Search, Recognition, and Visualization in Chess: Rebuttal to Gobet's Critique of Chabris & Hearst (2003)

Christopher F. Chabris
Department of Psychology
Harvard University
Cambridge, MA 02138 USA

Eliot S. Hearst
Department of Psychology
College of Social and Behavioral Sciences
University of Arizona
Tucson, AZ 85721 USA

18 October 2005

In an article published in *Cognitive Science,* Chabris and Hearst (2003; hereafter C&H) used computer analysis of errors in grandmaster games to measure the effects of additional thinking time and the ability to see the board on decision quality in expert-level chess. By comparing games played by the same players at rapid (R) and slower "classical" (C) time limits, we concluded that additional thinking time substantially increases the quality of moves played. By comparing rapid games to blindfold (B) games between the same opponents, we concluded that the benefit of being able to see the chessboard and pieces during the game was surprisingly small. We discussed the implications of these results for theoretical claims about the relative values of pattern recognition and forward search in chess expertise, especially a claim by Gobet and Simon (1996a; hereafter G&S) that pattern recognition is the more important process. Here we respond to comments made in an unpublished critique of our work by Gobet (2003).[1]

We have great respect for Gobet's large body of work on chess expertise, and for his own expertise as an international master of chess. However, we believe his critique suffers from several deficiencies. Besides being vague in an empirical sense, it merely repeats many of the points that C&H brought up and that others have said before, seems to be inconsistent with Gobet's previous statements by trying to make it appear that our results present no problems for a model/theory that strongly stresses pattern recognition over forward search, and contains a good number of false statements and non sequiturs.

The Relationship Between Thinking Time and Decision Quality

One of Gobet's main points (e.g., in his Conclusions section, p. 12) is that the "message" from G&S is the same as that of C&H: a reduction in thinking time leads to a loss in the quality of play. The disagreement between his conclusions and ours lies in how much of a loss would cause problems for his theory and models. As we noted in the second paragraph of C&H, it seems clear that G&S's conclusion was that chess skill in top players does not deteriorate much when

1. Some of these critical comments were repeated by Gobet et al. (2004, pp. 122–123).

thinking time is reduced, a reflection of previous views from Simon's lab that "recognition, by allowing knowledge to be accessed rapidly, allows the slower look-ahead search to be greatly abridged or dispensed with entirely without much loss in quality of play" (G&S, p. 53). *We think almost anyone reading G&S would conclude that their message was not that thinking time leads to a loss in quality of play, but that amount of time makes very little (or as the their abstract and text said, "slight") difference.* C&H's conclusion was that a reduction in thinking time makes a considerable difference—the message is not the same! We argued that the whole issue of pattern recognition vs. forward search is currently unresolvable, which seems clearly true since Gobet himself cannot tell us how much of a difference we would have to find to undermine his theory.[2]

G&S analyzed then-world champion Garry Kasparov's performance in simultaneous play to estimate how much (or how little) his ability declined when his thinking time was divided among multiple opponents. However, G&S did not evaluate the uncertainty associated with their conclusion because they did not recognize the variability in their estimate of Kasparov's strength under the time limits in these displays.[3] A reanalysis of G&S's data using maximum likelihood estimation (see Chabris, 1999; Glickman & Chabris, 1996) indicated that Kasparov did play worse than his ELO rating during that period, but there is not enough information to determine how much worse; he may have played at a rating level 200 or more points below his normal tournament strength, which would be quite a large difference.[4]

Leaving aside issues regarding the accuracy of the FIDE rating model (based on Elo, 1986), we would argue that a more precise rating estimate than Gobet's could be obtained by considering a range of opposition on both sides of, or at least closer to, the target player's own strength; the method we used, since it compares the behavior of multiple players who compete against one

---

2. Gobet continues to refer to the Calderwood et al. (1988) experiment in support of his claims about recognition vs. search (e.g., pp. 3, 6). C&H pointed out on p. 644 that this study not only used subjective judgments of move quality, but that these judgments could not even distinguish between the moves of strong players and much weaker players in classical, slow chess (ratings of 2.97 vs. 2.96 on a 1–5 scale), let alone between fast and slow chess in general. Gobet offers nothing to rebut this point, which undermines any contribution the Calderwood et al. study might make on this issue.

3. Gobet's footnote 1 (p. 4) mentions updated results for Kasparov in simultaneous play, but gives no indication that he has ceased to use the poor-quality Elo (1986) linear approximation for calculating performance ratings (see Glickman & Chabris, 1996). Nonetheless, the interquartile range he reports for Kasparov's performance ratings is fully consistent with our claim that he may have played as much as 200 points below his tournament rating under these conditions.

4. A rating difference of 200 points is significant because it predicts a 3–1 victory margin for the superior player in a match, no matter where along the absolute rating scale the two players fall (because the rating scale is logarithmic; Elo, 1986; cf. Glickman, 1995; for caveats, see Glickman & Jones, 1999). This would be considered a decisive result; for comparison, all recent world championship matches have been decided by much smaller margins. Thus, if Kasparov lost 200 points of strength under the clock-simultaneous conditions analyzed by G&S, it would be fair to conclude that the lost thinking time affected his play significantly. It is interesting to note that research in computer chess (e.g., Thompson, 1982; Condon & Thompson, 1983; Newborn, 1985; Hsu, Anantharaman, Campbell, & Nowatzyk, 1990; Hyatt & Newborn, 1997; for a review, see Heinz, 1998) has equated a 200-point rating advantage to the approximate benefit derived from searching one ply (one move for one side) deeper in the game tree, and that this additional search typically increases the time spent by a factor of 4–6. Thus, under the same time constraints, Kasparov suffers the same performance decrement as a typical chess-playing computer, which of course has much less knowledge and poorer pattern-recognition ability than he does, and derives most of its skill from efficient tree-searching. One can at least conclude from this similarity that Kasparov and other grandmasters derive a significant part of their skill from slow thinking processes as opposed to rapid perceptual processes.

another rather than the performance of one player versus a separate, weaker group, does not suffer from this limitation. A more intuitive argument against the reliability of the G&S technique is that a strong player such as Kasparov, who faces only weaker opponents (which, as the top-rated player, he always does), might play only as well as he needs to in order to win, and thus his performance level will not reflect his true or optimal level of ability. In our dataset, by contrast, players only face opponents of similar skill levels.

Let us examine how Gobet draws the curious conclusion that our differences are "also" relatively small. He discounts the 36.5% increase in number of blunders between C and R (C&H, Table 1) as not being substantial, and he totally disregards the doubling or tripling of the number of blunders that was observed when we applied a more stringent standard for blunders in comparing C and R with the 3-, 6-, and 9-pawn criteria (C&H, p. 643). Again, if Gobet is the sole judge of whether a C–R difference is big enough to contradict his favored theory, then the predictions of the theory have become unfalsifiable matters of opinion.

When he engages directly with our data, on p. 9, Gobet performs statistical acrobatics to try to further minimize the effects we observed. He does this by dividing the number of blunders in each condition by the total number of moves played in the games in that condition. For C games this is 176/35036, or 0.5023%, and for R games this is 266/38816, or 0.6853%. By subtracting these Gobet arrives at a 0.183% difference between C and R, which seems very tiny indeed compared to our 36.5%. But this reasoning is completely inappropriate, as an analogy will show. In one component of the Physicians' Health Study, 22,071 doctors were randomly assigned to receive aspirin or a placebo regularly to test the hypothesis that aspirin would reduce heart attack risk (Steering Committee, 1988). In each group, there were very few heart attacks during the study period, but the 104 in the aspirin group was about half of the 189 in the placebo group. If Gobet were the data analyst for this study, he would have concluded that the results were unremarkable, since 0.9% of the aspirin users had heart attacks and 1.7% of the placebo users did, a difference of only 0.8%. But in fact, despite an effect that appears tiny when viewed through Gobet's statistical lens, the result was highly significant ($p < .00001$), caused the study to be stopped early for ethical reasons, and led to a widespread change in medical practice.

Returning to the case of chess expertise, the greater number of total moves in the R and B conditions compared to the C condition is most likely due to the fact that players don't give up as easily in R and B games … because there is a greater chance that their opponent may still commit a blunder! That is, players intuitively know that blunders are more probable in R and B games, which is the very conclusion we drew in our paper.[5]

---

5. In his footnote 3 (p. 8), Gobet criticizes C&H's use of the χ^2 statistic to test the significance of differences in blunder frequencies, on the grounds that individual blunders and games are not necessarily independent of one another. In effect, this nonindependence would imply that we have fewer "true observations" than we have claimed. For now, we address this point as follows: Suppose we have so much nonindependence that our sample was really half as large (i.e., 88 blunders in C, 133 in R, and 138 in B conditions). The χ^2 statistics that were significant in our analysis are still significant ($p < .005$ in each case). Reducing our sample sizes by half again still results in significant differences. This suggests that the pattern observed—a large increase in blunders in the rapid and blindfold conditions compared to the classical condition—is not a statistically fragile result. Furthermore, and perhaps even more important, is that Gobet's strongly-stated point about blunders leading to other blunders, and losses leading to more losses, is speculative; he offers no data from grandmaster games to suggest that this is actually the case.

Gobet also discounts the difference of about a half-pawn in average blunder magnitude between R and C. Like Gobet, we are both chess masters, and we are sure that we would all like to see our opponents make mistakes that large. In any event, the old problem returns; how do we decide when a difference is big enough to embarrass or invalidate Gobet's theory/model? Without precise predictions calibrated to human performance, the judgment is subjective, which is one reason why the controversy over recognition vs. search seems currently unresolvable.

In questioning number of blunders as a reasonable measure of chess performance, Gobet (p. 8) states that games among top-level players are probably lost more by the accumulation of small errors than by blunders. Our own experience analyzing and studying top-level games suggests that many are decided by definite blunders, sometimes large ones. But this is not an irresolvable matter of our opinion versus Gobet's opinion—we have actual data available in Table 1 of C&H. In rapid and blindfold games, using our conservative 1.5-pawn criterion, we found at least 2 blunders per 3 games (R: 266/396; B: 277/396), and even in classical games we found nearly 1 blunder per 2 games (176/396 = 44%). Gobet's claim is not supported.

Gobet later states (p. 9) that we need to know the rate of blunders by weaker players to evaluate the findings we presented. While it might be of interest to look at this measure, we were interested in comparing grandmaster play at different time limits, so the performance of non-expert players is not relevant. Does anyone doubt that they would make considerably more errors than the players in the Monaco tourneys, at any speed, with or without sight of the board? It is ironic that, in comparing Kasparov's play against strong (but inferior) opposition in simultaneous displays, G&S did not see the need to explore how well weaker players would do in such conditions compared to Kasparov. There is a further practical problem of obtaining databases of matched games played among weaker players in the three conditions we compared; the natural experiment afforded by the elite annual Monaco tournament is unique.

On pp. 9–10 Gobet concludes by offering four reasons why it is "fruitful to talk about the dominance of pattern recognition over search."[6] However, we cannot see exactly how any of the vague approaches he suggests will provide us with clear-cut empirical answers to relevant questions. How will "providing additional knowledge" to human players be accomplished, and what sort of "knowledge" is Gobet referring to? We agree that strong players can often choose optimal moves very quickly—but what experiments will determine how often this is the case, and whether it accounts for the majority, or some other fraction, of these players' skills? Verbal protocols of the De Groot (1946) type "will provide powerful means of disentangling the contribution of pattern recognition and search" according to Gobet, but these techniques, powerful and valuable as they are, have been used for decades without yielding clear answers to this particular question.[7] Gobet's fourth proposed approach, computational modeling, is

6. On p. 3 Gobet writes that Holding (1985) "denied the importance of pattern recognition." While Holding strongly stressed search over recognition-and-association, he admits a role for recognition numerous times in his book. The thrust of C&H is that Gobet and Holding have each taken a fairly extreme position and that our data reveal that both recognition and search are likely to be quite important.

7. It is worth noting a further potential complication in using verbal protocols to "disentangle." De Groot (1946) and virtually everyone since has used preset middlegame positions that the subject has to survey before starting to verbally report his analysis. In real chess you create, or build up, knowledge about the position yourself, during the course of play leading up to any given situation. This means that behaviors such as eye movements, for example, may be different in real chess. In De Groot's protocols subjects first gave a general impression of the position, and

potentially powerful, but it is not clear how its predictions will be compared with empirical data such as ours.

Finally, we note that G&S (1996) said virtually nothing about computer models, but Gobet criticizes us for the same omission. Our test of the recognition vs. search issue followed the same logic as theirs, but with more reliable and objective measures, and involving individual games between the best players in the world in real tournaments rather than in simultaneous exhibitions given by one world champion against much weaker opponents. An empirical question about human behavior cannot be answered by citing computer models. Wherever possible, we tried to cite available empirical results, e.g., the small number of relevant experimental studies of blindfold chess.

Blindfold Chess

On p.11 ("Blindfold Chess") Gobet describes C&H's results as "surprising," but elsewhere he seems to say that aspects of his theoretical approach(es) have no problem predicting (or perhaps more accurately, postdicting) them. He criticizes our results by repeating our own caveat (C&H, p. 646 that players may play more carefully in B than in R conditions, but offers no empirical data to resolve the issue.

Gobet seems to be unfamiliar with the history and general literature on blindfold chess, a subject one of us has been researching for a forthcoming book (Hearst & Knott, 2005). In the course of criticizing C&H for referring to expert opinion, Gobet appeals (p. 11) to "other experts" who claim that playing blindfold chess is "useless, if not dangerous," citing Saariluoma (1995). On p. 77 of his book, Saariluoma does state that simultaneous blindfold exhibitions against many players are so taxing that they were forbidden by law in the Soviet Union and that there is anecdotal evidence that somebody actually died while attempting to beat the world record.

First, careful research reveals that such displays were not forbidden by law but only discouraged. Second, no one ever died while trying to beat the world record. Bourdonnais did die weeks after giving a blindfold display but for a few years he had been in terrible health from dropsy and strokes; despite his health he needed money to support himself and his family and kept on giving various kinds of public exhibitions, etc. Pillsbury, one of the greatest of all simultaneous blindfold players (his record was 22 at once), died in his early 30s and *The New York Times* (doubtless from family sources) said in his obituary that the cause of death was an "illness contracted through overexertion of his memory cells." He actually died of syphilis, as stated on his death certificate, a disease he probably did not catch while playing blindfold chess. Two of the most recent world-record-setters in simultaneous blindfold play, Koltanowski and Najdorf, died at the old ages of 96 and 87. Hearst and Knott (2005) generally find that supposed examples of serious health hazards from playing simultaneous blindfold chess are unfounded. *The important point, which Gobet apparently fails to realize, is that any comments on the dangers of blindfold chess refer to playing many games at once—not the case in the Monaco tournaments where the contestants played one game at a time.*

---

almost everything else they said involved concrete analysis of the position, i.e., forward search. This pattern may not match perfectly what chess masters do during actual games.

As to whether blindfold chess is "useless," we know of no such statement by anyone, except those who consider playing many games at once to be a stunt rather than "real chess" (see above). It is true that many trainers, authors, and players neglect blindfold practice, but great players such as Lasker and Reti recommended visualizing the board, squares, and pieces (even, for Reti, practicing with just two pieces hunting each other around the board) at early stages of learning chess (e.g., Lasker, 1932). Similar training procedures were a commonplace part of instructional techniques in the Soviet Union, and even the three Polgar sisters, now among the best woman players in the world (Judit Polgar, the youngest, competes at male world championship level), have commented on how learning blindfold play at the age of 5 or 6 helped them to develop certain chess skills. However, we do agree with Gobet that the causality goes both ways: blindfold practice may improve chess skill, and the ability to play blindfolded improves as general chess skill improves. There is hardly any master we know who cannot play at least one or two games blindfolded, even without any special training at that form of chess.

Gobet states that the template theory (Gobet & Simon, 1996b) "explains the experimental results on blindfold chess reasonably well" according to Campitelli and Gobet (2005). But this latter article includes no studies of decision-quality under sighted and blindfold conditions, so it cannot refute our own results. And neither would Gobet's recommended comparisons of eye movements between the conditions prove conclusive, insofar as blindfold chess can be played with one's eyes closed.

Conclusion

In his critique, Gobet makes some points that are relevant to the issue of whether, as G&S claimed, recognition processes "dominate" search processes in chess expertise. He goes overboard, however, in attempting to totally discredit the results of C&H. In this rebuttal, we have shown that G&S's conclusion is not supported by their own data, and we have explained why C&H came to the entirely appropriate and reasonable conclusion that it makes little sense to speak of one process "dominating" another given the empirical data about skilled real-world chess performance. We would like to close by reiterating our belief that Gobet is doing excellent work on an important topic, chess expertise, and by sincerely wishing him well in this endeavor. We will be among the first to congratulate him if his work does prove, despite C&H's pessimistic prediction, to be fruitful in resolving the recognition-search controversy.

Acknowledgements

We thank D.J. Benjamin and T. Busey for helpful discussions.

References

Calderwood, B., Klein, G.A., & Crandall, B.W. (1988). Time pressure, skill, and move quality in chess. *American Journal of Psychology, 101,* 481–493.

Campitelli, G., & Gobet, F. (2005). The mind's eye in blindfold chess. *European Journal of Cognitive Psychology, 17(1),* 23–45.

Chabris, C.F. (1999). Cognitive and neuropsychological mechanisms of expertise: Studies with chess masters. Doctoral dissertation, Department of Psychology, Harvard University.

Chabris, C.F., & 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,* 637–648.

Condon, J.H., & Thompson, K. (1983). Belle. In P.W. Frey (Ed.), *Chess skill in man and machine* (2nd ed.) (pp. 82–118). New York: Springer.

de Groot, A.D. (1946). *Het denken van de schaker.* [The thought of the chess player.] Amsterdam: North-Holland. (Updated translation published as *Thought and choice in chess*, Mouton, The Hague, 1965; corrected second edition published in 1978.)

Elo, A.E. (1986). *The rating of chessplayers, past and present* (2nd ed.). New York: Arco.

Glickman, M.E. (1995). Chess rating systems. *American Chess Journal, 3,* 59–102.

Glickman, M.E., & Chabris, C.F. (1996). Using chess ratings as data in psychological research. Unpublished manuscript.

Glickman, M.E., & Jones, A.C. (1999). Rating the chess rating system. *Chance, 12(2),* 21–28.

Gobet, F. (2003). Forward search, pattern recognition and visualization in expert chess: A reply to Chabris and Hearst (2003). Unpublished manuscript.

Gobet, F., de Voogt, A., & Retschitzki, J. (2004). *Moves in mind: The psychology of board games.* Hove, UK: Psychology Press.

Gobet, F., & Simon, H.A. (1996a). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychological Science, 7(1),* 52–55.

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

Hearst, E.S., & Knott, J. (2005). *Playing chess blindfolded: Its history, psychology, techniques, world records, and important games.* Book in preparation.

Heinz, E.A. (1998). DarkThought goes deep. *ICCA Journal, 21(4),* 228–244.

Holding, D.H. (1985). *The psychology of chess skill*. Hillsdale, NJ: Erlbaum.

Hsu, F-H., Anantharaman, T., Campbell, M., & Nowatzyk, A. (1990, October). A grandmaster chess machine. *Scientific American, 263,* 44–50.

Hyatt, R.M., & Newborn, M. (1997). Crafty goes deep. *ICCA Journal, 20(2),* 79–86.

Lasker, E. (1932). *Lasker's chess manual.* London: Printing-Craft.

Newborn, M. (1985). A hypothesis concerning the strength of chess programs. *ICCA Journal, 8(4),* 209–215.

Saariluoma, P. (1995). *Chess players' thinking: A cognitive psychological approach*. London: Routledge.

Steering Committee of the Physicians' Health Study Research Group (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. New England Journal of Medicine, 318(4), 262–264.

Thompson, K. (1982). Computer chess strength. In M.R.B. Clarke (Ed.), *Advances in computer chess 3* (pp. 55–56). Oxford: Pergamon.