University of Zurich[UZH]

An empirical study of problem-solving in chess

Bachelor Thesis
December 5, 2021

Michael Blum
UZH, Switzerland

Student-ID: 17-717-232
michael.blum2@uzh.ch

Advisor: **Dr. Nicolò Pagan**

Prof. Dr. Anikó Hannák
Institut für Informatik
Universität Zürich
http://www.ifi.uzh.ch

Acknowledgements

I am grateful for the opportunity provided by Prof. Dr. Anikó Hannák to write my bachelor thesis in her Social Computing Group. In Dr. Nicolò Pagan, I had an excellent supervisor who provided helpful inputs and supported my journey with valuable pieces of advice.

I would also like to thank Prof Fernand Gobet for providing feedback on my experiment design and his inspiring academic contributions about chess expertise.

Special thanks go out to members of the ASK Réti chess club for beta-testing my application and of course all participants of my survey who took the time to contribute to this project. In addition, my appreciation goes out to friends and colleagues with whom I could share my ideas and feelings along the way.

Abstract

Research about the acquisition of expertise provides insights into the mental processes of the human mind and allows us to improve our learning methods. Cognition researchers often perform experiments with standardized tasks which allow them to observe and measure these mental processes. The game of chess proved to be of high value for such explorative experiments, leading to findings of pattern recognition, memory capacity, and problem-solving strategies. Our work provides empirical evidence about the importance of opening familiarity in relation to general calculation abilities in chess. We recruited 297 chess players via social networks who solved 32 purposefully selected chess tasks each in an online setting. Contrary to previous studies, which focused primarily on expert chess players, we study the tactical problem-solving abilities among amateur-level players. Our results show that opening selection significantly shapes the tactical pattern recognition of beginners in the early stages of chess development. This effect diminishes with an increased skill level to a point at which puzzles are solved equally well regardless of opening familiarity. These findings are in line with established theories of skill acquisition in chess.

Contents

1

Introduction

How do we perceive our environment, how do we learn and build up knowledge or how do we improve our mental abilities? The research field of cognitive psychology is constantly trying to provide answers to these questions and in doing so improves our understanding of the inner processes of the human mind. Thanks to this research field, we can develop more efficient learning methods, improve our decision-making or help people recovering from brain injuries.

As we will discuss in more detail later in this Introduction, cognition researchers are often performing experiments in a standardized environment. They observe and measure mental processes, develop theories about their observations and iteratively test and refine them [Barsalou, 2014]. One domain that prevailed as a prime environment for such experiments is the domain of chess. It has especially been used for research about problem-solving and the acquisition of expertise. The theoretical findings have been generalized and applied to other domains and have shaped our understanding of expertise. This thesis contributes empirical data that can be used to interpret the theories of expertise acquisition in chess.

In this introductory chapter, we will first explain our definition of the term *Expertise*, how researchers study it and how it can be scientifically measured. We will then segue into expertise in the domain of chess and how it can be explained with theories from cognitive psychology. Lastly, we summarize previous work done in this area before explaining our own research questions and study approach.

1.1 Expertise

Experts can be found in almost every professional and non-professional domain. Whenever someone is labelled as an expert, he or she usually possesses a large amount of knowledge accompanied by years of experience. This combination allows them to solve problems fast and with high accuracy compared to novices from their respective domain.

As outlined by Chi [Chi, 2006] there are two main approaches to the study of expertise. The first way is to study people at the very top of their respective domains. An example would be a Nobel prize winner or the number one ranked tennis player. The caveat of this approach is that those best-performing experts might be outliers in terms of their innate talent, genetics or might have achieved their success under special unreproducible environments (see also [Ericsson and Smith, 1991] for a distinction of outstanding performances). Therefore the findings might be of little use for the rest of the population due to the infeasibility of reproducing the observations. A more practical approach is to look at *relative expertise*. In this sense, a professional surgeon with 15 years of experience is considered an expert in relation to a medical apprentice. In this scenario, the observed differences lead to a clearer picture of what separates novices from experts and can ultimately generate guidelines on how to reproduce expertise. The term *expertise* in this paper refers to the second approach, namely *relative expertise*.

Chi [Chi, 2006] put together a generalized list of properties that manifest the experts' advantages compared to novices. Amongst other things, experts are better at identifying relevant features of a problem, they select more appropriate solution strategies and need to spend less cognitive effort due to higher automaticity. These advantages result in faster and more accurate solution finding in comparison with novices.

From a scientific point of view, we need to be able to measure these differences with concrete metrics. As Ericsson and Towne [Ericsson and Towne, 2010] discuss in detail, measuring expertise is no easy task and the difficulty of getting an objective measurement varies between domains. Even in physical sports where we would expect that objective measurements act as a reliable indicator of skill, we encounter problems. Measuring the strength and agility of a tennis player for example is not enough, because cognitive and psychological factors play a significant role as well. Similarly, comparing two surgeons is complicated as their knowledge bases, experiences and their operating environment likely differ.

In an attempt to solve this adversity, Ericsson and Smith propose their *expert-performance approach* framework [Ericsson and Smith, 1991] for the observation of expertise. This framework suggests that the performance of participants should be observed under standardized conditions (i.e. standard tasks that are representative for the problem space). By observing and analyzing the cognitive processes during the achievement of outstanding results, proposals about the acquisition of those abilities can be derived. The foundation for this expert-performance approach was laid by de Groot and later Chase and Simon in their original work on chess expertise [de Groot, 1946, Chase and Simon, 1973]. Chess in general has been used extensively in cognitive psychology research. This is primarily because chess is an intellectual game, pushing the cognitive abilities of the human mind to their limits. In addition, the rules of chess are well defined and universal and the outcome is not prone to uncontrollable circumstances. We can see how this is an attractive field for experiments if we also consider that the competitive nature of chess has led to a high amount of regular players around the globe. Therefore, it has often been proposed that chess can act as a model organism for cognitive psychology, just like the drosophila (fruit fly) does for genetics [Charness, 1992]. With respect to domain-specific limitations, these findings can then be transferred to other domains or even generalized.

1.2 Expertise in Chess

Chess masters are often portrayed as people with above-standard intelligence [Holding, 1985]. This view might be influenced by the use of chess in culture or the display of chess skills by masters through simultaneous exhibitions where they play against multiple opponents at once. This impressive feat can understandably lead to the impression of remarkable abilities. However, research was not able to show a conclusive relation between chess skills and intelligence [Bilalic et al., 2007, Gobet and Charness, 2018], mainly due to the superior impact of other factors like deliberate practice over innate abilities. In fact, deliberate practice, as in most skill-based domains, seems to be the most important indicator of chess skill [Charness et al., 2005]. Several works promote a certain amount of practice needed to achieve a certain skill level. Ericsson et al. talk about a general threshold of 10 years to achieve true expertise [Ericsson et al., 1993]. A more concrete example states that it is hard to achieve the level of a chess master with less than 1000 hours of serious study [Charness et al., 1996]. Regardless of the precision of those numbers, literature agrees on the fact that chess skill is a function of deliberate practice.

Given the established fact that practice raises chess skills, we need to investigate which cognitive processes cause this increase. Two distinct explanations can be found in literature: One theory explains the increase in skill occurring due to the improvement of analytical reasoning (faster and deeper calculations), while the other claims that pattern recognition is the key factor. As Chabris and Hearst [Chabris and Hearst, 2003] pointed out, the truth lies most likely in the middle of the two extremes. Calculation skills and pattern recognition complement each other and therefore their independent study might be a fruitless endeavour. Let's look at these two concepts separately, as they are crucial for the present study.

1.2.1 Analytical skills

With the term *analytical skills*, we are referring to the cognitive processes involved in looking ahead from a given position. In an oversimplified analogy, we can also think about it as the "cognitive muscle strength" of a chess player. It encapsulates the process of visualizing possible moves and the potential responses by the other player. This task is - depending on the complexity of the position - very demanding because of the exponential increase in continuations with increasing search depth. Being able to look further ahead is advantageous, as the precision for evaluating candidate moves increases. And faster calculation skills allows for a broader search in the same amount of time.

There has been some controversy about the differences between experts and novices when it comes to their search ability [Bilalic et al., 2008], mainly due to the non-linear nature of this skill. As summarized by Bilalic and colleagues [Bilalic et al., 2009] the main consensus nowadays is that the average search of depth follows a power function. The calculation skills increase rapidly in early stages of chess improvement, but plateau at later stages. This is also in line with the initial findings of de Groot, that search strategies differ only slightly between grandmasters and candidate masters (both experts in comparison to novices) [de Groot, 1946].

Looking at it from a different angle, it becomes obvious that raw calculation skills can not be the only indicator of chess expertise. The complexity of chess positions and their exponential tree of possible continuations would overwhelm the capacity of the human mind. Even early computers outperformed humans in terms of calculation, but none of them reached master-level performance. We had to wait until the computer program "Deep Blue" in 1997 to defeat world champion Garry Kasparov. The calculation resources of this machine were several magnitudes higher than the human limits, but it still barely managed to edge out Kasparov.

1.2.2 Memory and pattern recognition

But then how are humans able to play chess on a very high level despite their capacity constraints? The answer is based on the concept of highly selective search. Humans do not consider every possible move in a position, but rather rely on knowledge and heuristics to generate *candidate moves* that are then calculated in depth. The selection of candidate moves is triggered by pattern recognition, which is rooted in the accumulated knowledge of chess players. The fundamental theory by Chase and Simon gave insights into how chess players perceive and store chess positions in their minds [Chase and Simon, 1973]s. They found that players are able to reconstruct a position by splitting them up into pieces of information. Chase and Simon referred to these pieces of information as *chunks*. As an example a position with 25 pieces is split up into 5 chunks. Pieces inside the same chunk are related by their proximity, a relation of attack or defence or other variables. This way we can also understand how the natural limits of short-term memory capacity, which allows around seven things to be held in memory at the same time, are seemingly bypassed when it comes to chess.

The framework provided by Chase and Simon has later been extended by Gobet and Simon with their template theory [Gobet and Simon, 1996]. Besides chunks, they introduce the concepts of *retrieval structures* and *templates*. A template is a pattern of pieces that are regularly encountered by a player. It is not as rigid as a chunk, which is usually restricted to around 5 pieces in close proximity. Templates can include around a dozen pieces with the interesting property of variable slots. These slots can correspond to the positioning of single pieces or other special information about the structure. A template is manifested in long-term memory by repeated exposure to a certain structure. These structures then form the core of the template, with the details filling the variable slots. Upon sight of a new position, a chess player may then recognize that the presented position resembles a template in memory and can quickly access common plans and candidate moves related to this structure. Following the example of Bilalic et al., we will combine the different terminologies like chunks and templates under the umbrella term of *knowledge structures* for the rest of this work.

1.2.3 The angle of specialization

We have now seen two potential explanations for expertise in chess. On the one hand the improvement of analytical skills, on the other hand the accumulation of knowledge structures in memory. There are studies indicating the importance of both, but researchers struggled to pin down their relative importance. In their paper from 2009, Bilalic et al. conducted an experiment in which they tried to circumvent those struggles by approaching the problem from the angle of specialization. They recruited highly skilled chess players (each with international titles ranging from Candidate Master to Grandmaster) who either specialized in the French Defense or the Sicilian Defense (two famous opening choices for black). All participants then solved strategical chess puzzles from within and outside their area of specialization. A puzzle in chess refers to a presented position of a chess game, from which the participants have to find the best continuation. They were able to show that participants inside their specialization area matched the performance of stronger players outside their specialization area. These results indicate that accumulated knowledge structures and the related pattern recognition skills outweigh general (analytical) problem-solving abilities for strong players in strategical positions. These findings make sense if we consider that general analytical abilities plateau on high levels of skill.

1.3 Research Goals

The goals of this work can be divided into two concrete research questions about expertise in chess, one concerning general problem-solving performance across skill levels, the other one concerning an exploration of the specialization effect as introduced in the above chapter.

RQ1 How do chess players of different skill levels perform on tactical problem-solving tasks with varying attributes?

We will let chess players from a continuous rating range solve typical chess puzzles. Their task will be to find the game-winning move of a given position as fast as possible. The time used and the success rate is tracked and will allow us to draw conclusions about tactical puzzle-solving abilities.

RQ2 Can the specialization effect observed by Bilalic et al be replicated on tactical positions and with players of lower general skill?

Our second research question focuses on replicating the specialization effect observed by Bilalic et al. Compared to the related study, our population consists of amateur level players and we observe their tactical problem-solving abilities instead of strategical decision-making. In particular, we research if opening familiarity influences tactical problem-solving.

On a meta-level, the design and implementation of a user-friendly and stable application that allows the data collection via an online study constitutes the main task of this work.

2

Experiment Design

Most experiments in the area of chess expertise are performed in a lab setting, allowing for a standardized environment across all participants. Diverging from this standard, we designed our experiment as a web-based application, allowing chess players to participate over the internet. This decision was driven by the fact that the present study was carried out at the Institute of Informatics and the implementation of a user-friendly and stable application represented the main contribution of this work. It also allowed us to leverage social networks for recruiting participants from all over the world.

In this chapter, we first provide a holistic overview of the experiment flow, before focusing on the content of the experiment and the made design choices.

2.1 Overview of the application

The experiment flow follows a linear sequence consisting of five steps (see figure 2.1). The participant is guided through all steps by the application, unable to skip a step or jump between pages.

Figure 2.1: Experiment flow

The first screen the participant sees upon visiting the page communicates the general purpose of the study. After confirming that they are in a non-disturbing environment and accepting that the collected data may be published in an anonymized form, the participant may start the experiment.

The questionnaire, in the beginning, queries the user for their chess experience and some demographic data (age group and gender). To assess the participants' chess level, we ask for their official chess rating (international or national rating) and their online accounts ([lic, 2021b, che, 2021a]). All these questions are optional to answer. A control question is included to filter out subjects who are not participating for the first time.

In the next step, the participant has to indicate their familiarity with specific openings. There are four levels of familiarity, which are listed in table 2.1. In a pilot study, we observed that people have drastically different understandings of what the familiarity labels mean. As a consequence, we added reference texts for each label. As an example, *advanced* knowledge of an opening implies having played it in at least 25 long time control games.

Table 2.1: Familiarity table

Familiarity label	**Definition**
Unfamiliar	I never studied this opening and played it less than 5 times.
Basics	I know the basics of this opening but have played it less than 25 times.
Advanced	I studied this opening and used it more than 25 times.
Expert	I spent extensive time learning this opening and have played it more than 75 times.

The fourth page consists of the main task of the experiment, where a set consisting of 32 tactical chess puzzles needs to be solved by the participant. The puzzles are presented in a randomized sequence. After every puzzle there is a break of two seconds, allowing the participant to mentally reset before the next task is shown. An option to prolong the break at any point is included. Additionally, they may skip a puzzle in case they are unable to find the solution. For the display of the chessboard and the pieces, we implemented an adapted version of the open-source project *chessground* ([che, 2021b]). We assumed that all subjects are familiar with this kind of interface, as it is common in online chess. An example puzzle is included to accommodate the user before the test puzzles are presented. All performed moves were tracked and timed for later analysis.

Once the user finished all puzzles, the end screen is presented. On there, the participant can leave optional feedback and investigate how well they managed to solve the puzzles from the previous step.

Example screenshots of the user interface can be found in the Appendix.

2.2 Opening and Puzzle selection

The biggest challenge of the experiment design was the selection of appropriate tactic puzzles. As a starting point, we relied on the publicly available puzzles provided by lichess ([lic, 2021a]) consisting of over 1.7 million puzzles. Using lichess puzzles brought several benefits. Firstly, all puzzles are generated in such a way that there is one clear best solution for each position. Secondly, we were able to rely on the provided puzzle difficulty labels. This difficulty is calculated based on people of various skill levels solving the puzzles on the lichess platform. The third and most important benefit is the fact that all the puzzle positions occurred in actual games. We exploited this fact by tracing back the original game and extracting the opening from which the position originated.

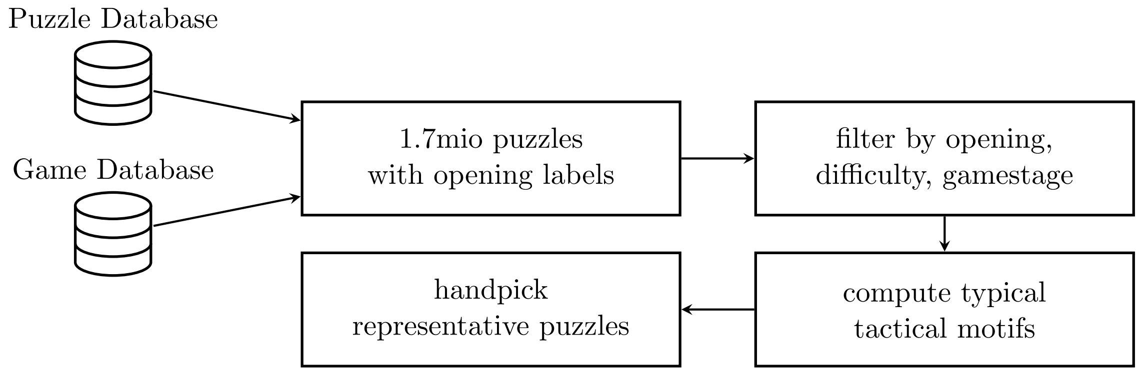
To investigate the specialization effect based on opening knowledge, we optimally want to be able to let every participant solve puzzles originating from openings they are familiar with and from openings they are unfamiliar with. Because it is unfeasible to ask the participant about their complete opening knowledge, we only look at the six most common first moves from the black side after white opens up the game with pawn to e4. This restriction is justified because every player that invested just minimal time in their opening repertoire will have learned about at least one response against the most common opening move from white (pawn to e4). Assisted by game databases we selected six of the most occurring opening variations (see table 2.2).

Table 2.2: Opening Table

Opening name	**% (after 1.e4)**	**moves**
Sicilian Defense	24.25	1. e4 c5
French Defense	12.22	1. e4 e6
Scandinavian Defense	9.26	1. e4 d5
Caro-Kann Defense	7.64	1. e4 c6
Italian Game	5.98	1. e4 e5 2. Nf3 Nc6 3. Bc4
Spanish Opening	3.92	1. e4 e5 2. Nf3 Nc6 3. Bb5
Others	36.73	N/A

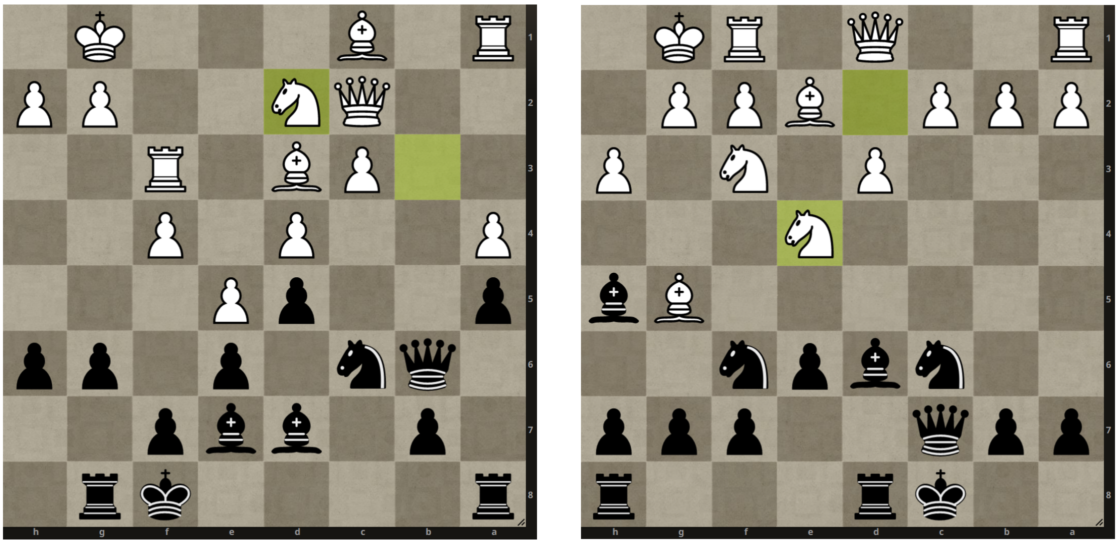
We further distinguished between two ranges of difficulties, utilizing the provided rating labels from lichess. Puzzles with a rating of 1350 up to 1550 are labelled as *easy*, and puzzles in the range of 1850 to 2050 are labelled *hard*. Another distinction was made in terms of the number of moves that were played until the puzzle position was reached. We use the terminology *early* if the position arises before the fourteenth move, and likewise *late* if the position occurs between moves 16 and 20 in the game.

Figure 2.2: Puzzle Selection



The cartesian product of these three attributes (6 openings, 2 difficulties and 2 game stages) yields 24 possible combinations. Our goal was to find 4 representative puzzles for each combination. For this purpose, we implemented an application that allowed us to filter the 1.7 million puzzles based on the above-mentioned criterias. After applying this filter we had on average a couple hundred potential puzzles per combination. Because we want to get the most representative puzzles, we algorithmically computed the most frequently occurring moves for each subset. We then proceeded to manually select puzzles based on this ranked list of typical moves. As a concrete example, we can look at the left position of figure 2.3. This is one of the four selected puzzles that originates from the *French Defense* (note the typical pawn formation) with an *easy* difficulty level and from the *late* opening stage. For this category, our algorithm ranked the moves Nxd4 and Qxd4 (the Knight, respectively the Queen capturing a piece on the central square d4) as typical motifs. As a consequence, we selected the presented puzzle, because its solution consists of both these moves (the winning combination starts with Nxd4 and wins material by force). Figure 2.2 displays a high-level overview of the puzzle selection process.

Figure 2.3: Two representative puzzles



The above positions are two samples from the total of 96 puzzles. Like in all positions there is an immediate tactical win for black. Whites last move is highlighted in yellow. The left position represents an easy puzzle that originated from a French Defense after 19 moves. The right position represents a hard puzzle originating from the Scandinavian Defense after 11 moves.

2.3 Population and Recruitment

Contrary to the study by Bilalic et al [Bilalic et al., 2009] our main target population consists of amateur chess players. In particular, we wanted to observe the specialization effect and puzzle-solving abilities on a continuous skill range. At the lower end of this range, we have people who show affection to the game of chess but have still little experience. They should have developed at least basic opening knowledge and puzzle-solving abilities. We did not target people who show no interest in chess or do not know the basics of it. The upper end of our population range aimed to observe strong chess players with multiple years of experience and a thoroughly developed opening repertoire.

Our recruitment process relied on personal contacts and promotions on various social media platforms. Direct affiliations to chess clubs and personal relationships allowed recruitment from local chess communities. This approach was mainly used in the early data collection phase to ensure the stability of the application. Later data collection phases targeted big online chess communities via social media channels. We heavily relied on the spread of the application on social networks, primarily *Reddit*, *Twitter*, *Discord* and *WhatsApp*. This approach proved to be efficient, as an initial creation of a promoting message was enough to gain traction on the respective platform. The temporal proximity of new experiment sessions allowed us to roughly judge the network reach on the various social media channels. Message-based networks (e.g. Discord, WhatsApp) generated quick but relatively few responses. Thread-based platforms (i.e. Reddit and Twitter) provided an increasing inflow of participants for about 36 hours. Reddit turned out to be by far the most efficient way of getting the attention of the chess community.

There were no kinds of compensations associated with participation.

2.4 Technical Implementation

The holistic structure of the application follows the traditional server-client model. The backend is written in the programming language Python and leverages the Django REST framework for processing requests. A connected Postgres database acts as the data storage of the application. The frontend is built utilizing the popular Javascript library Vue. The whole interface is designed in a mobile-first approach, meaning that the user interface is suitable for smaller handheld devices as well as large desktop screens. The communication between the server and client is established via the traditional web protocol HTTPS. Interested readers may contact the author for implementation details or access to the source code.

3

Analysis

In this chapter, we first perform a descriptive analysis about our observed experiment population in regards to their demographic data, chess experience and opening knowledge. Secondly, we investigate which factors influence puzzle-solving speed and accuracy. We focus on how the puzzle-solving speed and success percentage is influenced by the attributes of the puzzle, the players rating and ultimately the players opening familiarity.

3.1 Population

From the initial 455 people who started the experiment, about 20 per cent dropped out before completion, resulting in 306 complete data sets. Nine data sets had to be discarded because they did not pass the control question or due to other obvious invalid behaviour. This resulted in 297 complete and valid data sets on which the following analysis is based on.

The large majority (66%) of all participants were between 19 and 34 years old. 17.8% were 18 years old or younger and 15.1% were aged 35 or older, with three participants that did not want to disclose their age.

In terms of chess experience, a big portion of our population have started playing chess regularly in the last 2 years (47.8%). The median is at 3 years of experience with a standard deviation of 9.48 (see figure 3.1).

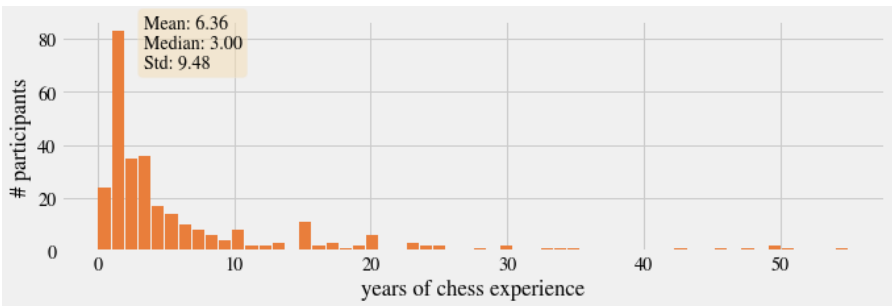


Figure 3.1: Chess experience distribution

We asked all participants for their "over-the-board" (OTB) chess ratings provided by official chess federations as well as their online accounts (lichess.org and chess.com). We received the OTB rating of 104 people and were able to retrieve reliable online ratings from 146 lichess accounts and 101 chess.com accounts. Combined we had some kind of rating metrics for 211 out of the 297 participants. All ratings are calculated using the "Elo" rating system, which is used to measure the relative skill of players. Due to different distributions of the Elo rating system between the time control formats, we needed to standardize all data points. For that purpose, we used an iterative imputation algorithm based on a bayesian ridge estimator. The resulting (standardized) ratings are displayed in figure 3.1. The mean and the median are around 1850 with a standard deviation of 366. Note that interpretations of the absolute rating numbers have to be taken with care because all numbers rely on either the precision of online ratings or the trust of user inputs. The ratings should therefore only be interpreted in relation to each other and not with rating scales from other studies or OTB ratings.

We divided all users into three rating groups. *Novices* are participants with a rating up to 1600, *Intermediates* range from 1600 to 2000 and all users with a rating above 2000 are categorized as *Experts*.

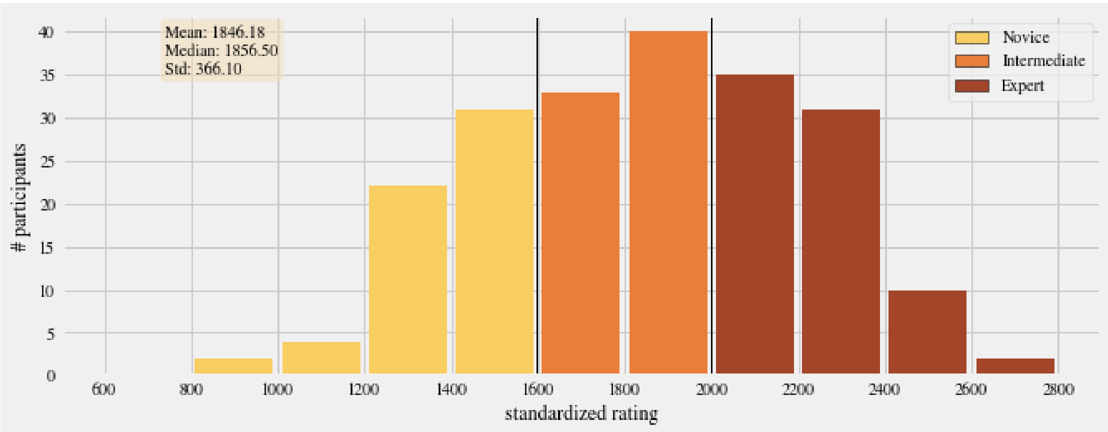


Figure 3.2: Rating distribution

3.2 Opening Familiarity

In table 3.1 we can see that there are clear differences in terms of the popularity of the various openings. 81 per cent are completely unfamiliar or have only basic knowledge of the Scandinavian Defense for example. The Sicilian Defence seems to be the most popular with over a quarter of the subjects considering themselves as experts.

Every participant got assigned one opening in which they are the most proficient and one in which they are most unfamiliar. The ideal scenario, in which we were able to assign an opening from the *Expert* and an opening from the *Unfamiliar* category, was

Table 3.1: Opening Popularity

Opening	Familiarity			
	Unfamiliar	Basics	Advanced	Expert
French Defense	48.5% (144)	30.0% (89)	9.8% (29)	11.8% (35)
Caro-Kann	43.4% (129)	26.0% (77)	15.5% (46)	15.2% (45)
Scandinavian Defense	55.6% (165)	26.0% (77)	15.5% (46)	15.2% (45)
Spanish Opening	35.4% (105)	33.3% (99)	19.5% (58)	11.8% (35)
Italian Game	28.6% (85)	30.0% (86)	22.9% (68)	19.5% (58)
Sicilian Defense	24.9% (74)	29.3% (87)	18.5% (55)	27.3% (81)

possible in 113 cases. In 80 cases the combination Unfamiliar-Advanced was assigned and in 34 cases we could assign openings from the Basics and the Expert category. That accumulates to 227 (76.4%) cases in which the familiarity gap of a participant was large enough to expect an effect. In the other cases (e.g. if a participant was Unfamiliar with all openings), we assigned them two openings at random.

3.3 Solving Speed and Accuracy

When looking at the median times spent per puzzle (see table 3.2) and the success rate (see table 3.3), we see clear differences between the three rating groups. More skilled players solve puzzles faster and more accurately. This holds regardless of the puzzles' difficulty or game stage. It is also obvious that the puzzles we labelled as hard were indeed solved slower and less accurately. The time spent on hard puzzles is about twice the time needed for easy puzzles across all rating groups. Positions from later game stages were also solved slower than early positions.

Table 3.2: Success Percentage

Rating Group	Difficulty		Gamestage	
	easy	hard	early	late
Novice	81.25%	41.74%	59.53%	63.45%
Intermediate	91.67%	65.41%	76.46%	80.65%
Expert	96.88%	89.02%	92.79%	93.11%

Table 3.3: Solving Time (Median)

Rating Group	Difficulty		Gamestage	
	easy	hard	early	late
Novice	11.82	25.58	13.61	15.19
Intermediate	9.56	20.59	11.56	13.52
Expert	5.40	10.50	6.58	8.22

In figure 3.3 we can observe how opening familiarity (Unfamiliar, Basics, Advanced, Expert) influences the solving times for the various rating groups (Novices, Intermediates and Experts). The comparison of the two familiarity extremes (*Unfamiliar* and *Expert*) shows that the median solving time for Novices decreases by 2.37 seconds on easy puzzles. On hard puzzles, the effect is even more significant, as Novices' solving time drops from above 25 seconds to below 20 seconds. The same trend is visible for Intermediates, although to a lesser degree. Their median solving times drop by 1.56 on easy puzzles, respectively almost 3 seconds on hard puzzles. Experts solving time stays more or less constant regardless of opening expertise.

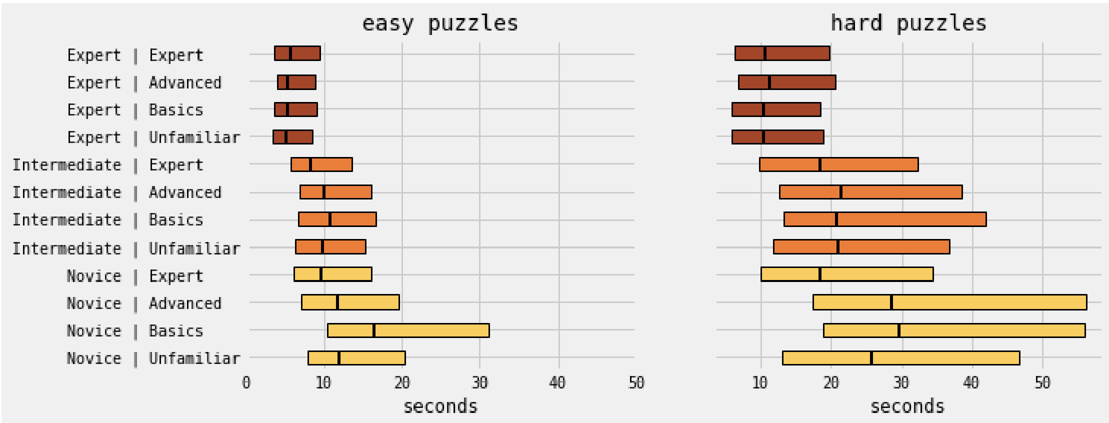


Figure 3.3: Specialization Effect on Solving Times

Looking at the success percentages (see figure 3.4) we can similarly observe that opening familiarity improves performance for Novices. Their success rate goes from 78.5% to 86.5% (+ 8%) on easy puzzles and likewise increases by 7.46% on hard puzzles. Intermediates and Experts have a considerably higher base success percentage even if they are Unfamiliar and their increase is not as high (+1.24% resp. +1.47%) on easy puzzles. For hard puzzles, experts profit from familiarity (+4.46%) but the success rate for Intermediates drops by 5.17%.

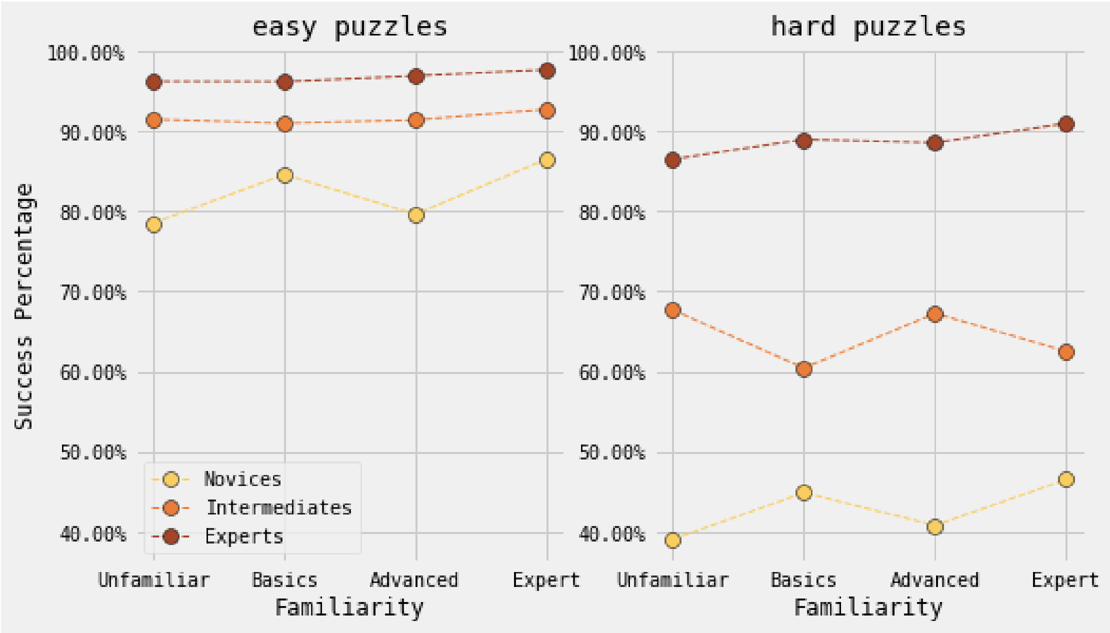
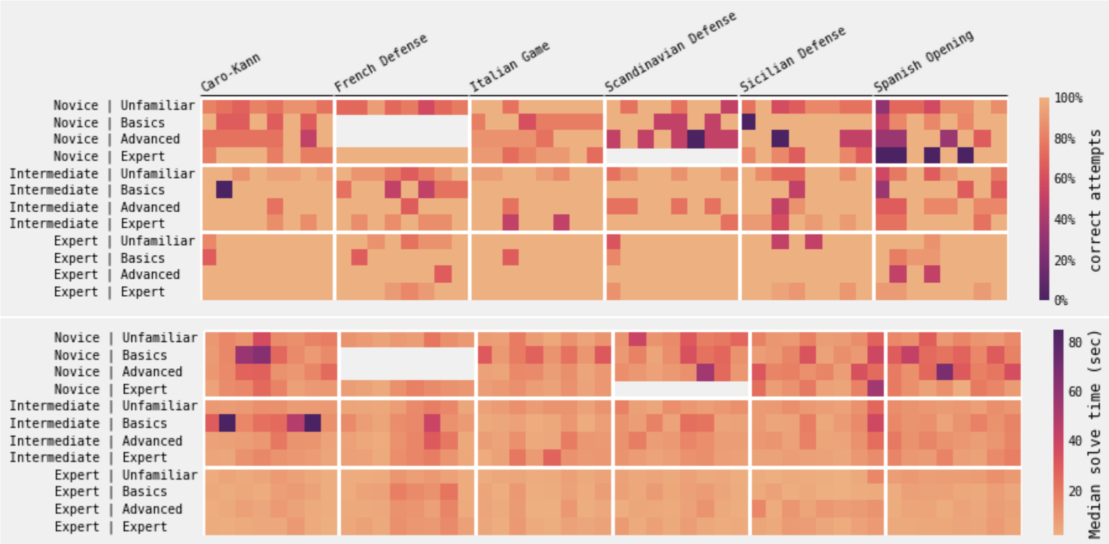


Figure 3.4: Specialization Effect on Success Percentages

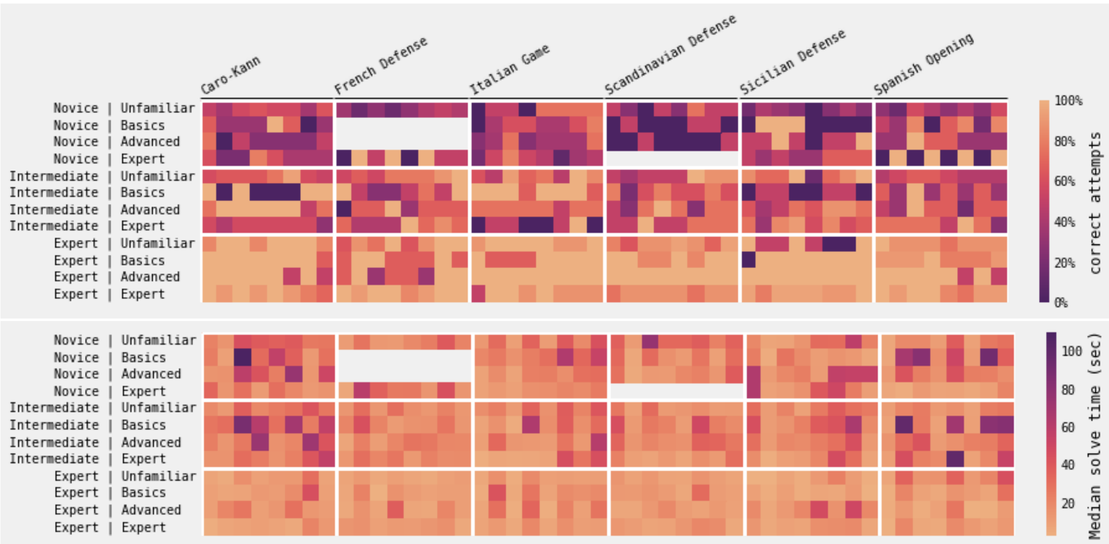
In the analysis so far we aggregated all puzzles belonging to a certain difficulty or game stage and implicitly assumed that the differences across puzzles and openings are not biasing the results. In figures 3.5 and 3.6 we show a visual comparison of each of the 96 individual puzzles (columns), sorted by opening affiliation. Furthermore, we separated all data points by the rating group of the participant and on a second level by their familiarity. This low aggregation level leads to very sparse data points which are not suited for statistical reasoning. But by visually interpreting the figures we can observe that both the success rate and median solve time do not differ drastically across openings. Some anomalies can be found (e.g. the easy puzzles from the Spanish Opening and the Caro-Kann appear to be more difficult compared to the other easy puzzles) but not to an extent that it would harm the results.

Figure 3.5: Cross-Puzzle comparison of Success Percentage and Solving Time of Easy puzzles



Each column represents one specific puzzle, sorted by its opening. Each row represents the subset of participants belonging to the specific rating group (e.g. Novice) and familiarity (e.g. Unfamiliar). Reading example: The top left square contains all data points of Novices being unfamiliar in the Caro-Kann solving the same puzzle.

Figure 3.6: Cross-Puzzle comparison of Success Percentage and Solving Time of Hard puzzles



4

Conclusions

In this chapter, we answer our two initial research questions based on the findings presented in the Analysis. The interpretation is based on the theories introduced in the first chapter. In particular, we try to pinpoint the effects of general analytical abilities and pattern recognition.

RQ1 How do chess players of different skill levels perform on tactical problem-solving tasks with varying attributes?

The analysis proves that **stronger players solved puzzles faster and more accurate regardless of difficulty and game stage**. Furthermore, **puzzle difficulty correlates strongly positive with problem-solving time and strongly negative with accuracy**. Although intuitive, these facts and especially the exact metrics provide room for interpretation about the cause of this effect. It is noteworthy that across all rating groups the time needed for hard puzzles roughly doubles in comparison to easy puzzles. The presence of this constant factor across rating groups could indicate that analytical calculation skills are the dominant factor and that hard puzzles required about twice as many calculations as easy ones.

Comparing the times needed to solve puzzles that originate from early game stages and later game stages, the data shows that the **earlier positions were solved faster but less precise than the later ones.** The time aspect can be explained by the fact that fewer pieces have been moved in early positions and the positions tend to be less complex. This lower degree of complexity could make pattern recognition easier and make calculations easier due to narrower possibilities of moves. It is counter-intuitive that the success percentage increases with later positions because improvements in solving time and success rate are usually coupled.

Experts were able to solve easy puzzles with a median time of 5.4 seconds and an accuracy of almost 97%. Such low times are a strong indicator that the search was highly selective and the solution was part of the first candidate moves, pointing towards fast and precise pattern recognition. When successful, Novices needed almost 12 seconds to solve easy puzzles. It is unclear if this time difference is caused by a worse generation of candidate moves (i.e. the participant looked at wrong moves first) or due to slower calculation speed.

RQ2 Can the specialization effect observed by Bilalic et al be replicated on tactical positions and with players of lower general skill?

Our data suggests a strong specialization effect in lower-rated players, but the same effect is relatively smaller or non-existent for higher-ranked players. Opening familiarity had the biggest effect on Novices, by increasing their success rate and dropping their solving time. The same effect can be observed to a lesser degree for intermediate players, but expert players' solving time was not affected by opening familiarity but still showed a slight increase in success rate.

These results act as proof for theories about the accumulation of knowledge structures and the importance of pattern recognition. A possible explanation for the differences across rating groups can be found in the way a player builds up his pattern recognition base. Unexperienced players naturally have encountered fewer patterns due to lack of exposure to them. Therefore, their knowledge gap between patterns from unfamiliar and familiar openings might be bigger. On the other side, we have experienced players whose knowledge base is wider and even patterns of unfamiliar openings have been encountered several times and can be recognized. This is rooted in the non-strict delimination of tactical patterns (i.e. a tactical combination that occurs regularly in one specific opening will occur from time to time in other openings as well). Another reasoning could be that strong players training methods often include solving tactical puzzles, which are typically not based on ones opening repertoire, leading to a wider knowledge base.

It is also interesting to see that **the solving time of a Novice player in familiar positions drops below the solving time needed of an intermediate player facing an unfamiliar position**. But in the same scenario, the success percentage stays lower. This may lead to the interpretation that familiarity can bridge the general skill gap in regards to candidate moves generation, but Novices still make more calculation mistakes after spotting the right candidate moves, leading to a lower success rate.

Furthermore, it is striking that the specialization effect between the familiarity levels *Unfamiliar*, *Basics* and *Advanced* is almost neglectable. It is only when a player considers himself an *Expert* in an opening, that the effect manifests itself. **That could mean that it is not enough to play an opening for less than 75 times to ingrain the related knowledge structures.**

5

Limitations and Future Work

Here we identify the limitations of our work. We categorize them into constraints regarding the experiment design and constraints about the analysis and interpretation of the results. The last section outlines what future work could contain and how our methodology can be improved.

5.1 Experiment Design

We think the main limitation of our study design is in the selection of appropriate tasks. We tried to make the puzzle selection process as objective and automated as possible, but in the end, subjective judgements had to be made. Given the fact that the author of this work does not possess expert knowledge for all openings, some selected puzzles might not have been representative of a specific opening.

In hindsight, we would have changed our categorization and description of the familiarity levels. There was some room for interpretation about opening expertise and this certainly introduced some noise.

Although deliberately chosen, the online experiment environment is of course not ideal for scientific observations. We had no control over the environments of the participants and the hardware they used.

5.2 Analysis

We assumed two main premises to be true before performing our analysis. On the one hand, we assumed that players with a similar rating (i.e. they were categorized to the same rating group) have similar problem-solving abilities. There are individual differences across subjects, which we could not account for. Secondly, we assumed that puzzles from the same difficulty range require about the same time for solving. We only performed a shallow visual analysis of cross-puzzle differences and concluded that the variance is reasonably low.

How we standardized the skill ratings from the different platforms might have led to incorrect rating classifications. We manually checked the standardized rating of each subject, but a more scientific method could have been chosen for confirmation.

In general, our analysis is lacking tests for the significance of the results. The main reason for this lack of scientific soundness is the knowledge and time constraints of the author. In addition, we aggregated data points and used the median solving times as one of our main metrics for analysis. Only looking at the median might conceal the true distribution of the data and more fine-grained aggregations might have allowed for more meaningful results.

The recruitment process of our study might have led to a sample population that is not representative of the general chess population. Our subjects were mainly young adults who are frequent users of chess-related online platforms. Their playing habits might be biased towards faster time-controls and tool-assisted practice. We consider this a minor limitation for the generalization of our study.

5.3 Future Work

Our study contributes empirical data that sheds light on theories concerning calculation and knowledge structures. We assume that our gathered data set (a total of 9504 solved puzzles by 297 participants) does provide more information than we were able to extract with our limited analysis methods. In particular, the application of linear regression models could provide more insights. One could also consider training machine learning models on the data set and derive generalized prediction models for problem-solving.

Future researchers may adapt our puzzle selection process and improve on the noted limitations. The usage of the lichess puzzle database in combination with their game database is an excellent source that we can highly recommend.

Furthermore, we showed that chess-related experiments can be done in a remote setting using a web application. Our implementation for the data collection on the server-side and the adapted chessboard interface on the client side proved to be well-designed for this purpose. Readers interested in the implementation details may contact the author.

6

Declaration of Independence

I hereby declare that this thesis has been composed by myself autonomously and that no means other than those declared were used. In every single case, I have indicated parts that were taken out of published or unpublished work, either verbatim or in a paraphrased manner, as such through a quotation. This thesis has not been handed in or published before in the same or similar form.

M. Blum

Michael Blum

Zürich, 05.12.2021

Location, Date

References

[che, 2021a] (2021a). chess.com. *https://chess.com/*. Accessed: 2021-11-07.

[che, 2021b] (2021b). chessground project. *https://github.com/ornicar/chessground*. Accessed: 2021-11-07.

[lic, 2021a] (2021a). lichess puzzles. *https://database.lichess.org/#puzzles*. Accessed: 2021-11-07.

[lic, 2021b] (2021b). lichess.org. *https://lichess.org/*. Accessed: 2021-11-07.

[Barsalou, 2014] Barsalou, L. W. (2014). Cognitive psychology: An overview for cognitive scientists.

[Bilalic et al., 2007] Bilalic, M., McLeod, P., and Gobet, F. (2007). Does chess need intelligence? a study with young chess players. *Intelligence*, 35(5):457–470.

[Bilalic et al., 2008] Bilalic, M., McLeod, P., and Gobet, F. (2008). Expert and "novice" problem solving strategies in chess: Sixty years of citing de groot (1946). *Thinking & Reasoning*, 14(4):395–408.

[Bilalic et al., 2009] Bilalic, M., McLeod, P., and Gobet, F. (2009). Specialization effect and its influence on memory and problem solving in expert chess players. *Cognitive science*, 33(6):1117–1143.

[Chabris and Hearst, 2003] Chabris, C. F. and Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed and sight of the position on grandmaster chess errors. *Cognitive Science*, 27(4):637–648.

[Charness, 1992] Charness, N. (1992). The impact of chess research on cognitive science. *Psychological research*, 54(1):4–9.

[Charness et al., 1996] Charness, N., Krampe, R., and Mayr, U. (1996). The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition. *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games*, pages 51–80.

[Charness et al., 2005] Charness, N., Tuffiash, M., Krampe, R., Reingold, E., and Vasyukova, E. (2005). The role of deliberate practice in chess expertise. *Applied Cognitive Psychology*, 19(2):151–165.

[Chase and Simon, 1973] Chase, W. G. and Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, 4(1):55–81.

[Chi, 2006] Chi, M. T. (2006). Two approaches to the study of experts'characteristics. *The Cambridge handbook of expertise and expert performance*, pages 21–30.

[de Groot, 1946] de Groot, A. D. (1946). *Het denken van den schaker: een experimenteel-psychologische studie*. Noord-Hollandsche Uitgevers Maatschappij Amsterdam.

[Ericsson et al., 1993] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363.

[Ericsson and Smith, 1991] Ericsson, K. A. and Smith, J. (1991). Prospects and limits of the empirical study of. *Toward a general theory of expertise: Prospects and limits*, page 1.

[Ericsson and Towne, 2010] Ericsson, K. A. and Towne, T. J. (2010). Expertise. *WIREs Cognitive Science*.

[Gobet and Charness, 2018] Gobet, F. and Charness, N. (2018). Expertise in chess.

[Gobet and Simon, 1996] Gobet, F. and Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive psychology*, 31(1):1–40.

[Holding, 1985] Holding, D. H. (1985). *The psychology of chess skill*. Routledge.

A

Appendix

A.1 Frontend implemtation

Figure A.1: Start Screen

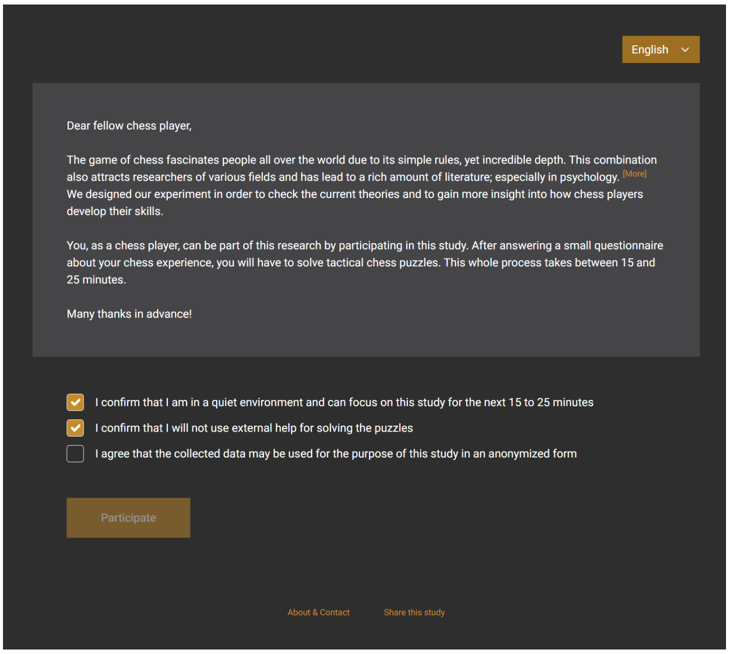


Figure A.2: Questionnaire

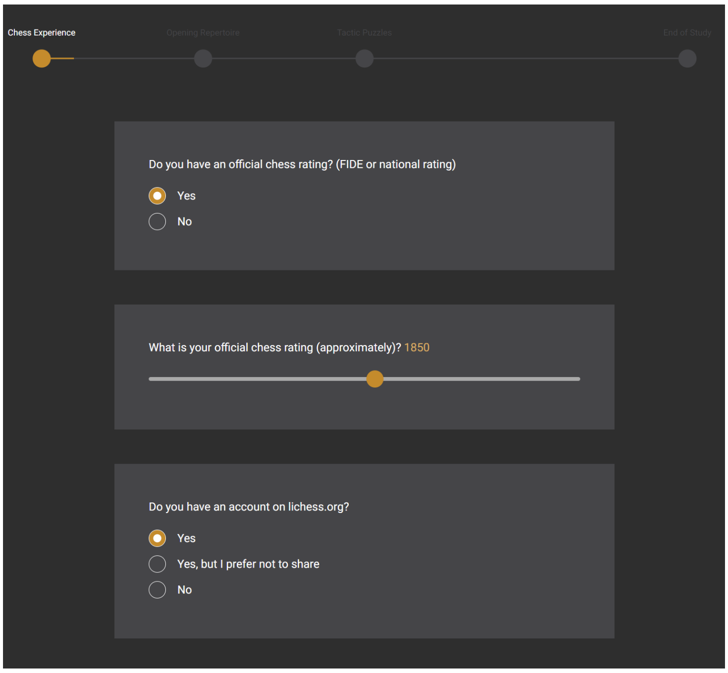


Figure A.3: Opening familiarity

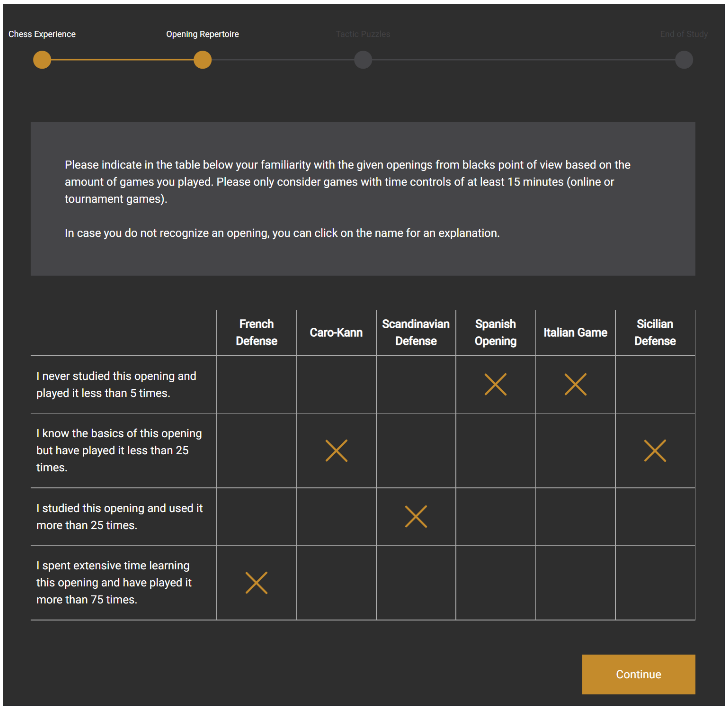


Figure A.4: Tactic Puzzles briefing

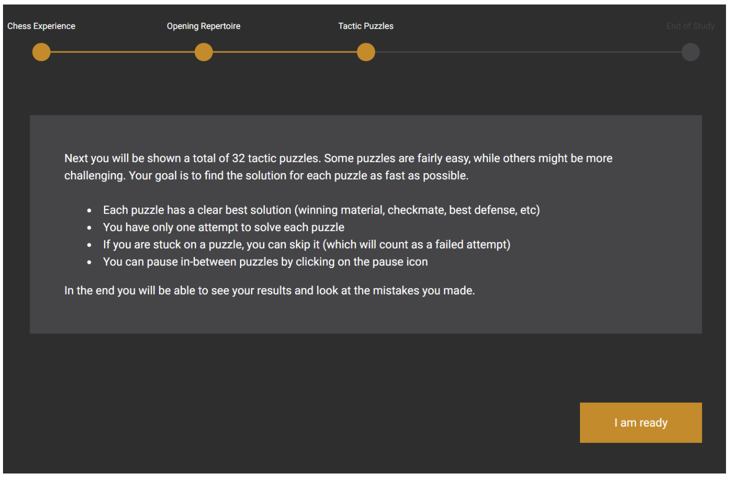
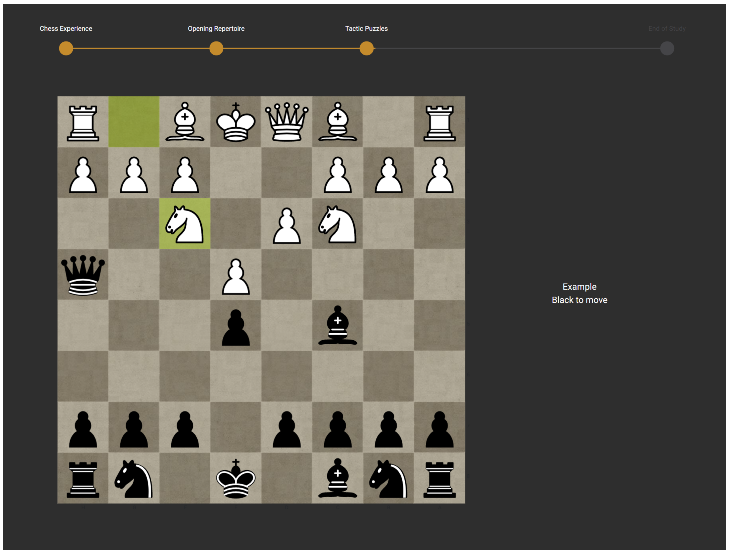


Figure A.5: Tactic Puzzles example



List of Figures

List of Tables