Tracing the Trajectory of Skill Learning With a Very Large Sample of Online Game Players
Tom Stafford and Michael Dewar
Psychological Science published online 30 December 2013
DOI: 10.1177/0956797613511466

The online version of this article can be found at:
http://pss.sagepub.com/content/early/2013/12/30/0956797613511466

A more recent version of this article was published on - Feb 7, 2014

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Association for Psychological Science

Additional services and information for Psychological Science can be found at:

Email Alerts: http://pss.sagepub.com/cgi/alerts
Subscriptions: http://pss.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

Version of Record - Feb 7, 2014
>> OnlineFirst Version of Record - Dec 30, 2013
What is This?
Tracing the Trajectory of Skill Learning With a Very Large Sample of Online Game Players

Tom Stafford1 and Michael Dewar2
1Department of Psychology, University of Sheffield, and 2The New York Times Research & Development Lab, New York, New York

Abstract
In the present study, we analyzed data from a very large sample (N = 854,064) of players of an online game involving rapid perception, decision making, and motor responding. Use of game data allowed us to connect, for the first time, rich details of training history with measures of performance from participants engaged for a sustained amount of time in effortful practice. We showed that lawful relations exist between practice amount and subsequent performance, and between practice spacing and subsequent performance. Our methodology allowed an in situ confirmation of results long established in the experimental literature on skill acquisition. Additionally, we showed that greater initial variation in performance is linked to higher subsequent performance, a result we link to the exploration/exploitation trade-off from the computational framework of reinforcement learning. We discuss the benefits and opportunities of behavioral data sets with very large sample sizes and suggest that this approach could be particularly fecund for studies of skill acquisition.

Keywords
skill acquisition, learning, game, perceptual motor coordination

Received 5/14/13; Revision accepted 10/7/13

The investigation of skill learning suffers from a dilemma. One horn of the dilemma is this: Experts in real-world skills can be brought into the lab and their performance tested, but it is difficult to reliably recover comprehensive details of their training. This makes it impossible to be certain of exactly how features of the history of their practice are related to the skilled performance that can be observed. The other horn of the dilemma is this: Researchers can test different training regimes rigorously, but they are restricted to measuring performance on trivial or unnatural skills and often using performers without extended training of the order that experts in complex real-world skills engage in.

Computer games offer a partial resolution to this dilemma. Even simple computer games are not trivial in terms of the cognitive abilities they test. In fact, these abilities are often the staples of cognitive science: perception, decision making, and motor responses. Computer-game playing is a real-world skill in which many people choose to become expert, devoting hundreds of hours of practice. Unlike most skills, computer games allow a potential record of every action in the history of that practice—allowing, for the first time, detailed investigation of the connection between features of practice and level of final performance. In the current investigation, we took detailed records of practice activity from an online game and related the amount of practice and features of practice to levels of eventual performance. Using the large data sample from this game, we confirmed and quantified established findings from experimental studies of learning at unprecedented levels of confidence. In addition, we confirmed a recent result based on the theoretical framework of reinforcement learning (Stafford et al., 2012). Use of online games to collect very large samples offers a new method for the investigation of skill acquisition, we argue, and the work reported here showcases just some of the possibilities opened up by this approach.
Practice Amount and Spacing

We first consider two well-established results against which we validated our data set as a model of skill acquisition: the effects of practice amount and of practice spacing on performance. Studies of learning have shown a lawful relation between practice amount and performance. If performance is gauged in terms of some measure of efficiency (e.g., time taken to make cigars by experienced cigar manufacturers; Crossman, 1959), then it is possible to express the relation between practice extent and performance in a power law of learning (Newell & Rosenbloom, 1981; Ritter & Schooler, 2001). The exact nature of the mathematical law has been questioned (Heathcote, Brown, & Mewhort, 2000), but the fundamental observation remains that learning slows as it progresses, and the rate of performance increase displays a regularity that holds across very different domains (Rosenbaum, Carlson, & Gilmore, 2001). The power law of practice demonstrates that important regularities in learning exist across a wide range of domains and that such regularities can be uncovered by a suitably abstract level of analysis.

For practical reasons, studies of the effect of extensive practice have typically looked at different learners possessing differing amounts of practice rather than the same learners at different stages (i.e., such studies have usually had cross-sectional rather than longitudinal designs). Experimental studies of learning that have followed learners longitudinally have predominantly focused on lab-based tasks that can be mastered in one session or a small number of sessions (although there are, of course, notable exceptions, such as the work looking at the automatization of visual search performance; e.g., Czerwinski, Lightfoot, & Shiffrin, 1992; Neisser, Novick, & Lazar, 1963).

Highlighting the importance of practice quantity in skill development, Ericsson and colleagues stress that the highest levels of performance are never reached without skill development, Ericsson and colleagues stress that the highest levels of performance are never reached without skill development, Ericsson (1993). Addition-ally, they report that the nature of that practice matters—effortful, directed, deliberate practice is what distinguishes elite performers, even among those who appear to have performed similar quantities of practice.

Experimental studies of learning have focused on another factor that defines the nature of practice—spacing. The distributed-practice effect denotes the finding that if time devoted to practice is separated rather than massed, or if the spacing between practice sessions is larger rather than smaller, retention tends to improve (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010). The distributed-practice effect is surely one of the most solid findings in learning and memory research. It holds for both motor skill and declarative learning (Adams, 1987). Because of the limitations of experimental methods, there is a dearth of evidence on longer spacing intervals (Cepeda et al., 2006), a dearth that we hope the present study offers a method of addressing.

Next, we review an area in which the approach adopted in the present study affords particular traction—for looking at how the history of skill acquisition affects performance.

Exploration Versus Exploitation

The computational framework of reinforcement learning (Sutton & Barto, 1998) outlines a fundamental trade-off in decision making: Every decision forces people to choose between taking the action that they estimate will yield the best long-term consequence (highest value) or trying out an action of unknown or less certain value. This is known as the exploration/exploitation dilemma. Every choice is an opportunity to receive the outcome from only one action and therefore to update one’s estimate of the value of a single option. Too much exploitation leads an agent to rely on suboptimal actions and seldom discover better valued actions. Too much exploration, in contrast, leads to an agent wasting time exploring the space of actions without garnering the reward of frequently choosing the action with the highest known value. The implications for skill learning are that not maximizing performance during early practice may allow superior subsequent performance. Indeed, one might even expect that expert learners would adopt an early exploration strategy in order to maximize final performance.

We have already found evidence for such a strategy in humans and rats using an experimental task (Stafford et al., 2012). There is other evidence that variability in practice conditions can aid final performance (Roller, Cohen, Kimball, & Bloomberg, 2001), as well as generate benefits in learning that apply across tasks (Seidler, 2004). In the domain of motor control, cross-situational learning has been termed structural learning (Braun, Aertsen, Wolpert, & Mehring, 2009).

Method

Game designer Preloaded produced a game for the Wellcome Trust called Axon, which can be played at http://axon.wellcomeapps.com/. They inserted tracking code that recorded a machine identity each time the game was loaded and kept track of the score and the date and time of play. The game was played over 3.5 million times in the first few months of release (Batho, 2012).

The game involves guiding a neuron from connection to connection by rapidly clicking on potential targets. A
screenshot can be seen in Figure 1 (see the figure caption for a description of how the game is played). Cognitively, the game involves little strategic planning, testing rapid perceptual decision making and motor responses.

The analysis was approved by the Department of Psychology Ethics Sub-Committee at the University of Sheffield and carried out in accordance with the university’s and the British Psychological Society’s ethics guidelines. The data were collected incidentally and so did not require any change in the behavior of game players, nor did the method of data collection affect their game-playing experience. No information on the players, beyond their game scores, was collected, and so the data set was effectively anonymized at the point of collection. For these reasons, the institutional review board waived the need for written informed consent from the participants.

Because the data we recorded were indexed by machine identity, which is derived from the Web browser used to access the game, it was not possible to guarantee that a single individual was responsible for all the scores recorded against a single identity. Nor was it possible to guarantee that a single individual was responsible for only one set of scores. These uncertainties added noise to our analysis, but the data set was large enough to accommodate this. It is not clear what, if any, systematic distortions these caveats would introduce. For the remainder of this article, we will use the term player to refer to the set of scores associated with a single machine identity.

The data were extracted from Google Analytics using a Python library (Ecker, 2009). Data from between the 14th of March and 13th of May 2012 were downloaded and compiled into the source data set for the analyses presented here. This data set comprised a total of 854,064 players. Most played only a small number of times (the modal number of plays was 1), but some played up to 1,000 times. The data and code for producing the analysis and plots presented here are available at https://github.com/tomstafford/axongame.

Results

Practice amount

On average, scores were higher with each consecutive play. This pattern held for up to 100 plays, after which the drop-off in number of players reaching this point meant that a consistent pattern was less clear. Taking only players who played more than nine times \( n = 45,672 \), we calculated a high score for each player (i.e., the highest score they achieved, irrespective of which play it occurred on). The criterion of nine or more plays for subset selection is arbitrary, an attempt to balance the size of subset (which drops with a higher criterion) against the likelihood that practice effects will be reliable (which should be greater for higher criterion values). For this and all other analyses presented in this article, the results are not contingent on the particular values used to divide up the data (i.e., we got similar results if greater than 5, 8, 10, or 20 plays were used as the criterion. To confirm this, interested readers can run the analysis with altered parameters by visiting the data and analysis code repository referenced previously).

From this subset, players were then organized into five groups on the basis of the percentile ranking of their high score, and the average score was calculated for each attempt for all players in each percentile group. These calculations showed that the difference between higher and lower scorers was not merely the amount of practice. The difference in average score was present from the very first plays (Fig. 2).

Practice spacing

Taking only players who played more than nine times, we divided players into percentile groups according to their highest score, regardless of the play on which that
score was obtained. We also calculated the separation in time between their first and last play. The result shows a clear upward trend (Fig. 3, dots), with players who scored most highly spreading their first and last plays further apart. This is unsurprising, however, because even if there were no relation between practice and scoring, and scores were simply random on each attempt, players who had more attempts would tend to collect higher scores and have first and last attempts that were more separated in time. We used bootstrapping to estimate confidence intervals as if this were the case. Keeping the number of players and the number and time of the attempts constant, we generated 2,000 simulated data sets, sampling with replacement at random from the total record of all scores for all players. The observed data fell below this bootstrap data for low maximum-score percentiles and above for high maximum-score percentiles, which suggests that the scores really are distributed non-randomly and according to the spread in time of participants’ plays (Fig. 3). A one-sample t test was performed on the difference between the observed and expected values (recoded by reversing the sign of the differences for Percentiles 51 to 100, so that positive differences across the whole range reflect differences in favor of the spacing hypothesis). This test was highly significant, \( t(99) = 7.27, p < .0001 \), which confirms that the greater spacing was associated with higher than expected scores.

It is possible to examine this result further by slicing the data more finely. Taking only players who played more than 14 times (\( n = 21,575 \)), we calculated the spread in time between the 1st play (or 2nd play when these data were missing) and their 10th play (or 9th, when this number was missing). We also identified their best score on Plays 11 to 15. We then divided players into two groups, those who played their 1st 10 games within a 24-hr period (“goers”), and those who split their 1st 10 plays over more than 24 hr (“resters”). Resting between 1st and 10th plays appears to have a benefit on subsequent performance (“goers: \( M = 44,050, SD = 26,882 \); resters: \( M = 47,264, SD = 29,461 \)). The difference between the groups was highly significant, \( t(20354) = 6.219, p < .00001 \), albeit for a small effect size (Cohen’s \( d = 0.11 \)).

A third analysis revealed something of the difference in individual’s learning curves when they were categorized by spacing. We identified players with similar scores on their 1st play, who played their 1st through 6th games within a 2-hr window and their 15th to 20th games also within a 2-hr window. The motivation for this classification was to find players with similar habits who had comparable initial ability on the game. We then divided them...
into two groups: Those who had a 6-hr (or more) gap between their 6th play and their 15th play and those who did not. The result (Fig. 4) shows how our previous finding that more practice spacing is associated with higher performance reveals itself in the shape of the learning curves—the average learning curve for players of comparable ability diverges at the time that one group begins to space its practice.

Exploration versus exploitation

The variance of scores for each player in the first five plays was calculated, and this statistic for each player ranked to create percentile groups. The same was done for the average on Plays 6 to 10. Higher early variance was associated with higher subsequent performance (Pearson’s r = .59, p < .0001). Randomizing the scores for each attempt within the structure of the number of players and the number of attempts per players, it is possible to generate a bootstrap data set that gives a confidence interval for this correlation—in other words, answers the question “to what extent is a correlation between high early variance and high late scoring inherent in the distribution of scores and the structure of how players accumulate scores from that overall distribution.” These bootstrapped confidence intervals for correlation, at the 95% level, were 0.009 to −0.009. Thus, we concluded with a high degree of confidence that the observed correlation is both significantly different from zero and not a trivial consequence of the distribution of scores. Instead, the correlation results from the particular way that individual players’ early scores are related to their later scores.

Discussion

These results confirm, but also quantify, results from experimental psychology regarding the effects of practice quantity and quality on performance. As players practice, their average score improves. Dividing the players into percentile groups according to high scores appears to show that practice alone does not allow most players to achieve the highest scores. The best players have an advantage from the very first plays. This advantage is consolidated with practice, in that not only do they score more on their first plays, but also their rate of improvement is faster. This finding is in marked contrast to some popular (e.g., Gladwell, 2008) and academic (e.g., Ericsson et al., 1993) accounts of high performance that have denigrated the importance of talent with respect to practice.
We suppose that for each individual, there will be a range of sensory, motoric, cognitive, and experience-dependent factors that position him or her in terms of initial ability and potential to improve at the game (as with performance in any complex domain). These factors will include, for example, level of experience with computers and, specifically, level of experience with games of this general type. Our level of analysis asks, in effect, how do predisposing factors—of all kinds—play out in the shape of learning curves? The answer, which remains true regardless of the extent to which initial scores are influenced by prior experience, is that players with higher initial scores tend to progress faster. It is notable that the result that higher scorers got higher scores from the start and improved more quickly than players with lower initial scores held throughout the population of scorers (i.e., not just for the top 20%). Certainly there should be some advantage from factors such as previous game play, but the presence of the same basic effect across the distribution of players suggests that any such factors combine continuously with all the other factors that support high scoring.

The analysis of practice spacing confirms the wisdom from experimental studies of learning and memory that distributed practice is better than massed practice. It remains to be seen if there is an optimal amount of spacing, as has been reported for semantic knowledge (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008) or an optimal timing of spacing (Goedert & Miller, 2008). The exploration/exploitation result confirms a preliminary result from a recent experimental study (Stafford et al., 2012). Although bootstrapping confirms that this finding is not an incidental result of the distribution of scores, it still is not clear if the level of exploration (operationalized as score variance on early plays) per se causes the higher level of performance (“exploitation,” characterized as the average score on later plays). It is doubtful that lower scoring attempts in themselves cause higher subsequent performance. Rather, low scores may be the impetus for players to shift their playing style or tactics in ways that unlock higher subsequent performance (similar to the postulated freeing and freezing of degrees of freedom that have been thought to characterize changes in motor skill; Bernstein, 1967; Berthouze & Lungarella, 2004). The ultimate test of exactly if and how early exploration affects subsequent performance will be to intervene to make players explore and see how this affects later scores. In other domains, there have been suggestions that introducing guided mistakes or deliberate failure into early training may have benefits for overall performance (something for which there is some evidence; Lorenzet, Salas, & Tannenbaum, 2005).
Games

Games are a great opportunity for the psychological science of learning. They provide participants in high numbers who are engaged and willing to undertake extensive practice. Games can provide large amounts of detail on training conditions and actions in ways that other paradigms cannot. In the future, it may even be possible to introduce experimental manipulations into engaging games through partnership with games designers.

Big data

Because of the method of study adopted here, we lost experimental control over the factors involved in learning. However, advantages stemmed from the very large sample size we collected data from. Some of the emphasis on the importance of experimental control in cognitive science is due to the loss of statistical power than can result from uncontrolled measurement. With large sample sizes, loss of statistical power is not an issue. We need only concern ourselves with the ways in which lack of experimental control introduces systematic confounds into our data set. As well as providing high statistical power, very large sample sizes mean that data can be analyzed in new ways. One of these is “slicing,” by which we mean identifying individuals who meet certain conditions and comparing within that group. This is a substitute for the conventional experimental method of creating individuals that meet certain conditions in low numbers. In experimental design, you control potential confounds in advance (by attempting to remove them). With slicing, you attempt to account for potential confounds after the fact by selecting multiple different sub-data sets, each of which controls statistically for a potential confound—and, thus, by a process of elimination gathering support for your hypothesized causal variables. This is a less powerful method than experimental control, but it does offer some advantages. Another promising method of analysis is bootstrapping, which provides a way of testing observed patterns against sophisticated null hypotheses. Both bootstrapping and slicing are illustrated in our analysis of spacing effects.

Two modern crises of psychology are the apparent low replicability of effects (Pashler & Wagenmakers, 2012) and the use of inappropriate statistics (Meehl, 1978; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Very large sample sizes can help researchers sidestep both of these concerns. With a large enough sample size, you do not need to use inappropriate statistical techniques—small effects are easy to find. Furthermore, you have enough data to use techniques such as cross-validation to guard against false positives.

Analyzed in detail, very large data sets provide an observational playground in which scientists cannot just detect effects but compare the size of different effects against each other. For example, in the present data set, it can be seen that the benefit of spacing plays over more than 24 hr is about 3,000 points (Fig. 3). This is comparable to about five plays in the 10- to 15-play range (Fig. 2) or equivalent to an extra 50% practice at this stage of experience.

Obviously, nothing will replace controlled experimentation in terms of causal inference. For hypothesis testing, controlled experiments must remain the gold standard. However, there is space within the scope of investigation for studies with purposes other than theory-driven hypothesis testing (Rozin, 2009). This article has focused on characterizing the data and confirming effects discovered in traditional controlled experiments. We believe that the approach illustrated here can be complementary to experimental studies and has the potential to open up new avenues for investigation in the study of skill acquisition.

Author Contributions
T. Stafford conceived the study. T. Stafford and M. Dewar conducted the analysis. T. Stafford wrote the manuscript.

Acknowledgments
We thank Tony Barnes for the introduction to game designers Preloaded; Phil Stuart, Charles Batho, and Cameron Yule at Preloaded; the Wellcome Trust for allowing the data from its game to be passed to us; Stuart Wilson for help with Python; Ashvin Shah for discussion of reinforcement learning; and seven anonymous reviewers. Special thanks go to Edith Mary Cameron, whose late arrival and postbirth disposition allowed T. Stafford to carry out the bulk of the analysis and writing of this manuscript.

Declaration of Conflicting Interests
The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note
1. This pattern still holds if we use aggregate score rather than attempt number (as an index of total amount of practice, to account for higher scoring attempts conferring more experience), and if we calculate high score on a particular attempt number (e.g., on attempt 20) rather than across all attempts. We thank the reviewers for suggesting these control analyses.

References