Categorization and Abstract Similarity in Chess

Pablo León-Villagrá (pleonvil@uos.de) Frank Jäkel (fjaekel@uos.de)
Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany

Abstract

Chess experts remember meaningful chess positions better than novices (de Groot, 1978; Chase & Simon, 1973). This can be explained with a larger number of chunks in experts' long-term memory (Gobet & Simon, 1998). These chunks are mainly based on visual representations—that is, pieces on squares. However, a recent experiment highlighted that experts prefer to group chess positions by abstract similarities that cannot be explained purely visually (Linhares & Brum, 2007). Based on these data it was claimed that chess expertise, in addition to chunks, crucially relies on abstraction and analogies. These data and the conclusions were heavily criticized because the instructions strongly biased the participants to group positions in a certain way (Bilalić & Gobet, 2009). Here, we successfully replicated this experiment with less explicit instructions. In addition, we collected category labels for the groupings that allowed us to explore the abstract principles that participants used.

Keywords: Analogy, Categorization, Chess, Expertise, Pattern Recognition, Representations, Similarity

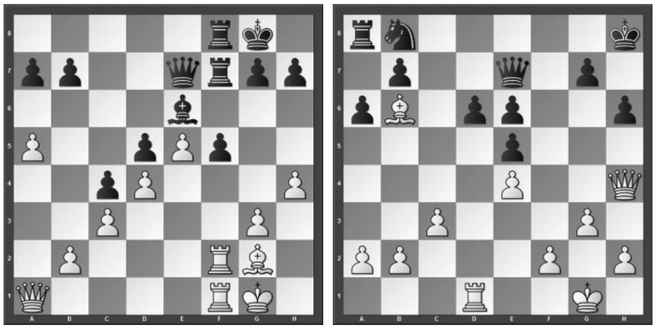

Figure 1: The left panel shows the game Rosentalis-Appel before white played Qa3. The right panel shows the game Smyslov-Reshevsky, World Championship 1948. Rosentalis commented that his game was similar to Smyslov-Reshevsky although there are few obvious visual similarities (Rowson, 2001).

Introduction

After a match strong chess players often comment that aspects of their game were similar to well-known classical games. For example, after his win against Aronian in January 2013 world champion Viswanathan Anand stated at the press conference: "It was *the same concept* [...], Rubinstein's version was even Rook takes c3 and Rook to h3, but essentially [it was] the *same idea* [...]." Or take another example, Rosentalis commented on one of his games: "When playing Qa3 the game Smyslov-Reshevsky came to my mind" (Rowson, 2001). The left panel of Figure 1 shows the position in Rosentalis' game which made him remember the position from Smyslov-Reshevsky (right panel). The two positions share no obvious visual similarity and differ considerably with regard to the pieces and their arrangement on the board. Nevertheless, Rosentalis perceived both positions to share some crucial aspects and based on this similarity he considered the move Qa3 (which allowed an exchange of queens). How do chess players represent chess positions and what kind of similarity do expert chess players perceive in positions that are visually very different?

The classical conception of expertise in chess is based on the idea that finding the right move is a process of recognition and association (Gobet & Simon, 1998). There are convincing data indicating that experienced chess players have access to a large database of stored patterns, called chunks, and these chunks are associated with plausible plans and ideas (de Groot, 1978; Chase & Simon, 1973). A chunk in chess is, hence, defined as a unit of information in long-term memory containing a meaningful grouping of pieces on squares, plus associated moves and ideas. Each chunk consists of up to five pieces and the size of the stored chunks is positively correlated with skill. Furthermore, under the assumption that experts and novices can both retain 7 ± 2 chunks in short-term memory, more skilled players can make better use of their short-term memory because they have the right chunks available. Hence, differences in skill are, to a great part, based on differences in the number and the size of the chunks stored in long-term memory. In order to accommodate various findings that were inconsistent with the original chunking theory the concept of a chunk was later expanded to more complex structures, so-called templates. Templates are formed if positions reoccur frequently and in addition to the template core (which is a classical chunk) can contain free variables (Gobet & Simon, 1996, 1998). Even though this notion expands classical chunking, in actual implementations of the theory templates are still accessed via discrimination nets and thus patterns of specific pieces on squares are fundamental for recognition.

In the anecdotal examples mentioned above strong players did not seem to rely on purely visual information, such as pieces on squares, to retrieve relevant information from memory. Linhares (2008) has argued forcefully that chess research should not focus too strongly on the visual aspects of the game. According to him, although visual similarities between stored chunks and presented positions surely play a role in chess expertise, abstract-level relations and analogies are more important for understanding expert performance (Linhares & Brum, 2007). Chess experts excel because their representations are on a high level of abstraction and these representations are not explainable by chunking alone.

In other areas of expertise it is well established that experts rely on abstract representations while novices concentrate on superficial aspects. Chi, Feltovich, and Glaser (1981) asked experts and novices to sort physics problems into groups. They found that the novices grouped problems based on superficial similarities (problems with pulleys vs problems with springs) whereas experts grouped problems based on physical, non-obvious principles (conservation of energy vs conservation of momentum). Inspired by this work, Linhares and Brum (2007) constructed a set of chess positions that could be grouped either by superficial, visual similarity or based on abstract principles. In their experiment they showed subjects 20 positions that formed 10 pairs based on 10 abstract ideas, like "material gain due to a double attack" or "endgame with bishops of the same color." These pairs are fairly natural as they consist of well-established categories in chess. Importantly, they constructed the material in a way that there were also 5 obvious pairings based on purely visual similarity. That is, the pieces and their respective positions on the board were extremely similar. But these 5 pairs were strategically or tactically very different situations due to small, but crucial, differences. Figure 2 shows an example of positions that can be grouped either visually or abstractly.

Linhares and Brum (2007) then asked chess players of varying strengths, from relatively strong masters to unrated amateurs, to pair their 20 chess positions based on strategical similarity. The expert players almost exclusively grouped the positions into abstract pairs while the novices only matched about half of the abstract pairs. Almost no visual pairs were chosen by the experts whereas novices often paired by visual similarity. Linhares and Brum concluded that multiple levels of encoding of chess positions exist, from surface representations of concrete piece relations to abstract semantic or conceptual representations consisting of abstract roles of pieces. Expert chess players perceive positions as global semantic arrangements and associate them with future developments and plans. Therefore, what differentiates experts and novices is the level of abstraction at which positions can be represented.

Bilalić and Gobet (2009) reproduced the study by Linhares and Brum introducing a condition in which participants were not asked to pair positions based on abstract similarity but on visual similarity. They did this because in the original study the instructions explicitly encouraged grouping by abstract similarity and discouraged grouping by visual similarity. They argued that this is a big methodological flaw of the original study and wanted to demonstrate that the explicit instructions simply biased the subjects to respond as the experimenter wished. In the abstract-similarity condition Bilalić and Gobet could replicate the original results—experts paired more than twice as many abstract pairs as novices. In the visual-similarity condition, in which players were instructed to group positions together based on visual similarity, both groups matched an equally low number of abstract pairs and a high number of visual pairs. The point that Bilalić and Gobet wanted to make with this experiment is that experts will group the material in any way they are instructed to. Hence, the strong conclusions that Linhares and Brum drew from their data are not warranted. Bilalić and Gobet argued

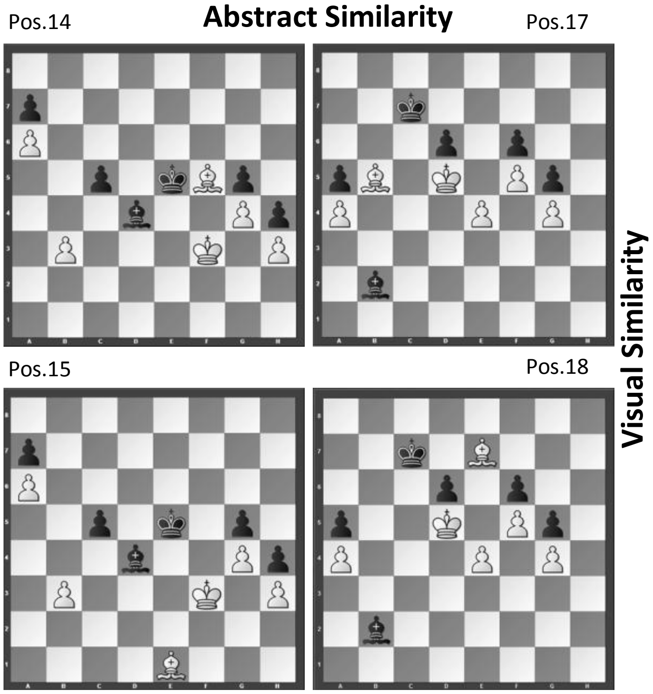

Figure 2: Four of the positions presented in the experiment by Linhares and Brum (2007). While positions 14 and 15 and positions 17 and 18 are visually almost identical they differ considerably in their abstract essence. On the other hand, positions 14 and 17 and positions 15 and 18 are very similar on an abstract level. Positions 14 and 17 are examples of endgames with opposite-colored bishops in which no side can make progress as the opponents pawns are fixed on the wrong color. In contrast, positions 15 and 18 consist of endgames with bishops of the same color and are easily winning for white because the black pawns are attackable. For example, in position 18 white can immediately win the pawn on d6 and proceed to win the game.

that the original experiment did not show any evidence that analogical reasoning or abstract similarity play an important role in chess expertise. According to them, experts did not group abstract pairs because they thought that was the natural thing to do but because they were explicitly instructed to do so.

Linhares and Brum (2009) responded to this criticism by stating that experts can of course behave like novices if told to do so. But there is an important asymmetry: "Novices cannot behave as experts." (Linhares & Brum, 2009, p. 750) The original experiment was meant to demonstrate that experts can group the stimuli by abstract similarity whereas the novices have to rely on superficial similarity, just like in the study by Chi et al. (1981). We agree with this observation but still think that Bilalić and Gobet had a valid point. The original experiment just shows that expert chess players can group by abstract similarity if told to do so, but this in itself does not show that noticing abstract similarities and making analogies is as crucial for chess expertise as Linhares and Brum claim. As participants were explicitly discouraged to pair by visual similarity we don't know whether expert players considered visual similarity relevant at all—however unlikely this may seem. Visual similarity, in any case, might still play the dominant role in memory retrieval during a game (despite the anecdotal evidence mentioned in the introduction). We think that even if the link between real-world expert performance and subjects' pairings in the experiment by Linhares and Brum is unclear, it is still interesting to try and directly probe experts and novices for their intuitions about the similarity of positions. This was also the first step in the work of Chi et al. (1981). But similarity is a difficult notion. Unless the "respects for similarity" (Medin, Goldstone, & Gentner, 1993) are precisely specified neither subject nor experimenter can be sure about what is meant by "similarity." The pairing experiment was cleverly designed to compare abstract and visual similarity against each other and thus the material probably biased the subjects to focus on these two aspects. In an actual chess game there might be even more respects for similarity than the two (abstract vs visual) that Linhares and Brum had in mind for their pairs. However, even in their experiment participants might perceive several notions of similarity in conflict with each other—but because of the clear instructions they use the one that was intended by the experimenters.

The present paper tried to replicate Linhares' and Brum's study once again. In our experiment less explicit instructions about the nature of pairs were provided. Therefore, every participant was able to pair the positions based on his or her individual, intuitive understanding of similarity in chess. We think that this is the "missing condition" in the debate. As we didn't give the subjects any obvious instructions on what we meant by similarity, potentially there could be other notions of similarity that participants might consider relevant or in conflict with the abstract similarities that Linhares and Brum had constructed. Visual surface similarity could be one of them—but not the only one. Hence, if participants, even under our fuzzy instructions, grouped the stimuli mainly as predicted by Linhares and Brum this would be somewhat stronger evidence for their claims. But if the experts spontaneously grouped by visual similarity or in any other way this would show that Bilalić's and Gobet's methodological concerns are indeed important. In addition, we decided to go beyond Linhares' and Brum's original study by also asking participants to provide a category label for each pairing. In the study by Chi et al. (1981) the category labels proved to be very helpful for understanding the difference in the representations of experts and novices. These category labels will allow us to see more directly what the participants deemed relevant for the task.

Methods

The design was based to a large extent on the original experiment and the same set of stimuli was used (Linhares & Brum, 2007). In contrast to the original experiment, the present study changed the instruction given to the participants and permitted a more differentiated evaluation of the positions in the orientation phase.

Participants

32 participants were recruited at local chess clubs, of which two participants aborted the experiment due to fatigue. The remaining 30 participants were all at least familiar with the rules of chess and basic strategies. The participants' skill is reflected in their DWZ rating. The DWZ rating system is an adaptation of the Elo rating system for the German national chess federation. Like the Elo scale the DWZ rating allows a fine differentiation of skill based on the players performance at chess tournaments. The mean DWZ rating was 1395.5 (SD = 750.3, min = 0, max = 2461), 5 players had no official rating. On average each player had performed about 2-3 hours of chess-practice per week in the last year. The mean age was 32.7 years (SD = 15.5, min = 12, max = 74). Only two participants were female and both female participants were unrated. The participants were divided into an expert and a novice group according to the mean playing strength of all registered German players (M = 1518). The expert group was composed of 16 participants with a mean DWZ of 1945.9 (SD = 268.2, min = 1569, max = 2461). The novice group had 14 members, with a mean DWZ of 766.4 (SD = 611.3, min = 0, max = 1490). Splitting the participants in this way is not ideal since both groups now contain players that are around the German average. This will blur the differences between novices and experts, making it harder to find an effect if there is one. However, this

grouping is consistent with the grouping that was used in previous studies and allows for a better comparison with these studies (Bilalić & Gobet, 2009; Linhares & Brum, 2007). In addition we also calculated and report correlations here.

Procedure

At the beginning of the experiment verbal instructions explaining the basic procedure of the survey were provided. The participants received a form of consent and after signing it proceeded to the first part of the experiment. Each participant received a questionnaire that asked for gender, age, and chess rating (DWZ). The final question asked for the amount of chess practice the subject had performed per week in the last year (Scale: 0-1 hours per week, 1-2 hours per week, 2-3 hours per week, 3-4 hours per week, more than 4 hours per week).

Familiarization Phase After these general questions the main experiment started. Participants received a short instruction on how to perform the survey. The twenty chess positions were presented in random order. The task consisted in giving an evaluation of the position and checking the particular response ("White has a winning advantage", "White has a minor advantage", "The position is equal", "Black has a winning advantage", "Black has a minor advantage", "No evaluation is possible"). For each position white was to move and unlimited time was granted to perform the task. Nevertheless, the participants were instructed to perform like in an over-the-board chess game and to use a reasonable time investment per position. After evaluating a position participants had to write down the next move for white. The familiarization phase served the purpose of activating, for each stimulus, the representations that would also be relevant in an actual game.

Pairing Phase The task in the second part of the experiment was to group the twenty chess positions from the familiarization phase into pairs. Participants received overview-sheets in which the twenty positions were displayed. The sheets were placed in front of the participants so that they could inspect all position simultaneously. There were three types of overview-sheets with different random arrangements of the positions. Positions were labeled with successive numbers (from 1 to 20) and each participant received one type of randomized overview-sheet. Instructions were to pair together positions which "intuitively seemed similar." Additionally, participants should perform "as if they were thinking about similar positions in a chess game." It can be argued that this additional instruction biased participants more than necessary. However, we decided to include it so that subjects understand that they should use the representations that they would use in an actual game.

Participants had to fill the position pairs into a designated table. Several of the expert players asked for clarification of the instructions, indicating that our instruction was indeed vague, as intended. They asked in what regard they were to interpret the similarity and were told that this was up to them and they should follow their intuitions.

Labels and Features The final part of the survey consisted of ten sheets with questions about the chosen pairs. Participants were instructed to use the overview-sheet and their response table as an aid. First of all, participants were asked to name a headline or topic for each chosen pair. After that, participants had to name attributes or features which made them select the pair. Finally, they had to rate the similarity of both positions and the prototypicality for each of the two positions for the chosen headline. The ratings are not easily interpretable as the labels were highly idiosyncratic and we usually did not have enough identical pairings for a statistical analysis. We still performed an exploratory and qualitative analysis (which had unclear results) but omit it here due to space limitations.

Stimulus Material

The twenty chess positions in the present study were the same as in the original study (Linhares & Brum, 2007). The positions were constructed in a way that 10 abstract pairs (i.e., all 20 positions) could be selected. At the same time, five control pairs (i.e., 10 positions out of 20) consisted of visually almost identical positions. The positions with a high visual similarity were very different on an abstract level, while the abstract pairs had no similarity on a visual level. An example is shown in Figure 2. All positions displayed a middlegame or endgame position and in all positions white was to move. Most positions were relatively easy to solve.

Results

Pairings and Evaluations

The two groups differed significantly in the number of abstract and control pairs that they chose. The expert group chose a significantly higher number of abstract pairs (M = 5.5, SE = .62) than the novice group (M = 3.1, SE = .54, $t(28)$ = -2.9, p = .007). The novice group matched significantly more visual pairs (M = 3.4, SE = .37) than the expert group (M = 1.7, SE = .38, $t(28)$ = 3.1, p = .004). As mentioned above, grouping participants in this way is not ideal, therefore we also calculated the correlation between DWZ rating and the choices, excluding the five unrated participants. DWZ rating correlated positively with the number of abstract pairs, $r = .52$, $p = .007$ and negatively with the number of control pairs, $r = -.49$, $p = .014$ (Figure 3). Not surprisingly, there was also a positive relationship between rating and correct evaluation of the position, $r = .77$, $p < 10^{-5}$.

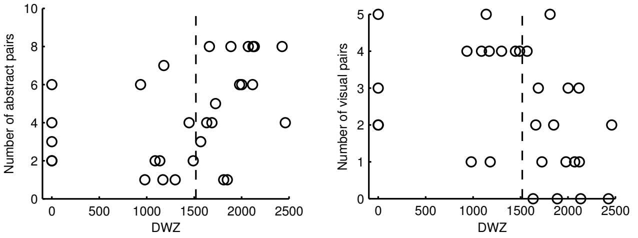

Figure 3: Number of chosen abstract pairs (left) and visual pairs (right) as a function of skill.

Category Labels

In order to compare the labels given by each participant to the labels of the other participants, the first author came up with a hierarchical classification scheme for the labels based on the abstract categories given in the original design of the material (Linhares & Brum, 2007). He then assigned the participants' labels to these classes. Some additional classes were added later based on labels that were given by several participants, for example "last pairing(s)" (the participant could not match all positions and some pairs had to be chosen randomly from the remaining material). One other unexpected label occurred relatively frequently: Situations described as deadlocked, stuck, or impenetrable were summarized in one label ("deadlocked"). This classification resulted in the hierarchical system of labels shown in Table 1.

The most common label for the novice group was "visual similarity" (34 out of 140 labels), followed by "last pairing" (12). This is in stark contrast to the experts who chose these labels rarely (5, 1). The most common labels for the expert group were "checkmate in one" (15 out of 160 labels), "endgame with opposite colored bishops" (13), "pawn endgame" (13), "passed pawn in the pawn endgame" (13) and "endgame with bishops of the same color" (12). Those labels were chosen much less frequently by the novices.

Overall, novices chose a higher number of pairs based on visual characteristics and often gave purely visual descriptions. In addition, novices often chose general levels of description ("check", "endgames" or "bishop endgame") while most experts used more specific labels (e.g., "passed pawn in the pawn endgame", "bishop endgame with bishops of the same color"). Finally, various unexpected labels were given which in many cases could not be classified using the descriptions (or more general instances of these descriptions) of the categories by Linhares and Brum (17 idiosyncratic labels in the novice group and 11 in the expert group).

1. Visual Similarity [Novices: 34, Experts: 5]
2. Tactics
   (a) Material Gain [N: 1, E: 3]
      i. Double Attack [N: 7, E: 11]
      ii. Discovered Attack [N: 2, E: 1]
   (b) Check [N: 6, E: 1]
      i. Checkmate [N: 4, E: 9]
         - Checkmate in One [N: 5, E: 15]
         - Discovered Checkmate [N: 1, E: 6]
         - Smothered Checkmate [N: 2, E: 10]
3. Endgames [N: 3, E: 0]
   (a) Pawn Endgame [N: 7, E: 13]
      i. Pawn Chain [N: 2, E: 6]
      ii. Passed Pawn [N: 2, E: 13]
      iii. Opposition [N: 6, E: 10]
   (b) Bishop Endgame [N: 8, E: 6]
      i. Bishops of the same color [N: 1, E: 12]
      ii. Bishops of different color [N: 1, E: 13]
4. Other Labels
   (a) Last Pairing [N: 12, E: 1]
   (b) No Label [N: 0, E: 2]
   (c) Incomprehensible Label [N: 5, E: 1]
   (d) Deadlocked [N: 7, E: 9]
   (e) Advantageous Positions [N: 4, E: 2]
   (f) Drawish Positions [N: 3, E: 0]
   (g) Idiosyncratic Labels [N: 17, E: 11]

Table 1: Hierarchical classification scheme used for participants' category labels.

Label-based Reanalysis of Pairing

The material consisted of pairs of visually very similar positions that could also be interpreted as clear and well-established instances of abstract categories. Even though the results in preceding studies (Linhares & Brum, 2007; Bilalić & Gobet, 2009) were very clear, in our study several unexpected pairings could be observed. Analysis of the labels given by the participants for particular pairings showed that unexpected abstract relations existed in the material. Therefore, several pairs originally designed as visual pairs allowed for a plausible and consistent abstract classification. One of the most striking examples of such an underlying unexpected abstract similarity was the pair consisting of positions 3 and 6 (not shown). Although designed as a visual pair, this pair was perceived as abstract by two of the strongest participants in this study, stating that both positions share similarity with the abstract concept of a fortress. Also, some abstract pairs were chosen for the wrong reasons. For example, if a participant chose an abstract pair but gave the label "last pairing without similarity" it is very improbable that the abstract pair had been chosen based on abstract similarity.

The labeling task we used in this study allows us to reconsider for each pair whether it should be considered as an abstract pairing or not, independent of the intended pairings by Linhares and Brum. In total, about half of the labels for abstract pairs in the novice group did not contain any sort of abstract information and most explicitly stated that they were not selected due to abstract similarity. A reevaluation of the pairings chosen by the subjects based on their labels resulted in a slightly higher number of abstract pairs for the expert group but did not change the general difference considerably (M expert abstract = 6.3, SE = .67, M novice abstract = 1.8, SE = .55). The reanalysis did increase the correlation between rating and number of abstract pairs ($r = .66$, $p = .0003$) considerably, while it did not change the correlation between rating and the number of visual pairs ($r = -.49$, $p = .01$).

Discussion

The present study replicated the results obtained by Linhares and Brum (2007) and Bilalić and Gobet (2009). On average experienced chess players chose considerably more abstract pairs than the novice group. On the other hand, novice players selected about twice as many visual pairs than the expert group.

As in our experiment no explicit instruction about the nature of expected pairs was given, the number of chosen abstract pairs was smaller than in the previous studies. The results obtained in this study weaken the methodological concerns raised by Bilalić and Gobet (2009). Experts did not simply do what they were instructed to do, but in our experiment freely chose to pair positions based on abstract similarity. Even though the material contained a considerable amount of visually almost identical positions, these possible pairings were not interpreted as relevant for the task by the experts.

As we have the category labels for the pairings as well, we have very direct evidence for what the participants deemed relevant for each pairing. The labels clearly show that novices very often spontaneously grouped by visual similarity whereas experts did not and preferred well-established abstract chess categories that, probably, simply weren't available to many of the novices. Furthermore, even though novices sometimes also used abstract categories, we could see that there was a tendency for experts to use better differentiated categories. This is a common finding in the study of expertise (Johnson & Mervis, 1997).

The present study showed that the categorization behavior observed in Linhares' and Brum's study was not simply based on the instruction but was a genuine, asymmetric characteristic of expert players. However, we still agree with the conclusion of Bilalić and Gobet that "it may well be that analogy is central to expert cognition [...]. This cannot, however, be demonstrated by asking experts to look for analogy in problems." (Bilalić & Gobet, 2009, p. 746). Future research should therefore follow more closely the example of Chi et al. (1981). While the stimuli that Linhares and Brum (2007) used were cleverly designed to contrast abstract and visual similarity, they may have biased participants to consider mostly these two aspects. The next step should be to have subjects group a representative selection of stimuli to avoid this bias. In addition, it remains to be demonstrated beyond anecdotal evidence that abstract similarity is important for actual chess playing. One way forward could be to look for abstract categories and analogical processes in think-aloud protocols that were collected during actual games.

References

Bilalić, M., & Gobet, F. (2009). They do what they are told to do: The influence of instruction on (chess) expert perception - Commentary on Linhares and Brum (2007). *Cognitive Science*, *33*(5), 743–747.

Chase, W., & Simon, H. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

de Groot, A. (1978). *Thought and choice in chess.* Mouton De Gruyter.

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

Gobet, F., & Simon, H. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*(3), 225–255.

Johnson, K., & Mervis, C. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, *126*(3), 248.

Linhares, A. (2008). *The emergence of choice: Decision-making and strategic thinking through analogies* (Tech. Rep.). Available from http://cogprints.org/6615/

Linhares, A., & Brum, P. (2007). Understanding our understanding of strategic scenarios: What role do chunks play? *Cognitive Science*, *31*(6), 989–1007.

Linhares, A., & Brum, P. (2009). How can experts see the invisible? Reply to Bilalić and Gobet. *Cognitive Science*, *33*(5), 748–751.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254.

Rowson, J. (2001). *The seven deadly chess sins.* Gambit Publications.