Prediction of Users' Learning Curves for Adaptation while Using an Information Visualization

Sébastien Lallé, Dereck Toker, Cristina Conati, Giuseppe Carenini
Department of Computer Science
University of British Columbia, Vancouver, Canada
{lalles, dtoker, conati, carenini}@cs.ubc.ca

ABSTRACT

User performance and satisfaction when working with an interface is influenced by how quickly the user can acquire the skills necessary to work with the interface through practice. Learning curves are mathematical models that can represent a user's skill acquisition ability through parameters that describe the user's initial expertise as well as her learning rate. This information could be used by an interface to provide adaptive support to users who may otherwise be slow in learning the necessary skills. In this paper, we investigate the feasibility of predicting in real time a user's learning curve when working with ValueChart, an interactive visualization for decision making. Our models leverage various data sources (a user's gaze behavior, pupil dilation, cognitive abilities), and we show that they outperform a baseline that leverages only knowledge on user task performance so far. We also show that the best performing model achieves good accuracies in predicting users' learning curves even after observing users' performance only on a few tasks. These results are promising toward the design of user-adaptive visualizations that can dynamically support a user in acquiring the necessary skills to complete visual tasks.

Author Keywords

Information visualization; learning curve; eye tracking; machine learning; user-adaptation

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Information visualization (Infoviz) is gaining importance as a means for analyzing large and complex datasets. There is evidence that user performance and satisfaction in working with visualizations can be influenced by individual differences such as cognitive abilities (e.g., perceptual speed, visual working memory, and verbal working memory [13,17,41]), or personality traits (e.g., locus of control [23]). Thus, researches have started investigating how information visualization tools could be made more effective and usable by enabling them to adapt in real-time to some of these user characteristics. In this paper, we extend this line of work by focusing on another user characteristic that can influence the user's experience with a new visualization: the user's ability to acquire through practice the skills necessary to work with the visualization at best. In particular, we investigate whether a user's skill acquisition ability can be predicted in real-time while the user is working with a visualization. If accurate prediction is possible early on during interaction, adaptive interventions could be devised to support users who appear to be slow in learning the necessary skills.

Learning curves are mathematical models well-studied in cognitive psychology to model skill acquisition. They represent the relationship between practice and the associated changes in behaviour, such as the evolution of users' proficiency over time [e.g., 37]. For instance, in the context of Infoviz, a user's learning curve captures the user's initial level of expertise with a given visualization as well as how fast the user will learn the set of skills required to perform tasks with it. If we could predict and track these characteristics (initial expertise, learning speed) while users acquire a new set of visualization skills, individualized support could be provided in order to improve user performance and engagement. In this paper, we investigate the feasibility of predicting a user's learning curve while the user is performing tasks with ValueChart, a relatively complex, interactive visualization to support decision making.

In educational systems, modeling and tracking domain skill acquisition over time based on users' observed proficiency have been used to design real-time adaptive strategies to support learning. For instance, Intelligent Tutoring Systems (ITS) can provide visual, verbal or textual help, or suggest exercises of adapted difficulty depending on user/task features and the user's estimated mastery of domain skills [8,31]. To the best of our knowledge, the only work in Infoviz done on modeling learning skills is preliminary work presented in Toker et al. [42], where users performed visualization tasks with simple bar graphs. In that work, Toker et al. predict in real time the user's skill acquisition level in a coarse binary way (during learning vs. after

IUI' 2015, March 29–April 1, 2015, Atlanta, GA, USA.

ACM 978-1-4503-3306-1/15/03…$15.00.
http://dx.doi.org/10.1145/2678025.2701376

learning). In contrast, in this paper, we intend to model the user's learning curve, i.e., a more detailed description of the user's learning experience with a visualization. We ultimately aim to design user-adaptive visualizations able to predict user's learning curves in real time and use this information to customize the visualization accordingly. For instance, we envision supporting users according to their predicted individualized learning curve in the following manner:

- If a user is predicted to have low initial expertise and low learning rate, support could be provided by simplifying the visualization in order to ensure that necessary basic skills are learned before the user becomes overwhelmed or confused.
- If a user is predicted to have high initial expertise and high learning rate, the system can engage the user with more advanced functionalities within the current visualization or possibly even offering more complex visualizations.

Predicting learning curves is particularly relevant at the beginning of the interaction with a novel visualization, as providing early adaptation can reduce disengagement and improve learning. In its most basic form, the simplest way to model learning curves is to observe past task performance. However, various data sources can also be leveraged to make these predictions in real-time. For example, Steichen et al. [41] highlighted the potential of using eye tracking to predict user's characteristics in Infoviz. Pupil dilation is also a data source worth investigating since it has been shown to be influenced by cognitive workload [27]. Other data sources that are plausible candidates for predicting user skill acquisition are long-term user characteristics, such as cognitive abilities and personality traits.

The goal of this paper is to compare the performance of models that leverage these various data sources, as well as basic information on past task performance, to predict users' skill acquisition in real-time. As a test-bed for our investigation, we use an interactive visualization for multiple-criteria decision making called ValueChart.

The rest of the paper is organized as follows: Section 2 gives an overview of the related work; Section 3 describes ValueChart and the study that generated the dataset used in our experiments; Section 4 describes how we build the learning curves that we aim to predict; Section 5 details the models we built to predict the curves; Section 6 discusses the performance of these models; and Section 7 concludes.

RELATED WORK

A typical method used in cognitive psychology for tracking how user performance improves with practice is by using a learning curve [40]. Learning curves are also frequently used in HCI for off-line comparison and evaluation of alternative interfaces, including information visualization systems, e.g., [35,39,45]. In contrast, in this paper we use the concept of a learning curve for building *predictive models* that can identify in real-time a user's evolving proficiency with an information visualization system.

Similar work has been extensively conducted in the field of Intelligent Tutoring Systems (ITS). In ITS, learning curves have been used to track and adapt help policies to a student's evolving skills in the target educational domain (e.g., performing one and two digit subtraction for a math tutor), based on her past interactions with the ITS tracked via action logs (e.g., [7,36,44]). In contrast, we aim to track and adapt to a user's evolving proficiency in using a visualization interface, leveraging gaze data as the main information source.

Some work exists in HCI on adapting to a user's level of familiarity with an interface. For example, [12] designed a mixed-initiative GUI-customization tool that provides suggestions on how to personalize the menus of a word processor based, among other factors, on the user's expertise with the word processor. However, the ability to track such expertise in real time was not implemented. Other work in HCI has focused on predicting user skills in a coarse binary way. For instance, Ghazarian et al. [21] built models to automatically classify a user's general skill in using different computer applications, based on mouse/keyboard logs and interface events such as menu selection, in order to adapt the complexity of the interfaces to novice/expert users. Hurst et al. [26] proposed a method to detect skilled vs. unskilled use of an image editing program by investigating menu usage and mouse logs, allowing them to design a user-adaptive menu based on the prediction of skilled behavior. In contrast in this paper, we look at user's learning with a visualization over time, by predicting continuous learning curves, rather than a binary categorization of users (i.e., skilled vs. unskilled).

Gaze data has been extensively used to detect different kinds of user states during interaction with an ITS, such as boredom, curiosity, disengagement [19,28], mind-wandering [10], as well as domain learning [11,29]. [9] has also used gaze data to predict users' problem-solving strategies as well as user performance while solving a visual puzzle. In addition, pupil dilation has been reliably shown to vary depending on changes in cognitive load [6,24]. In the context of building user-adaptive systems, [27] used pupil dilation measures to evaluate cognitive workload during route planning and document editing tasks in order to identify opportune moments for interrupting the user. Similarly, [37] monitored pupil dilation in order to predict user preferences when confronted with a choice of visually presented objects, and [33] tracked pupil dilation features in order to infer skills related to reading comprehension.

Within the Infovis community, eye tracking has been used off-line to understand how users with different domain expertise process relevant visualizations, e.g., [16,34]. Gaze data has also been investigated to predict long-term user traits (e.g., perceptual speed, visual working memory, verbal working memory, locus of control), as well as task type and task completion time [13,17,41]. Past studies have shown that user characteristics themselves (e.g., cognitive abilities, personality traits) can predict how well a user will perform on, or prefer, a given information visualization system [13,17,23,41].

The first attempt to predict interface skill acquisition in Infoviz is described in [42], where they evaluated users performing a series of low-level tasks using bar graphs. User performance on these tasks was categorized into two general phases of skill acquisition: *during* and *after*; indicating whether users were still in the process of acquiring, or had already acquired, the relevant visualization skills. These labels were derived based on performance across all users, as opposed to being customized to individual user performance. In contrast, in this paper, we extend that work not only by predicting actual continuous learning curves (as opposed to much coarser skill acquisition phases), but also by making predictions that are specific to each individual user (as opposed to being only based on pooled data). Furthermore, we look at skill acquisition within the context of a more complex and interactive visualization (ValueChart), as opposed to the simpler bar graphs used in [42].

VALUECHART AND USER STUDY

The dataset used in this paper was collected from a user study using ValueChart[1], an interactive visualization to support decision-makers in preferential choice, namely selecting the best option out of a set of alternatives characterized by a variety of attributes [14,15]. Figure 1 shows an example of ValueChart for selecting rental properties among ten available alternatives (listed in the leftmost column), based on a set of relevant attributes (e.g., location, bus distance, appliances, etc.). These attributes are arranged hierarchically in the top part of the ValueChart, forming the columns in the central part of the display. The width of each column indicates the relative weight assigned to the corresponding attribute (e.g., utilities is less important than rent). The available alternatives (i.e., rental homes) are represented as the rows in the display. Each cell specifies how the alternative in that row fares with respect to the attribute in that column, indicated by the amount of filled cell color. In the rightmost part of the ValueChart, all values for each alternative are accumulated and presented as horizontal stacked bars, displaying the overall value of each alternative (e.g., home4 is the best home in the example). Some of the interactive functionalities available to support the decision process include *inspecting* the specific domain value of each attribute (e.g., the rent of home1 being equal to $500), *sorting* the alternatives with respect to a specific attribute, *swapping* attribute columns, and *resizing* the width of an attribute's column to see how that would impact the decision outcome.

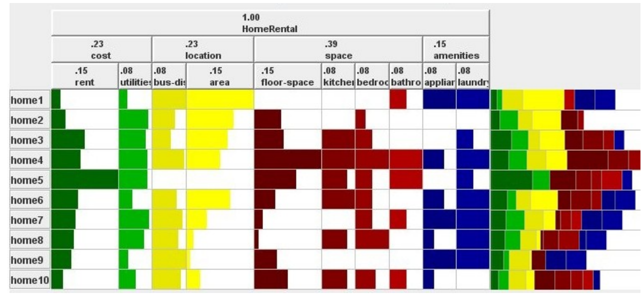

Figure 1: An example of the main elements of the ValueChart visualization, here displayed in a horizontal layout [17].

For the ValueChart user study (fully described in [17]), 95 participants were recruited (ages 16 to 40) to perform 5 different types of visualization tasks, chosen from a set of low-level data analysis tasks defined by Amar et al. [2]. These five tasks (shown in Table 1) require answering questions from different domains for preferential choice (i.e., rental homes, universities, cell phones, restaurants, and hotels) using functionalities of ValueChart (e.g., sorting, reordering, weighting attributes). For each of these tasks, Table 1 shows their definition from [2], a sample question from the study, and the conceptual operations involved in answering the question. These include both lower-level mathematical and cognitive actions (e.g. *generate aggregate value*, *compare values*) [2], as well as instances of the five Amar et al. tasks themselves (e.g., *compute derived value* typically requires multiple preceding *retrieve values*). The number of conceptual operations in Table 1 reflects this additional layer of actions per task type, and gives a conceptual measure of complexity (see [17] for more information). Thus, the quantities and types of conceptual operations in a task, in conjunction with the current level of skill a user has with these operations, will impact task performance. The specific operations shown in Table 1 are based on our study tasks, which always involved domains with 10 different alternatives and 10 attributes.

Participants repeated each task in Table 1 four times in a randomized fashion to account for within-user variability. For purposes outside the scope of this paper (described in [17]), two ValueChart layouts were evaluated: vertical vs. horizontal. The 20 study tasks were first performed with one of the two layouts for one domain, and then repeated with the second layout and a different domain, with order fully counterbalanced across users. Thus each participant performed a total of 40 tasks (5 task types x 4 repetitions x 2 layouts) equally divided in two sessions (one per layout). While performing these tasks, the use's gaze was tracked with a Tobii T120, a non-intrusive eye-tracker embedded in the study computer monitor. Each user also performed two decision making tasks in the study, but we will not consider

[1] Video demo: www.cs.ubc.ca/group/iui/VALUECHARTS

Task type	Task Definition from [2]	Sample task question from study	Conceptual operations	Mean response time (st.dev.)
Retrieve Value	Given a set of specific cases, find attributes of those cases.	Is the value of 'skytrain-distance' of home3 less than home6?	2 Retrieve values, 2 Compare values	15sec. (10)
Find Extremum	Find cases possessing an extreme valued attribute over its range within a data set.	What factor contributes the most towards the overall value of home4?	10 Retrieve values, 10 Compare values, 1 Retrieve labels	19sec. (14)
Sort	Given a set of cases, rank them according to some ordinal metric.	List the top 3 homes (in descending order) according to overall value.	10 Retrieve values, 10 Compare values, 1 Retrieve labels	17sec. (9)
Compute Derived Value 1	Given a set of data cases, compute an aggregate numeric representation of those data cases.	For how many homes is the 'rent' less than the 'rent' of home3?	10 Retrieve values, 10 Compare values, 1 Generate aggregate	20sec. (11)
Compute Derived Value 2		List the top 3 homes (in descending order) according to the aggregated value of 'cost' and 'space'.	20 Retrieve values, 10 Generate aggregates, 10 Compare aggregates, 3 Retrieve labels	42sec. (27)

Table 1: Descriptions of the five task types [17].

them in this paper because we don't have a sufficiently precise definition of users' performance for those tasks.

PREDICTION TARGET: LEARNING CURVES

Our goal is to predict a user's learning curve as they perform tasks with ValueChart, i.e., predict the curve parameters that model the acquisition of visualization skills relevant to work with this particular visualization.

A learning curve is a mathematical function that models changes in performance over time, as the amount of practice with a given activity increases. In psychology, a common approach to model learning curves is by fitting a *power law* function defined as follows [3,4,5]:

$$Y = a \times X^b \quad (1)$$

Where X is a variable ranging over the number of tasks performed by the user; Y is the observed performance of the user over X; a is the intercept denoting the initial expertise of the user; and b the slope representing the learning speed (a slope of zero indicates no apparent learning).

In Infoviz, two performance measures typically used for Y are: task accuracy (percentage of correctly/incorrectly performed tasks), and response time (i.e., time needed to complete a task). In this paper, we selected response time because there was a ceiling effect of accuracy over the study tasks.

In the next two subsections, we evaluate two alternative ways to define our target learning curves. The first assumes that the different types of tasks performed by the users are unknown when building the curve (*task-independent*), whereas the second requires this information to be available, and thus generates what we will call from now on *task-dependent* learning curves. Notice that both task–independent and task-dependent learning curves are models of the same data, i.e., the performance of the user on all the trials. The only difference between the two is in how the curves are fitted to the data. In the end, we chose the task-dependent model as our gold standard, because it yields a better fit of the data.

Task-independent learning curves

In order to get learning curve coefficients (i.e., the intercept a and the slope b in (1)), we simply fit a learning curve for each user using the power law function. A user's task-independent learning curve is fit by pooling together, with order maintained, the different types of tasks that were performed by that user. Due to the fact that users performed the same battery of tasks in each of the two study sessions (corresponding to using different ValueChart layouts, cf. supra), we average trials from the two sessions that were performed in the same order and use the resulting 20 data points per user to fit the learning curve for that user. Figure 2 shows an example of a learning curve for one user, where the red dots are the response times (in order of completion) for each of the 20 tasks performed by the user, and the black curve is the power law function fit over the red dots.

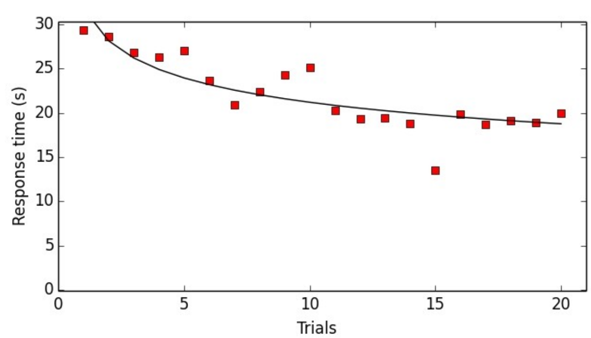

Figure 2: A sample learning curve for one user based on task response time in the ValueChart study.

The learning curves we obtained for the users in our study indicated that 88 users improved over time (negative slope), whereas 7 users showed no learning (zero or positive slope). The upper rows of Table 2 report the following summative statistics of the fitted task-independent learning curves: r^2 (a measure for how well a learning curve matches the underlying performance data); the fitted curve's intercept and slope (i.e., initial expertise and learning rate); and *final expertise*, i.e., the performance (here response time) expected to be achieved by the user at the end of the trial sequence. As Table 2 shows, for task-independent curves the fit is rather low (mean r^2=.33, std. dev= .25), indicating that the power law can only partially model a user's overall acquisition of visualization skills when information on user tasks is not taken into account. In our study, a plausible explanation for this low fit is that the complexity of the different task types varies considerably, as discussed in the previous section. Thus, different task types may generate different learning behaviors within users, which cannot be accurately captured by the task-independent approach described above. To address this problem, in cognitive psychology, learning curves are often fit for different activities, for instance based on the knowledge required to complete them [3,4], or using an additional parameter to model tasks difficulty [18]. In the next sub-section, we explore a similar approach by building individualized learning curves derived from fitting separate learning curves for each of the five different task types administered in the study (Table 1) and by then aggregating them.

Task-dependent learning curve

A task-dependent learning curve for a given user is computed by averaging the coefficient of her learning curves across all five task types as follows. First, we build five learning curves, one for each task type k:

$$Y = a_k \times X^{b_k} \quad k \in \{1..5\} \quad (2)$$

where, in each of the five learning curves, X is a variable ranging over the number of times the corresponding task type was performed by the user, Y is the observed response times, a_k is the intercept for task type k, and b_k the slope of task type k.

Next we compute: $a = avg(a_k)$ and $b = avg(b_k)$ with $k \in \{1..5\}$

Then, the task dependent learning curve can be again simply expressed as: $Y = a \times X^b$

The lower rows of Table 2 shows summary statistics for the learning curves obtained using this task-dependent approach. As expected (mean r^2 = .69, std. dev= .20) shows a better fit compared to the task-independent approach. We did find that the intercept coefficients for the task-independent and task-dependent curves are highly correlated (r^2 = .80), whereas slopes are not (r^2 = .45), confirming that learning speeds are different among tasks. Additionally, the correlation between values of final expected expertise (i.e., the predicted response time of the last trial) is high (r^2 = .75) indicating that the task-independent curves still provide an acceptable approximation for this measure despite the lower accuracy in predicting the learning trajectory that a user will follow to reach this expertise.

Given the overall better fit of the task-dependent curves, we use them as the target for our predictive models in the rest of the paper.[2]

		Min	Max	Mean	Std.dev
Task-independent	**r^2**	0.03	0.78	0.33	0.25
	Intercept	15	80	32.6	13.3
	Slope	-0.9	0.05	-0.18	0.15
	Final Expertise	3	40	18.6	5.8
Task-dependent	**r^2**	0.01	0.96	0.69	0.2
	Intercept	23	89	41.4	12.3
	Slope	-1.7	-0.02	-0.4	0.24
	Final Expertise	5	41	21.8	6.1

Table 2: Summary statistics of two approaches for defining individualized learning curve coefficients.

MACHINE LEARNING EXPERIMENTS

Our goal is to ascertain whether we can predict a user's skill acquisition process by predicting the intercept and the slope of that user's learning curve by using different data sources as predictors. This section describes these data sources, the machine learning models we built, and how they are evaluated.

Data Sources

Here we outline three feature sets that will be used as predictors for inferring learning curve parameters. One feature set consists of measures that summarize a user's gaze patterns, as tracked by the Tobii T-120 eye tracker during the study (*Gaze* feature set). A second set consists of measures that describe changes in a user's pupil size during tasks, also based on raw data provided by the eye-tracker (*Pupil* feature set). A third feature set models a variety of user's long-term characteristics that have been shown to impact visualization processing (*User Characteristics* feature set).

***Gaze* feature set.** Gaze data is captured by the Tobii eye-tracker in terms of fixations (gaze maintained at one point on the screen), and saccades (quick movement of gaze from one fixation point to another). We then processed this raw data using EMDAT, a gaze data analysis toolkit[3] to generate a battery of gaze-based features summarized in Table 3. Some of these features capture overall gaze activity on the screen (*Overall features*, top half of the table), while others do so for specific Areas of Interest (AOI) in the visualization (*AOI features*, middle part of Table 3). The seven AOIs defined for ValueChart in this study are shown in Figure 3. In total, we have 135 Gaze features.

[2] It should be noted, however, that we obtained results similar to those reported in the later sections of this paper when predicting task-independent learning curves.

[3] EMDAT: http://www.cs.ubc.ca/~skardan/EMDAT

Overall Gaze Features
Fixation rate
Mean & Std. deviation of fixation durations
Mean & Std. deviation of saccade length
Mean & Std. deviation of relative saccade angles
Mean & Std. deviation of absolute saccade angles
AOI Gaze Features (for each AOI)
Fixation rate in AOI
Longest fixation in AOI
Proportion of time, Proportion of fixations in AOI
Number & Prop. of transitions from this AOI to every AOI
Pupil Features
Mean, Std. deviation, Maximum, Minimum pupil width
Pupil width at the first and last fixation in a given trial

Table 3: List of Gaze, AOI, and pupil features.

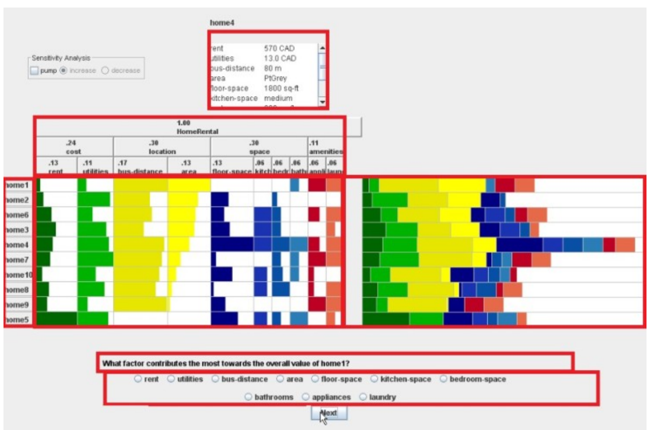

Figure 3: Areas of Interest (AOI) defined over the complete visualization interface for horizontal ValueChart.

***Pupil* feature set**. The Tobii eye-tracker records the user's pupil diameter (the horizontal width of each pupil) at each sample. We once more used EMDAT to compute a variety of features that describe the pupil diameter over the span of a task (bottom part of Table 3) for a total of 6 features. In order to avoid possible confounds on pupil size due to lighting changes, the study was administered in a windowless room with uniform lighting. To compensate for physiological differences in pupil size among individual users, we also collected pupil diameter baselines for each user by having them stare at a blank screen for ten seconds at the beginning of each session. Measured pupil dilation is adjusted using the *percentage change in pupil size* (PCPS), which is defined in [27] as:

$$\frac{\text{measured\_pupilsize} - \text{baseline\_pupilsize}}{\text{baseline\_pupilsize}} \quad (3)$$

***User characteristics* feature set**. Several long-term user traits were measured via standard tests administered to participants at the beginning of the study. We measured three cognitive abilities: *perceptual speed* (a measure of speed when performing simple perceptual tasks), *verbal working memory* (a measure of storage and manipulation capacity of verbal information), *visual working memory* (a measure of storage and manipulation capacity of visual and spatial information). We also measured the personality trait known as locus of control (a measure of the extent to which a person believes they are able to control events affecting them)[4]. We chose these particular user characteristics because other studies have shown that they impact the effectiveness of visualization processing [13,17,41], and thus they may likely also affect skills acquisition related to this processing.

Prediction models

In this subsection, we first present the method used to produce a baseline for predicting individualized learning curve coefficients (intercept and slope) based solely on task performance from previous tasks. Next, we describe the predictive models that leverage gaze data, pupil dilation data, and user characteristics as input data sources. It should be noted that all models in this section are built without considering information on which task type a user is performing, as this information may not always be available to a visualization system when a user is working with it.

Response-Time Baseline Model

A simple way to infer a user's learning curve is to track their task response times so far. To achieve this, as each task in our dataset is completed, we re-fit the learning curve function based on all the trials seen so far for that user. The result are two *temporary* learning curve coefficients which can be used as predictions of the actual final learning curve coefficients for that user. For instance at the end of trial 6, we fit a learning curve per user using only their 6 first completed trials. We use this approach as our baseline model since it requires only basic information regarding user task performance. At least two trials are required to fit a temporary learning curve this way.

Machine Learning Models

We want to ascertain whether we can achieve a better prediction of a user's learning curve than that afforded by the basic response-time baseline by using as predictors data sources not linked to task response time: gaze, pupil, and user characteristics. In particular, we evaluate each of the three feature sets (gaze, pupil, user char.) individually, along with a feature set with all three combined. Predictive models based on gaze and pupil features are built over increasing numbers of consecutive trials, from 1 trial to 20(all) in order to ascertain how model performance depends on the amount of evidence seen. This does not apply for the model based only on user-characteristics, because these features are static and do not change over trials. Thus, we built a total of 3 (feature sets) x 20 (trials) x 2 (learning curve coefficients) = 120 models utilizing gaze

[4] The standard tests are, for PS: Kit of Factor-Referenced Cognitive Tests-P3 [20]; for verbalWM: OSPAN test [43]; for visualWM: Luck & Vogel's test [32]; for Locus of Control: Rotter's test [38].

and pupil features, plus one model for user characteristics per coefficient, yielding a total of 122 models.

We used backward stepwise linear regression to build our models using Akaike information criterion (AIC) optimization [1] to fine-tune our models as well as handle the high number of gaze features. Models were trained and tested via 10-fold cross validation over users, namely for any given fold 90% of the users in the study are in the training set and 10% are in the test set. The model that generates predictions for learning curve coefficients after seeing n trails in the study trial sequence (where n varies from 1 to 20) is trained over features of these n trials pooled over all users in the training set. This model is then used to predict the coefficients of the individualized learning curve of each user in the test set. We used the R software environment for statistical computing to evaluate and compare models.

Model performance is measured via the root-mean-square error (RMSE) of the difference between the predicted learning curve coefficients and the actual ones. That is, the RMSE for each of our 122 models is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{u=1}^{|U|}(\hat{y}_u - y_u)^2}{|U|}} \quad (4)$$

where U is the set of users in the study, $\hat{y}_u$ represents the predicted coefficient for a given user u, and y_u represents the actual target coefficient for that user.

RESULTS

In this section, we report and discuss the performance of our models in predicting individualized learning curve coefficients (intercept and slope), compared to the baseline derived solely from tracking users' response-times. We then discuss the ten features with highest predictive power for the best performing model. Finally, we report accuracy of this model in performing binary classification over the target coefficients (i.e., low/high intercept and slope) in order to give a more practical measure of model's reliability when used to guide adaptive interventions.

Predicting learning curves' coefficients

Figure 4 shows the RMSE for predicting learning curve *intercept* when the models described in the previous section are trained over different number of trials. Figure 5 shows the analogous results for *slope*.

To formally compare the effectiveness of the different models, we use each model *over-time performance*, i.e., RMSE averaged across the twenty trails. We then run pairwise comparisons between models using Bonferroni-adjusted t-tests with over-time performance as the dependent measure. Table 4 summarizes the results of these comparisons by ordering models according to their overtime RMSE, bold underlining indicates models for which there are no statistically significant differences (here statistical significance is reported at $p < .05$). For example, the comparisons for predicting Intercept coefficients shown in Table 4 indicate that All-features is better than Gaze which is better than UserChar. UserChar is better than Pupil but this difference is not significant, and Pupil is also not significantly better than the baseline model.

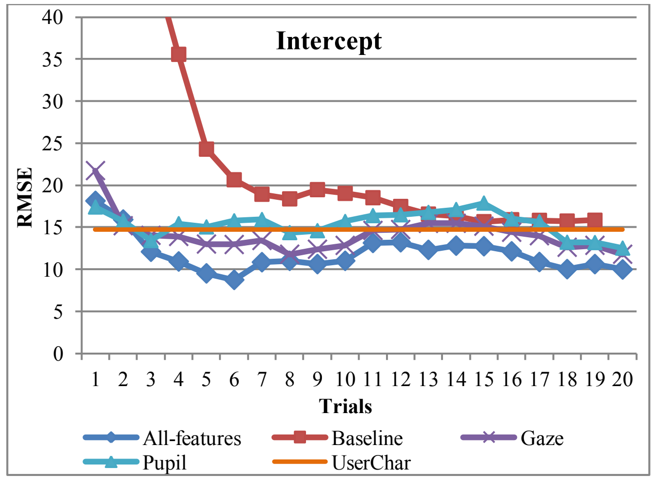

Figure 4: RMSE at each trial count for predicting learning curve intercept (lower values are better).

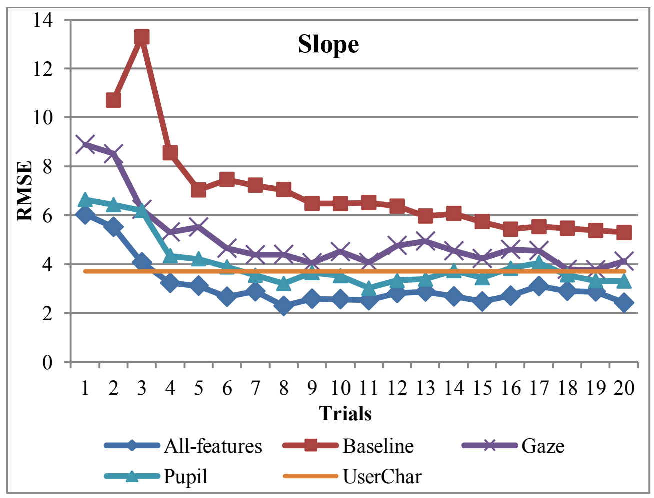

Figure 5: RMSE at each trial count for predicting learning curve slope.

Learning Curve	Comparison of models' performance
Intercept	All-features > Gaze > UserChar > Pupil > Baseline
Slope	All-features > Pupil > UserChar > Gaze > Baseline

Table 4: Effect of feature set on overall model performance across all trials.

The results for intercept and slope in Table 4 show that the *All-features* model has the best performance in both cases and is significantly better than the other models which consist of only one type of feature, including the baseline.

In fact all models statistically beat the baseline except for Pupil features for predicting intercept. These results indicate that eye tracking (gaze and pupil) as well as user characteristics are all valuable sources for the prediction of a user's learning curve and should all be considered together for optimal predictive performance. In terms of how early the prediction can be performed, the performance plots of the *All-features* model in Figure 4 and Figure 5 show that although the best performance (lowest RMSE) is achieved after 6 trials for intercept and 8 for slope, RMSE already drops considerably for both coefficients after seeing only 3 trials. These results indicate that eye tracking and user characteristics are valuable sources for early prediction of a user's learning curve. In practice early prediction of the slope can be used, for example, to support users who appear to be learning slowly in speeding up their learning process. Similarly, early prediction of the intercept can be used to support users with low initial expertise by, for instance, disabling advanced functionalities or recommending additional training examples.

If we look only at models built on one type of feature set, we can see that Gaze is better than both UserChar and Pupil for predicting intercept (i.e., initial expertise), whereas Pupil and UserChar are better than Gaze at predicting slope (i.e., learning rate). In other words, it seems that learning rate in visual tasks is better predicted by long-term cognitive abilities and pupil dilation features, whereas the initial expertise (intercept) can be captured best by gaze behavior. It is well established that pupil dilation is correlated to the level of a user's cognitive load [22], thus learning rate seems to be impacted by both long-term cognitive abilities (such as working memory or perceptual speed) as well as short-term cognitive load. In terms of adaptation, it means that customizing a visualization to reduce cognitive load could help slow learners.

Feature Selection: Exploring the Relative importance of features

As described in the previous section, we found that multiple data sources together can better infer a user's learning curve (i.e., *All-Features*), as opposed to any single source. In this section, we report the top selected features for this model in order to get a better sense of which features best predict the intercept and the slope of a user's learning curve, as well as the directionality of the relationships. To investigate the top 10 features, we averaged over all trials the relative importance of features at each individual trial. We applied the method described by Kruskal in [30] to get the relative importance of each feature, which consists in computing the R^2 contribution averaged over orderings among features for a regression model. Table 5 and Table 6 show the top ten selected features for intercept and slope respectively, where the relative importance is normalized so that the most important feature has a score of 100 and other scores express the relative importance proportionally to the top one. Additionally, a positive direction of the effect (D) indicates that initial expertise or learning rate is higher when the value of the feature increases. For the layout of the AOIs, refer to Figure 3.

Features predicting intercept	Set	D	Score
Fixation rate on 'data attributes'	Gaze/AOI	-	100
Perceptual speed	UserChar	+	71
Mean fixation duration	Gaze	-	68
Proportion of fixations on 'questions'	Gaze/AOI	-	41
Std.dev pupil size	Pupil	+	39
Proportion of time on 'questions'	Gaze/AOI	-	38
Fixations on 'items'	Gaze/AOI	-	32
Mean pupil size	Pupil	-	31
Std.dev saccade distance	Gaze	-	30
Transitions from 'data attributes' to 'data visualization'	Gaze/AOI	+	27

Table 5: Top 10 features for predicting intercept.

Features predicting slope	Set	D	Score
Perceptual Speed	UserChar	+	100
Std.dev pupil size	Pupil	+	84
Visual Working Memory	UserChar	+	74
Mean pupil size	Pupil	+	72
Std.dev fixation duration	Gaze	+	59
Transitions from the 'input' to the 'question'	Gaze/ AOI	-	49
Fixation rate	Gaze	+	29
Verbal Working Memory	UserChar	+	28
Proportion of fixations on the 'input'	Gaze/AOI	-	25
End pupil size	Pupil	-	18

Table 6: Top 10 features for predicting slope.

The preponderant features for predicting intercept are gaze related (7 out of 10) with five of these relating to AOIs, meaning that tracking gaze behavior on different parts of the interface is important to detect initial expertise. This result makes sense as users with different levels of initial expertise with ValueChart likely process the visual display differently up front. There are then two pupil-based features, mean and standard deviation of pupil size, both with positive directionality meaning that higher values for them correspond to higher initial expertise. Generally, increase in pupil size is correlated with higher cognitive load [22]. Thus the fact that higher mean pupil size relates to higher initial expertise may indicate that more advanced users are able to maintain an overall higher level of cognitive load, and higher std.dev pupil size indicates that more advanced users can more readily increase or reduce their mental processing depending on the tasks. Lastly, only one user characteristic, perceptual speed, appears in the top ten for intercept, although it has a prominent second place in the ranking.

In contrast, for slope three of the four user characteristics appear in the top 10 most predictive features: perceptual speed, visual working memory and verbal working memory. Since these are all cognitive abilities, it is not surprising that they are predictive of learning rate, and that they have positive directionality, namely that higher values of these abilities correspond to faster learning rate. These results suggest that measuring these user cognitive abilities prior to the interaction with a visualization is worthwhile if an adaptive system aims to predict learning rate. In particular, both perceptual speed and visual working memory are among the top three predictive features for learning rate, indicating that a big part of how quickly a user will learn depends on cognitive abilities relating to the capacity to process visual information. Given that these traits are typically considered fixed (i.e., a user has no control over them), we ought to design adaptive support which may 'ease' the visual load for users with low perceptual speed or visual working memory.

In general, the distribution of feature types in the top ten features for predicting learning rate is more balanced than for initial expertise: in addition to the 3 cognitive measures, there are 3 pupil-related features and 4 gaze-based features. Interestingly, only two of the four gaze features are AOI-related, suggesting that attention to specific areas of the visualization is not as predictive of learning rate as it is of initial expertise.

As far as pupil-related features are concerned, the two that appeared as top predictors for initial expertise also appear, with the same directionality, for learning rate suggesting that being able to maintain a high cognitive load, but have it vary with the demand of the task at hand, is predictive of faster learning rate. Interestingly, the third pupil-related feature in the top ten is one with a negative directionality with learning rate, namely end pupil size (i.e., size of pupil at the end of each trial). This finding suggests that this pupil measure provides an indication of excessive cognitive load that interferes with learning.

Binary classification

In this section we report the performance of the *All-Features* model in terms of binary classification accuracies for learning curves coefficients, to give a more practical measure of this model's predictive ability compared to RMSE. More specifically, we simulate the real-time classification of users into groups of fast/slow learners (slope) and high/low initial expertise (intercept).

Users are divided into two balanced groups for each coefficient using a median split: high/low intercept, and high/low slope. We compared different classifiers implemented in Weka [25] and selected the most promising one: Random forests tuned to 50 random trees. As we did for the prediction of the actual coefficients in a previous section, classification is carried out over incremental sequences of trials, from 1 to 20, using 10-fold cross validation. We report class accuracies and do not include a baseline accuracy since we have already shown in the previous section that the *All-features* model was always significantly better or equal to the baseline. Results for classification of intercepts are shown in Figure 6, those for slopes are shown in Figure 7.

T-tests between class accuracies across all trials shows no significant difference ($p > .05$) for either the intercept or the slope, indicating that both classifiers are well balanced (i.e., the random forest classifier can predict equally well high/low intercepts and high/low slopes). From a practical point of view, it means that adaptive strategies can be designed for each group of users.

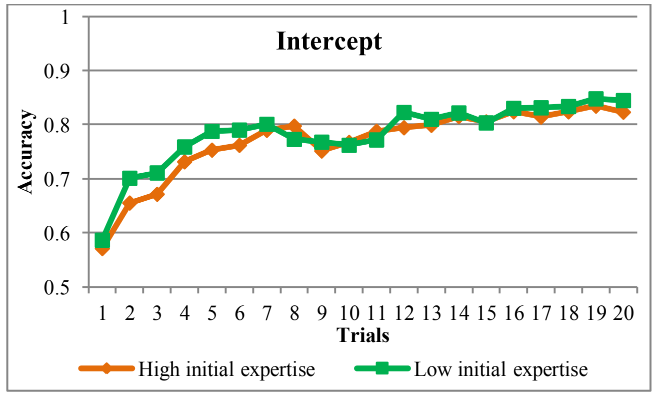

Figure 6: Class accuracies of binary predictions of intercepts using the *All-features* model.

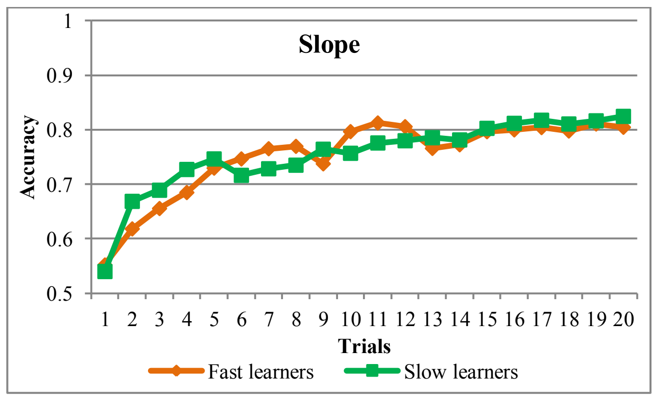

Figure 7: Class accuracies of binary predictions of slopes using the *All-features* model.

For the intercept, class accuracies range from 57% (after trial 1) to 83%, with an overtime accuracy of 77%. Regarding the slope, accuracies are slightly lower at the beginning (55% at trial 1) and reach as high as 82%, with overtime accuracy of 75%. We can notice that, for both coefficients, accuracies as high as 70% are reached after seeing only 4 trials. Very early predictions (i.e., trials 1-3) yield moderate accuracies ranging from 57% to 69% for the intercept, and from 55% to 67% for the slopes.

In terms of designing adaptive support for a visualization system, achieving accuracies of 70% after only 4 observed trials of data (which is on average 90 seconds of data) is promising in terms of inferring which users can benefit from tailored support.

CONCLUSION, DISCUSSION, AND FUTURE WORK

In this paper, we have studied the feasibility of predicting a user's learning curve while they perform a series of visual tasks using ValueChart, an interactive visualization for multiple-criteria decision making. Learning curves model a user's ability to learn new skills through practice, and in this paper we use them to model a user's initial expertise with a visualization, as well as their expected learning rate for the related skills. Our long-term goal is to leverage these predictions in user-adaptive visualizations that can tailor the interaction to a user's learning trajectory.

Whereas learning curves have been leveraged in the field of Intelligent Tutoring Systems to provide real-time adaptive support to learning a target educational domain, their usage in other areas of Intelligent User Interfaces has so far been limited. Toker et al. [42] have done preliminary work on using learning curves fit to population data to coarsely classify users as being either in the skill acquisition phase or having acquired necessary skills. We extended that work by predicting learning curves fit to individual users, thus achieving a finer-grained ability to track the user's skill acquisition process.

We showed that we can predict users' learning curve coefficients, i.e., the intercept (initial expertise) and the slope (learning rate) with substantial accuracy early on during the user's interaction with ValueChart. Our predictive models, which leverage different combinations of features including gaze behavior, pupil dilation, and cognitive abilities, significantly outperformed a simple but strong baseline model built on observed past performance. The best performing model overall leverages all of user gaze behavior, pupil dilation, and cognitive abilities to make its predictions. We described its performance both in terms of RMSE in predicting learning curve coefficients, as well as in terms of classifying users into binary groups (fast/slow learners; high/low initial expertise). On the latter measure, this model reached an accuracy of 70% after observing only 4 user tasks.

These results provide encouraging evidence that early prediction of user's skill acquisition is possible in information visualization, although it will of course be necessary to show the generality of these findings by replicating them with visualizations other than the ValueChart used in this study. Early prediction of skill acquisition, in turn, is important for our general goal of devising user-adaptive visualizations that can tailor information presentation to each user's individual needs. For instance, we plan to design and evaluate adaptation strategies for users predicted to have low initial expertise and low learning rate by either providing interventions that help these users identify and process the relevant elements of the visualization (for instance by highlighting relevant parts of the visualization), or by simplifying the visualization in order to ensure that necessary basic skills are learned before more advanced functionalities are available.

Another thread of future work relates to further improving our predictive models of skill acquisition. First, we expect that adding a measure of task performance to gaze, pupil, and user characteristics features can improve model accuracy. Second, we will study if stochastic models can reinforce predictions overtime. Lastly, we plan to investigate the addition of features based on interface actions, as ValueChart is an interactive visualization, and past work on combining gaze and action information has showed promising results [29].

ACKNOWLEDGEMENTS

This work was made possible thanks to funding from NSERC through the Grant: STPG381322-09.

REFERENCES

1. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 6 (1974), 716–723.
2. Amar, R., Eagan, J., and Stasko, J. Low-Level Components of Analytic Activity in Information Visualization. *Proceedings of the 2005 IEEE Symposium on Information Visualization*, IEEE Computer Society (2005), 15–21.
3. Anderson, J.R. *Cognitive Skills and Their Acquisition.* Psychology Press, 1981.
4. Anderson, J.R. *Rules of the Mind.* Psych. Press, 2014.
5. Bailey, C.D. and McIntyre, E.V. The relation between fit and prediction for alternative forms of learning curves and relearning curves. *IIE Transactions 29*, 6 (1997), 487–495.
6. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin 91*, 2 (1982), 276–292.
7. Beck, J.E. and Sison, J. Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education 16*, 2 (2006), 129–143.
8. Beck, J.E., Woolf, B.P., and Beal, C.R. ADVISOR: a machine-learning architecture for intelligent tutor construction. University of Massachusetts (2000).
9. Bednarik, R., Eivazi, S., and Vrzakova, H. A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data. In *Eye Gaze in Intelligent User Interfaces*. Springer, 2013, 111–134.
10. Bixler, R., Kopp, K., and D'Mello, S. Evaluation of a Personalized Method for Proactive Mind Wandering Reduction. *Proc. of the 4th Workshop on Personalization Approaches for Learning Environments, UMAP* (2014), 33–41.

11. Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R., and Bouchet, F. Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning. *Proc. of the 16th Int. Conference on Artificial Intelligence in Education*, Springer (2013), 229–238.
12. Bunt, A., Conati, C., and McGrenere, J. Supporting Interface Customization Using a Mixed-initiative Approach. *Proc. of the 12th International Conference on Intelligent User Interfaces*, ACM (2007), 92–101.
13. Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., and Enns, J.T. Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, ACM (2014).
14. Carenini, G. and Loyd, J. ValueCharts: analyzing linear models expressing preferences and evaluations. *Proceedings of the Working Conference on Advanced Visual Interfaces*, ACM (2004), 150–157.
15. Chamberlain, B.C., Carenini, G., Oberg, G., Poole, D., and Taheri, H. A Decision Support System for the Design and Evaluation of Sustainable Wastewater Solutions. *IEEE Trans. on Comp. 63*, (2014), 129–141.
16. Çöltekin, A., Fabrikant, S.I., and Lacayo, M. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science 24*, 10 (2010), 1559–1575.
17. Conati, C., Carenini, G., Steichen, B., and Toker, D. Evaluating the Impact of User Characteristics and Different Layouts on an Interactive Visualization for Decision Making. *Proceedings of the 16th Eurographics Conf. on Visualization*, (2014), 371–380.
18. Desmarais, M. Performance comparison of item-to-item skills models with the IRT single latent trait model. *Proc. of UMAP*, (2011), 75–86.
19. D'Mello, S., Olney, A., Williams, C., and Hays, P. Gaze tutor: A gaze-reactive intelligent tutoring system. *Int. J. of Human-Comp. Studies 70*, 5 (2012), 377–398.
20. Ekstrom, R.B. and Harman, H.H. *Manual for Kit of Factor-referenced Cognitive Tests, 1976.* Educational Testing Service, 1976.
21. Ghazarian, A. and Noorhosseini, S.M. Automatic detection of users' skill levels using high-frequency user interface events. *User Modeling and User-Adapted Interaction 20*, 2 (2010), 109–146.
22. Granholm, E. and Steinhauer, S.R. Pupillometric measures of cognitive and emotional processes. *Int. Journal of Psychophysiology 52*, 1 (2004), 1–6.
23. Green, T.M. and Fisher, B. Towards the Personal Equation of Interaction: The impact of personality factors on visual analytics interface interaction. *Proc. of the 2010 IEEE Symposium on Visual Analytics Science and Technology*, IEEE (2010), 203–210.
24. Hess, E.H. and Polt, J.M. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science 143*, 3611 (1964), 1190–1192.
25. Holmes, G., Donkin, A., and Witten, I.H. Weka: A machine learning workbench. *Second Australian and New Zealand Conference on Intelligent Information Systems*, (1994), 357–361.
26. Hurst, A., Hudson, S.E., and Mankoff, J. Dynamic Detection of Novice vs. Skilled Use Without a Task Model. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2007), 271–280.
27. Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., and Bailey, B.P. Towards an index of opportunity: understanding changes in mental workload during task execution. ACM Press (2005), 311.
28. Jaques, N., Conati, C., Harley, J.M., and Azevedo, R. Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System. *Proceedings of ITS*, Springer (2014), 29–38.
29. Kardan, S. and Conati, C. Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. *Proceedings of UMAP,* (2013), 215–227.
30. Kruskal, W. Relative Importance by Averaging Over Orderings. *American Statistician 41*, 1 (1987), 6–10.
31. Lallé, S., Mostow, J., Luengo, V., and Guin, N. Comparing Student Models in Different Formalisms by Predicting their Impact on Help Success. *Proceesings of the 16th International Conference on Artificial Intelligence in Education*, (2013), 161–170.
32. Luck, S.J. and Vogel, E.K. The capacity of visual working memory for features and conjunctions. *Nature 390*, 6657 (1997), 279–281.
33. Martínez-Gómez, P. and Aizawa, A. Recognition of understanding level and language skill using measurements of reading behavior. *Proceedings of the 19th International Conference on Intelligent User Interfaces*, ACM (2014), 95–104.
34. Ooms, K., De Maeyer, P., and Fack, V. Study of the attentive behavior of novice and expert map users using eye tracking. *Cartography and Geographic Information Science 41*, 1 (2014), 37–54.
35. Pascual-Cid, V., Vigentini, L., and Quixal, M. Visualising Virtual Learning Environments: Case Studies of the Website Exploration Tool. *Proceedings of the 14th International Conference on Information Visualisation*, IEEE (2010), 149–155.

36. Pavlik, P.I., Cen, H., and Koedinger, K.R. Performance Factors Analysis–A New Alternative to Knowledge Tracing. *Proc. of the International Conf. on Artificial Intelligence in Education*, IOS Press (2009), 531–538.

37. Prendinger, H., Hyrskykari, A., Nakayama, M., Istance, H., Bee, N., and Takahasi, Y. Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation. *Universal Access in the Information Society 8*, 4 (2009), 339–354.

38. Rotter, J.B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied 80*, 1 (1966), 1–28.

39. Saraiya, P., North, C., and Duca, K. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 443–456.

40. Speelman, C.P. and Kirsner, K. *Beyond the Learning Curve: The Construction of Mind.* Oxford University Press, 2005.

41. Steichen, B., Conati, C., and Carenini, G. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Trans. on Interactive Intelligent Systems 4*, 2 (2014).

42. Toker, D., Steichen, B., Gingerich, M., Conati, C., and Carenini, G. Towards Facilitating User Skill Acquisition: Identifying Untrained Visualization Users Through Eye Tracking. *Proceedings of the 19th International Conference on Intelligent User Interfaces*, ACM (2014), 105–114.

43. Turner, M.L. and Engle, R.W. Is working memory capacity task dependent? *Journal of Memory and Language 28*, 2 (1989), 127–154.

44. Yudelson, M.V., Koedinger, K.R., and Gordon, G.J. Individualized Bayesian Knowledge Tracing Models. *Proceedings of AIED*, Springer (2013), 171–180.

45. Zhu, Y. Measuring Effective Data Visualization. *Proceedings of the 3rd International Symposium on Advances in Visual Computing*, Springer (2007), 652–661.