PSCI-13-0838 Revision

Tom Stafford
Department of Psychology, University of Sheffield

29 July 2013

Dr Arkes

Attached is our revision of "Tracing the trajectory of skill learning with a very large sample of online game players" [PSCI-13-0838]. We have done a number of further analyses and re-analyses which we believe address the concerns of you and your reviewers. You will appreciate that the size of the sample meant that collecting more data was not possible within a short timespan. In keeping with our 'open science' position, we have updated the repository of the files which produce the analysis described in the paper and in this letter (referenced here for completeness only). They are available from `https://github.com/tomstafford/axongame` as before. References to figures in this cover letter are prefixed 'R' so as to distinguish them from references to figures in the original manuscript.

Your decision letter highlighted three main areas of concern. We address these first, before giving a point by point discussion of the issues raised by the reviewers.

Yours

Tom Stafford & Mike Dewar

1. Analysis of individual trajectories

further investigation of the learning curves of high and low scorers. Are their strategies different?

We have addressed this in two ways. Reviewer 2 made a specific request for more detailed analysis of learning curves with respect to the spacing analysis. We have done this (discussed below, and included an additional figure in our manuscript).

Secondly, we have tried to show how our previously presented analyses (and those done for this revision) reflect on individual learning curves. The data for individuals is very noisy, and so picking one individual at random from the sample is

not particularly informative. Rather, we have focussed on analysing average learning curves, so as to access the underlying patterns, and picking out individuals according to criteria so we can compare groups of individuals. All the analyses we present here reflect patterns that hold for individuals — the relationship with practice extent, spacing and variability.

2. Transfer effects from prior online game experience

If brief, we do not wish to refute this possibility, and acknowledge it as a factor which plays a role in producing the scores we analyse. 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 score is influenced by prior experience, is that those with higher initial scores tend to progress faster. We have added a paragraph to acknowledge the effects of prior experience and other factors predisposing to performance on the game, and tried to make our position on the relevance of these clearer.

We did carry out an additional analysis which should lay the reviewer's mind at rest over the possibility that there is some exception distortion introduced into our data by transfer effects (see below).

3. Is operationalising 'practice amount' as 'number of plays' distorting

plays on which the score is higher confer more practice experience

The reviewer makes a good point which we address by carrying out a re-analysis of the result linking practice to performance, but using total aggregate score as our index of practice extent, rather than number of plays. This allows us to gauge the total number of clicks a player has made and match it, on average, against their performance. The original result holds. We discuss this further below.

Reviewer 1

Re: The claim "that this is the first time such data has been collected"

Our argument is that this is the first time full details of training history can be connected to performance. By this we mean that not only is practice extent recorded, but also something about the nature of actions taken during practice (this data is

used in the analysis which links early variation with higher performance). We have amended the text to make this clearer.

Re: The claim that the participants were "engaged for a sustained amount of time in effortful practice"

The reviewer is correct that this is not directly supported by the evidence we collect — we did not directly assess the motivational state of player. We defend the assertion that on the grounds that playing the game is a voluntary activity associated with no extrinsic rewards. Players, some of whom choose to repeat play hundreds of times, improve in their abilities to get high scores. This obviously does not establish high levels of effortful engagement, but it attests to levels of engagement at least as high as in the majority of experimental studies in which undergraduate students participate as a obligatory part of their course. The review is correct that players could have engaged with the game with different levels of effort, as with lab based experimental studies. Different levels of effort would be an additional factor in our results, but would not - we believe - interact so as to confound our major results on the effects of practice extent, spacing and variability.

Re: lack of information on participant demographics

The reviewer points out that earlier research on skill acquisition has investigated demographic variables. Our method denied us this information. Whilst this leaves important questions unaddressed, we do not see that our neglect of demographic variables should confound the results we have presented.

Re: transfer effects

We do not deny the possibility, indeed likelihood of transfer effects. We suppose that for each individual there will be a range of sensory, motoric, cognitive and experience-dependent factors which position them in terms of initial ability and potential to improve at the game (as with performance in any complex domain). The existence of transfer from previous game and/or computer experience does not contradict the conclusions of our analysis. For example, the result that higher scorers have higher scores from the start, and improve more quickly, holds throughout the population of scorers (i.e. not just for the top 20%). Certainly there should be some advantage to previous game play, but the presence of the same basic effect across the

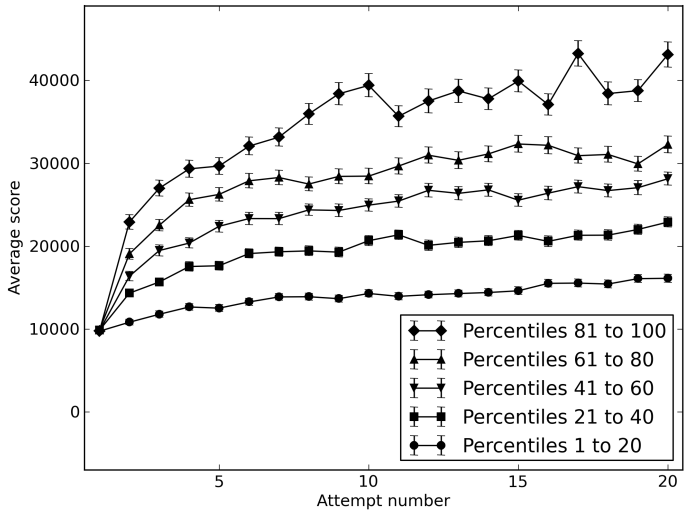

Figure 1. Re-analysis of original Figure 2, but selecting only players who scored between 7500 and 12500 on their first play (n=1824). File: `ps_fig2_equate.py`

distribution of players suggests that this factor combines continuously with all the other factors which support high scoring.

We have added a paragraph to make clearer our acknowledgement of the existence of transfer effects among other predisposing factors for performance, and to make clearer how these impact on our analysis.

An additional analysis that might help convince the reviewer of this point is if we re-perform the analysis that produced figure 2, but only for players who have closely matched scores on the first (or first 3) plays (See Figures R1 and R2).

This analysis allows us to the see that the tendency for higher scorers to improve more quickly is pervasive - it asserts itself even when we look at players who have similar first scores (albeit in a more noisy form). This shows that the effect of what players bring to the game isn't a "one off" effect that works solely on their first experience of the game. Neither is it the case that our dataset is contaminated by a subset of players who bring a high level of game experience to their performance and so distort out results.

Re: variable experience with plays of different length

The reviewer makes a good point. To address this, we perform our first analysis again, but using the total of all scores so far as the index of experience, rather than attempt number. The result is shown Figure R3. By comparing points which are

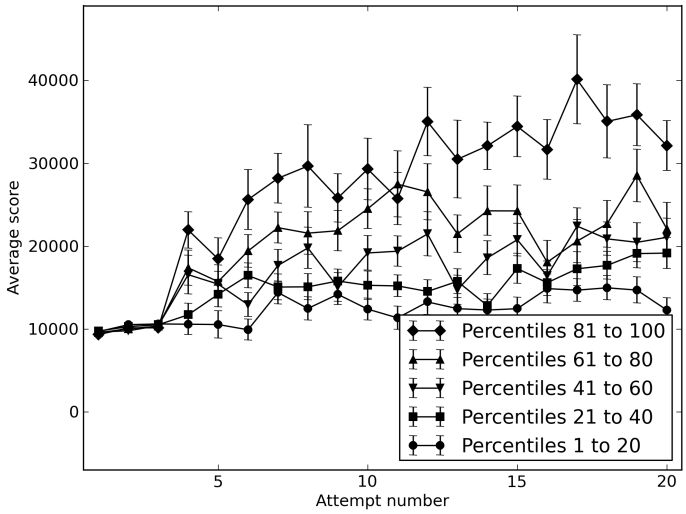

Figure 2. Re-analysis of original Figure 2, but selecting only players who scored between 7500 and 12500 on their first 3 plays (n=126). File: `ps_fig2_equate.py`

vertically aligned we can compare players in the five percentile groupings at points of equivalent game experience. The distinctive pattern remains - those who are the highest scorers begin by performing at a higher level and accelerate in their improvement of this performance compared to lower scorers. If the pattern seen in the original Figure 2 had been solely due to a benefit (among the high scorers) of additional experience (over the lower scorers) then we would see overlaid lines for the five percentile groups, with the line for the highest scorers merely extending along the x dimension. This is not the case. It is not that the highest scorers accelerate along the same experience-performance curve as the lower scorers, but rather that they trace a difference experience-performance curve. We have noted this in the text.

Re: analysing the different clicks of different players.

Unfortunately, this data does not exist, and so the exact analysis suggested by the reviewer is not possible. We have attempted to provide some insight into the nature of the learning between higher and lower performing players with our analysis of early variability in scores. Regarding the analysis of click-patterns (i.e. within game action strategies), see below.

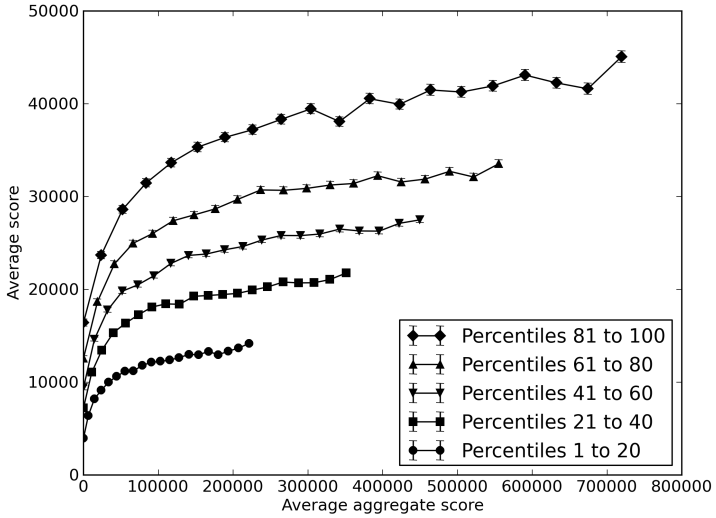

Figure 3. Re-analysis of original Figure 2, but using total score so far instead of attempt number on the x-axis. File: `ps_fig2_score.py`

Re: analysis of behaviour in game environment.

Our description of the game environment was minimal because we did not carry out a detailed task analysis of how high levels of performance are reached. To provide a thorough and valid account of the cognitive and behavioural strategies which support high performance would be a substantial task. We left this task aside, believing that the analysis of summary statistics of play could still provide insight into the nature of skill acquisition. An inspiration for this is the literature on the power law of learning, which shows that regularities exist across a wide range of domains and can be uncovered by an suitably abstract level of analysis. It is in this spirit which we wish to extend the analysis of how practice and statistics associated with practice are related to performance.

The detailed analysis the reviewer suggests would be best supported by lower n, higher resolution, experimental investigations.

Re: player backgrounds

We agree that it would be useful to conduct this analysis with more information on player background. Sternberg et al (2013) provide a recent example of just this type of approach. It was not our objective with this work to do this however, and we believe the claims we make are important and stand despite the lack of background information.

There is an important level of analysis which thoroughly characterises the strategies employed by the players and the attendant cognitive changes associated with an improvement in performance. Such work would be useful, but it was not our task here. The abstract level of analysis we employ here allows us to take advantage of the large dataset of learning behaviour that we have, and supports the theoretical claims that we make. We have, where appropriate, clarified the language we use so we do not risk making unjustified claims.

Reviewer 2

We are pleased to note that the reviewer recognises the novelty of the work and recommends publication.

Re: the effect of shared machines

Our analyses remain consistent when we exclude players with a high (>100) number of plays (i.e. those most likely to be from a high use shared machine). It is worth noting that shared machines, to the extent that they are present in the data set, are most likely to add noise to our analysis (by adding data in which there is no consistent practice-performance patterns) rather than be likely to produce false positives/type 1 errors. Further analysis of this issue is discussed below.

Re: could the 'fanatic' players (1000+ plays) be 200 players playing 5 times each

This is possible, but it is unlikely given that average score increases with number of players. Informal discussion with game designers suggests that from experience, although new players can play on old machines (and old players play on new machines) the rates of this are low enough that the vast majority of data reflects a "1 player - 1 machine" situation.

Furthermore, the effect of such cases would be to contaminate the data for the lowest scoring percentiles. The fact that the same pattern holds across all percentiles (e.g. when comparing 81st-100th percentile vs 61st-80th, or with 41st-60th vs 61st to 80th) suggests that our result is not contaminated by a set of multiple uses 'impersonating' single users and distorting one tail of the distribution

Looking at the graph of average score vs attempt number (file: `ps_machinechurn.py`) it does seem as if the relationship between practice and per-

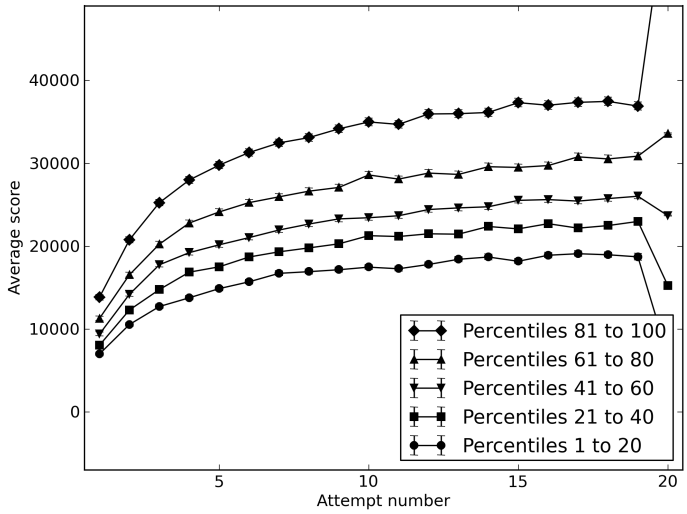

Figure 4. Re-analysis of original Figure 2, but calculating percentiles using players' scores on their 20th attempt. File: `ps_fig2_equate.py`

formance breaks down (i.e. becomes lost in noise) after 200 plays. A factor in this could be that some of the "players" who seem to have a high number of attempts are actually shared-use machines, which record the scores of many players having their first goes (as the reviewer suggests). To check that this isn't distorting our results, we re-ran our analysis with any players who made more than 100 plays excluded (`psy_fig2_exclude.py` and `StaffordFig2exclude.png`). The results are nearly identical (because the vast bulk of data is made up of people who play less), disconfirming this concern (we do not show the graph because it is so similar to Figure 2 in the original paper).

Re: dividing into percentiles without taking account on which play the highest score is obtained

We provide an alternative analysis to that shown in the original figure 2, except with the players divided into percentiles based on their score in their 20th play. The results are shown in Figure R4. The same qualitative pattern of results holds, but with the graph distorted for the points at the 20th play (since these are used for the percentile classification it is understandable that they should be most extreme). We have noted this analysis in the text.

Re: figure 3

Each dot is a percentile group. The n is all those who played more than 19 times (n=12604) so the n for each dot is data from 2520 or 2521 individuals.

The reviewer does not find the figure convincing. Although only 7/50 points in the top 50 percentiles are above the 95% bootstrapped confidence limits (compare 0/50 for the bottom 50 percentiles), there is a clear trend for nearly all points to be above the bootstrapped mean (in percentiles 51-100) and below the bootstrapped mean (in percentiles 1-50). An analysis which takes into account the related nature of the points (i.e. rather than considering 100 independent confidence intervals) strongly supports the claim that the observed values differ from the bootstrap (i.e. the expected values under the null hypothesis that spacing practice confers no advantage on performance). Graphically, the alternative hypothesis is reflected in the claim that the true curve traced by the data is below the bootstrapped mean in percentiles 1-50 and above the mean in percentiles 51-100. To test this we calculated the observed - expected difference so that a positive score reflects support for the experimental hypothesis (by calculating expected-observed for the bottom 50% of scorers and observed-expected for the top 50%). A one sample t-test on these differences confirms that they are highly significantly different from zero (mean = 7.59 hours, $t(99)=7.27$, $p<0.00001$) and so we are justified in rejected the null hypotheses.

We hope that this goes some way to nudge the reviewer's initial intuition. We have added details of the t-test to the main text.

Re: analysis of individual trajectories

In response to the reviewers request, we performed an additional analysis — identifying players with similar scores on their first play, who played their 1st-6th games with a 2 hour window and their 15-20th games also within a two hour window. The motivation for this classification is to find players with similar habits, who have comparable initial ability on the game. We then divide them into two groups: those who had a six hour (or more) gap at some point between their 6th play and the 15th play, and those who didn't. The result (Figure R5) shows our previous finding that practice spacing is associated with higher performance, as revealed in the shape of the learning curves - the average learning curve for players of comparable ability diverges at the point that one group begins to space its practice. We have added this figure to the paper.

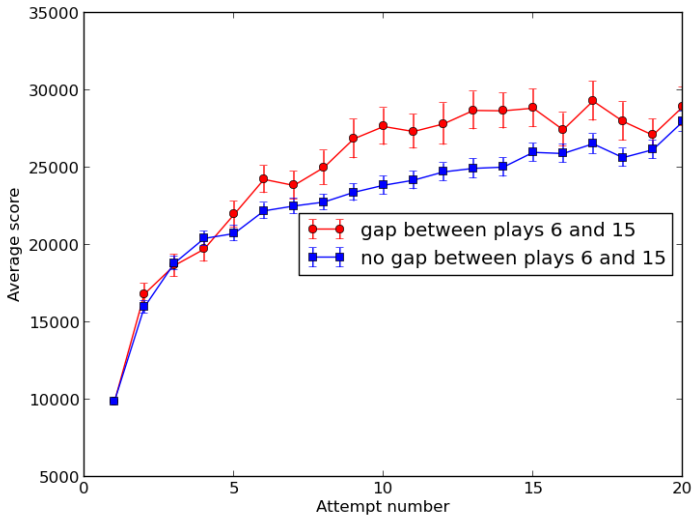

Figure 5. Players of comparable ability but grouped according to whether they left a six hour+ gap at some point between plays 6 and 15 File: `ps_fig2_tracegap2.py`

Re: generalisability

The reviewer raises an important point. Our strategy has been first to use our data set to validate results which are extant in the literature (i.e. which are known to generalise). Where we have new results (concerning the learning curves of high and low performers, and concerning practice variability) we hope that the level of abstraction we have adopted means that we are extracting features of learning which will generalise to other situations. The result on variability is itself a replication of the same finding in laboratory task (using undergraduates and rats, Stafford et al., 2012)).

Re: only 5% played more than 9 times, a seemingly small proportion

The "drop-out" rate is very high (50% at each play). In fact, the game developers told us that 50% of people who click on the link to play the game quit the browser in the three seconds the game takes to load. This supports the idea that the drop-out is something inherent to the nature of attention spans on the internet, rather than due to intrinsic qualities of the game.

Reviewer 2 has a number of brief requests:

page 3 line 36-37 typo: corrected

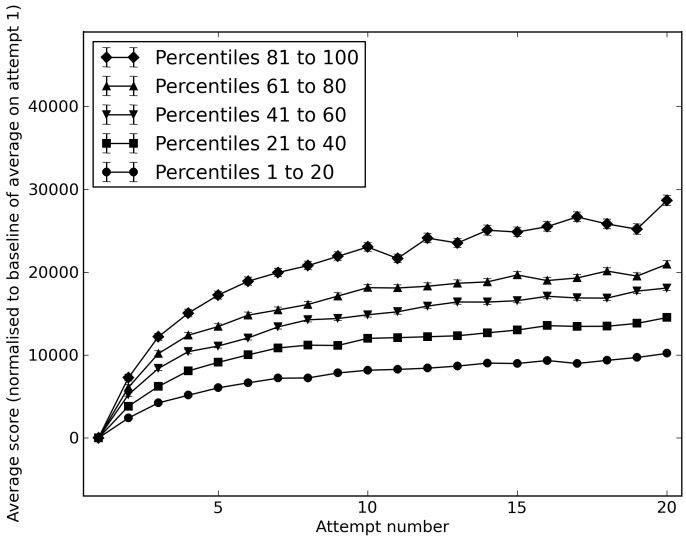

Figure 6. Re-analysis of original Figure 2, but using average score on attempt 1 in each percentile grouping as a baseline File: `ps_fig2_rebaseline.py`

page 4 line 22-23 query on citation style: corrected

page 5 line 24-25 request to expand on discussion of the power law of learning: done

page 8 line 8-15: Has there been any attempt to identify whether specific 'players' (shared machine, e.g.) show a significantly different profile than real players?

We didn't do this. The high levels of individual variability make this non-trivial. As mentioned above, our analyses remain consistent when we exclude players with a high (>100) number of plays (i.e. those most likely to be from a high use shared machine). It is worth noting that shared machines, to the extent that they are present in the data set, are most likely to add noise to our analysis (by adding data in which there is no consistent practice-performance patterns) rather than be likely to produce false positives/type 1 errors.

page 9 Figure 2: An alternative presentation of this figure would be to normalise the plot to a common starting score.

This is shown in Figure R6. This figure makes very clear the different learning rates between the groups, but hides the different initial scores, so we have kept the original in the manuscript.

page 10 line 22-23: Missing Figure reference: corrected

page 11 Caption of Figure 3: Which 'standard errors' are shown? :

no standard errors are shown, this line has been removed.

page 14 line 5-6: the bracket opened is never closed: corrected

References

Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., et al. (2012). A novel task for the investigation of action acquisition. *PloS one*, *7*(6), e37749.

Sternberg, D. A., Ballard, K., Hardy, J. L., Katz, B., Doraiswamy, P. M., & Scanlon, M. (2013). The largest dataset of human cognitive performance reveals insights into the effects of lifestyle factors and aging. *Frontiers in Human Neuroscience*, *7*, 292.