TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 522–532

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12265

This article is part of the topic "Game XP: Action Games as Experimental Paradigms for Cognitive Science," Wayne D. Gray (Topic Editor). For a full listing of topic papers, see http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-2/issuetoc.

Allen Newell's Program of Research: The Video-Game Test

Fernand Gobet

Department of Psychological Sciences, University of Liverpool

Received 5 January 2017; accepted 2 February 2017

Abstract

Newell (1973) argued that progress in psychology was slow because research focused on experiments trying to answer binary questions, such as serial versus parallel processing. In addition, not enough attention was paid to the strategies used by participants, and there was a lack of theories implemented as computer models offering sufficient precision for being tested rigorously. He proposed a three-headed research program: to develop computational models able to carry out the task they aimed to explain; to study one complex task in detail, such as chess; and to build computational models that can account for multiple tasks. This article assesses the extent to which the papers in this issue advance Newell's program. While half of the papers devote much attention to strategies, several papers still average across them, a capital sin according to Newell. The three courses of action he proposed were not popular in these papers: Only two papers used computational models, with no model being both able to carry out the task and to account for human data; there was no systematic analysis of a specific video game; and no paper proposed a computational model accounting for human data in several tasks. It is concluded that, while they use sophisticated methods of analysis and discuss interesting results, overall these papers contribute only little to Newell's program of research. In this respect, they reflect the current state of psychology and cognitive science. This is a shame, as Newell's ideas might help address the current crisis of lack of replication and fraud in psychology.

Keywords: Chess; Cognitive architecture; Cognitive modeling; Expertise; Strategy; Unified theory of cognition; Video games

Correspondence should be sent to Fernand Gobet, Department of Psychological Sciences, University of Liverpool, Liverpool L69 7ZA, UK. E-mail: fernand.gobet@liverpool.ac.uk

1. Introduction

When topiCS's editor Wayne Gray invited me to evaluate the papers of the current issue through the lenses of Allen Newell's (1973) "Twenty-Question" paper, I accepted without hesitation. I have always had great admiration for Newell, and in particular for this paper, which I think should be required reading for every psychology and cognitive-science student. In this paper, Newell reviewed the contributions from the Eighth Annual Carnegie Symposium on Cognition, papers that he described as "a sample of the best work in current experimental psychology" (p. 283). While excited by their quality, he was also struck by the slowness of theoretical progress in psychology as a whole. He highlighted several concerns: the extreme focus on empirical phenomena, as opposed to theoretical developments; a lack of attention to the methods (strategies) used by participants; the binary nature of the questions asked in psychology (e.g., nature vs. nurture; massed vs. distributed practice; preattentive vs. attentive processing); the dearth of formal modeling and the reliance on informal theories too weak to allow clear-cut predictions; and the lack of theoretical integration of findings in psychology. To address these concerns, he proposed three research strategies: (a) the development of complete computational models, able to carry out the task of interest (criterion of sufficiency); (b) the systematic analysis of a complex task, such as chess, examining it from different angles (e.g., perception, memory, search, planning, etc.), in order to develop a "sufficient theory of a genuine slab of human behavior" (p. 303); and (c) the development of a computational theory that accounts for many phenomena across different tasks—that is, a general theory of cognition. Newell developed this last idea fully in his last book (Newell, 1990), where he called for the development of unified theories of cognition (UTCs) expressed as computer programs. At the time of his untimely death in 1992, Newell was engaged in developing Soar, an exemplar UTC. Soar had been applied to several domains. In some cases (e.g., two-choice response task or skill acquisition), Soar's behavior was closely compared to human behavior.

2. Newell's Twenty-Question paper today

When rereading Newell's chapter in the preparation of this commentary, I noticed that I had, during my career, implemented all three research strategies proposed by Newell. (I guess it is why I was asked to comment of the papers of the current issue.) First, I developed a computer model of perception and memory in chess (Chunk Hierarchy and REtrieval Structures; CHREST) (De Groot & Gobet, 1996; Gobet, 1993; Gobet & Simon, 2000; Gobet & Waters, 2003). This program can perform the task at hand (recall of a chess position during after the brief presentation of a position) and has closely replicated several phenomena, such as the pattern of eye movements during the short presentation of a position, recall performance as a function of the presentation time and the type of position, and the way pieces are grouped in chunks during recall. A variation of

CHREST, called SEARCH (Gobet, 1997) was devoted to look-ahead search, simulating variables such as the average depth of search and the number of moves generated per minute. By Newell's criterion of sufficiency, SEARCH is weaker theoretically than CHREST as it does not carry out the task to explain; that is, it does not play chess at a high level by finding good moves, but "merely" makes predictions about behavioral variables such as average depth of search.

Second, I have carried out systematic empirical research on chess, the very task proposed by Newell, studying it from the viewpoints of perception, mental imagery, problem solving, memory, development, learning, transfer, deliberate practice, talent, intuition, emotions, personality, gender differences, clinical aspects, and even handedness and month of birth (for summaries, see Gobet, 2016; Gobet & Charness, in press; Gobet, de Voogt, & Retschitzki, 2004; Gobet & Simon, 2001). Different methodologies were used, including behavioral experiments, eye-movement recording, concurrent and retrospective protocol analysis, analysis of databases, questionnaires, and brain imaging (fMRI and EEG).

Finally, CHREST slowly evolved from a model of chess perception and memory into a more general theory of cognition, and it has simulated data in a number of domains beyond chess and board games, such as problem solving in physics (Lane, Cheng, & Gobet, 2000), concept formation (Lane & Gobet, 2005), and the acquisition of vocabulary (Jones, Gobet, Freudenthal, Watson, & Pine, 2014) and syntactic structures (Freudenthal, Pine, & Gobet, 2009). While Soar (Newell, 1990) took problem solving as its starting point, and ACT-R (Anderson, Matessa, & Lebiere, 1997) did so with memory, CHREST's foundations lay in perception, and in particular perceptual chunking.[1]

When considering the current state of psychology, it is clear that, unfortunately, Newell's call has not been heeded. While some domains have been systematically studied (e.g., typing, reading, and language acquisition), the focus has typically been on understanding these specific domains rather than on understanding cognition as a whole. Computational models have been developed to understand phenomena in these fields and others, but their scope has been relatively limited. In addition, their impact has been small compared to informal theories. A striking example might be mentioned here. The April 1995 issue of *Psychological Review* contained two theoretical articles dealing with expertise: Ericsson and Kintsch's (1995) article on long-term working memory, and Richman, Staszewski, and Simon's (1995) article on EPAM IV's simulation of expert performance in the digit-span task. The former article presented an overall theory of expertise and memory, covering a wide range of phenomena but expressed informally and containing several inconsistencies, as for example noted by Gobet (2000a,b). The latter article described, in great detail and with good fit to human data, a computer program accounting for the development of superior memory in the digit-span task. The program was able to perform the task, and therefore satisfied Newell's criterion of sufficiency. There is no doubt that Ericsson and Kintsch's theory is an instance of the kind of theories that Newell criticized, and Richman et al.'s (1995) theory is of the kind he was calling for. Newell would have been disheartened to see that the former has been highly influential (1,216 citations in the Web of Science), while the latter had limited impact (75 citations).

Even more disappointing is the state of affairs with respect to the idea of UTC. Very few cognitive scientists, and even fewer psychologists, have taken up Newell's call, and many actually fully reject this idea. Why is it so? They find the idea too ambitious, are more interested in uncovering (sometimes trivial) phenomena rather than developing theories, and believe (incorrectly) that applications are more likely to come from empirical rather than theoretical research. In addition, there is often a more implicit dislike for the kind of mechanistic theory Newell used as an exemplar UTC, Soar. Finally, there is the pragmatic issue that it is easier to generate papers based on experiments rather than on computational theories, as developing such theories is a very time-consuming endeavor.

Interestingly, one important potential application of UTCs has been fully overlooked in the climate of lack of replication and even fraud that is currently marring psychology. While the proposed solutions focus on replication, meta-analysis, and statistical analysis for identifying abnormalities, to my knowledge nobody has proposed to use UTCs to address these issues. The idea is simple: If a result is inconsistent with the predictions of a UTC—taking some margin of error into account—then it is worth while to further scrutinize it. The strength of this approach is that the evaluation is not based on subjective plausibility, but on an architecture validated by a large number of empirical results in different domains of psychology. For example, Carter, Ferguson, and Hassin (2011) found that "a single exposure to the American flag shifts support toward Republicanism up to 8 months later," as summarized by their title, a result that could not be replicated by the *Many Labs* project (Klein et al., 2014). However, studies in cognitive psychology have consistently shown that priming, while a genuine phenomenon, is of short duration (typically no more than hundreds of milliseconds; e.g., Neely, 1977). With hindsight, several authors have argued that the flag effect was implausible, given that its effect was larger by several orders of magnitude than the implicit effects identified in cognitive experiments. It is my contention that an UTC would have made it possible to identify such an anomaly a priori.

Of course, it is possible that an effect judged implausible by an UTC is in fact genuine, and this could be established empirically. In this case, the UTC will have to be modified to account for this new result, while still accounting for the previous empirical data, assuming that they are robust enough. However, the idea is that an implausible effect requires more empirical validation than a plausible one. In a similar way, UTCs could be used to quantify to what extent an effect is surprising or counter-intuitive—again based on simulations covering a large number of phenomena rather than on personal intuitions and rhetorical devices, as is often the case currently.

3. Newell's Twenty-Question paper and the contributions of this issue

So, how does research into video games (as reflected by the papers in this current issue) fare when evaluated with the criteria set forth by Newell in his Twenty-Question

Table 1
Summary of methodology and applicability of key criteria discussed by Newell (1973), for articles published in this topic

		Methodology	Strategy Analysis	Task: Doing It All	Computer Modeling	UTC: Theory Integration
1	Boot et al. (2017)	Verbal protocols	Yes	No	No	No
2	Huang et al. (2017)	VG archival data	Yes	No	No	No
3	Reeves et al. (2017)	Ethnomethodology, conversation analysis	No	No	No	No
4	Schrodt et al. (2017)	Artificial intelligence	No	No	No	Yes
5	Sibert et al. (2017)	Machine learning	Yes	No	Yes	No
6	Stafford and Haasnoot (2017)	VG archival data	No	No	No	No
7	Thompson et al. (2017)	VG archival data	No	No	No	No
8	van der Maas and Nyamsuren (2017)	VG archival data	Yes	No	No	No

Note. VG, video game.

paper? Before answering this question, it is worth briefly discussing the methodologies used in these papers (see Table 1).

3.1. Methodologies used

Four of the eight papers (Huang, Yan, Cheung, Nagappan, & Zimmermann, 2017; Stafford & Haasnoot, 2017; Thompson, McColeman, Stepanova, & Blair, 2017; van der Maas & Nyamsuren, 2017) used video-game archival data. The analysis of large databases for studying cognition, which was not anticipated by Newell, magnifies some of the problems he identified (in particular, the peril of averaging across strategies and tasks), but also provides means to address them (e.g., using sophisticated statistical and data-mining techniques for identifying strategies). This is a topic with considerable opportunities for future research.

The four other papers use a variety of methodologies. Schrodt, Kneissler, Ehrenfeld, and Butz (2017) use artificial-intelligence methods to build an ambitious cognitive architecture that comes close to an UTC; however, they do not compare the behavior of their program with actual human data. Boot, Sumner, Towne, Rodriguez, and Ericsson (2017) use the time-honored method of concurrent and retrospective protocols, used for examples by Geyser (1909) and De Groot (1965) to study problem solving and expertise. Interestingly, in sharp contrast with big-data analysis, Boot et al. (2017) analyzed the behavior of a single participant. Sibert, Gray, and Lindstedt (2017) apply optimization techniques from machine learning to identify the best combination of feature weights for selecting actions in Tetris. Finally, Reeves, Greiffenhagen, and Laurier (2017) use ethnomethodology and conversation analysis to describe game playing from different viewpoints. While offering interesting insights, this methodology rejects the

concept of cognition used by Newell and is almost at the opposite pole from Newell's mechanistic approach.

The different methods cover different levels of analysis—from mouse clicks to verbal protocols to furniture and computer arrangements in Internet cafés. Interestingly, while Newell's discussion focused on the experimental method, this approach is not represented in the eight target papers. But just like the experimental papers discussed by Newell (1973), the papers in this issue tend to focus on data collection and analysis and not on theory development. In particular, with the big data approach, there is a danger of uncovering more disjointed phenomena and explaining them by binary concepts, thus leading to more information as opposed to more understanding.

3.2. Identifying strategies

Newell (1973) warned against analyzing data without paying great attention to methods, and in particular against averaging data across participants using different methods. This topic is popular in the target papers, as four deal directly with it (see Table 1). Huang et al. (2017) and van der Maas and Nyamsuren (2017) use archival data to detect and validate strategies. Sibert et al. (2017) use an optimization algorithm to find the best weights for board features given four different goals; this provides important information for understanding strategies used by players. Finally, Boot et al. (2017) use a softer, but no less effective approach: protocol analysis. In all these papers, the focus on strategies is a goal in itself rather than a means to understand fundamental principles of cognition, as advocated by Newell. Nevertheless, this attention to strategies is welcome as they are typically neglected in most experimental psychology, with the consequences that results are difficult to interpret because the reported statistics conflate behaviors produced by different strategies. This very problem of averaging data across strategies, a deadly sin according to Newell, remains with several of the analyses reported in this special issue. There is a need to develop new analytic methods for avoiding this infelicity in video game data analysis. I will take up this point in the discussion.

3.3. Computer modeling, complete analysis of a task, and computational theory accounting for several tasks

(I deal with these last three points together, as they are hardly addressed in the target papers.) The means of theorizing in the target papers would have certainly disappointed Newell. With the exception of Schrodt et al. (2017) and Sibert et al. (2017), all theories are expressed verbally, where "too much is left unspecified and unconstrained" (Newell, 1973, p. 301). Schrodt et al. (2017) use a formal model, and indeed a potentially general model of cognition, but do not compare its behavior directly to human data—the real test of a model of human cognition. Sibert et al.'s (2017) cross-entropy reinforcement learning models are compared to human behavior

but are limited in their scope (optimization of feature weights used). For example, they do not make assumptions about memory capacity or learning rates.

4. Discussion

In his Twenty-Question paper, Newell (1973) made two central points. First, in order to analyze human data meaningfully, one must understand the methods participants use and not average data across them, as this would lead to spurious patterns. Second, modeling is a sine qua non condition for understanding human cognition, and several phenomena should be accounted for by a single model within a complex task, and ideally between several tasks. The aim of the current paper is to evaluate the extent to which the contributions in this issue, and more generally research into video game playing, satisfy Newell's ambitious program of research, which sets the bar very high.

It must be acknowledged that the aim of the current paper is a bit unfair for the authors of the target papers, as they did not carry out their research with the explicit goal of testing Newell's ideas. In addition, some of Newell's ideas are difficult to demonstrate in a single paper, as they constitute a research program rather than a single study. At the same time, video game playing is a domain where one could expect to see these ideas implemented. Strategies play an essential and obvious role, and, given that they are played on computers, tablets, smartphones, and the like, video games lend themselves naturally to computer modeling. In addition, the time seems ripe for such undertaking: Currently, computers are more efficient by orders of magnitude than those used by Newell and colleagues; new methodologies for building computer models have been developed (Lane & Gobet, 2003, 2012); and new technological developments such as deep learning (Mnih et al., 2015) can be used as a perceptual front end and thus offer new opportunities for modeling.

We have seen that, while four out of the eight papers dealt with players' strategies, they did not do it to develop a more general theory of cognition, even a partial one. In general, all eight papers focused on specific aspects of video-game playing, and none of them were aimed at developing a general theory of cognition, the closest to an exception being Schrodt et al.'s (2017) paper, which, however, presented an artificial-intelligence architecture rather than a cognitive architecture. The difference is relative, however, and this architecture could be used to make specific predictions about human play, predictions that could be tested empirically.

It is fair to say that the relatively narrow focus of the target papers (with the exception of Reeves et al., 2017) is representative of video-game playing research more generally. There are few, if any, attempts in this field to develop general theories of cognition, and researchers typically aim to answer specific questions—of the binary type that Newell (1973) criticized. For example, there is an extensive literature dealing with the question as to whether the benefits of playing action video games transfer to other domains, with mixed results (Gobet et al., 2014; Green, Li, & Bavelier, 2009) and a no less substantial literature on whether this kind of activity leads to more violent behavior generally, again with conflicting evidence (APA Task Force on Violent Media, 2015; DeCamp, 2015).

Compared to chess, the example used by Newell, which is a single game, there are different types of video games, as illustrated in the papers of this issue. Differences can be important, as for instance between a real-time strategy game like *StarCraft* and a first-person shooter video game like *Halo: Reach*. From this point of view, it seems harder to develop a full model of video-game playing than chess—not that the latter is easy! Of course, one could limit oneself to study one game, say *Tetris*, although it is unclear whether enough data would be available, given that research resources are divided into many games. Another approach would be to develop a computational model accounting for human behavior in several games. This would actually come close to developing a general theory of cognition, given the large variability of games studied. Games that share some similarity would be useful for cross-validation purposes.

4.1. Big data analyzed individually

The importance of methods used by participants is highlighted repeatedly in Newell (1973). I believe that he was right to argue this point: Different participants might use different methods in the same task; the same participants might use different methods over time; and strategy use interacts with individual differences. Theories must be able to capture this level of flexibility and adaptation. Therefore, normal approaches to science are unlikely to be sufficient of understand human cognition. However, I also believe that Newell underestimated the difficulty of identifying methods, which is due to participants being able to modify their methods and even create new ones. This underestimation actually just gives more support for the need of using computer modeling to address the richness and variability of human behavior.

Several years ago, together with Frank E. Ritter, I proposed a methodology directly aimed at tackling the problem posed by strategies in human behavior (Gobet & Ritter, 2000). The starting point of this methodology, called individual data analysis (IDA), was Newell's (1973, 1990) idea of using the same cognitive architecture for simulating data in different tasks. The new idea was that these data should be analyzed individually. For example, a model would be developed for participant #14 and applied to simulate her performance in different experiments, such as a visual working memory task, an implicit learning task, and solving the tower of Hanoi. Modeling different experiments makes it possible to constrain the model's parameters, which include both numerical values (e.g., span of short-term memory) and, critically for the current discussion, strategies and their distribution. Only after parameters have been set for all participants are values averaged across participants. Thus, while nearly all research in psychology and cognitive science averages results across participants for each experiment, and uses averages for developing models, IDA uses all the information present in experimental data for setting the model's parameter values for each individual. Averaging across participants is done only at the end of the process, using theoretical parameters rather than empirical data.

As noted by several contributors, one advantage of studying video games is that they come with very high-density data, already coded in a way amenable to analysis, in particular when they are played over the Internet. Thus, they offer data that could be used for

large-scale, abundant IDA—AIDA. This approach takes the best of two worlds: the detailed and fine-grained analyses of single-subject designs, and the large amounts of data made possible by big data.

4.2. Conclusions

Just like in 1973 when Newell wrote his book chapter, psychology relishes in sophisticated methodology and produces exciting empirical results, as illustrated by the articles in this issue. However, authors have repeatedly expressed their disappointment of how little theoretical progress has been made in psychology in the last decades (Dar, 1987; Meehl, 1978; Miller, 2004), and there is no doubt that psychology as a whole, like the field of video game studies, falls short of Newell's exacting research standards.

It is of course possible that Newell's (1973) analysis was wrong and that psychology—then and now—is heading a correct course. However, I do not believe this to be the case, and the current crisis in psychology tends to support his analysis. Even brain imaging, which was widely seen as the way forward for psychology, has met with its share of problems, such as lack of replicability and poor data reliability (Eklund, Nichols, & Knutsson, 2016; Gobet, 2014; Uttal, 2012). The reasons for the slow progress in psychology are multiple, and some of them were anticipated by Newell. An unkind but fair conclusion would be that psychology, as a field, is not intellectually ready for a focus on theoretical developments and the hard fundamental questions that this raises. This article has taken a more positive perspective and pointed to some possible avenues of research, in line with Newell's analysis. In particular, for addressing his central point that the presence of different strategies limits the usefulness of "normal" scientific methods for studying the human mind, it has proposed AIDA, a methodology combining single-subject analysis with big data analysis, for the benefit of both.

Note

1. The term "chunking" has a different meaning in ACT-R, Soar, and CHREST. For a discussion, see Gobet, Lloyd-Kelly, and Lane (2016).

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*, 439–462.

APA Task Force on Violent Media (Producer). (2015). Technical report on the review of the violent video game literature. Available at: http://www.apa.org/pi/families/violent-media.aspx. Accessed December 26, 2016.

Boot, W. R., Sumner, A., Towne, T. J., Rodriguez, P., & Ericsson, K. A. (2017). Applying aspects of the expert performance approach to better understand the structure of skill and mechanisms of skill acquisition in video games. *Topics in Cognitive Science*, *9*(2), 413–436. doi:10.1111/tops.12230.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, *22*(8), 1011–1018.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist*, *42*(2), 145–151.

De Groot, A. D. (1965). *Thought and choice in chess (first Dutch edition in 1946)*. The Hague, The Netherlands: Mouton Publishers.

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

DeCamp, W. (2015). Impersonal agencies of communication: Comparing the effects of video games and other risk factors on violence. *Psychology of Popular Media Culture*, *4*, 296–304.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. doi:10.1073/pnas.1602413113

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

Freudenthal, D., Pine, J. M., & Gobet, F. (2009). Simulating the referential properties of Dutch, German and English root infinitives in MOSAIC. *Language Learning and Development*, *5*, 1–29.

Geyser, J. (1909). *Einführung in die Psychologie der Denvorgänge*. Paderborn: Schöning.

Gobet, F. (1993). *Les mémoires d'un joueur d'échecs [Chess players' memories]*. Fribourg: Editions universitaires.

Gobet, F. (1997). A pattern-recognition theory of search in expert problem solving. *Thinking and Reasoning*, *3*, 291–313.

Gobet, F. (2000a). Long-term working memory: A computational implementation for chess expertise. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the third international conference on cognitive modelling* (pp. 150–157). Veenendaal, The Netherlands: Universal Press.

Gobet, F. (2000b). Some shortcomings of long-term working memory. *British Journal of Psychology*, *91*, 551–570.

Gobet, F. (2014). William R. Uttal: Mind and brain: A critical appraisal of cognitive neuroscience. *Minds and Machines*, *24*(2), 221–226.

Gobet, F. (2016). *Understanding expertise: A multidisciplinary approach*. London: Palgrave.

Gobet, F., & Charness, N. (in press). Expertise in chess. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbook of expertise and expert performance* (2nd ed.). New York: Cambridge University Press.

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

Gobet, F., Johnston, S. J., Ferrufino, G., Jones, M. B., Johnston, M., Molyneux, A., & Weeden, L. (2014). 'No Level Up!': No effects of video game specialization and expertise on cognitive performance. [Original Research]. *Frontiers in Psychology*, 5:1337, pp. 1–9. doi:10.3389/fpsyg.2014.01337

Gobet, F., Lloyd-Kelly, M., & Lane, P. C. R. (2016). What's in a name? The multiple meanings of "chunk" and "chunking." *Frontiers in Psychology*, *7*, 102. doi:10.3389/fpsyg.2016.00102

Gobet, F., & Ritter, F. E. (2000). Individual data analysis and unified theories of cognition: A methodological proposal. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the third international conference on cognitive modelling* (pp. 150–157). Veenendaal, The Netherlands: Universal Press.

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

Gobet, F., & Simon, H. A. (2001). Human learning in game playing. In J. Fürnkranz & M. Kubat (Eds.), *Machines that learn to play games* (pp. 61–80). Huntington, NY: NOVA Science.

Gobet, F., & Waters, A. J. (2003). The role of constraints in expert memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *29*, 1082–1094.

Green, C. S., Li, R. J., & Bavelier, D. (2009). Perceptual learning during action video game playing. *Topics in Cognitive Science*, *2*(2), 202–216.

Huang, J., Yan, E., Cheung, G., Nagappan, N., & Zimmermann, T. (2017). Master maker: Understanding gaming skill through practice and habit from gameplay behavior. *Topics in Cognitive Science*. doi:10.1111/tops.12251. [Epub ahead of print].

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. *Developmental Science*, *17*(2), 298–310. doi:10.1111/desc.12111

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Lane, P. C. R., Cheng, P. C. H., & Gobet, F. (2000). CHREST+: Investigating how humans learn to solve problems using diagrams. *AISB Quarterly*, *103*, 24–30.

Lane, P. C. R., & Gobet, F. (2003). Developing reproducible and comprehensible computational models. *Artificial Intelligence*, *144*, 251–263.

Lane, P. C. R., & Gobet, F. (2005). Discovering predictive variables when evolving cognitive models. In S. Singh, M. Singh, C. Apte, & P. Perner (Eds.), *Pattern recognition and data mining, pt 1* (pp. 108–117). New York: Springer.

Lane, P. C. R., & Gobet, F. (2012). A theory-driven testing methodology for developing scientific software. *Journal of Experimental and Theoretical Artificial Intelligence*, *24*(4), 421–456.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Miller, G. A. (2004). Another quasi-30 years of slow progress. *Applied and Preventive Psychology*, *11*(1), 61–64. doi:10.1016/j.appsy.2004.02.010

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi:10.1038/nature14236

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory—Roles of inhibition-less spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Reeves, S., Greiffenhagen, C., & Laurier, E. (2017). Video gaming as practical accomplishment: Ethnomethodology, conversation analysis, and play. *Topics in Cognitive Science*, *9*, 308–342.

Richman, H. B., Staszewski, J. J., & Simon, H. A. (1995). Simulation of expert memory with EPAM IV. *Psychological Review*, *102*, 305–330.

Schrodt, F., Kneissler, J., Ehrenfeld, S., & Butz, M. V. (2017). Mario becomes cognitive. *Topics in Cognitive Science*. doi:10.1111/tops.12252. [Epub ahead of print].

Sibert, C., Gray, W. D., & Lindstedt, J. K. (2017). Interrogating feature learning models to discover insights into the development of human expertise in a real-time, dynamic decision-making task. *Topics in Cognitive Science*. doi:10.1111/tops.12225. [Epub ahead of print].

Stafford, T., & Haasnoot, E. (2017). Testing sleep consolidation in skill learning: A field study using an online game. *Topics in Cognitive Science*. doi:10.1111/tops.12232. [Epub ahead of print].

Thompson, J. J., McColeman, C. M., Stepanova, E. R., & Blair, M. R. (2017). Using video game telemetry data to research motor chunking, action latencies, and complex cognitive-motor skill learning. *Topics in Cognitive Science*, *9*, 467–484.

Uttal, W. R. (2012). *Reliability of neuroscience data: A meta-meta-analysis*. Cambridge, MA: MIT Press.

van der Maas, H. L. J., & Nyamsuren, E. (2017). Cognitive analysis of educational games: The number game. *Topics in Cognitive Science*, *9*, 395–412.