In Search of Templates

Fernand Gobet (frg@psyc.nott.ac.uk)
Samuel Jackson (jacksonnumber5@hotmail.com)
School of Psychology
University of Nottingham
Nottingham NG7 2RD, UK

Abstract

This study reflects a recent shift towards the study of early stages of expert memory acquisition for chess positions. Over the course of fifteen sessions, two subjects who knew virtually nothing about the game of chess were trained to memorise positions. Increase in recall performance and chunk size was captured by power functions, confirming predictions made by the template theory (Gobet & Simon, 1996, 1998, 2000). The human data was compared to that of a computer simulation run on CHREST (Chunk Hierarchy and REtrieval STructures), an implementation of the template theory. The model accounts for the pattern of results in the human data, although it underestimates the size of the largest chunks and the rate of learning. Evidence for the presence of templates in human subjects was found.

Introduction

There has been widespread research into experts' remarkable memory for domain-specific material. Much interest stems from how experts apparently overcome normal cognitive limitations, such as limits in short-term memory (STM). Research has covered a wide variety of areas, including games, music, academic domains, mnemonics, and sports. In developing theories of expertise, focus has been almost entirely centred on comparing high performers, such as Grandmasters in chess, with intermediate individuals and novices. Relatively little is known, however, about the details of the very early stages of learning in complex domains. This study aims at helping to bridge this gap, both by collecting new empirical data and by carrying out computer simulations.

Chase and Simon's (1973) Chunking Theory

In studying strong and weak chessplayers in a problem-solving situation, De Groot (1965) found that there was no real difference in type of heuristics used, depth of search, or number of positions searched. However, in a recall task for briefly-exposed positions, he found a clear difference in performance. Masters and Grandmasters achieved near perfect recall, while performance dropped off dramatically below Master level. De Groot concluded that expertise is not dependant on superior information-processing skills, but on the acquisition, over years of dedicated practice, of a large amount of domain-specific information, which can be rapidly accessed during problem solving.

Chase and Simon (1973) gathered further experimental data and developed an influential theory of expertise, the chunking theory. A *chunk* is defined as long-term memory (LTM) information that has been grouped in some meaningful way, such that it is remembered as a single unit. Each chunk will only take up one slot in STM, in the form of a 'label' pointing to the chunk in LTM. Using Miller's (1956) estimate, Chase and Simon proposed that 7±2 chunks can be stored in STM (this estimated has later been revised to four for visual material; Zhang & Simon, 1985, Gobet & Simon, 2000). In chess, a chunk may consist of up to 4-5 pieces, which are related to each other in any number of different ways, such as colour and proximity. Therefore, while a novice may only be able to recall around 7 single-piece chunks, a master can recall around 7 multi-piece chunks, more than 30 pieces. Even though recall performance of *random* positions is a great equaliser between Masters and weaker players, the former still maintain a small but reliable advantage over the latter. The chunking theory accounts for this superiority in terms of the small patterns that will appear by chance in random positions (Gobet & Simon, 1998).

Chase and Simon's (1973) study included a copy task of positions in full view, as well as a recall task of briefly-presented positions. Glances at the board being copied, and latencies greater than 2 seconds between the placements of pieces during recall, were used to analyse the size and nature of chunks. They found that the size of chunk increased as a function of skill level. Additional support for the chunking theory was found in several studies where the concept of chunk was studied in detail (see Gobet & Simon, 1998, for review).

Aspects of the chunking theory were implemented in a computer program by Simon and Gilmartin (1973), who proposed that LTM is accessed via a discrimination net. Identification of a chunk in LTM results in a pointer to that chunk being placed in a limited-capacity STM. Expertise requires the acquisition of a large database of chunks, with the appropriate discrimination net.

From Chunks to Templates

Although the chunking theory has explained many of the phenomena discovered in expertise research (Gobet, 1998), a few problems were later uncovered. One of its assumptions was that insufficient time is available during the brief presentation time of a position for any LTM encoding. Therefore, recall depends only on labels in STM pointing to LTM chunks. However, several experiments using interfering tasks have shown that LTM encoding does in fact happen (e.g., Charness, 1976; Gobet & Simon, 1996).

The template theory (Gobet & Simon, 1996), which is in part implemented in CHREST (Chunk Hierarchy and

REtrieval STructures, Gobet & Simon, 1998, 2000), was proposed to account for these data, while keeping the strengths of the original chunking theory. The most important improvement over the chunking theory is the presence of templates, which are larger and more sophisticated forms of retrieval structure than chunks. Like traditional schemas in cognitive science, templates have a *core* that remains unchanged, and a set of *slots*, perhaps with default values, whose value can be rapidly altered. CHREST incorporates mechanisms explaining how chunks evolve into templates through extensive experience, using frequent but variable information to create slots. The rapid encoding leaves the information safe from interference in STM, and so the template theory overcomes the problems created by the interference studies.

A Shift to Early Learning

The importance and influence of the chunking theory is clearly evident in the literature, and certainly not limited to the domain of chess. However, the research to date has been almost entirely focussed on the higher skill levels, as it naturally should in the study of *expertise*. But surely, when studying the acquisition of a skill, the first few hours of learning can be equally informing on the mechanisms involved. An important shift towards the early stages of expertise came from Fisk and Lloyd (1988), who studied novices' acquisition of skilled visual search in a chess-like game. They found that learning followed a negatively accelerating learning curve, in which improvement was very rapid at first but quickly became much slower. They could not, of course, have seen this so clearly by studying later stages of skill acquisition alone. The presence of this learning curve, which has also been found in other domains (Rosenbloom & Newell, 1987), could provide an explanation of why so many more years of practice are needed to become a Master than to become a good amateur.

In a similar study, Ericsson and Harris (1990) trained a novice chess player to the point when she could recall briefly-presented game positions to the standard of a Master player. However, performance on random positions did not reach that of Masters'. Saariluoma and Laine (2001), extending Ericsson and Harris' (1990) study, had two novices learn a set of 500 positions over the space of a few months. The participants were tested intermittently with a brief (5 s) presentation task. They had to recall 10 game and 10 random positions in each testing session. The results showed a clear improvement in percentage correct, from about 15% to 40-50% for game positions. The learning curve also looked like a power function, as found by Fisk and Lloyd for skilled visual search, with the greatest recall percentage increase within the first 100-150 positions learned. In addition, a slight increase was seen in percentage correct for random positions.

Saariluoma and Laine (2001) compared their human data to two computer models. Their aim was to differentiate between two possible methods for constructing chunks, both emphasising the flat (as opposed to hierarchical) organisation of chunks in LTM. From their simulations, they concluded that frequency-based associative models fit human data better than those based on spatial proximity of pieces. However, Gobet (2001) shows that CHREST, which uses a proximity-based heuristic for chunk construction, accounts for Saariluoma and Laine's human data equally well as their frequency-based heuristic. CHREST also accounts for the subtle effect found for random positions, which none of Saariluoma and Laine's models could do.

Preview of the Experiment

The present study differs from Saariluoma and Laine's in three important ways. First, while their participants had some experience with chess prior to the experiment, our participants were selected on the criteria that they knew as close to nothing about chess as possible. Second, the diagnostic power of Saariluoma and Laine's results is weakened by the lack of indication about how well the participants had learned the positions during the training sessions. In the present study, the participants are tested after every position in the learning phase. This helps keep motivation going, and keeps tabs on when concentration may have faltered. Third, presentation and reconstruction of positions was done on the computer, which allows precise and detailed data collection. In particular, our apparatus records latencies in piece placement, which can be used to infer chunks (Gobet & Simon, 1998).

With regard to the computer simulation, the present study is fundamentally different to Saariluoma and Laine's. While these authors were interested in comparing general learning algorithms, the present study aims at exploring how well a computational model that had already been well validated with experts' data could account for novices' data.

Human Data

Method

Subjects

There were 2 subjects, CE and JD, both female Psychology Undergraduates at the University of Nottingham, who had never taken any interest in chess and didn't know the rules. They were paid £6 per session and were told that they would be paid a bonus of between £5 - £15 at the end, depending on performance.

Materials and Stimuli

Positions, taken from a large database of Masters' games, were presented on a Macintosh 2cx, and subjects used the mouse to reconstruct positions. The software was the same as that used by Gobet and Simon (1998), to whom the reader is referred for additional detail.

Each session started with a *training* phase and ended with a *testing* phase. During training, 20 positions were presented for 1.5 minutes each. All positions were after the 20th move of Black. Of the 20 positions, 12 were game positions selected randomly from the database. The remaining 8 were pairs of game positions selected from 4 specific types (or 'families') of positions, which were used to help induce the putative learning of templates.

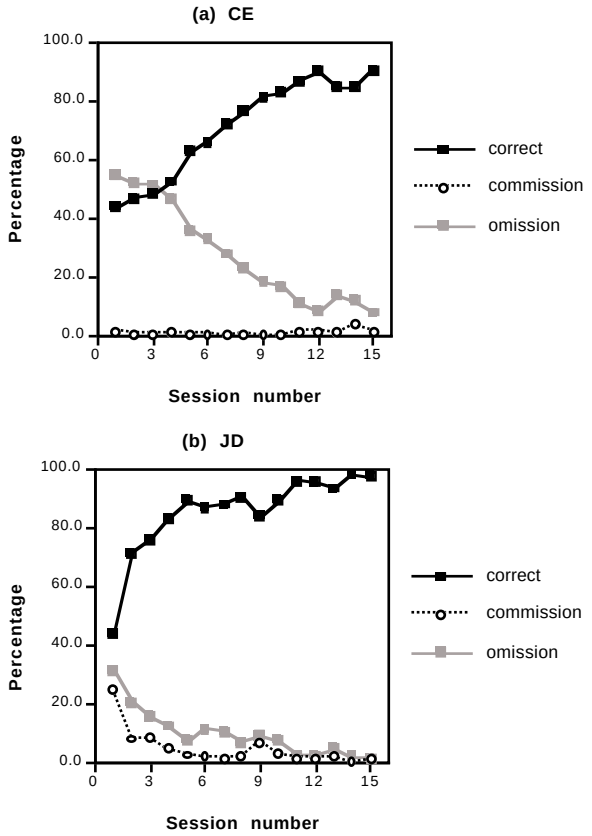

Figure 1. Training phase: Average percentage of correct placements, omissions and commissions for game positions against session number. Each position could be studied for 1.5 minutes.

Twenty positions were used in testing, with each presented for 5 seconds. Four were 'old' positions taken from the training phase, 2 of the game and 2 of the family positions. Four new game positions were selected randomly from the database. A new position was selected from each of the 4 family positions used in training. A position from each of 4 new families was also selected. The remaining 4 were random positions, created by shuffling the location of pieces from a game position. The order in which the positions were presented in both training and testing was randomised and different for each subject, as a control for any systematic effects of presentation order.

Procedure

At the start of both training and testing, the subjects were presented with an empty board on which they could familiarise, or re-familiarise, themselves with the placement/removal of pieces. This also gave them control over when the first position was to be presented, by clicking an "OK" button, as they did with each successive position after reconstruction. There was a pause between the training and testing phases for as long as the subjects wanted, which was never more than 5 minutes.

Table 1. Testing phase: Power functions ($y = ax^b$) computed for percentage correct against session number.

		CE			JD		
		a	*b*	r^2	a	*b*	r^2
Human data	Game	18.1	*.24*	**.86**‡	17.7	*.32*	**.95**‡
	Rand.	13.0	*.10*	**.16**	12.2	*.10*	**.13**
Model	Game	19.4	*.19*	**.71**‡	28.5	*.10*	**.66**‡
	Rand.	8.7	*.18*	**.35**	14.2	*.00*	**.00**

Note: ‡ $p < .001$

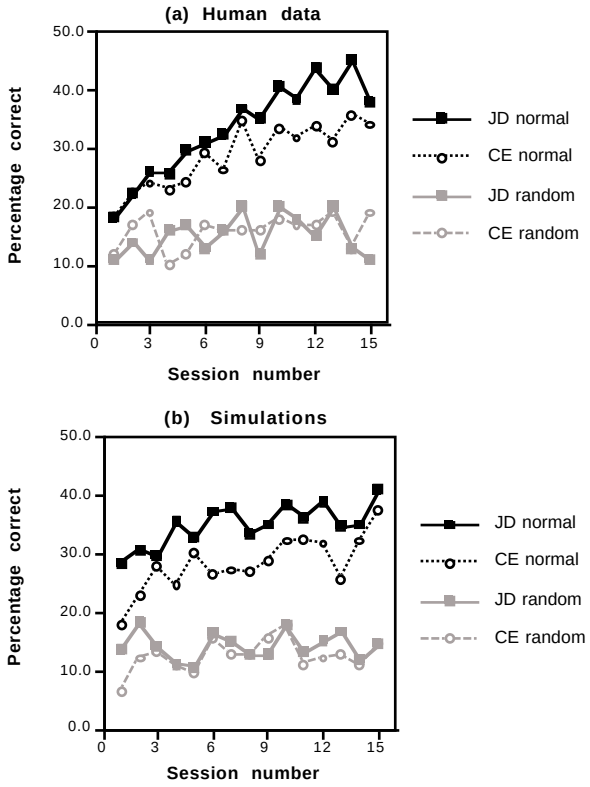

Figure 2. Testing phase: Average percentage correct for game and random positions as a function of session.

Results

To keep the data presentation short and highlight the contrast with random positions, we have grouped all the non-random positions into a single category called 'game' positions, both for the human data and the simulations, and both for training and testing.

The subjects varied greatly on the amount of time they spent on recall in both the training and testing phases, in

Table 2. Testing phase: Power functions ($y = ax^b$) computed for the size of the largest chunk against session.

		CE			JD		
		a	*b*	r^2	a	*b*	r^2
Human data	Game	4.2	*.24*	**.84‡**	5.6	*.18*	**.77‡**
	Rand.	2.5	*.23*	**.54***	4.4	*.02*	**.01**
Model	Game	3.5	*.28*	**.86‡**	5.0	*.11*	**.54***
	Rand.	2.4	*.18*	**.53***	4.0	*-.01*	**.01**

Note: * $p < .01$ ‡ $p < .001$

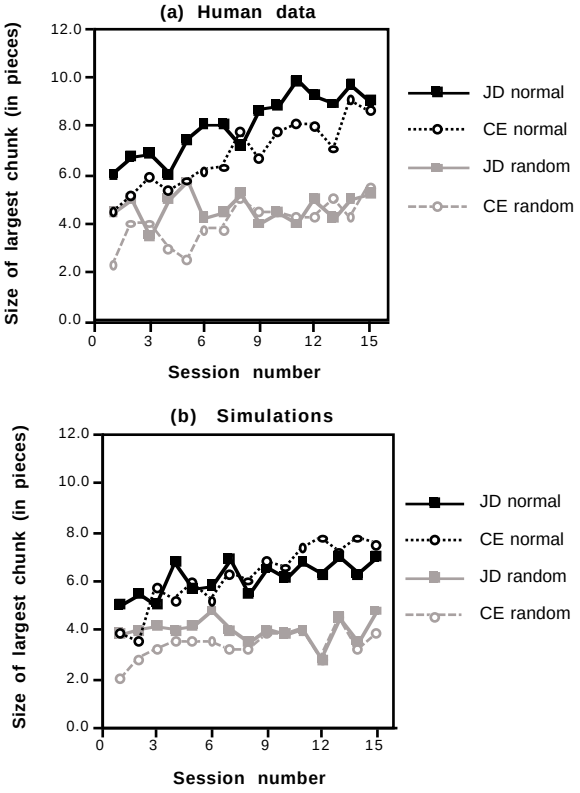

Figure 3. Testing phase: average largest chunk for game and random positions as a function of session number.

that JD spent consistently more time than CE. For example, during training, CE used consistently between 2500 s and 3000 s, while JD used between 3000s and 4000 s.

Performance with Training Positions

Figure 1 shows the relationship between the percentage of correctly placed pieces and that of errors of commission (pieces placed incorrectly) and omission (pieces not placed) in the training phase, over the fifteen sessions. For both participants, a power function accounts for the percentage correct well (for CE: $39 * N^{.31}$, $r^2 = .93$, $p < .001$, and for JD: $55 * N^{.23}$, $r^2 = .81$, $p < .001$, where N stands for the session number). Both subjects were averaging above 90% correct towards the end, and JD was often recalling full positions, as shown by the higher average.

CE made a small percentage of commission errors across the sessions, especially at the start, leaving omissions as an almost direct reflection of the percentage correct. JD made more commission errors in the first sessions than CE. In both cases, omissions were more frequent.

Performance with Testing Positions

Figure 2a shows the percentage of pieces correctly replaced during testing for the two subjects, for game and random positions. The graph suggests the presence of power functions for performance against session number with game positions, but not with random positions. This impression is confirmed in Table 1. Figure 2a and Table 1 also show that there was a slight improvement for the recall of random positions, although this does not result in statistically significant power functions.

Size of Chunks

The average median size of chunk for each type of position was extracted from the data (counting single-piece placements as chunks). Pieces placed in succession with a latency of less than 2 seconds were classified as belonging to the same chunk (Chase & Simon, 1973; Gobet & Simon, 1998). Analysis showed that the average median size of chunk increased as a function of session number, often with a significant power function. However, the nature of the data, which was often highly skewed due to many single-piece chunks (particularly by JD), meant that the median was a particularly noisy and volatile measure of the average chunk size.

The average size of the largest chunk within a position was a more robust measurement of the increase in chunk size. The size of the largest chunk increased reliably during training, both for CE ($7.42 * N^{.21}$, $r^2 = .77$, $p < .001$) and JD ($7.09 * N^{.22}$, $r^2 = .97$, $p < .001$). Towards the end, the average largest chunk contained up to 10 pieces. A similar increase was observed for testing (Figure 3).

Table 2 shows the results of power functions computed on the average size of largest chunk against session number, for testing. Significant power functions are displayed across all types of position, except random for JD.

Number of Chunks Recalled

Two measures of the number of chunks recalled were used. One measure counted the average number of all chunks, including single-piece chunks. The other measure only took the average number of chunks that contained at least 2 or more pieces. The first measure tended to overestimate the average number of chunks, as placements of single pieces may indicate guessing. The second measure tended to under-estimate the number of chunks recalled, as single-piece placements are just as likely as not to be valid chunks, particularly for beginners. Due to the large number of single chunks recalled by JD, as compared to

CE (probably as a result of guesswork), we only discuss the total number of chunks *excluding* single pieces.

During training, the two subjects showed an opposite pattern: CE recalled increasingly more chunks as a function of session number, while JD recalled increasingly fewer. By the end, however, both were recalling around 3 chunks. Given the long presentation time, however, the number of chunks reflects mostly subjects' strategies and their readiness to guess.

Testing, where positions are presented for only 5 seconds, offers a better way of measuring STM capacity, as there is little time to encode information into LTM. As predicted by the template theory, the number of chunks recalled during testing by CE and JD was consistently 3 or less, even when counting single pieces as chunks. Only occasionally did either subject exceed 3 chunks.

Computer Simulations

Methods

We essentially used the same version of CHREST as that described in Gobet and Simon (2000), to which the reader is referred for additional detail. There were two differences. The first was that eye-movement heuristics based on attack and defence were disabled. The second relates to how much CHREST knows about piece location. In previous simulations, it was assumed that players could readily encode the piece and its location as a single chunk (e.g., "Pawn on g4"). Here, given that we dealt with absolute beginners, we assumed that they had to construct such chunks. They would first learn a chunk for "Pawn", then one for the column "g", and finally one for the row "4". They would then combine these bits of information by chunking to learn the chunk "Pawn on g4".

For the simulations, CHREST was run twice, using the same order of positions and the same time per training session as each subject (see De Groot & Gobet, 1996, and Gobet & Simon, 2000, for details about the time parameters used in CHREST).

Results

We focus on the results from the testing phase, analysing in turn the percentage correct, the size and number of chunks, and the correspondence between humans' chunks and CHREST's templates. After 15 sessions, the model had acquired 4,772 chunks and 107 templates for the simulation of CE, and 5,811 chunks and 111 templates for the simulation of JD.

Percentage Correct

Figure 2b shows the results for the average percentage correct for the simulations, and Table 1 gives the power-function analysis. The overall percentage correct and the fluctuations of performance from session to session are reasonably similar to that of CE and JD, both for game and random positions. With both CHREST and the human subjects, power functions account for the recall of game positions, but not of random positions. However, learning was slower in CHREST in the simulation of JD, although, in this case, performance after one training session started at a higher percentage correct than with JD. The correlation between model and human data is .80 for the estimated *a*, and .22 for the estimated *b*.

Size and Number of Chunks

Figure 3b shows the results for the size of the largest chunk for the model, and Table 2 gives the results of the power functions used to fit the data. The correspondence between human data and model is good for CE, and a bit less for JD. Interestingly, the model captures the differences in parameters between CE and JD—which is due either to the order of positions or to the difference on time spent on task by the two subjects. However, although the absolute values are not far off, the model underestimates the size of the larger chunks, in particular with JD. The correlation between model and human data is .99 for the estimated *a*, and .93 for the estimated *b*.

Given its limited-capacity STM and the relatively small number of templates it possesses, especially at the early stages of learning, CHREST predicts that the number of chunks should not exceed three. As we have seen, this prediction was beautifully borne out by the data.

Templates

The templates that were formed by CHREST in the course of the simulation were compared to the human data. A search was carried out to match these templates to the groups of pieces recalled by the subjects, as defined by latencies. Only groups containing 4 or more pieces were included in the search, and a match was made only when at least 4 pieces were the same in both the template and group. The types of position searched were those belonging to the 'position families' used during the training phase (see section 'Material and Stimuli').

For training, out of the total number of chunks recalled by the subjects containing 4 or more pieces, about half were explained by CHREST's templates. During testing, this drops to nearer 30% of the total number of chunks for both subjects. Of the pieces accounted for by the templates, most of these were in the core. This is particularly true of CE, who for the majority of templates, placed no pieces predicted to be in the slots. JD, however, placed about 25% of the template pieces in the slots.

Discussion

The testing phase showed an increase in recall performance across sessions, an increase that is captured by power functions for game positions. This replicates the negatively accelerating learning curve found by Ericsson and Harris (1990), Fisk and Lloyd (1988) and Saariluoma and Laine (2001), and confirms predictions made by the chunking and template theories. The slight, but not significant, improvement for random positions is also predicted by the theory.

Another phenomenon predicted by both theories is the stability in chunk number across testing sessions; the number of chunks recalled was consistently 3 or below, for both subjects. Because of the limits of STM, only a certain number of chunks can be stored for recall of

briefly-presented positions. Even if a template is used in the later stages of the experiment, its contents would be output as a single sequence of rapid placements, which would not inflate the number of STM chunks.

The substantial size of the largest chunk was predicted by the template theory, but not by the original chunking theory. Towards the end of the testing sessions, the subjects were recalling chunks containing an average of 10 pieces, meaning that some individual chunks were much larger still. The chunking theory assumed chunks containing at most 4-5 pieces.

Overall, the simulations accounted for the human data reasonably well, especially when one considers the fact that no parameter of the model was varied to improve the fit of the simulation. The differences were that learning was somewhat slower and chunks smaller than with the human data. Power functions captured the same dependent variables in the humans and in the simulations, although the estimated parameters differed somewhat. Templates formed by the simulation matched a substantial number of piece groupings by the subjects. The matching method was rather simple, however, and better methods should be developed to assess the presence of templates.

The exact method of chunk construction could be an underlying factor towards explaining some of the differences between humans and the model. As noted above, CHREST first learns the piece (e.g. 'P'=white pawn), then the location horizontally (e.g. 'Pg'), and finally the exact position (e.g. 'Pg4'). The subjects in the present study did not show any sign of using such notation, and subjects did not have to know the name of the piece (and often did not), or the exact location. The subjects were probably more likely to recognise shapes and patterns of pieces, like chains of pawns, which they both mentioned during the course of the experiment. However, the fact that about 50% of subjects' placements in training were explained by CHREST's templates suggest that the two types of representation are not fundamentally different.

The data from training highlight marked differences between the two subjects, which are reflected to some degree in their performance during testing, and are worth some discussion. One difference is that JD made almost as many errors of commission as errors of omission near the beginning of the experiment. This suggests that she was guessing a lot more than CE, who made almost no errors of commission. Indeed, the number of chunks JD recalled decreased as a function of session (despite her increased performance), suggesting that she may have been guessing numerous, small chunks, possibly incorrectly. We speculate that the extra time spent on recall by JD is the result of time spent deliberating over whether she had recalled all that she knew. CE spent no such extra time before moving on to the next position, and so time spent simply increases as a function of the number of pieces being placed (hence the opposite trends between the two subjects). JD did appear to be especially highly motivated to perform to the best of her abilities (reflected in time spent and guesses).

In spite of these individual differences, the predictions of the template theory proved robust with regard to chunk size, STM capacity, and the shape of learning. As a first trial at comparing the simulation data to detailed human data for complete novices, the results are promising, and suggest that the same cognitive mechanisms operate with novices and experts.

Acknowledgements

This study was supported by the Economic and Social Research Council of the United Kingdom. We thank Daniel Freudenthal and Peter Lane for helpful comments.

References

Charness, N. (1976). Memory for chess positions: Resistance to interference. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 641-653.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

De Groot, A. D., Gobet, F. (1996). *Perception and memory in chess.* Assen: Van Gorcum.

Ericsson, K. A., & Harris, M. S. (1990). Expert chess memory without chess knowledge. A training study. *Poster presentation at the 31st Annual Meeting of the Psychonomics Society, New Orleans*.

Fisk, A. W., & Lloyd, S. J. (1988). The role of stimulus to rule consistency in learning rapid application of spatial rules. *Human Factors, 30*, 35-49.

Gobet, F. (1998). Expert memory: A comparison of four theories. *Cognition, 66*, 115-152.

Gobet, F. (2001). Chunk hierarchies and retrieval structures: Comments on Saariluoma and Laine. *Scandinavian Journal of Psychology, 42*, 149-157.

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

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

Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24*, 651-682.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production systems models of learning and development*. Cambridge, MA: The MIT Press.

Saariluoma, P., & Laine, T. (2001). Novice construction of chess memory. *Scandinavian Journal of Psychology, 42*, 137-147.

Simon, H. A., & Gilmartin, K. J. (1973). A simulation of memory for chess positions. *Cognitive Psychology, 5*, 29-46.

Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and acoustical loop hypothesis. *Memory and Cognition, 13*, 193-201.