Adaptive eLearning Powered by DeckChair Tutor: Measuring Response Fluency to Customize the Curriculum and Improve Outcomes

Jeff Graham[1], Allan Sura[2]
Department of Psychology, University of Toronto Mississauga[1], Canada
DeckChair Learning Systems[2], Canada

Abstract

This research examines an online curriculum's effectiveness, tests new ideas in curriculum delivery and focuses on providing adaptive training to targeted students. The data were collected from 200 of 1400 first-year undergraduates in an introductory psychology course. They performed three 30 to 60 minute tasks (History of Psychology, Classical Conditioning, and Descriptive Statistics) over an eLearning platform developed by the authors called DeckChair Tutor. DeckChair's mobile eLearning platform measures viewing time and reaction time during the learner's online experience. It adjusts content to remediate weaknesses and adapts questions to match the student's knowledge level. We collect performance data from our online assignments and monitor their effects for teaching and quality assurance research evaluating the contribution of the eLearning assignments to success in the course. We measure response fluency (fast and accurate application of knowledge) and use that information to customize and adapt the content based on the skill level of the student. We compare the merits of a new measure of fluency (SkillScore) to the traditional measures of accuracy and response time. Evidence for long-term retention was a 5% to 17% grade advantage for those students who did the Classical Lab and the History Lab. Very strong correlations were found between our fluency measures and midterm and exam grades in the course. In multiple regressions, BOTH SkillScore and critical time independently account for up to 50% of the variance in the final exam scores.

1. Introduction

The potential benefits of online education are enhanced when we can measure a hybrid curriculum's effectiveness, test new ideas in curriculum delivery, and incorporate adaptive training to targeted students. These are clearly important for any plan to establish an effective hybrid course in any teaching institutions or corporate training program. Often the roadblock to adoption is the required support of implementing quality online curricula and complementing the face-to-face teaching methods [1].

2. Quality assurance of online tasks

We present an example from higher education, in a very large class of using all of the component grades already in the LMS, to help evaluate the contribution of the eLearning assignments to success in the course. We demonstrate how we collect performance data from our online course components and monitor their effects for teaching and research purposes. This is made manageable by the learning and assessment engine we developed, called DeckChair Tutor (deckchairlearning.com), which measures students' behaviours and tracks their performance.

One emphasis of our instructional design is to measure response fluency (fast and accurate application of knowledge) and to use that information to customize and adapt the content based on the skill level of the student. There is a rich history in the cognitive and performance training literatures of using reaction times to infer memory retrieval processes [2], skill acquisition [3], [4], and learning and practice effects [5]. These concepts have found application in special needs curricula with ADHD, and autism spectrum populations [6]. Mathematics education research has shown that the combination of fluency training and other cognitive instruction strategies improves mathematics achievement [8], [9].

This paper describes a study of three concept mastery assignments with first-year undergraduate psychology students using DeckChair Tutor's internet technology to implement research on curriculum effectiveness in hybrid course setting. These three labs and four others are published as

AdapTrack Psychology Labs by Nelson Education [7]. Students were familiar with the software platform and used it throughout the course. In addition to two-hour lectures every week for 24 weeks, there are ten two-hour computer laboratories where students learn research methods and conduct hands-on experiments. The four other Adaptrack labs are psychology experiments in the areas of perception, personality, memory and cognition. In our laboratory course, students act as subjects and as researchers collecting and analyzing data.

One of the many goals of adaptive instruction is to customize the learning paths for each student or groups of students to offer timely and targeted layers of help and feedback as needed. One of the methods we use in DeckChair Tutor is to stream the good students forward through the content as quickly as possible, while directing weaker students to additional materials that review core-concepts with examples and diagnostic feedback. We hope to convince you that such procedures can be incorporated into any course content, for students of all abilities. Fluency-based mastery criteria are used to allow students to jump ahead (or get more help) and we discuss the instructional design issues that drive our research on fluency.

The three concept mastery tasks (History of Psychology, Classical Conditioning, and Descriptive Statistics) were designed to help students master core-concepts in psychology and to evaluate the course curriculum after students have completed those sections of the course. We use pre-tests and post-tests at the start and end of each lab to measure their understanding before and after taking the review section in the lab that can remediate and augment student learning. Similar designs could be implemented for any subject area, skill set or discipline.

While DeckChair Tutor does have resources for human grading of more complex written responses, this paper focuses on performance with traditional multiple-choice test items (MC) and more complex multiple response items, such as multiple fill-in-the-blank scenarios and tables (MFB). The software times the students' responses (called critical time) and automatically grades the accuracy for each item. There are typically 4 to 6 single MC items, and 3 to 5 MFB items in the pre-tests and post-tests to collect a sample snapshot of the students' academic skill in the specific topic of interest.

2.1. Our fluency measure called SkillScore

In any learning domain the goal is to master the facts, rules, and problem solving techniques well enough to perform adequately in real life settings (e.g., language, mathematics, social skills) and efficiently in time sensitive areas (e.g., emergency services, commercial activity, and performance sports). The data reported were collected from first-year undergraduates in an introductory psychology course. They performed these 30 to 60 minute tasks in a supervised computer lab over an eLearning platform developed by the authors called DeckChair Tutor.

The three tasks we present (History of Psychology, Classical Conditioning, and Descriptive Statistics) involve fact retrieval and concept application skills with scenario-based questions, and rule-based calculations in descriptive statistics. Each task has pre-test and post-test questions for which fluency is measured and a multimedia training or review section designed to review core concepts and examples. Typically, there are very few questions in the review slides so we measure total viewing time to determine how much effort was directed to the review.

Measuring fluency requires timing the student responses and the time between clicks (or touches on mobile touch-screens). DeckChair Tutor employs a patented algorithm to measure the learner's speed and accuracy during placement, assessment, teaching, and remediation questions. Each question has predefined correct answers and speed requirements for the target population. While we are studying first-year university students in an introductory psychology course, the processes employed in this course can generalize to any course domain and any student population to improve performance and maintain proficiency – we put the "Pro" in proficiency.

The following formula outlines the performance measure of fluency we call SkillScore. The inputs are **Accuracy** out of 100, Speed out of 1, and Ratio out of 1. Accuracy, for a typical MC question is right or wrong (100 or 0) but could have part marks. **Speed** is defined by three inputs: the student's response time to the specific stimuli (called critical time), the author-defined minimum time (**Min** - the time expected for fluent performers) and the maximum time (**Max** - the slowest time expected for beginners). **Ratio** defines the relative weighting of accuracy vs. speed; a ratio of .5 weights them equally. We use a ratio of .8 for accuracy and .2 for speed. Thus, SkillScore can be interpreted like a grade out of 100.

$$\mathbf{SkillScore = Accuracy * Ratio + Accuracy * Speed * (1 - Ratio)}$$

For example, students should be able to recall simple math facts in 2 to 5 seconds. If they are 2 seconds or faster their speed score is 1. If they are slower than 5 seconds their speed score is 0, and for those inside the minimum to maximum range, the closer they are to the minimum the higher the speed score. If accuracy is zero then SkillScore is zero.

The Min and Max setting will vary depending on the difficulty and the complexity of the question, and the performance goals for the student population. It is best to determine these empirically as each class conducts the tasks. In fact, one of the goals of this norming project was to verify the min and max thresholds, and adjust them if required for the next cohort of students, which was required in a number of cases.

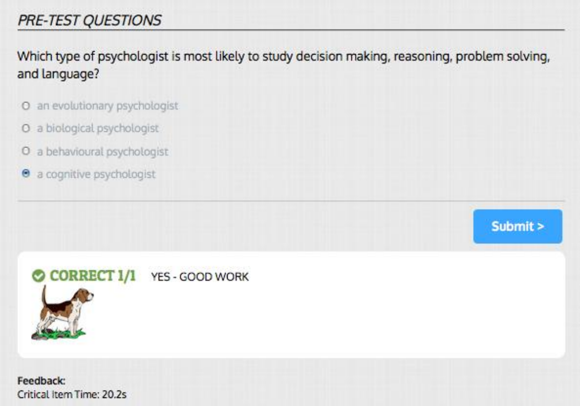
In general, the lab's author (i.e., the instructional designer of the task) specifies the min and max thresholds, often at the 5th and 95th percentiles (or 20th and 80th). They also specify the relative weights of accuracy and speed using the ratio. We have used a ratio of .8 in our tasks, implying that accuracy contributes 80% and speed contributes 20% to the final SkillScore. Other learning situations may warrant larger or smaller ratios.

We compare the merits of SkillScore to the traditional measures of accuracy and response time in detail below. SkillScores are always lower than the Accuracy grades, unless the student has become fluent at an expert level. The larger the difference between SkillScore and Accuracy the more room that student has to improve their fluency. SkillScore is used as the grade for a student's work in order to motivate more fluent performance.

3. Method: Collecting normative data for AdapTrack Psychology Labs

To illustrate the adaptive learning features incorporated within Adaptrack labs and how they might promote behavioural fluency and retention, three sets of data are presented from a large introductory psychology course at the University of Toronto Mississauga (AdapTrack Psychology Labs, Nelson Education). These results are the consequence of a quality assurance program within the course rather than an independent research study, and may not generalize to other settings. However, areas of further research can be suggested from the trends observed.

In each of the three tasks (History, Classical, and Statistics) response data are collected for typical textbook test-bank **MC questions** shown in Figure 1, and for a more complex question format like multiple fill-in-the-blank **MFB Scenarios** and **MFB Tables** as well as self-paced 9-option MC questions is a series called **MC Quickfire** format (as shown in Figure 2). There are many additional measures available in the results export files. The time stamp data produced by DeckChair Tutor's engine allows ways to track reading time, viewing time, and the feedback effects from the time-stamps collected during the learner's online performance.



PRE-TEST QUESTIONS

Which type of psychologist is most likely to study decision making, reasoning, problem solving, and language?

○ an evolutionary psychologist
○ a biological psychologist
○ a behavioural psychologist
● a cognitive psychologist

YES - GOOD WORK

Feedback:
Critical Item Time: 20.2s

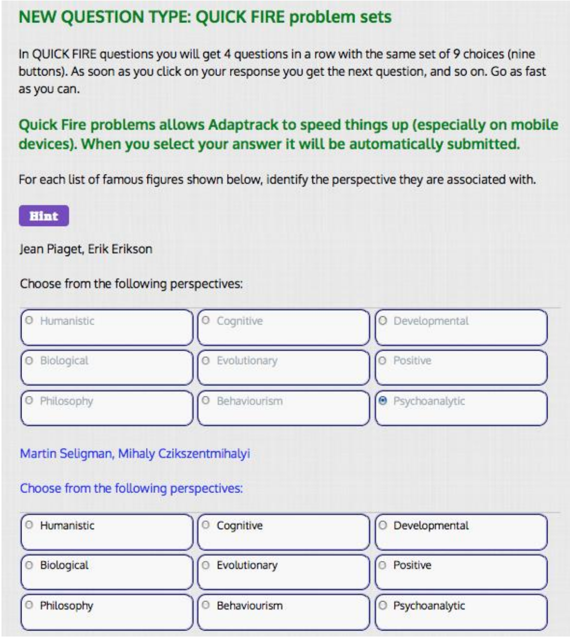
Figure 1. Multiple-Choice question (MC)

To measure **Accuracy** we sum the scores for each question (1 for MC, 4 to 9 for MFB and Quickfire] and convert the sum to a percent correct. To measure speed we record the **Critical Time** in seconds spent on the critical part of the question which is then compared to the Min (expert speed) and Max (beginner speed) thresholds to calculate a speed score between 1 (fast enough] and 0 [too slow]. To calculate **SkillScore** we take a weight combination of the Accuracy and Speed scores to produce a score out of 100.

While short-term learning effects may not translate into long-term skill retention it is important to establish whether short-term learning is occurring and then measure how long it lasts. In the next section we review short-term and long-term benefits of the labs.

4. Results for the History of Psychology lab (HISTORY)

The content of the labs' test and review sections was derived from the textbook and study guide materials. The MC questions come from the textbook test-bank and a more complex question format called QuickFire (QF) questions were designed by the authors to mirror the textbook content. The motivation for more complex question types like Quickfire (QF) and multiple-fill-in-the-blank (MFB) in the other two labs, is driven by the need to make more engaging tasks, and by the hypothesis that multiple response items produce better estimates of academic performance.



NEW QUESTION TYPE: QUICK FIRE problem sets

In QUICK FIRE questions you will get 4 questions in a row with the same set of 9 choices (nine buttons). As soon as you click on your response you get the next question, and so on. Go as fast as you can.

Quick Fire problems allows Adaptrack to speed things up (especially on mobile devices). When you select your answer it will be automatically submitted.

For each list of famous figures shown below, identify the perspective they are associated with.

Jean Piaget, Erik Erikson

Choose from the following perspectives:

○ Humanistic ○ Cognitive ○ Developmental
○ Biological ○ Evolutionary ○ Positive
○ Philosophy ○ Behaviourism ● Psychoanalytic

Martin Seligman, Mihaly Czikszentmihalyi

Choose from the following perspectives:

○ Humanistic ○ Cognitive ○ Developmental
○ Biological ○ Evolutionary ○ Positive
○ Philosophy ○ Behaviourism ○ Psychoanalytic

Figure 2. The first 2 of 4 questions in a successive Quickfire item

In the Mastering History Lab, which has the most content of the three concept mastery labs there were six multiple choice and five quick fire questions (as in Figure 2) in the pre-test and the post-test. In addition, there were three recall items during the training/review section where students were asked to memorize and recall the 12 historical schools of psychology, the 7 unifying themes of psychological research, and the 12 modern research areas. Students (N=67) completed the task in 27 to 70 minutes (5th and 95th percentiles) with an average of 49 minutes. In a second replication, the summer class sample (N=34) had a wider range (17 to 96 minutes) with an average of 51 minutes.

4.1. Does student performance improve from the pre-test to post-test?

In Table 1 the three dependent variables are presented for MC and QF items in the normative sample and in the summer replication sample. The data were analyzed by paired-sample T-tests, and the two-tailed significance levels are shown with asterisks in the last column.

Table 1. HISTORY LAB pre-test and post-test speed, accuracy and SkillScores

HISTORY N=67	Pre-test	Post-test
MC Critical Time (s)	8.8	8.1
MC Accuracy (%)	54.7	69.1***
MC SkillScore (%)	51.9	65.9***
QF Critical Time (s)	61.9	53***
QF Accuracy (%)	50.6	60.4***
QF SkillScore (%)	42.8	51.9***
HISTORY summer N=34	Pre-test	Post-test
MC Critical Time (s)	9.1	8.9
MC Accuracy (%)	66.2	83.3***
MC SkillScore (%)	62.9	79.1***
QF Critical Time (s)	61.4	56.1*
QF Accuracy (%)	67.5	68.1
QF SkillScore (%)	57.1	57.9
t-test (2-tailed) *p<.05, **p <.01, ***p<.001		

There were very clear improvements in performance from the pre-test to the post-test. Both MC and QF problems showed accuracy improvements (of 10% and 14% respectively), and only the QF items showed significant speed improvement. However, both MC and QF problems showed SkillScore improvements (of 14% and 10% respectively). Students were also more confident after training (31% vs. 10% reported they were fairly or very confident) and made those confidence decisions much faster after training (12.9 seconds vs. 4.8).

4.2. Does History lab performance predict follow-up test or exam scores?

The HISTORY lab provided the first of two opportunities to test the hypothesis that task performance would lead to long-term retention (the CLASSICAL lab provided the second). In the winter study (N=67 participants, with 800 peers) students did the History task for a credit up to 2% based on task SkillScore about 6 months after last being tested on the topic, and about one month before the final exam. Then in the summer replication study (N=34 participants, vs. 30 peers) students did the History task voluntarily (no credit, just a study aid) the same week they first studied the material, and within 1 day of the term test.

The comparisons for the two HISTORY samples are potentially confounded. The winter subjects were participating for needed credits and were generally no different academically (with slightly lower grades if anything) than their peers in the rest of their class. In contrast, the summer participants were consistently stronger than their peers on all term tests (8% more on MC and Short Answer tests, and 6% better on the Exam). This reflects the common notion that good students are motivated to do more of the study preparation materials than weaker students who tend to embrace extra work only if there are marks associated with it. Thus, volunteers who did the labs without any credit tended to be better students overall.

The participants who tried harder on the History Task did better specifically on the History questions in the exam one month later, relative to the participants who spent less effort studying. This effect was not observed in the much smaller summer sample where the test was on the following day and the exam data were not available for the study.

Specific long-term retention effects were tested with the exam data for each of the 16 chapters (i.e., topics) tested. There were 10 questions (in most cases) for each chapter, and a percent correct was calculated for each student in each topic. In the first analysis, there were no significant differences between the 66 participants and their 794 peers on any topic, varying +/- about 0 to 3 percent across topics. Thus, there is no evidence that those who did the History lab would do better than their peers on the exams history questions. However, there was a clear long-term advantage for the more diligent half of the students who did the History lab, as measured by the time spent studying the review material.

There is very strong evidence that those who put more effort into the History Lab did much better specifically on the history items one month later compared to the lab participants who spent less effort memorizing. We measured the time spent memorizing material on the **12 historical schools** (11 to 86 seconds for 120 words), the **7 unifying themes** (25 to 205 seconds for 494 words), and the **12 modern subject areas** (25 to 220 seconds for 686 words). Students were instructed to “Commit these SUBJECT AREAS to memory. Be prepared to list them on a test, and more importantly, be able to describe in your own words the main ideas, subject matter and contributors for each.”

Very similar results were found when using a split-half based on the subject AREAS study time (median 88 seconds) and a split-half based on unifying THEMES study time (median 97 seconds). There are two main observations to point out about performance on the relevant exam questions one month later. The half who studied longer (perhaps reflecting motivation or persistence) were about 8% better on the exam overall than those who spent less time studying. That advantage was evident on 11 of the 16 topics. The more important result is that the advantage was more than double (17%) on the HISTORY topic specifically. Surprisingly, the extra study time did not have an effect on immediate recall. Both groups scored about 68% correct on the subject area recall test, just after studying the material.

Table 2. Exam topic scores for the median split on study time for AREAS

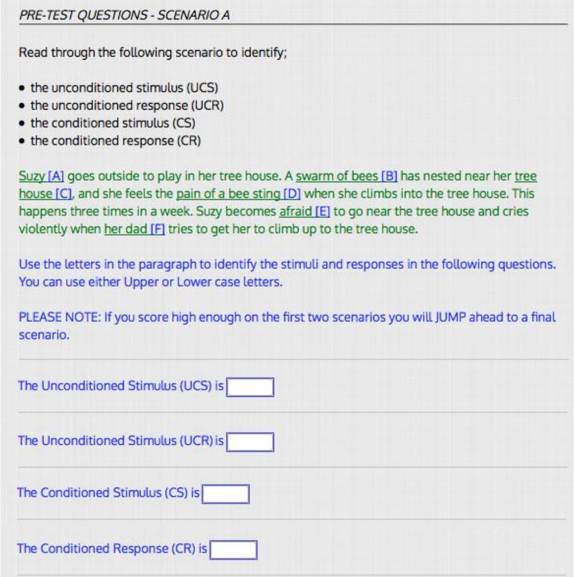
Study Time (median split)	Hi Effort	Low Effort
Final Exam Total	65.1	57.4**
HISTORY	71.0	53.7***
RESEARCH METHODS	80.0	73.8
BIOLOGICAL	61.6	52.1
SENSATION/PERCEPTION	60.0	60.9
CONSCIOUSNESS	65.9	55**
LEARNING	66.9	60.3
MEMORY	62.5	52.1*
LANGUAGE	57.5	47.9*
INTELLIGENCE	55.3	50.9
MOTIVATION	75.3	68.5
DEVELOPMENTAL	55.9	51.5
PERSONALITY	58.1	49.1
HEALTH	73.6	68.7
ABNORMAL	59.0	50.3*
TREATMENT	65.6	58.2
SOCIAL	71.6	63.8
t-test (2-tailed) *p<.05, **p <.01, ***p<.001		

The statistical evidence suggest that working diligently to memorize content in the History lab leads to a 10% higher grade in History topics one month later compared to their peers and 17% higher than the other participants who put less effort into the task. This speaks directly to the importance of motivation, effort and/or diligence, and introduces the concept that DeckChair Tutor’s online measure of speed and accuracy do more than predict grades, but also allows researchers to build a student’s learning profile that may generalize across topics and tasks.

5. Results for the Classical Conditioning lab (CLASSICAL)

Response data were collected for standard textbook test-bank **MC questions** shown in Figure 1, and for a more complex multiple fill-in-the-blank **MFB Scenario** questions, shown in Figure 3. We measure how long students take to read the scenario paragraph, and then present four successive FB questions asking them to identify the UCS, the UCR, the CS, and the CR from the options tagged in the paragraph. The critical time is from when the question is presented until the student submits their response.

The content in the test and review sections of this lab was derived from the textbook and study guide materials. The MC questions come from the textbook test-bank and the scenario MFB questions were designed by the first author to mirror the textbook content. Part of our agenda is to develop more engaging tasks for students to apply their new knowledge to real world cases. Just like the Quickfire questions in the HISTORY lab, we hypothesize that that multiple response items like the Scenario MFB, will produce better estimates of academic performance than traditional 4-option MC questions.



PRE-TEST QUESTIONS - SCENARIO A

Read through the following scenario to identify;

- the unconditioned stimulus (UCS)
- the unconditioned response (UCR)
- the conditioned stimulus (CS)
- the conditioned response (CR)

Suzy [A] goes outside to play in her tree house. A swarm of bees [B] has nested near her tree house [C], and she feels the pain of a bee sting [D] when she climbs into the tree house. This happens three times in a week. Suzy becomes afraid [E] to go near the tree house and cries violently when her dad [F] tries to get her to climb up to the tree house.

Use the letters in the paragraph to identify the stimuli and responses in the following questions. You can use either Upper or Lower case letters.

PLEASE NOTE: If you score high enough on the first two scenarios you will JUMP ahead to a final scenario.

The Unconditioned Stimulus (UCS) is □

The Unconditioned Stimulus (UCR) is □

The Conditioned Stimulus (CS) is □

The Conditioned Response (CR) is □

Figure 3. The 4 successive questions in a scenario MFB item

The response time is the average time in seconds over all four FB questions. The scenario reading time (indicated by clicking continue to get the first question) is quite variable and is often shorter than expected. Many students take a long time to answer the first question, and are spending time reviewing the paragraph, which remains up for all four questions. On the 3rd and 4th scenarios which stronger students jump over due to their fluency scores, extra help is provided including branching to a short review of the Pavlov's findings.

5.1. Does student performance improve from the pre-test to post-test?

The Classical Conditioning Lab had the same pre-test/train/post-test format as the History and Descriptive Statistics labs but the emphasis was on scenario-based training on the identification of the UCS, UCR, CS, and CR in real world examples. There were five scenarios in the pre and post-tests and if subjects' performance was good enough on the first two, then they would advance to the 5th one, and thus "jump over" two that had extensive feedback and coaching branches. About 44% of students jumped in the pre-test and 37% in the post-test. Those who were fluent enough to jump in the pre-test spent 40% more time studying scenarios (43.4 vs. 30.9 seconds). This was also true in the post-test: those who jumped spent 60% more time studying scenarios (35.2 vs. 21.9 seconds) than those who did not meet the jump fluency criterion.

Students completed the task in 11 to 33 minutes (5th and 95th percentiles) with an average of 21 minutes. Item analyses indicated a problem in the 2nd scenario of the post-test, which proved to be too difficult for our introductory psychology students. This item was dropped from the post-test analyses. Response times for MC and Scenario questions (Critical Times) did not include the time spent reading the MC question or the scenario text, but consisted only of the time to select among the foils offered.

Table 3. CLASSICAL LAB pre-test and post-test speed, accuracy and SkillScore

CLASSICAL N=75	Pre-test	Post-test
MC Critical Time (s)	8.8	10.2*
MC Accuracy (%)	83	75.3*
MC SkillScore (%)	80.7	72.8*
4FB Critical Time (s)	13.4	9.7***
4FB Accuracy (%)	49.8	67.5***
4FB SkillScore (%)	47.4	65.7***
t-test (2-tailed) *p<.05, **p <.01, ***p<.001		

In the Classical Lab there were clear speed and accuracy improvements from Pre to Post but only for the 4 FB scenario questions. In fact, the traditional MC questions were significantly slower and marginally less accurate than the pre-test MC questions. MC accuracy went down from 83% to 75% but the Scenario questions went up from 47% to 66%.

5.2. Does the Classical Conditioning lab performance predict follow-up exam scores?

The CLASSICAL lab provided the second opportunity to test the hypothesis that task performance would lead to long-term retention. In the winter study students did the Classical task for a credit up to 2% based on task SkillScore, about 6 months after last being tested on the topic, and about one month before the final exam. Thus, the pre-test measure how much they remembered from studying 6 months earlier. The exam data measures retention one month after doing the classical lab.

To test the hypothesis that task performance would lead to long-term retention we compared those students who took the Lab (for course credit one month before the exam) with their peers (N=75 participants, with 792 peers).

Table 4. Exam topic scores

CLASSICAL N=75 vs. 785	CLASS	PEERS
Final Exam Total	62.1	60.6
HISTORY	61	61.1
RESEARCH METHODS	76.4	74.9
BIOLOGICAL	59.1	55.6
SENSATION/PERCEPTION	58.5	59.5
CONSCIOUSNESS	59.6	60.6
LEARNING	65.5	61.4*
MEMORY	57.2	57.1
LANGUAGE	53.3	53.3
INTELLIGENCE	55.1	55.2
MOTIVATION	72.5	70.6
DEVELOPMENTAL	57.5	54.9
PERSONALITY	53.3	55.4
HEALTH	74.8	72.8
ABNORMAL	56.4	57.4
TREATMENT	60.4	61
SOCIAL	70.7	70.4

t-test (1-tailed)
*p<.05, **p <.01, ***p<.001

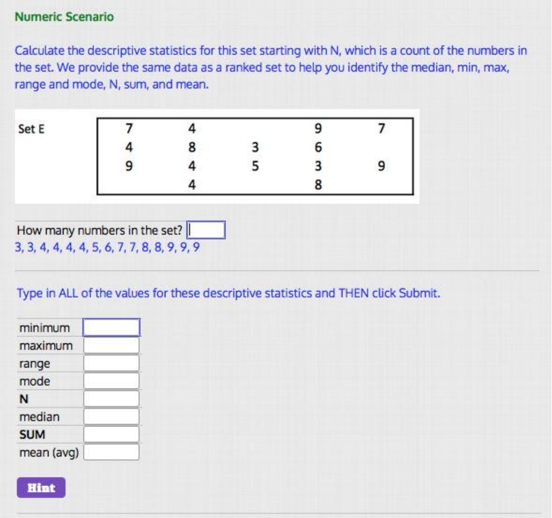
Specific long-term retention results were tested using the exam data for each of the 16 chapters (i.e., topics) tested. There were about 10 questions for each chapter producing a percent correct for each student in each topic. There were no significant differences between the 75 participants and their 792 peers on any topic, except for the Learning chapter which covers classical conditioning. There was a specific advantage for those who took the lab over their peers of 4.1 % (p<.05) on the Learning topic (which covers classical conditioning) although they were also marginally better on the Biological and Developmental chapters (P<.1).

6. Results for the Descriptive Statistics lab (STATISTICS)

The Descriptive Statistics Lab was designed to establish whether or not students have a firm grasp of the basics in statistics and to give hands-on practice to apply simple definitions to sets of data presented as numeric list or graphical histograms. Students completed the task in 21 to 47 minutes (5th and 95th percentiles) with an average of 31 minutes.

There was a pre-test/post-test design with four MC questions on basic definitions and 2 data sets (one numeric, and one graphic) asking students to report the minimum, maximum, range, mode, N, the median, sum, and mean of 15 to 20 numbers. After the pre-test there was a review of core concepts, and then practice with 5 numerical data sets, and 5 graphical data sets.

Response data were collected for standard textbook test-bank **MC questions** shown in Figure 1, and for a more complex multiple fill-in-the-blank **MFB Table** questions, shown in Figure 4 (Numeric data sets) and Figure 5 (Graphic data sets).



Numeric Scenario

Calculate the descriptive statistics for this set starting with N, which is a count of the numbers in the set. We provide the same data as a ranked set to help you identify the median, min, max, range and mode, N, sum, and mean.

Set E

7	4		9	7
4	8	3	6	
9	4	5	3	9
	4		8	

How many numbers in the set?
3, 3, 4, 4, 4, 4, 5, 6, 7, 7, 8, 8, 9, 9, 9

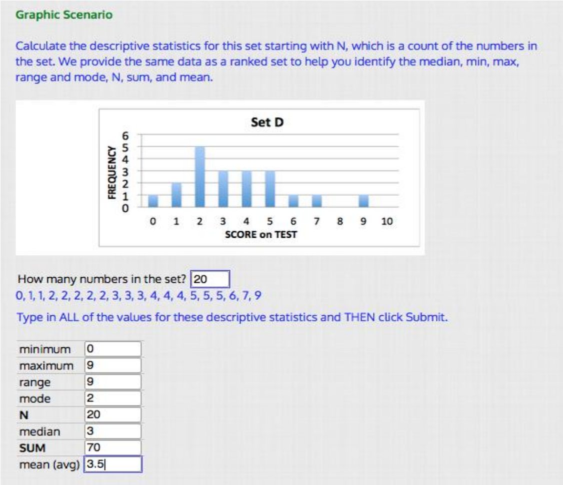
Type in ALL of the values for these descriptive statistics and THEN click Submit.

minimum
maximum
range
mode
N
median
SUM
mean (avg)

Figure 4. The 8 FB questions in a Numeric MFB item

Each of the training sets had five scenarios (i.e., sets of 15 or 20 numbers) and student's speed and accuracy were checked after the first two sets. If they demonstrated mastery by meeting the fluency threshold, they would jump over the next two sets, and finish with the fifth set. Data are averaged only over sets, 1, 2 and 5, since most students did meet the performance criteria.

We measure how long students take to read the scenario, view the data, and then report the count. The critical time is the total time for calculating and entering the eight FB boxes, which are all on the screen at the same time, not including the time to read and count the numbers in the set. The eight response boxes appear only after they enter how many numbers are in the set, and that starts the clock for critical time.



Graphic Scenario

Calculate the descriptive statistics for this set starting with N, which is a count of the numbers in the set. We provide the same data as a ranked set to help you identify the median, min, max, range and mode, N, sum, and mean.

How many numbers in the set? 20

0, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 7, 9

Type in ALL of the values for these descriptive statistics and THEN click Submit.

minimum	0
maximum	9
range	9
mode	2
N	20
median	3
SUM	70
mean (avg)	3.5

Figure 5. The 8 FB questions in a Graphic MFB item

The content in the test and review sections of this lab was derived from the textbook and study guide materials. The MC questions come from the textbook test-bank and the number set scenario MFB questions were designed by the first author to mirror the textbook content. On the 3rd and 4th MFB data sets (which stronger students jump over due to their fluency scores) additional feedback is provided.

6.1. Does student performance improve from the pre-test to post-test?

We were pleased to see that most students have already mastered these core concepts averaging over 90% correct on the graphical and numeric data sets. Very few students failed to jump over the remedial trials even though nearly all of the students were slower on the data sets than anticipated (over 60 to 120 seconds vs. the min of 30 and the max of 60 seconds). There were significant improvements in speed (from 130 to 63 seconds) as measured by the time to complete all eight fill-in-the-blank fields for each data set.

There were also significant increases in accuracy (from 90.7% to 95.8%) with most of the increase on the relatively more difficult concepts of mean, median, sum, and range (which went from about 85% to 94%) as compared to the very basic concepts of minimum, maximum, mode, and N (from went from about 97% to 98%). The mean, median, sum, and range calculations were more accurate in the Numeric practice sets (88.1%) than in the Graphical practice sets (83.8%).

Table 5. STATISTICS LAB pre-test and post-test speed, accuracy and skillscores

STATISTICS N=61	Pre-test	Post-test
MC Critical Time (s)	11.9	10.6*
MC Accuracy (%)	67.6	72.5
MC SkillScore (%)	61	66.6
MFB Critical Time (s)	129.9	63.3***
MFB Accuracy (%)	90.7	95.8*
MFB SkillScore (%)	75.8	88.5***
TRAINING SETS	GRAPHIC	NUMERIC
MFB Critical Time (s)	79.7	68.4***
MFB Accuracy (%)	89.8	93.6***
MFB SkillScore (%)	80.1	85.6***
t-test (2-tailed) *p<.05, **p <.01, ***p<.001		

The MC definition questions were solved about a second faster in the post-test but the improvement in accuracy (68% to 73%) was not significant. On the other hand, the numeric (NUM) and graphical (GRA) questions did show significant increases in speed (about 50% faster) and accuracy (about 5%) after the training sets had been completed.

Apparently the graphical sets (which also had the numeric data in a sorted array) proved to be about 11 seconds slower and 4% less accurate overall.

6.2. Does student lab performance predict follow-up exam scores?

There are many interesting relationships between speed and accuracy performance in the Statistics lab and overall performance in the course, and final exams. We will review these effects in the next two sections concentrating on our proposed metric of skill called SkillScore, which is a weighted combination of accuracy and reaction time.

There were very few descriptive statistics questions on the final exam and so we could not do the same sort of analyses that showed long-term retention effects for diligent students in the History and Classical labs. It is clear that these sorts of long-term retention studies can be designed into any curriculum by carefully editing of midterms and exams, and other assessments to cover the same learning taxonomy as the online labs.

7. Adaptive eLearning: Customized learning pathways and targeted remediation (using JUMPS)

There are many definitions in the literature for adaptive training and we are concentrating on the main idea that the content provided to the student should match their skill in the domain. Thus we can customize the training and assessment for each student by making sure the material is not too hard or too difficult during training. Teachers can use their task data to flag struggling students early, and see which students are not being challenged enough.

To customize the learning pathway and provide targeted help we have implemented checkpoints within a series of trials such that if the student is good enough, i.e., meets the speed and accuracy criteria, they can move more quickly ahead in the task (i.e., **jump**). However, if the student is not good enough at that check point, then they are provided with additional questions with all of the feedback, branching, and remediation tools available. Thus, our learning platform adjusts the content to remediate weaknesses and adapts questions to match the student's knowledge level.

The designer of the tasks needs to decide upon a **JUMP mastery criterion** based either on accuracy scores or SkillScores. Some designers may decide that their target populations may be intimidated by speed requirements and so timing is done unobtrusively and jumps are based entirely on accuracy. We used an easy SkillScore jump criterion of 64% and an accuracy-speed Ratio of .8. Very slow students could meet the criterion if they got at least 80% correct (which when multiplied by .8 is 64% SkillScore). Very fast students could jump over if they got at least 64% correct with critical times under the minimum speed threshold.

Table 6. Percentage of students who jumped ahead based on mastery

			TIME percentile	
HISTORY	% JUMP	Accuracy	20th	80th
MC	--	61.9	3.7	12.7
QuickFire	7*	62.3	36.4	81
CLASSICAL	% JUMP	Accuracy	20th	80th
MC	--	77.9	4.8	13.6
MFB	45	56.6	2.4	18.6
STATISTICS	% JUMP	Accuracy	20th	80th
MC	--	70.1	6	14.8
MFB	--	93.2	89.3	172
Graphic MFB	87	89.8	61.1	93.1
Numeric MFB	93	90.5	50.5	83.4

We learned a few things about our jump settings. Very few people jumped in the HISTORY task's QuickFire questions, because the material was difficult (62% average) and the minimum speed threshold was set too low. The empirical data recommends that we set Min to 40 seconds and Max to 80 seconds (instead of 10 and 60). About half the people jumped in the CLASSICAL task, which is about what we were hoping for. They were faster than expected and so we would adjust the Min speed threshold down to 5 seconds to reflect that. Finally, nearly everyone jumped in the STATISTICS lab because they know the core concepts very well (80% of the sample were over 90% correct, and only 5% had less than 80% correct). In this case we severely underestimated the speed thresholds since most students took 30 seconds longer then expected. The new threshold would be better set at 30 and 120 seconds rather then at 30 and 60 seconds. The SkillScores were recalculated for the STATISTICS MFB items reported above.

For a fluency based SkillScore to be effective the Min and Max speed thresholds for each class of problems can afford to more extreme (with a lower Min and an higher Max) in order to a make sure we capture the relevant speed window. This is to avoid floor and ceiling effects where everyone responds faster than the Min or slower then the Max. As more data is collected annually, the Min and Max values can be revised to better match the target learning population. We are exploring other algorithms for converting speed to SkillScore, such as non-linear SkillScore gains for improvement either at the slow end or the fast end of the speed range. Will students notice or be motivated by bigger increases in scores for successive improvements in speed?

In the next section we will show how the raw data collected from the online users is exported for a quantitative analysis using SPSSX (Statistical Package for the Social Sciences). As reported above, we start with the within group comparisons of the average pre-test and post-test averages (critical time, accuracy, and SkillScore). To track the relations among specific knowledge, skills, and course-related grades, a set of multiple-regressions is performed using task data from DeckChair Tutor integrated with course data from the LMS gradebook. Can we understand how task performance is related to academic performance in general?

8. Predicting course outcomes: performance data analytics

Conducting a quantitative analysis of item performance for each subject will generate large data sets that can be mined for quality assurance purposes. Are there weak questions, or problematic lessons? Are there students or groups of students having more or less success? Instructional designers who want valid measures of learning can use the performance history on items to make fair tests, or to personalize difficulty levels, and manage the students total task time.

As teachers we are interested in these analytics to measure effectiveness of course components, and predict final grades in the course. Our course's LMS gradebook contains all the course component grades over 6 midterm tests as well as 2 term tests from the 12 Introductory Psychology computer laboratories. The student data is averaged over the tests to create independent measures of academic performance for 1) multiple-choice text-book questions, 2) short-answer lecture questions, and 3) paragraph answers on Labtest1 and Labtest2. Finally, there are data from the three-hour final exam that has 10 questions on each of the 16 chapters. The regression results summarized in this report used the Final Exam as our best measure of academic ability. It is used as the predicted (dependent) variable in the multiple regressions to represent either overall academic ability (how good is the student) or probability of success course outcome.

In addition to determining whether an online lab has short and long term beneficial outcomes for the student, another useful endeavor is to ask whether there are covert measures of academic ability that can give early warning signs to the teacher about the students' likely success in the course. This facilitates early interventions for weaker students, and early indications about which students will master the material. Such indicators are available from the monthly midterm test scores which take time and effort to collect. Can the same information may be obtained from a single online lab that contains a pre-test, a review of content, followed by a post-test in DeckChair Tutor.

Table 7. Correlations (r) and regression coefficients predicting final exam

Speed (s)	HIST	CLASS	STATS
Pre MC	-.416**	-.317**	-0.216
Post MC	-.412**	-.230*	-.359**
Pre QF/MFB	-.320**	-0.063	-0.136
Post QF/MFB	-0.183	-0.226	-0.046
Accuracy (%)			
Pre MC	.270*	.381**	0.082
Post MC	0.111	.276*	0.208
Pre QF/MFB	.458**	.319**	-0.063
Post QF/MFB	.532**	.386**	0.159
SkillScore (%)			
Pre MC	.328**	.408**	0.103
Post MC	0.164	.307**	.262*
Pre QF/MFB	.496**	.337**	-0.024
Post QF/MFB	.559**	.404**	0.184
Study Time	.461**	-0.089	0.248

Multiple R	HIST	CLASS	STATS
Speed (s)	0.732	0.615	0.488
Constant	45.189	48.624	66.676
Pre MC		-0.766	
Post MC	-0.563		-1.019
Pre QF/MFB			
Post QF/MFB			
Accuracy (%)			
Pre_MC			
Post MC			
Pre QF/MFB		0.099	
Post QF/MFB			
SkillScore (%)			
Pre MC		0.097	
Post_MC			
Pre QF/MFB	0.18		
Post QF/MFB	0.142	0.113	
Study Time	0.052		0.015

In this study we look at the **three concept mastery labs** and use data collected from **two types of questions** in the pre-tests and post-tests: simple 4-option multiple choice questions (MC), and more complex compound questions (QF - quickfire 9-option multiple choice and MFB - 4 to 8 multiple-fill-in-the-blank questions). The **three performance**

measures (Speed, accuracy, and SkillScore) were collected for both types of questions and these are the data used to predict final exams scores.

Do response time measures predict the final exam grade in the course? Do they predict better than accuracy for both types of questions formats? Moreover, is SkillScore, our weighted speed and accuracy measure, a better predictor than speed or accuracy alone?

The regression analyses are summarized in Table 7. To begin, I present the simple correlations in the top half, indicating in red font those variables that enter stepwise into the regression equation. In the bottom half, I present the regression coefficients that can be used to predict students' final grades in the course. About 50% of the variance in final exam scores can be explained by the regressions for the History tasks, and a little less for the other two tasks. To be able to tell that from a 30 to 50 minute online task is pretty impressive. The role of response fluency is a major part of understanding academic skill.

Response times are negatively correlated with final exam grade, and this pattern is evident for both the pre-test and the post-test MC questions (r = - .23 to -.42). Students who take longer to answer the MC questions do less well on the exam. Accuracy is also a good predictor of final exam grade, and it is significantly and positively correlated in all tasks except the Statistics lab. This holds true for both types of questions, particularly for QF or MFB post-test questions (r= .27 to r = .53).

Importantly, our SkillScore measure of fluency was a better predictor of exam grade than accuracy in 11 of the 12 comparisons (pre and post for 2 question types in three 3 tasks). SkillScores on MFB questions were recalculated based on 30 and 120 min-max thresholds for the Statistics lab since they were underestimated during the experiment. SkillScore was also a better predictor of exam grade than speed in 6 of the 8 comparisons (excluding the Statistics lab). The significant correlations of SkillScore with Exam grade ranged from .31 to .56. Thus our fluency measure, SkillScore, was most often the best measure of academic performance, lending support to our position that fluency matters and must be taken into account when designing adaptive tasks that predict future performance.

There are a number other exciting observations in our data. There is very strong evidence that those students who put more effort in mastering the study material during the online tasks are the same students that tend to do well in the exam. In two of the tasks (History and Statistics), students who spent more time on the critical study slides did better across all of topics in the final exam. We interpret this as a possible diligence measure. Notice that it is the opposite of the response time correlation in which weaker students are slower to respond. Positive correlations with Study Time indicate that stronger students are more willing to put more time and effort into study materials, and we can infer who they are from a three or four minute study episode during an online task.

The final analysis in our approach was to do multiple regressions using a stepwise free entry model. We consider this to be an exploratory exercise letting the data inform us of the relative unique contributions of speed, accuracy, and SkillScore. Pre-test and post-test averages for each question type generated the 12 predictors for each task. We also added a study time measure, slightly different for each sample. The stepwise free entry procedure simply enters the variable that captures the largest variance at each step.

SkillScore predictors accounted for unique variance in two of the three tasks excluding Statistics (perhaps due to the ceiling effect on accuracy). Surprisingly, both SkillScore and speed (critical reaction times) accounted for unique variance and entered independently in the regression equation. We are not sure how to interpret this, but it suggests that an accuracy score weighted by speed (i.e., SkillScore) measures academic skills that are different than simple reaction time. Further research is need to detangle these effects but it is very clear that fluency plays a significant role in performing efficiently online and in predicting general academic ability.

9. Conclusions

The three concept mastery labs explored in this study show both short-term and long-term benefits in performance. The grades (i.e., accuracy) improved 5% to 17% over the course of the lab. In some cases the task was too easy and showed only 5% improvement (Statistics) due to a ceiling effect. Given our interest in adaptive itineraries, feedback branches, and jumping thresholds it was good to have a range of task difficulty levels, to have real world data with students across all academic skill levels running labs of varying difficulty.

The evidence for long-term (1 month) retention was a 5% to 17% grade advantage for those students who did the Classical Lab and the History Lab. This was stronger for those who spent more time of the critical study slides in AdapTrack. Statistical analyses suggest that working diligently to memorize content in the History lab leads to a 10% higher grade in History topics one month later compared to their peers, and 17% higher than the other participants who put less effort into the task. This speaks directly to the importance of motivation, effort and/or diligence and that we can develop ways of measuring this from time-stamps data in the performance logs.

We learned that the advance criteria used for adaptive jumping or branching need to be calibrated to the learning population. Only 7% of the History subjects earned a jump, while 90% did so in Statistics. About 47% jumped in Classical and that was the desired result. DeckChair Tutor was built in part to be able to automate curriculum design research by varying threshold option across different groups of students, and evaluate optimal jump criteria.

We compared traditional multiple-choice questions with more complex question types and found that both speed and accuracy improve for both from pre-test to post-tests, with few exceptions. There were much bigger speed improvements for the complex items, which represent composite tasks and we argue that this provides more opportunity for students to see improvement in speed (and therefore their SkillScore). Other research could address the motivational and academic benefits or costs of gameifying the tasks with SkillScore-related badges or certificates.

Very strong correlations were found between our fluency measures (critical time, and SkillScore) and midterm and exam grades in the course. In multiple regressions, BOTH SkillScore and critical time independently account for up to 50% of the variance in the final exam scores. In the simple correlations, critical times predict the final exam grade better than accuracy scores. The best predictors do seem to vary for the two question types: for multiple-choice items (MC) critical time is better than accuracy, and for QuickFire (QF) or multiple-fill-in-the-blank (MFB) items, accuracy is better than critical time.

10. Acknowledgements

The authors would like to thank the students, and lab research assistants at the University of Toronto Mississauga, Introductory Psychology Computer Laboratory. The data reported were collected within the framework of curriculum quality assurance rather than research. Informed consent was obtained from all students whose data were included. The authors can be contacted at jeffrey.graham@utoronto.ca and/or info@deckchairlearning.com.

11. References

[1] Aguti, B., Walters, R. J., and Wills, G. B. (2014). Effective Use of E-Learning Technologies to Promote Student-Centered Learning Paradigms within Higher Education Institutions. International Journal for e-Learning Security, 4 (3/4), pp. 391-398.

[2] Unkelbach, C. (2006). The Learned Interpretation of Cognitive Fluency. Psychological Science, 17(4), pp. 339-345.

[3] Binder, C. (2003). Doesn't Everyone Need Fluency? *Performance Improvement*, *42*(3), 14-19.

[4] Teodorescu, T. M., Binder, C. (2004). Competence Is What Matters. Performance Improvement, 43(8), pp. 8-12.

[5] Peladeau, N. (2003). Effect of Paced and Unpaced Practice on Skill Application and Retention: How Much Is Enough? American Education Research Journal, 40(3), pp. 769-801.

[6] Brady, K.K., and Kubina Jr., R.M. (2010). Endurance of Multiplication Fact Fluency for Students With Attention Deficit Hyperactivity Disorder. Behavior Modification, 34(2) pp. 79–93.

[7] Graham, D. J. and Sura, A. A. (2015). AdapTrack Psychology Labs, 1st Edition. Toronto,
Ontario, Canada: Nelson Education.

[8] Carr, M.,Taasoobshirazi, G., Stroud, R., and Royer, J.M. (2011). Combined Fluency and Cognitive Strategies Instruction Improves Mathematics Achievement in Early Elementary School. Contemporary Educational Psychology, 36 (2011) pp. 323–333.

[9] Singer-Dudek, J., and Greer, R.D. (2005). A Long-Term Analysis of the Relationship Between Fluency and the Training and Maintenance of Complex Math Skills. The Psychological Record, 55(3) pp. 361-376.