical problems with respect to their difficulty.

In our experiment the strongest player was actually the best predictor of difficulty, by high margin. Furthermore, when ranking the extremes: 1 – easiest and 12 – hardest position; most players' estimations were close to the respective end of the difficulty spectrum (± 2 positions). However, participants showed high variability in their estimation. There were even cases of positions that were at the same time rated as a 1 and a 12 by the different players.

Until we finish data collection and are able to calculate the final results, and check whether there is a statistically significant correlation between Elo ratings and Kendall's coefficient for each participant, it is crucial to take positions of the experiment as case studies. This will allow us to identify aspects that have so far influenced participants' problem solving and estimation of the task difficulty.

B. *Eye tracking data*

A crucial part of the eye tracking data processing is the analysis of fixations and saccades in relation to the squares of the chessboard, defined as interest areas (IAs). We analyzed what percentage of the fixations fall on a particular interest area for both cases: 1) for each individual; 2) for all fixations of all participants. For the purpose of the analysis, we focus on the following phases: 1) the first 10 seconds after presentation; 2) the last 5 seconds preceding the input of a solution; 3) overall duration of the trial. The first two time phases are important for the data analysis as the first is conceptualized by Bilalić *et al.* [5] as a perceptual phase, and the second- as a conclusive decision making phase [5].

In the sequel, we analyze position N4 (Fig. 2) – one of the positions that was systematically estimated as more difficult than its ChessTempo rating (1861) indicates. Only 33% of the participants in our experiment inputted the correct solution, as opposed to a 50% Standard success rate in ChessTempo. According to both ChessTempo and to chess program Houdini the problem has only one good solution – Nc2-a1.

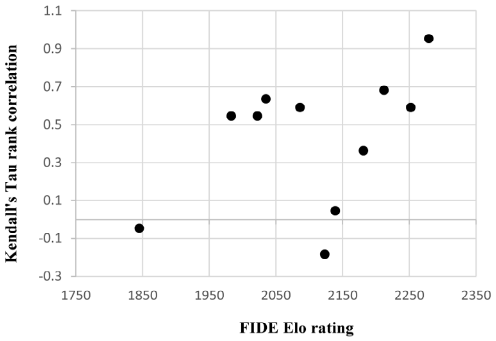

Figure 1. Correlation between Kendall's τ and FIDE Elo rating

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 and the retrospection data confirm that all participants spotted the first motif. The players considered different variations aiming at exploiting this motif (Fig. 2, solid arrows): 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 respectively 9.5%, and 9.3% of the fixations. 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 – trapping the Black Queen on b3 – has also been reported in the retrospections by all participants. As shown on Fig. 2 (with dashed arrows) three moves were considered by participants: Re4-b4, Nc2-d4 or Nc2-a1. The percentage of fixations recorded on a1 is low – 0.3% of the whole trial. However, this may also be influenced by the fact that a1 is a corner square. 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. During the perceptual phase only the Knights on c2 (2.9%) and c3 (8.9%), of all squares surrounding the Qb3 were among the fixations with highest attendance. However, during the decision phase, in addition to the knights (c3 –

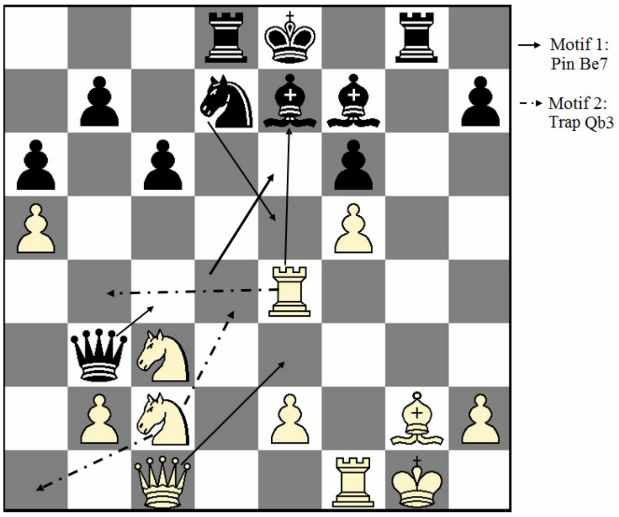

Figure 2. Position N4. The two main motifs: 1) pinned Bishop on e7; 2) trapped Queen on b3.

11.7%; c2 – 7.4%), players were also fixating more on other squares relevant to the second motif, such as: b3 (4.9%), a2 (4.3%) and b2 (3.7%).

C. Discussion of prelimiary results

Our data shows that despite of the differences in strength, participants' line of thought focused on the above two motifs. This position has only one good solution (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 strong precisely in the center of the chessboard. 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 counter intuitive 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%.

66% of the participants identified the Knight on c2 as the piece that should be used in the first move of the winning variation in this tactical position. However, half of these players were simply unable to see the move Nc2-a1 because all chess players are taught not to move a night into a corner. For a good player, a move like Nc2-a1 is almost "unethical". For the same reason, the incorrect alternative Nc2-d4, putting the night in the center, is so natural that it makes the correct Nc2-a1 practically invisible to many players**.** This is an example of a mistake made due to negative transfer [6] when the player oversees the solution of the problem as a result of their training. In other words, seemingly good moves can increase the difficulty of the chess position due to simple (but misleading) heuristics that people may use in order to solve the problem.

The number of feasible options for the player, as well as for her opponent, defines the number of nodes that are searched by the player. Then, more importantly, the chess player, according to his ability, can distinguish between relevant and not relevant moves. A player's search tree, i.e. a set of moves and variations the player might consider as reasonable candidate moves, depends on cutting off branches that do not seem promising. Often violent moves, such as capturing or mating, are favored as they lead to immediate improvement in one's situation. However, as in the case of position N4, the moves that seem to be the most promising are distracting the player from the correct solution.

V. Concluding remarks

The preliminary data does not offer statistically significant results supporting or disproving our hypothesis that the ability to predict the ChessTempo ratings correlates with the player's Elo rating. More conclusive results are expected upon completion of the data gathering. However, the mismatches between ChessTempo ratings and experts' ranking of problems according to their difficulty shed light on aspects of tactical chess positions that influence the estimation of difficulty. One of them is the properties of player's search tree. Furthermore, our case study in one of the problems also highlighted the impact of negative transfer [6] on problem solving and hence on the perception of difficulty.

Taking into account the limitations and the specificities of human problem solving, is a challenge for attempts to model the difficulty of chess problems. However, using performance and psychophysiology measures can provide the basis for modeling difficulty. This will enable the automatic detection of difficulty [2] of tactical positions that will be instrumental in the development of tutoring systems for chess. Since chess has proven itself in cognitive science research as a domain with high external validity, we hope that our work will be beneficial for modeling the difficulty of problems in other domains.

Acknowledgment

Sincere thanks to Grega Repovš, Anka Slana, Christian Armeni and Gregor Geršak for their support during the preparation of experiment this study is based on.

References

[1] R. Hunicke and V. Chapman, "AI for dynamic difficulty adjustment in games," in Challenges in game artificial intelligence, D. Fu, S. Henke and J. Orkin. Eds. Paper from the 2004 AAAI workshop, Technical report Wj–04–04, Menlo Park, CA: AAAI Press, July 2004, pp. 91–96.

[2] G. Dosi and M. Egidi, "Substantive and procedural uncertainty: An Exploration of Economic Behaviours in Changing Environments" Journal of Evolutionary Economics Volume 1 (2), June 1991 , pp. 145–168, doi:10.1007/ BF01224917

[3] E. Reingold and N. Charness, "Perception in chess: Evidence from eye movements," in Cognitive processes in eye guidance, G. Underwood, Ed. New York: Oxford university press, pp. 325– 354, 2005

[4] M. Guid and I. Bratko, "Search–Based Estimation of Problem Difficulty for Humans". Artificial Intelligence in Education. Lecture Notes in Computer Science (AIED 2013), Memphis, USA, vol. 7926, July 2013, pp. 860–863, doi: 10.1007/978–3–642–39112–5_131

[5] M. Bilalić, P. McLeod and F. Gobet, "Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect" Cognition.;108(3), September 2008, pp. 652–661. doi: 10.1016/j.cognition.2008.05.005

[6] R. J. Sternberg and K. Sternberg, "Cognitive psychology" Wadsworth: Cengage Learning, 2012