Behavior Research Methods manuscript No.
(will be inserted by the editor)

The joint contribution of participation and performance to learning functions: Exploring the effects of age in large-scale data sets

Mark Steyvers · Aaron S. Benjamin

Received: date / Accepted: date

Abstract Large-scale data sets from online training and game platforms offer the opportunity for more extensive and more precise investigations of human learning than is typically achievable in the laboratory. However, because people make their own choices about participation, any investigation into learning using these data sets must simultaneously model performance–that is, the learning function–and participation. Using a data set of 54 million gameplays from the online brain training site *Lumosity*, we show that learning functions of participants are systematically biased by participation policies that vary with age. Older adults who are poorer performers are more likely to drop out than older adults who perform well. Younger adults show no such effect. Using this knowledge, we can extrapolate group learning functions that correct for these age-related differences in dropout.

Keywords Dropout · Skill acquisition · Learning functions · Missing Data · Naturalistic Environments · large-scale data sets · Bayesian modeling

Mark Steyvers
Department of Cognitive Sciences, University of California, Irvine
2316 Social & Behavioral Sciences Gateway Building
Irvine, CA 92697-5100
(949) 824-7642 Phone
(949) 824-2307 Fax
E-mail: mark.steyvers@uci.edu

Aaron S. Benjamin
Department of Psychology, University of Illinois at Urbana-Champaign
603 East Daniel Street
Champaign, IL 61820
(217) 333-6822 Phone
E-mail: asbenjam@illinois.edu

Learning in the real world involves many choices. We decide when to study, how to study, and when to stop studying and turn to something else. In the history of research on learning, particularly within psychology, the vast majority of scientific approaches attempt to control for these and other sources of variation in self-control, with the hope that what will emerge is an uncontaminated view of learning and memory (Koriat & Goldsmith, 1994; Benjamin, 2007; Nelson & Narens, 1994)

Whatever the merits of this approach, it is unsatisfactory for large ecologically situated data sets in which learners come and go at their leisure. Understanding learning and memory in tasks in which learners exert considerable control over aspects of their learning requires an explicit consideration of metacognitive factors that determine participation and influence performance.

Here we present learning data from the online "brain-training" platform Lumosity. Lumosity provides a number of different games for users that are intended to tap memory, attention, flexibility, speeded processing, and problem solving. Many of these games are based on well-worn tasks from cognitive psychology. Millions of people play these games, providing a very rich platform on which to study learning (Donner & Hardy, 2015). However, unlike lab studies, where individuals follow a strict regime and can be coerced to provide a sufficient number of data points to fit functions to that individual's performance, participants in online platforms decide when to play, how often to play, and when to quit. A joint consideration of participation and performance allows us to use these large-scale data sets to evaluate theories of learning and of metacognition. Generally, the use of online platforms for investigating skill learning has grown in the past few years (Donner &

Hardy, 2015; Huang, Yan, Cheung, Nagappan, & Zimmermann, 2017; Stafford & Dewar, 2014), and is part of a welcome new trend of using naturally occurring large-scale data sets to develop and test theories of cognition (Goldstone & Lupyan, 2016; Griffiths, 2015).

The lesson we draw here is that any model of skill learning from an uncontrolled source like an online learning platform must jointly deal with questions of performance and of participation. When individuals drop out of the task randomly, like they often do in the lab (say, due to computer problems), then dropout behavior increases variability and the potential for heteroskedasticity at more distant points in the learning function. However, when individuals drop out for reasons that are related to their current or future performance, learning functions are directly biased. Averaging across individuals has long been known to influence the shape of learning functions (Estes, 1956), but the effects of voluntary participation on group learning functions has not, to our knowledge, previously been considered. This is not a statistical problem: only a model of the process by which individuals choose to stay or go can debias such effects.

In this paper, we present a theoretical and empirical investigation of the effects of voluntary participation and withdrawal on aggregated learning functions. We start with an empirical analysis of learning functions for individuals and for groups that differ in age. We show that individuals who drop out earlier lie on a different learning trajectory than those who continue, indicating that group learning functions will be biased by differential participation. Specifically, older adults who withdraw early exhibit a slower rate of improvement than older adults who continue with the task. Younger adults do not reveal this systematic pattern of withdrawal. In addition, we apply models of learning to individual performance functions and estimate the trajectory of those functions for a subset of users of different ages. Using these individual fits, we show that the slopes of the learning functions are typically shallower for individuals who drop out early. We then use the fits to extrapolate performance for those who withdrew to trials that they never actually completed. Doing so, we show that age-related group learning functions corrected for differential withdrawal are markedly different than the uncorrected functions.

As a starting point for considering the effect of systematic dropout on learning curves, Figure 1 shows simulated data under a number of scenarios. The left panel shows simulated learning curves that vary in learning rate and asymptote. The red curve shows the aggregate learning function. The middle panel uses the same learning curves and simulates the effect of dropout when individuals drop out for reasons unrelated to performance. In this case, the aggregate learning curve is unbiased by the dropout. In the right panel, the probability of dropping out is negatively related to (latent) asymptotic performance. Here it can be seen that the aggregate learning function is considerably biased from the original.

One of the advantages of using large-scale naturalistic data sets for cognitive research is the diversity of users such platforms attract. Here we use the large age range of participants to examine the learning curves and dropout rates of users across the lifespan. Older users might have different motivations for using Lumosity than younger users, and those motivations might influence participation policies. Older adults may be motivated to combat cognitive decline and thus be more inclined to stick with tasks that they find difficult. Alternatively, they may be more sensitive to the stereotype threat posed by poor performance and thus be quick to quit tasks that they perform poorly on. Age effects on memory and attention tasks are well documented (Park & Schwarz, 2000) but can only be fairly interpreted in naturalistic data sets by seriously considering participation policies.

Method

The Lumosity platform provides a number of games that tap memory, attention, flexibility, speeded processing, and problem solving. In the Lumosity program, users are given a recommended daily training session of five different cognitive training games. One five-game session takes approximately 15 minutes to complete. Outside of the training sessions, Lumosity users can also opt to select and play games directly from the entire library of over 50 available games. As of 2018, over 90 million users from 182 countries had signed up to participate. The data set that we are working with includes the gameplay event history for three cognitive games. This data set includes 194,695 users, 584,077 individual learning curves, and 54,224,152 single gameplay events.

Tasks

The tasks included *Lost in Migration*, *Ebb and Flow*, and *Memory Match*. Screenshots of these games are shown in Figure 2.

Lost in Migration. This is a selective attention game inspired by the Eriksen flanker task (Eriksen & Eriksen,

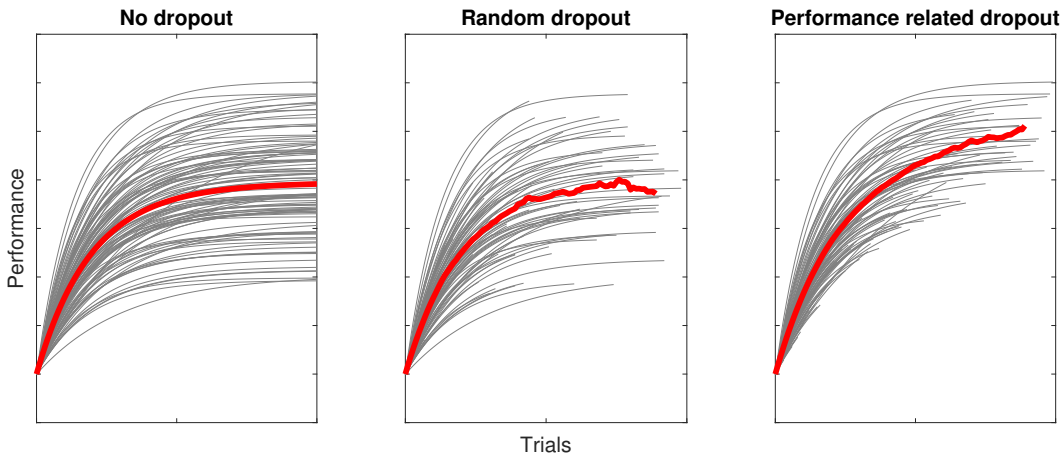

Fig. 1: Illustration of the effect of dropout under different scenarios. A random set of learning curves, shown in the left panel, was generated using the exponential learning function in Eq. 1 for 100 simulated users that varied in asymptote and learning rate but not intercept. In the middle panel, learners drop out at random. In the right panel, learners are more likely to drop out when the (latent) asymptote of the function is lower. The red line is the aggregate learning function.

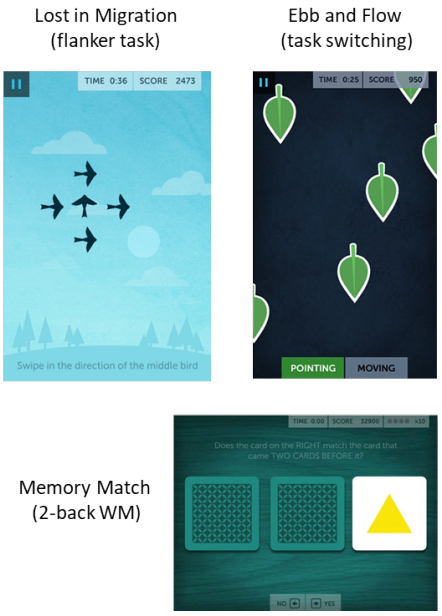

Fig. 2: Screenshots of the three cognitive games and their correspondence to classic cognitive tasks

1974). The goal is to respond to the direction of the target (a bird) and ignore the direction of distractors that flank the target. During each trial, the target and distractors are arranged in different spatial layouts. Users are asked to use the arrow keys to indicate which direction the target is pointing; the layout and orientation of the distractors varies from trial to trial.

Ebb and Flow. This is a game designed to test the ability to switch between different tasks. Users have to shift focus between two different rules depending on the color of the leaves. When the leaves are green, the user has to determine the direction in which the leaves are pointing and respond accordingly. When the leaves are orange, the user has to respond based on the direction that they are moving. "Inhibition" trials occur when the orientation and direction of movement are different, requiring the user to express behavior associated with one rule and inhibit behavior associated with the other rule. On "no-inhibition" trials, the orientation and direction of movement of the stimuli lead to the same response.

Memory Match. This is is a 2-back working memory task (e.g., Kane, Conway, Miura, & Colflesh, 2007) where sequential stimuli are presented one at a time. The user holds each stimulus in short-term memory while new stimuli are presented. In the 2-back task, users determine whether the stimulus currently presented (the card on the far right in the bottom display of Figure 2) matches the stimulus presented two trials earlier. If users make a mistake, the users are given hints by revealing the previous two stimuli in the sequence, which allows the user an opportunity to re-learn the current history for subsequent trials.

Scoring. Each gameplay event has a fixed duration: 45 seconds for Lost in Migration; 60 seconds for both Ebb and Flow and Memory Match. At the end of each gameplay event, users are provided feedback on mean response time per trial, mean accuracy, and a score that is based on the total number of correct trials completed within the fixed time period as well as bonus points based on a variety of factors (e.g. streaks of correct responses). The total score is the focal point on the feed-

back screen, so it can be assumed that the conditions foster a combination of speed and accuracy.

Data Processing

The raw data is described at the individual trial level (i.e., individual decisions within a particular gameplay event) and include response time, accuracy, as well as the type of condition associated with the trial. In the raw data, any trial with a response time higher than 5 seconds was coded as 5 seconds. For the purpose of this research, we analyzed the data summarized at the gameplay event level. Specifically, we focused on the *number of correct trials completed per gameplay*, a value that is closely related to the inverse of the mean response time for correct decisions. It is also closely related but not identical to the point score shown to the user because we omitted any bonus points that are part of the game scoring. For Memory Match, we did not include the hint trials in the total trial count (even if the decision was correct) because hint trials reveal part or the full history, making these trials substantially easier.

In total, the data set contains the full gameplay history for 194,695 users across these three games spanning a period from Dec 18, 2012 to Oct 31, 2017. Users spent a median of 2.2 years on the platform. Some of the gameplays had timestamps but lacked any recorded gameplay data. After removing these missing records, the data set reduced to 194,682 users, 572,825 individual learning curves, and 44,204,431 single gameplay events. Because a key aspect of our work involves an analysis of the timepoint at which users voluntarily stop playing, users that were still active within the 100 days prior to the last recorded event were removed from analysis. The final data set contained 163,160 users, with a total of 400,874 learning curves and 22,477,188 gameplay events. The games Lost in Migration, Ebb and Flow, and Memory Match were played a median of 69, 67, and 9 times, respectively, by individuals. The lower number of game plays for Memory Match could be due to differences in user interest and engagement but also because the game shows up less frequently (relative to Ebb and Flow and Lost in Migration) in the suggested training program sent to users.

User Demographics

Basic demographic information is available based on information provided when signing up for Lumosity. The majority of users are female (57%), with 36% males and 7% of users who did not provide gender information. The majority of users are older than 50 (65%), reflecting the appeal of these cognitive games to older players. We coded the age of users in 7 bins leading the following breakdown of the user sample: 1-20 (1.58%), 21-30 (9.43%), 31-40 (8.7%), 41-50 (14.1%), 51-60 (26.8%), 61-70 (27.6%), and 71-80 (11.4%). The youngest age group (1-20) is omitted from all analyses because of the relatively small sample size and the heterogeneous nature of this age group. Most users live in the United States (63%), with substantial populations from Canada (9.6%), Australia (9.1%) and Great Britain (2.2%). Consequently, it is a sample heavily biased towards the West.

Results

A subset of the model analysis scripts are publicly available on the Open Science Framework (https://osf.io/ymkhb/). For our analyses, we utilize Bayes factors (BFs) to determine the extent to which the observed data adjust our belief in the hypothesis that are differences between two groups and the null hypotheses (no difference between groups). There are numerous advantages of BFs over conventional methods that rely on p-values (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Jarosz & Wiley, 2014; Wagenmakers, 2007), including the ability to detect evidence in favor of a null hypothesis and a straightforward interpretation. In our notation, $BF > 1$ indicates support of the alternative hypothesis while $BF < 1$ indicates support of the null hypothesis. For instance, $BF = 5$ means the data are five times more likely under the alternative hypothesis than the null hypothesis. In some instances, we report log BF factors, such that $\log BF < 0$ indicates support of the null hypothesis and $\log BF > 0$ indicates support for the alternative hypothesis.

Aggregate Learning Curves. Figure 3 shows aggregate learning curves for the three cognitive tasks for six age groups (grouped by decade of life). The effects of age are readily apparent and robust across the tasks: older users start at a lower performance level and continue on a lower trajectory. It is noteworthy that the decrements are consistent throughout the decades–there is no point at which performance decrements accelerate. However, the point here is that caution needs to taken when interpreting these aggregated curves. The aggregate learning curves obscure not only individual differences in learner characteristics (Heathcote, Brown, & Mewhort, 2000), but also in participation. As training progresses, more users drop out, and the aggregate curves reflect only the progress of the self-selected users who remain. The effect of this dropout can be observed in Figure 3 by

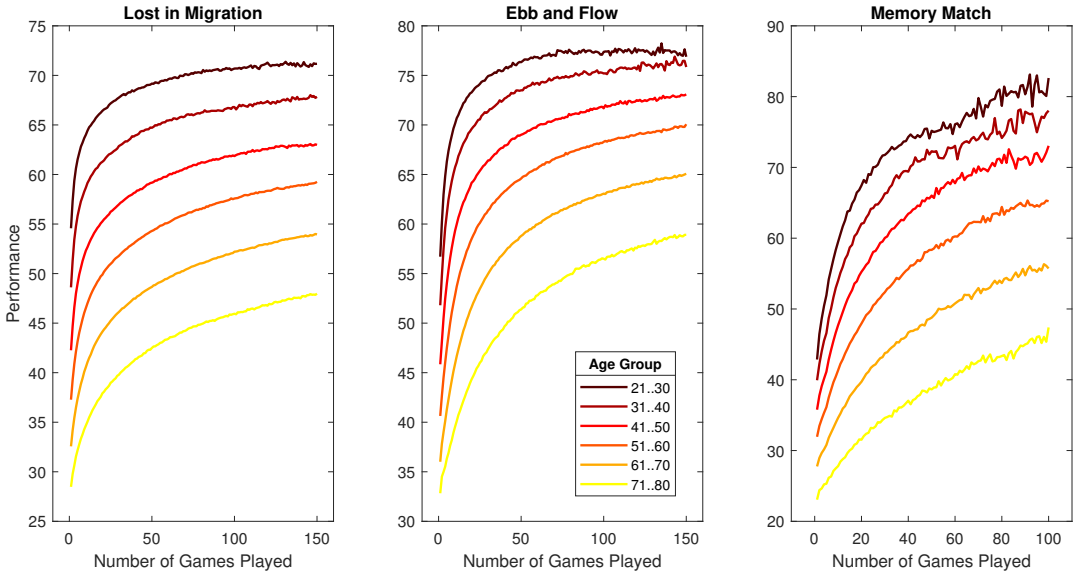

Fig. 3: Aggregate learning curves for three cognitive tasks separated by age groups. Performance is assessed by the number of correct decisions per game play. The learning curves are restricted to first 150 game plays for Lost in Migration and Ebb and Flow, and 100 games for Memory Match. Note that no curve smoothing was applied to obtain these results

the increase in the noise in the more distant points on the function.

The relationship between performance and participation. One way to address the effects of voluntary withdrawal is to examine whether the trajectory of learning curves differ for users who drop out early and late. Figure 4 shows the learning curves disaggregated into two groups: those who drop out early (after 30-50 games for Lost in Migration and Ebb and Flow, or 15-25 games in Memory Match) and those who continue and play at least 100 games in Lost in Migration and Ebb and Flow, or at least 60 games in Memory Match. The lower cutoff points for Memory Match were chosen because people have fewer plays of the game.

Across all three tasks, a pattern is clear: the trajectories of the learning curves for older subjects differ, depending on when they choose to stop playing. Yet this is not apparent for younger subjects. This effect confounds interpretation of the age-related learning functions shown earlier, and must be accounted for to gain an unbiased picture of age-related differences on these tasks. We return to this point shortly.

Table 1 shows performance and provides statistical tests of differences in performance between participants who drop out early and those who drop out late. We examine these differences at the start of learning (gameplays 1-3) and at a later stage of learning (gameplays 28-30 for Lost in Migration and Ebb and Flow; gameplays 13-15 for Memory Match). This later stage of learning corresponds to latest gameplays for which we have complete information from the users who dropped out early. In general, people who drop out early exhibit poorer performance (the average log BF in the table is 37.15). This effect is detectable even on the very first 3 trials that the subject experiences (average log BF = 37.19). The effect is considerably stronger in older adults: users 51 and over reveal an average log BF of 63.64, compared with 10.67 for those 50 and younger.

Applying a model of learning

To better understand the relationship between dropout and learning, we applied simple models of skill acquisition to the data from individual subjects. The goal is to fit these curves to individual users and to assess the properties of the learning curve as a function of dropout. If learning and dropout are related, as Figure 4 suggests, individual learning curves will be different for early and late dropouts. Here we will focus on the slope of the learning function, a parameter that captures the rate at which performance is increasing at one moment in time. Although the slope is not an individual parameter of any of the models we fit, it can be easily computed and directly compared across the models we evaluate. If users drop out because of slow acquisition, then the slope of the acquisition function for subjects who drop out early will be lower than the

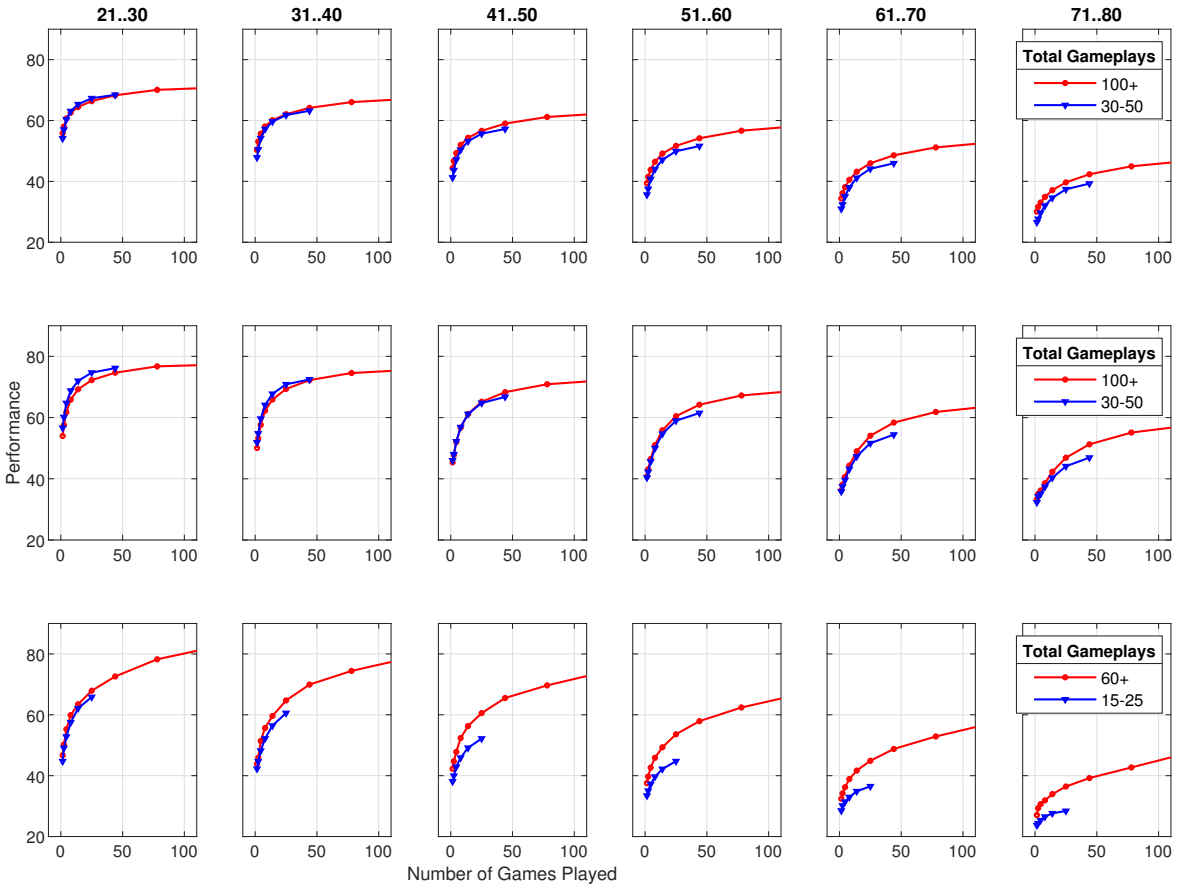

Fig. 4: Aggregate learning curves separated into early and late dropout users. Top, middle and bottom rows correspond to the cognitive tasks Lost in Migration, Ebb and Flow, and Memory Match. Columns correspond to different age groups.

slope of subjects who drop out later, conditional on number of gameplays to that point.

Once the individual learning curves are estimated, performance can be extrapolated to later gameplays for subjects who dropped out early, simulating the effects of continued play. This enables a direct comparison between the *aggregated* learning curves that do and do not take the effect of dropout into account. At that point we can make an assessment of the effects of learning in the absence of participant-related bias.

Because of the computational challenges of working with the large Lumosity data set, we subsampled the data for the purpose of modeling. We sampled for each game and age group a random set of 2400 users, creating a data set with 38,113 unique users across age age groups and 46,200 learning curves, involving a total of 2,197,964 gameplays.

Learning functions. Many different modeling approaches have been proposed to model the improvement of performance as a function of practice, including descriptive models such as exponential and power law learning functions (Newell & Rosenbloom, 1981; Evans, Brown, Mewhort, & Heathcote, 2018; Heathcote et al., 2000), and cognitive architectures such as SOAR (Laird, Newell, & Rosenbloom, 1987) and ACT-R (Anderson & Lebiere, 2014).

We focus on two simple learning functions that are sufficiently accurate to capture the overall characteristics of learning curves. We use a three parameter exponential (Heathcote et al., 2000), and a three parameter power function (Newell & Rosenbloom, 1981):

$$\begin{aligned} y_t &= u - ae^{-ct} \quad \text{Exponential} \\ y_t &= u - at^{-c} \quad \text{Power} \end{aligned} \tag{1}$$

These learning models describe performance y as a function of t, which in our case corresponds to the number

Table 1: Differences in performance across early and late dropouts as a function of age group. *Initial games* refer to the first three games played by the learner, *Late games* are based on games 28-30 for Lost in Migration and Ebb and Flow and games 13-15 for Memory Match, close to the point of withdrawal for the early-dropout group. y_{early} and y_{late} are the average performances across the early and late dropouts, respectively. N_{early} and N_{late} are the number of users that were used to compute the average performance levels. The log Bayes factor assesses the evidence for a difference between the early and late dropout performance levels at equivalent stages of practice.

		Initial Games					Late Games				
Game	Age Group	y_{late}	y_{early}	N_{late}	N_{early}	logBF	y_{late}	y_{early}	N_{late}	N_{early}	logBF
Lost in Migration	21..30	58.3	56.7	1487	2574	8.56	67.1	67.7	1259	2703	-0.96
	31..40	53.2	50.6	1729	2195	30.78	62.8	62.4	1461	2309	-2.13
	41..50	46.9	43.9	3648	2835	66.33	57.3	56.4	3165	3022	9.15
	51..60	42.0	38.0	8286	4011	211.45	52.4	50.7	7619	4647	79.58
	61..70	37.0	33.2	9023	3187	188.01	46.9	45.2	9433	4038	70.65
	71..80	32.8	28.7	3325	1027	79.11	40.9	38.7	3923	1525	40.24
Ebb and Flow	21..30	57.8	60.3	844	2096	19.83	73.3	75.3	795	1885	11.62
	31..40	53.4	54.7	1034	1630	2.39	70.4	71.8	1158	1566	6.43
	41..50	47.6	48.3	2209	1916	-1.11	66.2	65.5	2552	1968	-0.70
	51..60	42.8	42.2	5227	2846	1.05	61.7	59.6	5925	2989	49.94
	61..70	38.2	37.5	5668	2307	1.68	55.3	52.3	7023	2741	87.66
	71..80	35.0	34.6	1936	783	-2.21	47.8	44.8	3072	1180	37.41
Memory Match	21..30	52.9	51.2	233	1483	-1.27	65.4	63.9	297	2012	-1.70
	31..40	49.7	47.5	209	1229	-0.55	61.6	58.2	289	1840	2.66
	41..50	47.0	43.3	358	1578	8.06	57.5	51.2	555	2658	34.64
	51..60	42.5	38.4	742	2503	27.08	51.3	44.4	1190	4573	100.96
	61..70	37.2	33.6	740	1698	21.32	43.6	37.7	1390	3785	97.13
	71..80	32.5	28.0	230	405	8.95	37.2	30.6	509	1022	45.51

Note: the Log Bayes factors are based on JZS Bayesian two sample t-tests with default prior scales (Rouder et al., 2009). Negative values provide support towards the null hypothesis of an absence of differences between groups whereas positive values provide support for the alternative hypothesis of a difference between groups.

of gameplays. The models have three parameters: learning rate c, asymptotic performance u and learning gain parameter a, the difference between initial and asymptotic performance. The learning rate captures the speed of learning relative to the learning gain and does not allow for a simple comparison across learners who differ in their learning gain. To compare the rate of learning across users with different learning gains, we will estimate a slope parameter s_t, based on the derivative of the learning functions:

$$\begin{aligned} s_t &= (ac)e^{-ct} \quad \text{Exponential} \\ s_t &= (ac)t^{-c-1} \quad \text{Power} \end{aligned} \tag{2}$$

Before we explain how we estimate these parameters, we first describe our procedure for assessing model fit and the ability of the model to generalize to new data.

Model evaluation: predicting future performance

In order to choose among competing learning models, many methods have been used. A standard approach is to use fit statistics that quantify the balance between model fit and model complexity (Myung, 2000; Myung, Kim, & Pitt, 2000) such as AIC and BIC (Donner & Hardy, 2015), MDL (Pitt, Myung, & Zhang, 2002), WAIC (Evans et al., 2018), and Bayes Factors (Lee, 2004).

Here we follow a different approach based on generalization and cross-validation. Such tests are not widely used in psychology but they have many appealing properties (Yarkoni & Westfall, 2017) and have been used successfully in perceptual decision-making (Cassey, Gaut, Steyvers, & Brown, 2016) and memory modeling (Robinson, Benjamin, & Irwin, under review). Here we use a specific approach that is similar to one previously used to evaluate different models of forgetting (Wixted, 2004). To motivate the approach, it is important to consider that the goal of the model is to assess the characteristics of learning functions for users that drop out at different times. By definition, the learning curves for users who quit sooner will have fewer observations than the learning curves for subjects who continue to play. To compare across these groups, it is important that models estimated from smaller number of observations generalize accurately to future performance. Therefore,

one critical test for a model is whether it accurately extrapolates the learning function.

An example of our evaluation approach is illustrated in Figure 5. It shows a learning curve from one user in the Lumosity data. The Exponential and Power functions are estimated for different amounts of observed data. When the full performance history is observed, the power and exponential models produce very similar model fits (solid black lines). However, the differences between these models become more clear when the models have to extrapolate beyond the observed data. When the model is only given a portion of the learning history and is extrapolated beyond that limited training set, the two learning functions show clear differences. The extrapolated functions are shown by the dotted lines, in which darker shading indicates a training set with more gameplay events. The results for this particular user's learning curve shows that the exponential model consistently underestimates future performance, dramatically so when only a small part of the learning curve is observed. The power model also becomes more accurate as it is trained on more data, but there is no systematic bias. The exact results of the model comparison vary from subject to subject, but this tendency of the exponential model to underestimate asymptotic performance is quite general and is also consistent with an analysis of forgetting functions by Wixted (2004).

We will employ this model selection approach by withholding data from a sample of users and assessing the ability of each model to predict the withheld data. Specifically, we partially withheld data from 1134 randomly selected learning curves with the restriction that these learning curves included at least 150 gameplays. These learning curves were randomly assigned to three different types of generalization tests. In the three tests, the model observed either the first 20, 40 or the first 100 gameplays, and the remaining performance history was withheld from the model. The goal for the model is to predict the withheld performance between 100 and 150 gameplays.

Hierarchical Bayesian Model

We associate each individual learning curve with its own learning parameters (u_i, a_i, c_i). The learning curve models the latent learning state x_t after t gameplays. This leads to the following models:

$$\begin{aligned} x_t &= u_i - a_i e^{-c_i t} \quad \text{Exponential} \\ x_t &= u_i - a_i t^{-c_i} \quad \text{Power} \end{aligned} \tag{3}$$

We assume that the actually observed performance outcome is based on a sample from a positively truncated Normal distribution around the predicted learning state x_t at trial t:

$$y_{i,t} \sim \text{TN}(x_{i,t}, \sigma) \tag{4}$$

where σ is standard deviation which captures the performance variations around the latent learning state. We place a half normal prior $\text{TN}(0, 5)$ on σ. With this observation model, deviations from the theoretical learning curve can be explained in part by the noise model, a characteristic that is consistent with historical and more recent models of learning curves (Evans et al., 2018).

To define the hierarchical model, we need to specify how the individual learning parameters are sampled from population distributions. Normally, all available data is pooled in some fashion in a hierarchical model. However, because of the relatively large data size, the substantial performance differences we observed across games and age groups, and the different number of users within each age group, we apply a separate hierarchical model to each age group within each game (i.e., the model is applied to the subset of 2400 users for each age group and game). Within each hierarchical model, the learning parameters associated with all learning curves for a particular age group and game are sampled from a single set of population distributions.

The non-linearities in these learning models can be challenging for model inference if no restrictions are placed on model parameters. For our data, we can use knowledge of human limitations to place a priori constraints on parameters. For example, the fixed time periods placed on each gameplay imposes strong constraints on the number of correct trials that can be completed. It is very unlikely that any user will ever be able to complete 200 correct trials within the 45 or 60 second time limit (only 18 out of 22M gameplays led to a score higher than 200 and these scores are likely due to recording errors). Therefore, it is convenient to place bounds of [0,200] on the learning parameters u and a. In addition, it is useful to constrain the learning parameter c. While a low value of c is indicative of slow learning (learning stays close to the starting point), a very high value is also consistent with slow learning because the transition to asymptote is made very quickly and stays there. We imposed bounds of [0,0.5] on the learning parameter to facilitate inference and interpretation (Note that a learning rate of 0.5 captures even the fastest learners, achieving over 90% of their learning in six gameplays). With these a priori restrictions, it is useful to reparametrize the individual learning param-

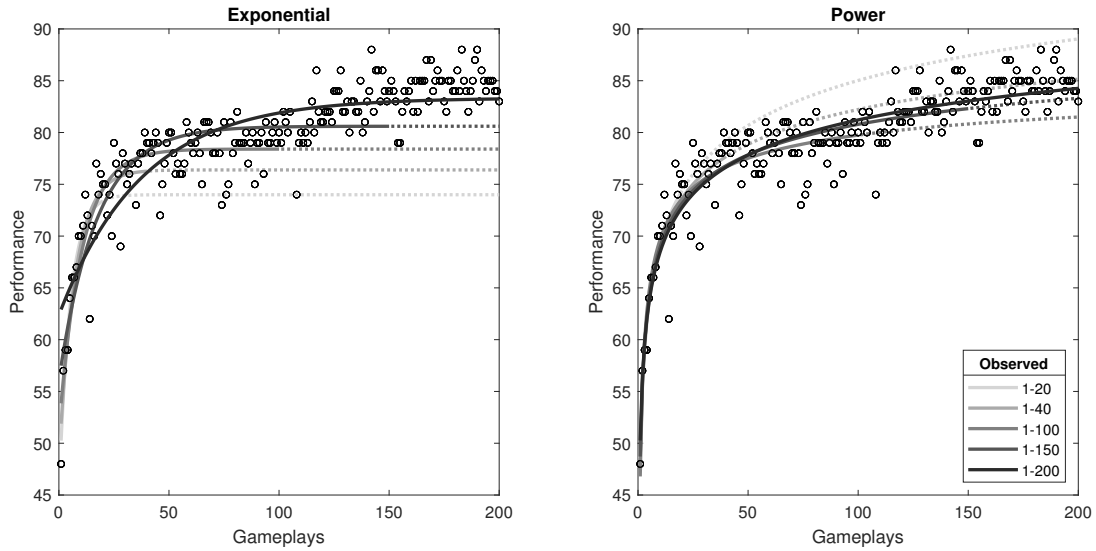

Fig. 5: Maximum likelihood fits of the Exponential (left) and Power (right) models to a learning curve from a single user playing Ebb and Flow. The lines correspond to model fits based on different amounts of observed data, varying from the first 20 gameplays only to the full learning curve up to 200 gameplays. The dashed lines show how the model extrapolates to future learning performance that it has not been trained on. results of extrapolating learning curves.

eters a, u, and c using scaled inverse-logit transforms:

$$\begin{aligned} u_i &= 200\text{f}(u'_i) \\ a_i &= 200\text{f}(a'_i) \\ c_i &= 0.5\text{f}(c'_i) \end{aligned} \tag{5}$$

where $f(x) = 1/(1+\exp(-x)$ is the inverse logit transform. This formulation insures that any value for u', a' will map to values in the restricted parameter range of the original parameters.

Within the hierarchical model, each transformed parameter u_i, a_i, and c_i is sampled from a Normal population distribution with a mean and standard deviation sampled from a Normal and Half Normal prior:

$$\begin{aligned} u'_i &\sim \text{N}(\mu_u, \sigma_u) \quad \mu_u \sim \text{N}(0, 1.5) \quad \sigma_u \sim \text{TN}(0, .75) \\ a'_i &\sim \text{N}(\mu_a, \sigma_a) \quad \mu_a \sim \text{N}(0, 1.5) \quad \sigma_a \sim \text{TN}(0, .75) \\ c'_i &\sim \text{N}(\mu_c, \sigma_c) \quad \mu_c \sim \text{N}(0, 1.5) \quad \sigma_c \sim \text{TN}(0, .75) \end{aligned} \tag{6}$$

The 1.5 standard deviation for the prior mean of the population distribution was chosen such that the learning parameters are roughly uniformly distributed in the original scale.

Parameter Inference

Two procedures were used for parameter inference. Because the data include 46,200 learning curves, and each learning curve has up to three parameters, model inference involves more than 138,000 parameters, making statistical inference computationally challenging. To facilitate initial model exploration and testing, we used the L-BFGS optimization procedure from Stan (Carpenter et al., 2017) to find MAP estimates. These are point-estimates of the parameters that maximize the posterior probability of the model parameters given the observed data. This procedure is not fully Bayesian, as it does not take the uncertainty in model parameters into account. However, the optimization allowed us to explore to space of models within a reasonable time frame (typically a few minutes).

The second procedure on which all results in this paper are based involved Markov chain Monte Carlo using JAGS. For each combination of age group and game, we ran the hierarchical model with 7 chains for 2000 iterations and obtained 100 samples from each chain. This procedure was repeated for both models. The Gelman-Rubin (Gelman, Rubin, et al., 1992) convergence diagnostic led to $\hat{R}$ values below 1.1 across variables suggesting that the chains converged. Posterior predictives were calculated for all withheld observations of the partially observed learning histories: for each sample s of the learning parameters for learning curve i, Equation 3 was used to generate predicted performance levels. These predictions were then averaged across samples to generate point predictions for extrapolated performance levels.

Modeling Results

Generalization Results. Figure 6 shows the results of the generalization tests. The Power model has lower absolute prediction errors overall than the Exponential model, a difference that is not significant when only the first 20 gameplays of the learner's performance are observed (N=366, t=1.79, BF=.286 in a Bayesian paired sample t-test), but becomes more pronounced when the models are trained on 40 and 100 observations (N=376, t=4.76, BF=100+ and N=392, t=4.86, BF=100+ respectively). In addition, the Power model is less biased overall than the Exponential model, with smaller mean deviations than the Exponential model (N=366, t=-2.75, BF=2.39 and N=376, t=-10.93, BF=100+, N=392, t=-11.51, BF=100+ for 20, 40 and 100 observed gameplays respectively). Overall, the exponential model tends to under-predict future performance levels, confirming the generality of the example result shown in Figure 5. Therefore, even though we will report modeling results for both models, these generalization results suggest that any model extrapolations are likely more accurate for the Power than the Exponential model.

Analyzing slopes of individual learning curves. We assessed the slope s of the learning function for both the early and late dropouts at the time of dropout for the early dropout users (using Equation 2). In this comparison, we can evaluate the rate at which performance is increasing at the time that the early dropouts stop playing. Table 2 shows the mean inferred slope across the early and late dropouts. In addition to Bayes Factors, Cohen's d values are shown to indicate effect sizes. The pattern across tasks is clear, and consistent with the analysis of performance shown in Table 1. In 14 of 18 comparisons, the slope for people who drop out early is lower than the slope for people who drop out later.

Among the older adults (61-80), this pattern is evident in 6 out of 6 comparisons, and the Bayes Factors are definitive in 5 of those 6 cases. Older adults who drop out sooner consistently exhibit slower acquisition at the point of withdrawal than older adults who continue. These is little evidence that younger adults show this pattern, especially when using the power model (which provided a better assessment of extrapolated performance in the majority of cases). This finding confirms our concern that the age effects apparent in the original learning functions shown in Figure 3 are compromised by differential participation.

Predicted effect of dropout on learning curves across age groups. With the predicted learning functions in hand for those who drop out early, we are in a position to debias the original learning functions. Figure 8 shows the aggregate learning curves as predicted by the model. Dashed lines show the aggregate learning curves when we simulate the effect of continued learning regardless of dropout. Solid lines simulate the effect of dropout and individual model learning curves contribute only to the aggregate at observed data points in the corresponding user data. The latter aggregate curves are closely related to the empirical learning curves from Figure 3 that are not corrected for differential participation. The results show that the aggregate empirical learning curves are biased and show increases in performance throughout gameplay that are exaggerated by dropout. These effects are more pronounced for older adults, and are evident for all age groups for the Memory Match task.

Discussion

Online training platforms like *Lumosity* provide an incredibly rich source of data on cognitive skill acquisition. A visual analysis of the aggregate, uncorrected learning functions shown in Figure 3 reveals this immediately. However, they also pose new challenges for the development of cognitive theories and computational models. Lumosity is designed to keep users engaged and to increase engagement, and grants users control over many aspects of their own learning. The complex role of voluntary participation and withdrawal cannot be ignored. We have shown here that decisions about participation affect learning functions. In these data, older adults who experienced difficulties with the task were more likely to quit than older adults who performed more ably. Younger adults showed this effect less clearly and less dramatically, if at all. Consequently, performance functions were biased differently by dropout as a function of age.

Current models of skill acquisition and learning are mostly designed to explain empirical data that is collected under carefully controlled laboratory circumstances where participants have limited or no control over the task and training schedule and where participants are trained for the same number of sessions. Models of skill acquisition and learning will have to be expanded to take into account the many metacognitive control processes that affect when and how learning takes place, as well as when learning ceases. Self-regulated learning is important in both applied and theoretical circles (Bennett, Benjamin, Mistry, & Steyvers, 2018; Bennett, Benjamin, & Steyvers, 2017; Gureckis & Markant, 2012; Lieder & Griffiths, 2017; Merkle, Steyvers, Mellers, & Tetlock, 2017) but the majority of models of learning eshew such concerns. We have demonstrated that there

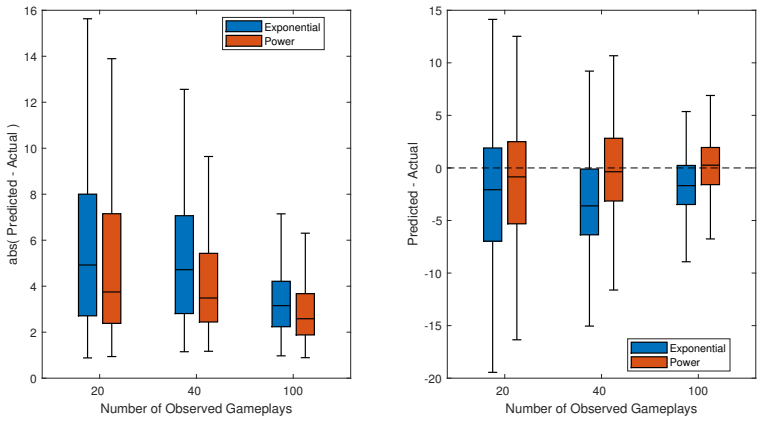

Fig. 6: Generalization performance for different learning functions across different levels of observed gameplays. For each of partially observed performance history, the model is used to predict future performance between 150-200 gameplays. Prediction error is assessed by the mean absolute deviation (MAD), shown left, as well as the mean deviation (MD), shown right. The prediction errors are averaged over the 150-200 gameplays, separately for each of the learning curves. The bars show the interquartile range of the prediction errors across learning curves. Horizontal lines correspond to the median.

Table 2: Mean slope of the individual learning functions across models, games, age groups and early and late dropout groups. The learning slope is assessed at gameplay 30 for Lost in Migration and Ebb and Flow and gameplay 20 for Memory Match, when the early dropout users start to drop. The Bayes factors are based on JZS Bayesian two sample t-tests with default prior scales (Rouder et al., 2009). Bayes factors above 5 are bolded and values above 100 are truncated. Additionally, d shows the effect sizes as assessed by Cohen's d.

Model/Game	Age Group	Early	Late	N_{early}	N_{late}	d	BF
Exponential							
Lost in Migration	21-40	0.073	0.129	1106	645	0.703	**100+**
	41-60	0.100	0.166	836	1183	0.710	**100+**
	61-80	0.142	0.180	723	1556	0.415	**100+**
Ebb and Flow	21-40	0.103	0.194	1151	577	0.862	**100+**
	41-60	0.167	0.250	851	1280	0.720	**100+**
	61-80	0.216	0.277	745	1636	0.504	**100+**
Memory Match	21-40	0.397	0.512	396	104	0.524	**100+**
	41-60	0.303	0.420	453	152	0.744	**100+**
	61-80	0.218	0.295	515	203	0.670	**100+**
Power							
Lost in Migration	21-40	0.122	0.109	1106	645	-0.228	**100+**
	41-60	0.140	0.133	836	1183	-0.121	1.80
	61-80	0.133	0.143	723	1556	0.155	**17.53**
Ebb and Flow	21-40	0.166	0.165	1151	577	-0.025	0.06
	41-60	0.186	0.193	851	1280	0.109	1.00
	61-80	0.164	0.207	745	1636	0.574	**100+**
Memory Match	21-40	0.342	0.426	396	104	0.555	**100+**
	41-60	0.241	0.336	453	152	0.803	**100+**
	61-80	0.158	0.249	515	203	0.872	**100+**

is a relationship between drop out and performance but the causal direction of this relationship is not yet clear. Some users might drop out because of changes in performance. Alternatively, users who are about to quit might be less motivated, try less hard and improve their performance less. To fully account for this data, a joint account of learning and dropout is required. Recently, a number of modeling approaches have been proposed to look at quitting times (Okada, Vandekerckhove, & Lee, 2018) as well as the relationship between perfor-

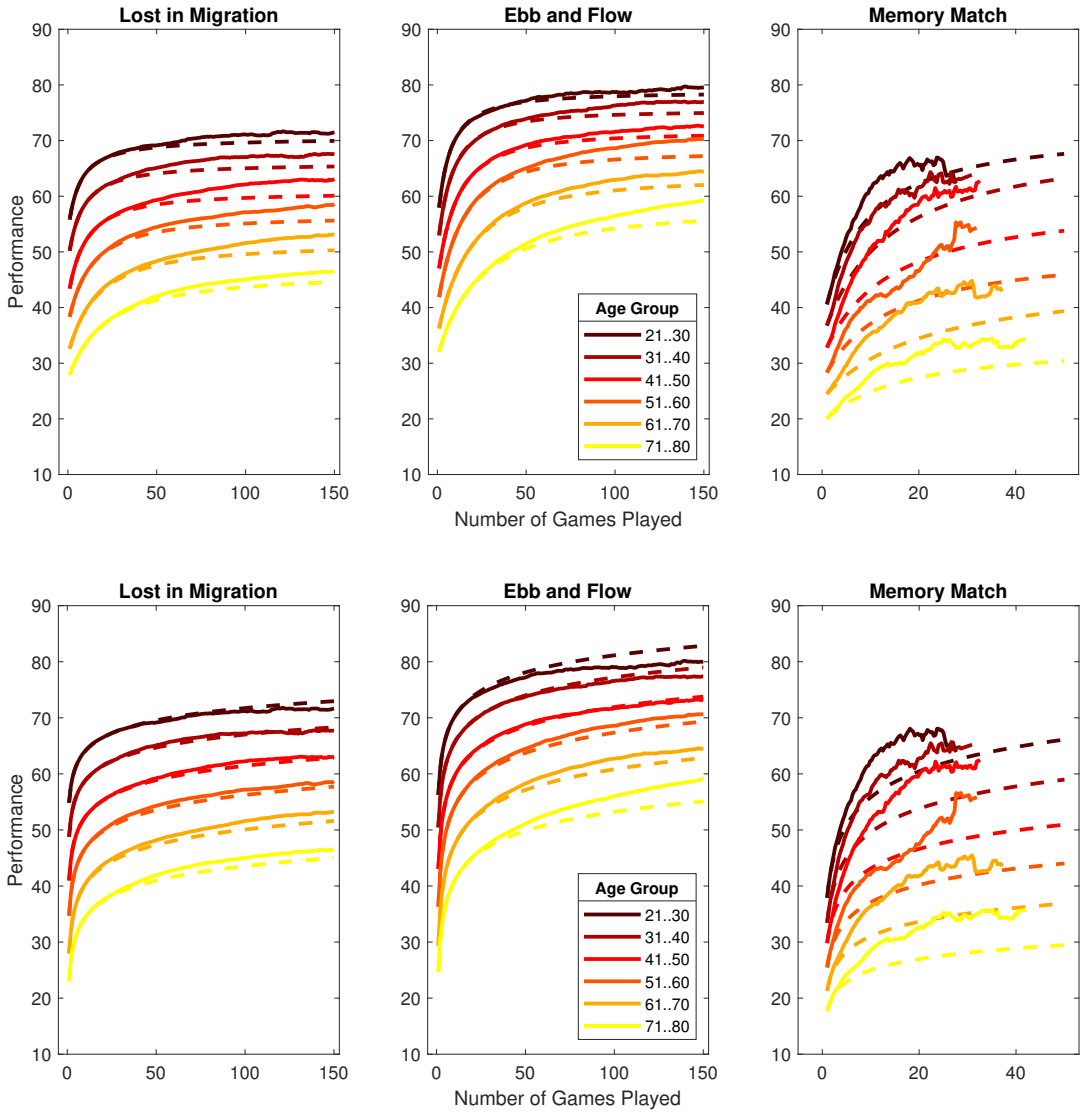

Fig. 8: Aggregate of individual model-based learning curves across age groups based on Exponential model (top) and Power model (bottom). Dashed lines indicate the predicted learning curves aggregated across all users, regardless of dropout. Solid lines represent the empirical aggregated learning curves that do not take dropout into account. For the latter curves, averages based on fewer than 20 observations were omitted from the visualization.

mance and quitting (Agarwal, Burghardt, & Lerman, 2017). The Lumosity data will provide a challenging data set to develop a broad computational framework for learning and dropout.

In this work, we used simple learning functions to capture the global changes in performance over time. It is possible that the model results depend on the definition of the performance measure. Currently, we focused on the number of correct decisions per game play but previous learning curve analyses (e.g. Heathcote et al., 2000) have focused on response time which is closely related to the inverse of our current performance measure. Future modeling work will have to investigate whether the model selection results are sensitive to the choice of performance measure.

The Lumosity data can be used to develop and test extensions of learning models to capture more complex aspects of learning dynamics. For example, additional parameters could be added to the learning functions to capture delays in the onset of learning (Evans et al., 2018). In addition, learning could be characterized by multiple piece-wise learning functions to capture different phases of learning (Donner & Hardy, 2015). The current learning functions represent time discretely but it is likely that a more complex learning model that explains learning as a function of actual elapsed time will

explain some variability in performance currently unaccounted for in our models. For example, such a model could explain learning dynamics at short time scales (e.g., within individual sessions when users play several games consecutively) as well as longer time scales (e.g. between sessions). Finally, future modeling work will have to investigate the learning dynamics across different games. Lumosity users typically interweave their practice of various games and a full account of learning will need to explain the joint performance over all games as a function of time.

One important challenge when analyzing naturally occurring data sets such as Lumosity is understanding the causal relationships between the uncontrolled factors that relate to behavior (Goldstone & Lupyan, 2016). For example, the results of Figure 4 show not only that early and late dropouts lie on different learning trajectories, but also that their initial performance differs as well. There are a number of causal interpretations for these initial performance differences. One explanation is that early performance feedback affected later participation decisions. Users who are adept at the game and feel reinforced by the feedback might stick with the game longer. In this case, there is a direct causal link between initial performance and participation decisions. Another explanation is that an unobserved variable mediates initial performance and dropout. Users who have better self-control or grit (Duckworth & Gross, 2014) are more likely to stick with the game, and these cognitive and motivational factors have made them better at related skills which enables them to perform better initially on a new task. Disentangling these causal relationships might require a combination of approaches. Obviously, conducting laboratory experiments that control some of the underlying factors could shed some light on the underlying effects. However, a number of modeling techniques could be pursued on the existing data to test the adequacy of these different causal assumptions. For example, if user self-control and grit are the causal forces that are responsible for extended learning and subsequent improved generalization across different games, we should be able to observe dependencies across games – users who finish playing one game after extended practice should be at an advantage at the start of practice for another game.

In addition to testing different causal relationships, the use of computational models can be helpful to test what-if scenarios. For example, the results in Figure 8 were used to simulate the counterfactual scenario in which users who dropped out actually continued to learn. These simulations can be used to predict the amount of training necessary for one individual to surpass another, or to reach a predetermined goal. Although these model predictions cannot be confirmed without actual data collection, the use of models is helpful to explore the space of possibilities for future empirical investigations and decide which experiments are most informative.

Generally, we believe that naturally occurring data such as the Lumosity data set will push the development of cognitive theory and computational modeling to exciting new directions of self-regulated learning, metacognitive control and self-assessment. It may also naturally lead to connections to other fields in psychology in order to understand individual difference factors related to motivation, effort and self-control.

Acknowledgements We would like to thank Bob Schafer, Noah Schwartz, and Elisbeth Cordell from Lumos Labs for providing the data and helpful discussions, and Tom Stafford and Andrew Heathcote for comments and suggestions that helped to improve the paper.

References

Agarwal, T., Burghardt, K., & Lerman, K. (2017). On quitting: Performance and practice in online game play. In *Proceedings of 11th aaai international conference on web and social media.* AAAI.

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought.* Psychology Press.

Benjamin, A. S. (2007). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. *Psychology of learning and motivation*, *48*, 175–223.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: benefits of individual metacognitive control on crowd performance. *Computational Brain and Behavior*, *1*, 90-99.

Bennett, S. T., Benjamin, A. S., & Steyvers, M. (2017). A bayesian model of knowledge and metacognitive control: Applications to opt-in tasks. In *Proceedings of the 39th annual conference of the cognitive science society* (p. 1623-1628). Austin, TX: Cognitive Science Society.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, *76*(1), 1–32.

Cassey, P. J., Gaut, G., Steyvers, M., & Brown, S. D. (2016). A generative joint model for spike trains and saccades during perceptual decision-making. *Psychonomic bulletin & review*, *23*(6), 1757–1778.

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review*, *22*(5), 1308–1319.

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current Directions in Psychological Science*, *23*(5), 319–325.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134.

Evans, N. J., Brown, S. D., Mewhort, D. J., & Heathcote, A. (2018). Refining the law of practice. *Psychological review*, *125*(4), 592.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in cognitive science*, *8*(3), 548–568.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Gureckis, T., & Markant, D. (2012). A cognitive and computational perspective on self-directed learning. *Perspectives in Psychological Science*, *7*, 464–481.

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185–207.

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*, *9*(2), 437–466.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*(3), 297.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*(1), 1–64.

Lee, M. D. (2004). A bayesian analysis of retention functions. *Journal of Mathematical Psychology*, *48*(5), 310–321.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review*, *124*(6), 762.

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of mathematical psychology*, *44*(1), 190–204.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & cognition*, *28*(5), 832–840.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing about knowing*, 1–25.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1–55.

Okada, K., Vandekerckhove, J., & Lee, M. D. (2018, Feb 01). Modeling when people quit: Bayesian censored geometric models with hierarchical and latent-mixture extensions. *Behavior Research Methods*, *50*(1), 406–415.

Park, D. C., & Schwarz, N. (2000). *Cognitive aging: A primer*. Psychology Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological review*, *109*, 472.

Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (under review). Is there a k in capacity? evaluating the discrete-slot model of visual short-term memory capacity.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

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.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wixted, J. T. (2004). On common ground: Jost's (1897)

law of forgetting and ribot's (1881) law of retrograde amnesia. *Psychological Review*, *111*(4), 864-879.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.