TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 485–496
Copyright © 2016 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12232

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.

Testing Sleep Consolidation in Skill Learning: A Field Study Using an Online Game

Tom Stafford,[a] Erwin Haasnoot[b]

[a]*Department of Psychology, University of Sheffield*
[b]*Department of Electrical Engineering, Mathematics and Computer Science, University of Twente*

Received 6 July 2015; received in revised form 4 April 2016; accepted 11 July 2016

Abstract

Using an observational sample of players of a simple online game (n > 1.2 million), we are able to trace the development of skill in that game. Information on playing time, and player location, allows us to estimate time of day during which practice took place. We compare those whose breaks in practice probably contained a night's sleep and those whose breaks in practice probably did not contain a night's sleep. Our analysis confirms experimental evidence showing a benefit of spacing for skill learning, but it fails to find any additional benefit of sleeping during a break from practice. We discuss reasons why the well-established phenomenon of sleep consolidation might not manifest in an observational study of skill development. We put the spacing effect into the context of the other known influences on skill learning: improvement with practice, and individual differences in initial performance. Analysis of performance data from games allows experimental results to be demonstrated outside of the lab and for experimental phenomenon to be put in the context of the performance of the whole task.

Keywords: Consolidation; Skill acquisition; Practice; Sleep

1. Introduction

1.1. Consolidation

It is widely accepted that memories are consolidated after acquisition (McGaugh, 2000)—that is, the organization and strength of habits, associations, and skills can

*Correspondence should be sent to Tom Stafford, Department of Psychology, University of Sheffield, Western Bank, Sheffield, S10 2TP, United Kingdom. E-mail: t.stafford@sheffield.ac.uk improve in the gap between acquisition or practice and subsequent testing, even without active rehearsal. Sleep is thought to be intimately involved in this consolidation process. A first basic demonstration was by Jenkins and Dallenbach (1924), who showed that retention of memories of nonsense syllables (following Ebbinghaus, 1885) was less degraded after a delay which involved sleep rather than a delay of equivalent time which did not involve sleep. Subsequent results have even shown that, for motor skills, performance can improve after a delay involving sleep (e.g., Karni, Tanne, Rubenstein, Askenasy, & Sagi, 1994). More recently, well-controlled experiments have demonstrated that sleep conveys a crucial benefit, beyond mere disengagement from the task for a comparable delay, and controlling for the known effects of practice spacing (Cohen, Pascual-Leone, Press, & Robertson, 2005; Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002; Walker et al., 2003).

Although the most consistent evidence for memory consolidation concerns procedural memories (Stickgold, 2005; Walker & Stickgold, 2004; Walker & Stickgold, 2006), there are good reasons to suspect this is not a phenomenon restricted to motor skills (Ellenbogen, Hu, Payne, Titone, & Walker, 2007), with there being a complex interaction of sleep and wakefulness in consolidation and reconsolidation of memories across procedural and declarative domains (Walker, Brakefield, Hobson, & Stickgold, 2003). Other evidence suggests that sleep may provide greater benefit for the most difficult aspects of a skill (Kuriyama, Stickgold, & Walker, 2004).

1.2. Games

Whereas sleep consolidation has been rigorously demonstrated in experiments, it has been difficult to validate outside the lab. We approach this problem by using a large naturally occurring dataset (Goldstone & Lupyan, 2016) collected from people who play a simple game of skill online (Stafford & Dewar, 2014).

Previously Stafford and Dewar (2014) have shown that observational data from this game can be used to validate and extend the analysis of phenomenon previously established in the experimental literature on skill acquisition. They show how practice amount and practice spacing contribute to skill development.

Our interest here is to build on this analysis, using an estimate of the players' time-zones. The time-zone of a player, combined with the time of each play, allows us to calculate the local time of each play and so compare comparable practice histories which are likely to contain, or not to contain, a night's sleep. This allows us to interrogate our dataset for the existence of the phenomenon of sleep consolidation. Our study allows us to use a large sample to quantify the magnitude of the effect as it manifests among those who are intrinsically motivated to learn an arbitrary task. It also allows us to put the phenomenon within the context of other factors affecting skill development.

The analysis of data from games has particular advantages and disadvantages for the cognitive scientist. Unlike so many of our experimental tasks, games are played for their intrinsic enjoyment rather than out of obligation or for external reward (Baldassarre et al., 2014). This allows us to look at skill development in a context where motivation plays as large a part as ability. This supports an expectation of generalization to skill development outside the lab and avoids the normal confound of large variation in participant motivation (and the attendant high degree of satisficing which occurs within traditional experiments Maniaci & Rogge, 2014; Oppenheimer, Meyvis, & Davidenko, 2009). Data from games allow us to measure skill development as it occurs in a naturalistic setting, over the course of days and weeks, rather than the mere minutes of most typical lab experiments.

Games also present a skill development domain in which automated data collection at a large scale is plausible. Unlike other skill development domains—for example, spoken language, playing the violin, soccer—each action taken during a game is conducted through a computer and so may be easily and unobtrusively recorded.

Games involve complex task performance. Further, they contain many elements which exist to facilitate enjoyment of play, rather than being strictly relevant to the operations which a cognitive scientist may be interested in. Because of this, the use of games in cognitive science requires, and will benefit from, analysis of the whole task (as encouraged by Newell, 1973).

2. Data acquisition

We used anonymized-at-source data from “Axon,” an online game developed for the Wellcome Trust by Preloaded. The game can be played at http://axon.wellcomeapps.com/. The game involves guiding a neuron from connection to connection, through rapid mouse clicks on potential targets. A screenshot can be seen in Fig. 1 (see figure caption for description of game dynamics). Cognitively the game involves little strategic planning, instead testing rapid perceptual decision making and motor responding.

The analysis was approved by the University of Sheffield, Department of Psychology Ethics Sub-Committee, and carried out in accordance with the University and British Psychological Society (BPS) ethics guidelines. The data were collected incidentally and so did not require any change in the behavior of game players, nor impact on their experience. Individuals were identified by cookie stored in their browser. For our analysis, we have assumed a one-to-one mapping between machine and player. No identifying information on the players was collected and so the data were effectively anonymized at the point of collection. Location information was approximate, to the city-block level at maximum. For these reasons the institutional review board waived the need for written informed consent from the participants. For further details of the dataset, see Stafford and Dewar (2014).

The data were extracted from Google Analytics using a Python library written by Nick Mihailovski. In contrast to Stafford and Dewar (2014), we were able to extract data for the longer period of between March 2012 and February 2015. The original data and code for coding, filtering, and analyzing it are available at https://osf.io/fckq8/.

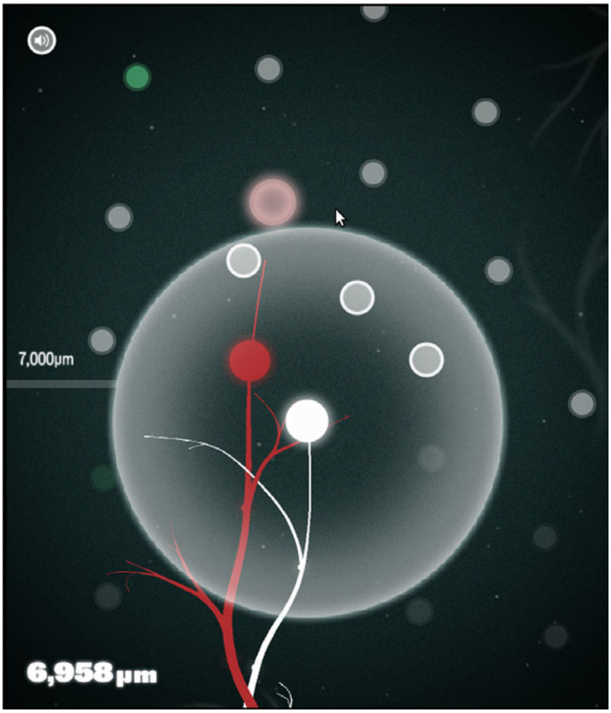

Fig. 1. Game screenshot. Players control the axonal branching of the white neuron. At each point, possible synaptic contacts (the other dots) are those within the zone of expansion (the larger transparent circle), which shrinks rapidly after each new contact is made. Non-player neurons (in red here) compete for these synaptic opportunities. Score is total branch length in micrometers (shown bottom left).

This data set comprised a total number of 1,201,515 players, the vast majority of whom played fewer than five times. The data and code for producing the analysis and plots presented here are also available from https://osf.io/fckq8/.

3. Analysis 1: Spacing and sleep consolidation

3.1. Aim

Our aim with this analysis was to compare subjects who took a break in their practice of the game, against those who played a comparable number of games without a break. This reproduces the analysis done in Stafford and Dewar (2014), which showed the benefits of practice spacing, and extends it to ask if activity during gaps in practice may influence subsequent performance. To do this, we wish to compare those for whom the timing suggests that they had probably slept between bouts of practice (e.g., someone who plays between 8 pm and 9 pm and then again between 8 am and 9 am) against those for whom the timing suggests that they probably did not sleep between bouts of practice, but nevertheless did take a comparable break (e.g., someone who plays between 8 am and 9 am and then again between 8 pm and 9 pm).

3.2. Filtering

First, we only analyze players who complete a minimum of 15 games, leaving 26,727 players. Additionally we filter the data for players on which we are unable to calculate valid longitude data or valid timing for their practice attempts. This leaves 26,291 players.

3.3. Coding

The local time for each play was calculated using the formula *localtime* = *UTCtime* + (*longitude* × 24 ÷ 360), *modulo* 24. This formula gives a local time which is correct in the majority of cases and almost always true within 2 h; the exceptions are due to irregularities in time-zone/national borders. Since our location information is approximate anyway, there is a limit to the possible level of accuracy regardless of the method of calculating the local time.

Next, we categorise players into four types, according to the nature of the timing of their first 15 attempts at the game. Players who play their first 15 games with a gap of less than 15 min between each game we categorise as “no gap” (9,388 players). Players who have a single gap of between 7 and 12 h are categorised as resting, either in the “sleep” or “wake” categories depending on the timing of the gap (761 and 423 players respectively). A break that finished between 5 am and 12 pm is categorized as a “sleep” gap (since gaps are 7–12 h, this means that the earliest rising player last played before 10 pm). A break that finished between 5 pm and 12 am is categorized as a “wake” gap. All other players are categorized as “no category” (15,719 players). This includes people who have medium length gaps, longer gaps, and multiple gaps.

4. Results

Results are shown in Fig. 2. We show the median scores, not means (inspection of score distribution showed that there were a small number of very high scores which made the results—although qualitatively the same—less consistent).

The 95% and 99% confidence bounds shown are calculated using a bootstrap analysis: scores from all categories resampled in sample sizes as large as the smaller category of the “no gap,” “sleep,” and “wake” categories (for 10,000 iterations). This gives an indication of how likely it is that samples of these sizes (or larger) would provide medians outside of the range predicted if the scores for players in these categories were all drawn from a common distribution. As can be seen, the “no gap” scores fall below the level predicted by the “no category” scores, and both the “sleep” and “wake” scores fall above.

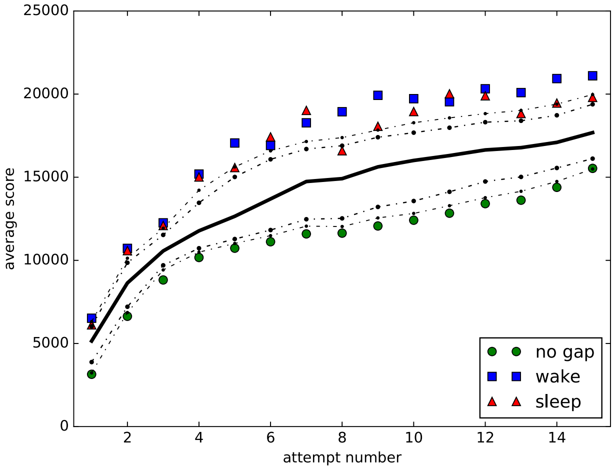

Fig. 2. Improvements in performance with practice for those who do not take breaks ("no gap") and those who have long breaks, either overnight ("sleep") or during the daytime ("wake"). Uncategorized players not shown. Black line shows median for all players and 95% (dashed line with large dots) and 98% (dashed line with small dots) confidence limits based on samples the size of the smallest of "no gap," "sleep," and "wake."

Subtracting the average score for "sleep" category players at each attempt from the corresponding score for "wake" players shows there is no advantage of the "sleep players" (indeed, the scores of the "wake" group are slightly, but significantly, higher; difference = 669.6, $t(14) = 2.81$, $p = .014$).

5. Analysis 2: Putting the effects into whole task context

5.1. *Aim*

Following Newell's (1973) injunction to study a whole task, we were interested to put the effect of spacing into the context of other effects which manifest in game performance. A disadvantage of observational data is that multiple different factors, both measured and unmeasured, simultaneously influence outcomes, but a corresponding advantage is that the data afford the chance to gauge the importance of different factors against each other. Hence, we ask, having established that the effect of spacing is statistically significance, if it is also a meaningful difference.

Secondarily, the quantity of data available makes it possible to analyze in more detail the functional "shape" of how various factors affect performance. In conventional experimental work we typically compare a small number of points, typically a control and experimental group, and analyze the contrast to reveal the effect of the manipulated factor. Here we can show how performance changes with many different levels of the factor. This "parametric analysis" shows more than just whether a factor has an influence on performance, but it has the potential to show something about how a factor influences performance.

One parametric analysis of the impact on performance that is already familiar is that of practice, specifically in the form of the learning curve. In this same domain, Stafford and Dewar (2014) showed that practice amount had the expected effect on performance of a relatively rapid initial increase which slowed down as practice amount increased (this can also be seen in the curves shown here in Fig. 2). That analysis also showed that early performance on the task was predictive of both rate of increase and asymptotic level of performance. We do not wish to commit on what constitutes these differences between players—probably it is influenced by a large variety of factors, including motivation, prior experience with online games, sensory-motor function, playing environment, and equipment as well as neuro-cognitive readiness for skill acquisition.

Here we compare three factors: practice spacing, practice amount, and initial performance for both the size and shape of influence over performance. We note that the comparison is inherently limited by the arbitrary bounds of the range over which the factors are analyzed. The effect of practice is bounded by the potential improvement in performance due to skill (and hence also by the range over which we assess practice). The effect of initial performance is bounded by the range within the population from whom data are gathered. The effect of spacing is bounded by the observed delay between some initial practice and subsequent attempts. Nonetheless, we believe it is instructive to see the comparison, and we wish also to highlight it as an example of the way larger data sets allow different analyses.

5.2. Filtering

As with Analysis 1, we remove all players who played fewer than 15 games, and those for which we could not calculate longitude or timing information.

5.3. Coding

First, to perform a categorical comparison with which to gauge the size of different effects, we split our data into high and low groups for each of the three factors we considered: spacing, practice, and initial performance.

To gauge the effect of spacing, we compared the average score on plays 11–15 for those who had no gap in their first 15 plays (i.e., the "no gap" group from Analysis 1, n = 9,388), with those who had a single gap of between 7 and 12 h (i.e., the "wake" and "sleep" groups from Analysis 1 combined, n = 1,184). To gauge the effect of practice, we compared the average score, over all players, on plays 1–5 and on plays 11–15. To gauge the effect of initial performance, we compared the average score on plays 11–15 of those who scored in the bottom third on plays 1–5 with the average score on plays 11–15 of those who scored in the top third on plays 1–5.

Second, we sought to make a "parametric" comparison of the effect of changes in these three factors. By this, we seek to show the way in which average scores change at each point along the range for which each factor can change. For practice amount, we calculated the average score, across all players, for each of the plays numbered 1–15. For initial performance, we calculated the average score on plays 1–5 for range from lowest to highest scorers (using 16 consecutive windows, covering the 100 percentiles). For spacing we calculated the average score on plays 11–15 according to the total gap time between plays 1 and 10 (using 16 consecutive windows, covering the range 0–60 min. The range was restricted to 0–60 min because average score does not change significantly for larger gaps). We used the median rather than the mean for all averages, since the score distribution contains a proportion of very high scores, which disproportionately skews mean scores.

5.4. Results

Fig. 3 shows the effect of the three factors when binary categorized. Fig. 4 shows the parametric comparison of the three factors. Note that there is no sense in which the range of the three factors may be compared absolutely. The initial performance line captures all the variation present in the population, the practice line captures the variation over the range of number of plays analyzed in this paper (1–15), while the spacing line shows a relatively short range compared to that used for the analysis shown in Figs. 2 and 3. This is because the spacing effect doesn't change significantly at cumulative gaps beyond 60 min.

The comparison of Figs. 3 and 4 illustrates that effects which appear to be of a comparable size from a "two point" analysis can be produced by underlying functions which have very different shapes. Practice affects performance with a decelerating function; initial performance has the opposite effect, such that the largest changes come at the high-

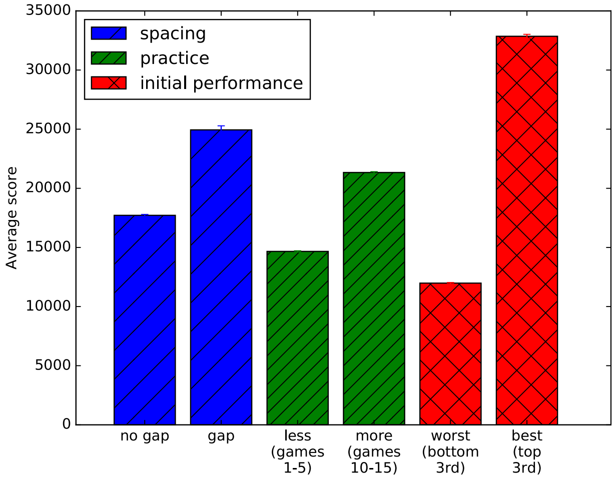

Fig. 3. Two-category comparison for the effects of spacing, practice, and initial performance. Standard error bars are shown.

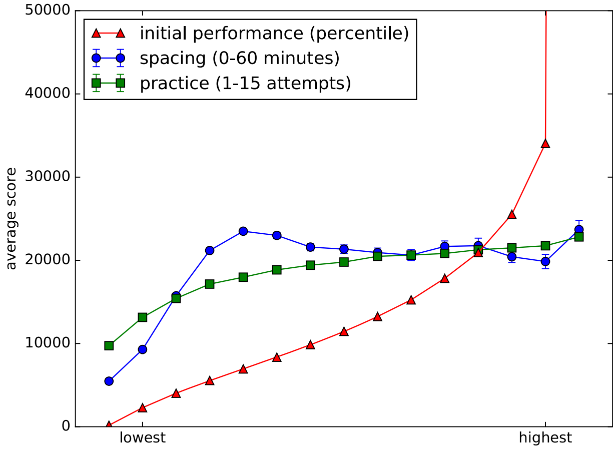

Fig. 4. Parametric comparison for the effects of least to most spacing, shortest to longest practice, and lowest to highest initial performance. Standard error bars shown for practice and spacing curves.

end of the distribution of that variable. The effect of spacing is a non-monotonic function, with an optimal point in the middle of the range (presumably reflecting a trade-off between the memory benefits of spacing-based consolidation and the memory costs of forgetting).

6. Discussion

These analyses show that there is a clear spacing effect. The psychological mechanisms by which this is produced may be assumed to be some combination of rest/recovery and active consolidation of memory. Analysis 1 suggests that, contrary to experimental results, breaks in training which contain sleep do not provide a superior benefit to equally long breaks which do not contain sleep. There could be many reasons for this. One possibility is that our task and/or analysis is insensitive to any additional effect of sleep consolidation. Although our large data set suggests this would not be due to a lack of statistical power, it might be that the nature of our task, or the ranges over which we conducted our analysis, fall outside the operating realm of the effect (in contrast to experimental results, which we might presume are carefully designed to capture the effect). If this is so, it is interesting to note that, whereas other learning phenomena such as practice or spacing effects do manifest, sleep consolidation does not here.

Other results suggest that the benefit of sleep consolidation is larger for more complex tasks (Ellenbogen et al., 2007; Kuriyama et al., 2004). It may be that our task was not complex enough for a sleep consolidation effect to manifest. Fig. 4 could be viewed as lending support to this idea—there is no additional benefit on performance of gaps longer than 15 min, with the spacing effect appearing as gap in practice lengthens from no gap to 15 min. This is a relatively short window compared to the size of many spacing effects

(Cepeda et al., 2009) and compared to the duration over which benefits of sleep consolidation are typically seen.

The lack of experimental control over players' behavior may be involved in the failure to observe sleep consolidation. Suppose that the phenomenon operates in concert with some other factor such as fatigue and amount of information needing consolidation.[1] Individual players may automatically calibrate their practice so that they are resting as and when they need to with respect to these factors, so that there is no additional benefit of sleep consolidation. In contrast, experimental studies dictate when participants practice and when they rest, which both controls for spacing effects and which may allow a benefit of sleep consolidation to be isolated.

It is striking that the benefit that comes from spaced practice is comparable to the benefit of players tripling their amount of practice (Fig. 3). Both of these effects are swamped by the range in aptitude for the game, as measured by initial performance (this importance of initial aptitude has been found elsewhere; Destefano, 2010; Huang, Yan, Cheung, Nagappan, & Zimmermann, in press; Stafford & Dewar, 2014). Two important caveats are, first, that although the amount and nature of our practice can be brought under an individual's control, it is less clear how initial performance can be controlled. This means that while differences in acquisition due to initial performance may be larger, it is not clear that they are more important for anyone wishing to infer how to improve rate of acquisition. Secondly, in this study we define aptitude entirely phenomenologically—that is, it is a simple effect read off from the data by dividing players according to their initial scores. Although this shows how players vary in the initial scores, it leaves completely unexplored *why* players vary. No doubt a constellation of factors contribute to initial ability, some of which are indeed mutable (for discussion of the contribution of initial ability to expertise development, see Detterman, 2014).

Games offer an opportunity to investigate learning in a naturalistic context, under conditions of intrinsic motivation, as well as bringing with them the advantages of easy collection of large data sets. We attempted to show here how one particular game can be used to study long established phenomenon. In particular we show ordered effects of practice amount, and a predicted effect of practice spacing, in a simple game. In contrast, the predicted benefit of rest periods that involved sleep was not observed. We also attempted to put these effects into mutual context, contrasting both the "size" of the effect—admittedly with arbitrarily defined ranges—and the parametric "shape" of the effects. In this way we hoped to show that the large amount of data available in the study of games does not just augment statistical power but makes possible new ways of analyzing behavioral data.

Acknowledgments

We are very grateful for the feedback and discussion provided by Wayne Gray, Walter Boot, and two anonymous reviewers.

Note

1. Although we note that Stafford and Dewer (2014, Fig. 4) provide evidence for a true consolidation effect in these data, and not just a “relief from fatigue” effect

References

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., & Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: An overview. *Frontiers in Psychology*, *5*.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236–246.

Cohen, D. A., Pascual-Leone, A., Press, D. Z., & Robertson, E. M. (2005). Off-line learning of motor skill memory: a double dissociation of goal and movement. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(50), 18237–18241.

Destefano, M. (2010). The mechanics of multitasking: The choreography of perception, action, and cognition over 7.05 orders of magnitude. Unpublished doctoral dissertation, Rensselaer Polytechnic Institute.

Detterman, D. K. (2014). Introduction to the intelligence special issue on the development of expertise: Is ability necessary? *Intelligence*, *45,* 1–5.

Ebbinghaus, H. (1885). *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Leipzig: Duncker & Humblot.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, *104*(18), 7723–7728.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*, 548–568. doi:10.1111/tops.12212

Huang, J., Yan, E., Cheung, G., Nagappan, N., & Zimmermann, T. (in press). Master maker: Understanding gaming skill through practice and habit from gameplay behavior. *Topics in Cognitive Science*.

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*, 605–612.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J., & Sagi, D. (1994). Dependence on rem sleep of overnight improvement of a perceptual skill. *Science*, *265*(5172), 679–682.

Kuriyama, K., Stickgold, R., & Walker, M. P. (2004). Sleepdependent learning and motor-skill complexity. *Learning & Memory*, *11*(6), 705–713.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

McGaugh, J. L. (2000). Memory–A century of consolidation. *Science*, *287*(5451), 248–251.

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.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*(2), 511–518.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*(7063), 1272–1278.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron*, *44*(1), 121–133.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory and plasticity. *Annual Review of Psychology*, *57*, 139–166.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: sleep-dependent motor skill learning. *Neuron*, *35*(1), 205–211.

Walker, M. P., Brakefield, T., Seidman, J., Morgan, A., Hobson, J. A., & Stickgold, R. (2003). Sleep and the time course of motor skill learning. *Learning & Memory*, *10*(4), 275–284.

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*(6958), 616–620.