Journal of Statistical Science and Application, April 2015, Vol. 3, No. 5-6, 63-73
doi: 10.17265/2328-224X/2015.56.001

Design and Analysis of Experiments Linking on-line Drilling Methods to Improvements in Knowledge

Gunnar Stefansson, Anna Helga Jonsdottir
University of Iceland Science Institute, Dunhaga 5, 107 Reykjavik, Iceland

An on-line drilling system, the tutor-web, has been developed and used for teaching mathematics and statistics. The system was used in a basic course in calculus including 182 students. The students were requested to answer quiz questions in the tutor-web and therefore monitored continuously during the semester. Data available include grades on a status exam conducted at the beginning of the course, a final grade and data gathered in the tutor-web system. A classification of the students is considered, using the data gathered in the system; a **G**ood student should be able to solve a problem quickly and get it right, the "diligent" hard-working **L**earner may take longer to get the right answer, a guessing (**P**oor) student will not take long to get the wrong answer and the remaining (**U**nclassified) apparent non-learning students take long to get the wrong answer, resulting in a simple classification **GLUP**. The (**P**oor) students were found to show the least improvement, defined as the change in grade from the status to the final exams, while the **L**earners were found to improve the most. More detailed analyses indicate that improvements in knowledge are best predicted as quadratic responses to the number of items requested and the time spent on each item. The results are used to demonstrate how further experiments are needed and can be designed as well as to indicate how a system needs to be further developed to accommodate such experiments.

Introduction

With the increasing number of web-based educational systems several types of educational systems have emerged. These include learning management system (LMS), learning content management system (LCMS), virtual learning environment (VLE), course management system (CMS) and Adaptive and intelligent web-based educational systems (AIWBES).[1]

The LMS is designed for planning, delivering and managing learning events, usually adding little value to the learning process nor supporting internal content processes [3]. A VLE provides similar service, adding interaction with users and access to a wider range of resources [5]. The primary role of a LCMS is to provide a collaborative authoring environment for creating and maintaining learning content [3].

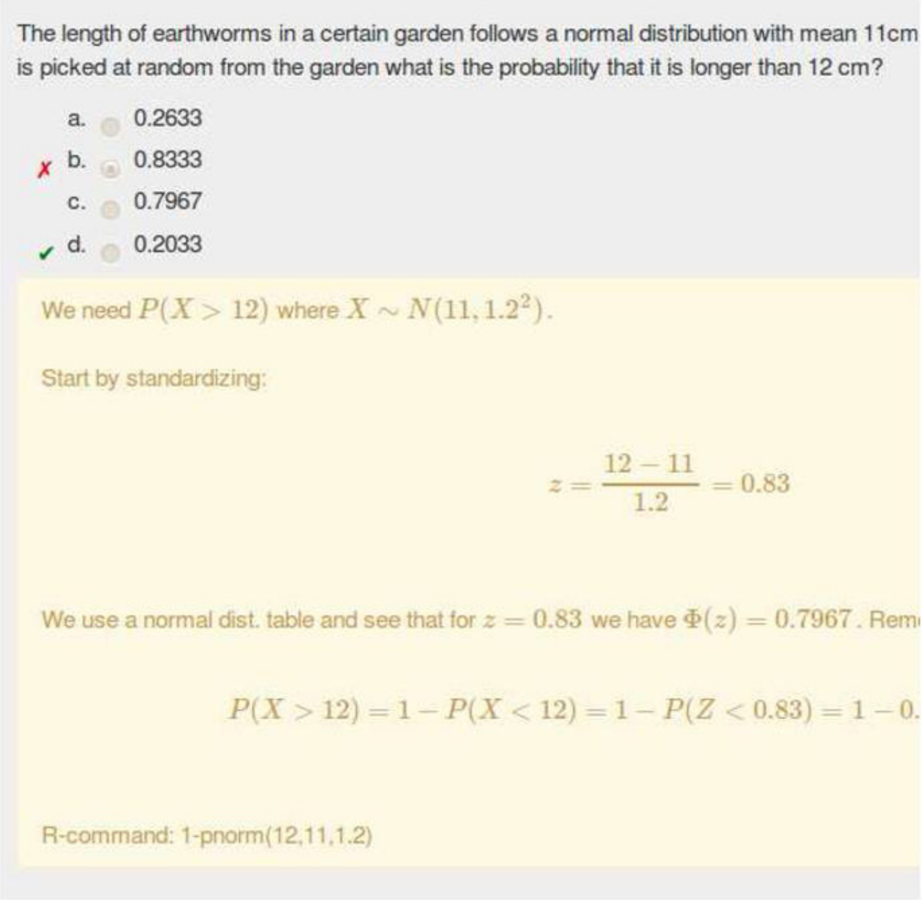
Many systems are merely a network of static hypertext pages [1] but adaptive and intelligent Web-based educational systems (AIWBES) use a model of the goals, preferences and knowledge of each student and use this to adapt to the needs of that student [2]. These systems tend to be subject-specific because of their structural complexity and therefore do not provide a broad range of content.

The tutor-web (at *http://tutor-web.net*) used here is an open and freely accessible AIWBES system, available to students and instructors at no cost. The system has been a research project since 1999 and is

[1] The terms VLE and CMS are often used interchangeably, CMS being more common in the United States and VLE in Europe.

completely based on open source computer code with material under the Creative Commons Attribution-ShareAlike License. The material and programs have been mainly developed in Iceland but also used in low-income areas (e.g. Kenya). Software is written in the Plone[2], CMS (content management system), on top of a Zope[3] Application Server.

In terms of internal structure, the material is modular, consisting of departments (e.g. math/stats), each of which contains courses (e.g. introductory calculus/regression). A course can be split into tutorials (e.g. differentiation/integration), which again consist of lectures (e.g. basics of differentiation/chain rule). Slides reside within lectures and may include attached material (examples, more detail, complete handouts etc). Also within the lectures are drills, which consist of quiz items. The drills/quizzes are designed for learning, not just simple testing. The system has been used for introductory statistics, mathematical statistics, earth sciences, fishery science, linear algebra and calculus in Iceland and Kenya, with some 2000 users to date.



The length of earthworms in a certain garden follows a normal distribution with mean 11cm
is picked at random from the garden what is the probability that it is longer than 12 cm?

a. ○ 0.2633
✗ b. ● 0.8333
c. ○ 0.7967
✓ d. ○ 0.2033

We need $P(X > 12)$ where $X \sim N(11, 1.2^2)$.

Start by standardizing:

$$z = \frac{12-11}{1.2} = 0.83$$

We use a normal dist. table and see that for $z = 0.83$ we have $\Phi(z) = 0.7967$. Rem

$$P(X > 12) = 1 - P(X < 12) = 1 - P(Z < 0.83) = 1 - 0.$$

R-command: 1-pnorm(12,11,1.2)

Figure 1. Typical drill item, after the student has responded (incorrectly).

A fundamental aspect of the system is that students can continue requesting and answering *ad infinitum*. They receive immediate feedback, usually including a detailed solution (see Fig. 1). In-class surveys indicate

[2] *http://plone.org*
[3] *http://zodb.org*

that students like the system as an addition to traditional homework, particularly the detailed solutions (see below). Naturally, students can monitor their own progress. Several grading schemes can be implemented, but using the last 8 answers, as in the analysis presented here, was the norm until late 2013.

An Item Allocation Algorithm (IAA) is used to choose drill items (questions) for learning, within each lecture. Important aspects of an IAA include the desire to start with easy items and increase difficulty with increasing grade (within a lecture). Given that repetition is known to enhance learning, the IAA also occasionally chooses an item from earlier material (lectures). It is likely to be useful to choose again from earlier mistakes or go to prerequisites if there is no learning, but these have not been investigated to date.

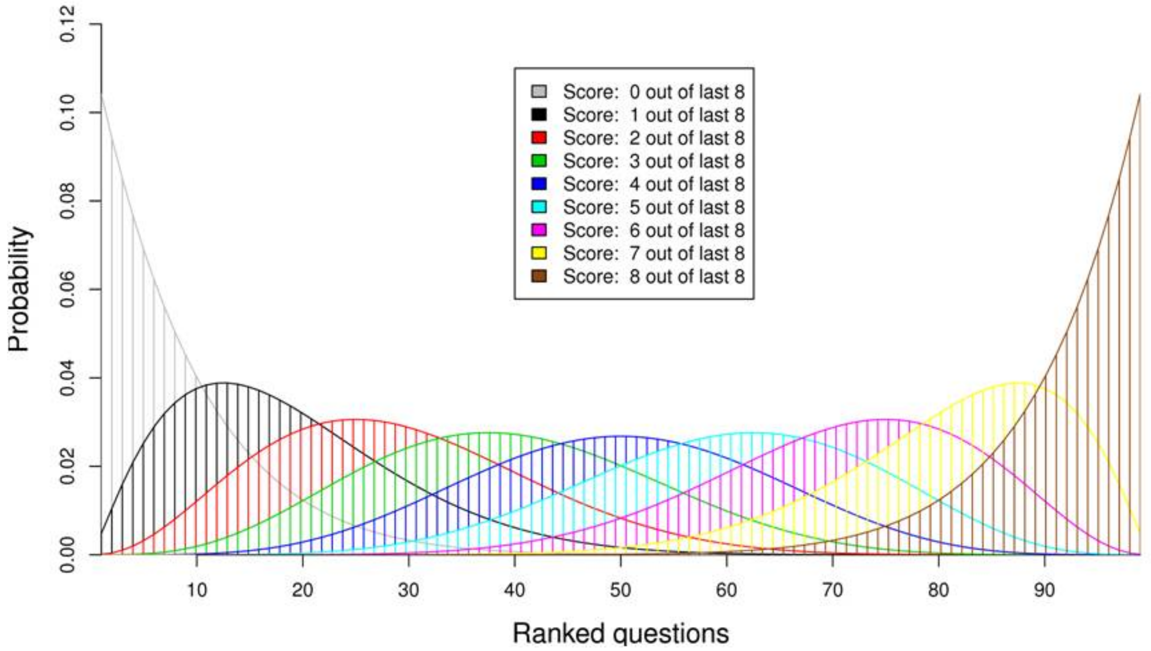


Figure 2. Tutor-web probability mass function used by the item allocation algorithm. The x-axis indicates the ranked item difficulty and the y-axis gives the probability of the next item, where the p.m.f. choses depends on the grade of the student.

The IAA is simply implemented as a probability mass function (p.m.f., Fig. 2), which is a function of difficulty. In addition, the p.m.f. depends on the grade, thus implementing personalized education appropriate for the student in question.

Student surveys are conducted in most courses using the system. A typical example of results is given in Fig. 3, where it is seen that students strongly prefer a combination of on-line exercises and traditional homework. Although it is useful to know that students appreciate a drilling system, more concrete evidence is needed in order to justify its use. One such is provided using an experimental design which compared groups of students using the system or using traditional homework in a crossover design [4]. The basic conclusion from this experiment was that the difference between the groups was insignificant, both statistically and from the point of view that the confidence interval for the two groups was very tight. From the above results it follows that the system can be used to reduce regular homework considerably, but not replace it completely.

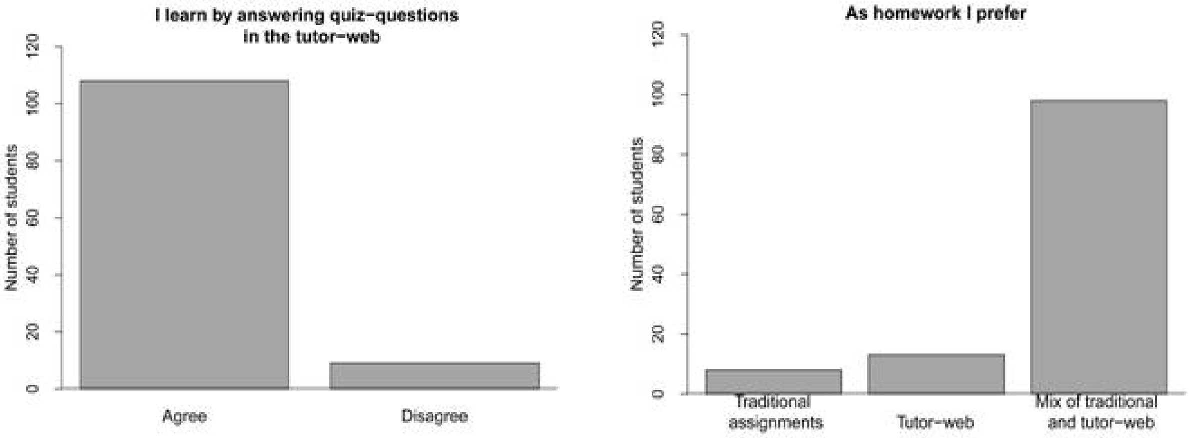


Figure 3. Student satisfaction survey results. Note how the tendency to like web-assisted methods (left panel) does NOT imply that regular homework can be dropped (right panel).

Monitoring Students

Consider next the data available to the system and how this may relate to "actual knowledge", as determined by exams, either an initial status exam or a final exam. A calculus course with complete data for 182 students is used for these analyses for the remainder of this paper. A status exam was submitted in the second week of the course. The problems on the exam covered numbers and functions, basic algebra, equation of a straight line, trigonometric, differentiation and integration, vectors and complex numbers. The performance on the status exam was poor, with an average score of 35%. Students were also evaluated multiple times during the semester, and monitored continuously using the tutor-web. In the following, summaries of the tutor-web grade and response times along with grades from an initial status exam and the final exam are used. In the tutor-web, the response time for each item is measured, along with a 0/1-grade. The items are grouped in lectures as described in section 1, with 34 lectures belonging to this particular course. As an example of data internal to the web-system, consider the average grade and average time spent on the first item in each lecture. This provides 182 pairs. Each of these can now be labelled in 4 ways, according to whether the student passed the status exam and/or the final exam. These results are given as points in Fig. 4. Notice how it is not at all clear from the figure whether there is a link between t-w performance and grades on either exam.

Since data are available both from a status exam at the beginning of the semester and a final exam, student progress can be defined as the grade improvement, i.e. the change in grade from the status exam to the final exams. Consider therefore a simple linear regression of grade improvement, on the tutor-web grade and the time used per item within the tutor-web. This regression reveals that those are indeed important variables, but relationships to performance on exams may be nontrivial. For example, one would expect the time taken to solve a problem to be a complex combination of the student's expertise and diligence. Thus a "**G**ood" student should be able to solve a problem quickly and get it right, but the "diligent" hard-working **L**earner who may not know the material very well may take longer to get the right answer. A guessing (**P**oor) student will not take long to get the wrong answer. The remaining (**U**nclassified) apparent non-learning students take long to get the wrong answer. This **GLUP** classification is derived from Fig. 4 and used below.

Relating On-Line Monitoring Results to Other Performance Measures

Relating On-Line Monitoring Results to Learning

Although there is no trivial grouping seen in the figure, consider using the **GLUP -** classification to predict actual learning, or "improvement", using a regular ANOVA. The "improvement" is defined as the change in grade from the status to the final exams, where the grade of both exams has been scaled to cover the interval from 0 to 100. The ANOVA was performed using the lm function in R [6]. The results are shown in Table 1.

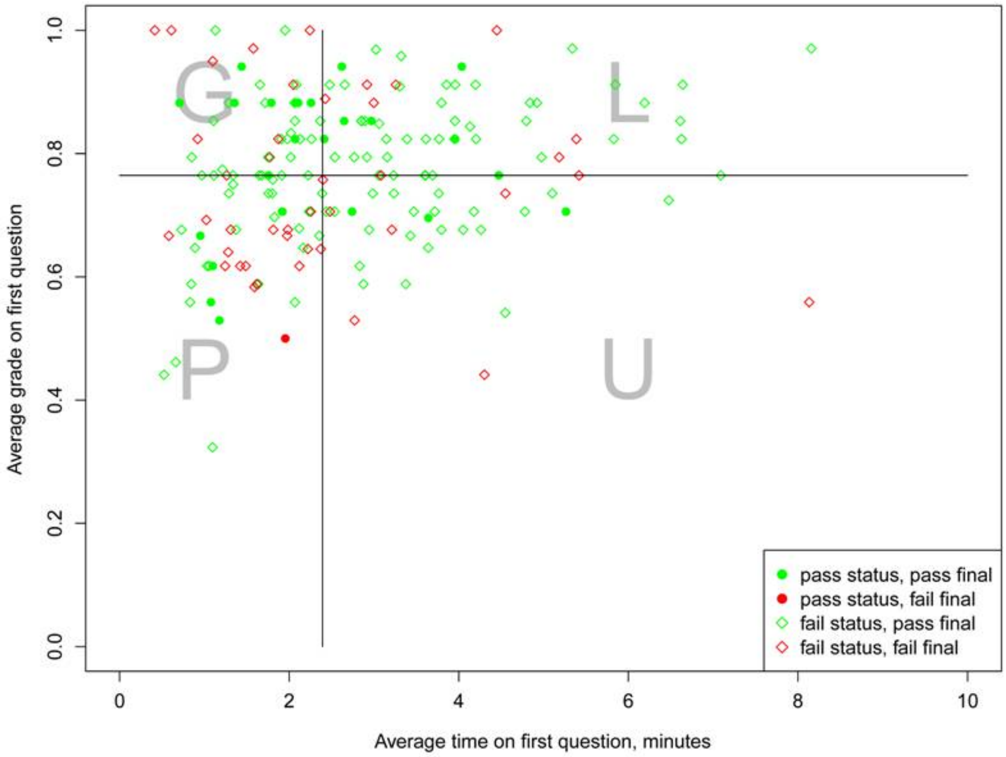


Figure 4. Plot of average grade and timing for first item request within each lecture. Vertical and horizontal lines indicate classification of students according to time and grade (using medians). Color: green/red=Pass/Fail on final exam. Shape: Circle/Diamond=Pass/Fail on status exam.

Table 1

Predicting improvement (final-status) from GLUP classification. Baseline: "Poor" students who tend to guess incorrectly and quickly. Note that although the apparent "Good" students (who quickly answer correctly) show greater improvement (class1G), the "Learners" (who spend more time to get the correct answer) show the greatest improvement, and even the "Unclassified" students who just spend a long time on each question even if they get it wrong, also show apparently greater improvement than the "Good" students (this difference is not significant)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.5582	2.7223	3.51	0.0006
class1G	6.9671	4.3282	1.61	0.1092
class1L	11.4772	3.9247	2.92	0.0039
class1U	9.0109	4.1953	2.15	0.0331

In this linear model the **P**oor students form a baseline and the estimates for the other groups can be interpreted as gain in improvement. It is therefore seen that all the other groups perform better on average than the baseline. The differences are both statistically significant and considerable.

The main results from this analysis are that the point estimate for the poor performers is the lowest among the four groups. The greatest increase from P is amongst the learners, L, but this is not significantly different from e.g. the Good students. It is interesting to note that the unclassified group (U) shows considerably (and significantly) more improvement than the poor performers (P). The only difference in their classification is the average amount of time spent on the items. Thus, although both groups perform poorly at the outset, the individuals who spent more time on each item outperformed the others by a considerable margin in terms of improvement.

Linking to Absolute Performance

Predicting the improvement during a semester, or the "value-added" is done directly above by fitting to the improvement in grade, from the initial status exam to the final exam. For several reasons it is also of interest to consider predictions of the final exam grade (finalG) directly, including the status exam as a regular explanatory variable (statusG). Many variables can in principle be defined and used. Here the average grades from different stages within the tutor-web are included (g1, g5 and gn, gn being the average grade on the last item requested in each lecture), as is the average time spent per item at different points (T1, T5 and Tn), the squared time spent per item (T1.2, T5.2 and Tn.2), an indicator variable of whether students spend more or less time on the last (usually most difficult) item compared with the first one (Tn>T1), the GLUP class variable (class1), number of items requested (twnattl) and finally the squared number of items requested (twnattl2). The model was fitted using the lm function and reduced using the step function in R [6]. The results are shown in Table 2.

Table 2

Final model selected using the AIC as a stepwise selection criterion.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-46.7506	11.0469	-4.23	0.0000
twnattl	2.4488	0.7809	3.14	0.0020
statusG	0.5211	0.0609	8.55	0.0000
g5	54.3603	10.1337	5.36	0.0000
T5	6.0022	3.6711	1.63	0.1039
Tn	2.9232	2.0771	1.41	0.1611
‘Tn > T1‘TRUE	10.3281	4.2538	2.43	0.0162
twnattl2	-0.0462	0.0182	-2.54	0.0119
T5.2	-1.0496	0.6013	-1.75	0.0826

Of the variables selected here, one has a slightly different status from the others: statusG is defined on data outside the tutor-web whereas other variables are defined completely with the on-line learning system.

As can be seen in the table, the GLUP class variable is not included, considered with the grade and time at different stages, which is not surprising since the classification is defined by those variables. It should also be noted that the squared number of attempts is significant, implying that there tends to be a reduction in grade for students who give more than 27 answers on average (per lecture). Earlier attempts at quantification of the effect of the number of attempt have given mixed output. For example, one might surmise that the number of attempts is like the time spent per item, i.e. be a measure of diligence, but there are also guessers and in fact the analyses in [8] showed a net negative linear relationship with the number of attempts. The greater number of students in the present study may be the reason why it apears to be possible to accomodate both effects using a quadratic response curve.

Note also how the (linear) effect of the time spent per item is positive (both T5 and Tn), i.e. the longer the student spends on an item the higher the final grade. As above, this is a measure of the effect of "diligence" but one must also note that the squared time spent on the 5th item was selected in this model, implying that the estimated response curve has a maximum. The point estimate corresponds to a reduced performance for students who use on average more than $T_5^* = 3$ minutes on the fifth item.

Finally, the positive coefficient on `$\mathrm{Tn} > \mathrm{T1}$`TRUE implies (conditionally on the other variables in the model) that those students who on average spend longer on the last item compared to the first one receive on average a 10.3% higher grade on the final exam (but note the large standard error). If all items had the same difficulty this would be counterintuitive, but here the difficulty goes up as the students progress and their grades increase.

Conclusions

It is clear from a number of student surveys, that students from Iceland to Kenya like an on-line drilling system, they feel they learn from it and, based on the results given here, one can statistically demonstrate this learning.

It is seen in the above results that web-assisted education can be used to enhance student learning, but this is already fairly well established. In addition, the system contains internal information on student groupings, and this is seen to including indications of whether and how much students are likely to improve their knowledge during the semester. It also appears to be possible to indicate when students have requested too many items or are too quick at answering (implying e.g. guessing or lack of attention given to detail). As this is the first time such quadratic responses have been estimated, more research is needed to fully understand the implications and statistical power of such studies.

The analysis of response times gives the intriguing result that what would seem an obvious classification of good students does not yield the students who show the greatest gains or highest results on a final exam. Although a formal experiment is needed to extract more detail, it is likely that there is confounding within the data obtained from such measurements: (a) What seems to be a "good" student in an elementary classification will also include a number of lucky guessers. (b) More difficult is the distinction between concepts actually being measured, i.e. a quick student may well be a good student showing expertise, but a student taking long on each item may simply be diligent (i.e. not slow). The quadratic response obtained would seem to illustrate this point, i.e. greatest improvements are seen for the students who spend an intermediate amount of time on each item (correcting for other variables). Future experiments designed to shed more light on this are discussed below.

Research reported elsewhere [4] implies that student learning is almost the same, regardless of whether an on-line system or traditional homework is used. As in-class surveys consistently indicate that students prefer to also get graded homework, it is not possible to replace all homework by computerised drills, but one can easily replace half the homework by on-line multiple-choice questions.

Since the instructor can make the drills form a part of the final grade and can set minimum return requirements as criterion for passing, this gives considerable potential for changes in emphases or reductions in instructor workload.

Discussion: Avenues of Research

For an actual real-world course, formal requirement are set by the instructors, not the system. Applications of the tutor-web system have therefore varied in student requirements. The above results imply, however, that it may be beneficial to incorporate features which drive the students towards certain behavior or performance.

In the course studied here, as well as in other courses where this system has been tested [7] students tend to work towards a fairly high grade (median $g_n = 0.94$ and median of last 8 is 0.92 in the present course). Hence changes to either the item allocation algorithm or the grading scheme will likely lead to a change in behavior where the students still work towards a goal of a high grade, assuming it is still a feasible goal. Similarly, a timeout option is also likely to lead to changes in student behavior. A generic positive system change has benefits over an instructor-defined criterion since it will affect all students at all times, not just the course in question.

The students appear to gain (in terms of exam grade) through requesting more items (up to 27) than normally required (8 for this course) or normally taken (median=15, upper 75% quartile=20). It would therefore seem reasonable to encourage an increase in the number of items requested by students.

The current “last 8” internal tutor-web grade assumes incorrect answers until at least 8 questions have been answered in a given lecture. Most students therefore answer at least 8 questions in each lecture. This scheme, however, implies that if the 8th answer is incorrect after a run of 7 answers, the grade will not increase unless a new run of 8 correct answers is obtained. Many students stop at this stage and this behavior is contrary to the goal of positive reinforcement. A simple change would be to use the most recent 30 answers, or, more generally, to use for grading the most recent

$$n_g = \max(8, \min(n/2, 30))$$

answers, possibly tapered, where n is the total number of answers given. This will penalise the guesser by introducing a longer tail and simultaneously give reduced weight to the accidental 8th incorrect response. A next-generation mobile-web version of the tutor-web will evaluate multiple grading schemes, including these. This will facilitate a simple experiment to investigate the relationship between the grading scheme and the number of attempts per lecture.

Although the tutor-web is a significant predictor of the final grade, it is not a very good one. For example, of the 113 students who obtain a grade of over 90% on the tutor-web work, 34% do not attain a grade of 50% on the final exam. The main problem with this is that the tutor-web grade is not a reliable indicator for the students themselves. The students with full marks, 100% on the tutor-web, have an 83% chance of passing the exam however. From this it is seen that the tutor-web grade is “too high” in the sense that it indicates more knowledge than is estimated using traditional exams. Future work therefore needs to investigate whether changes in the grading scheme, to the effect of lowering most grades, can provide better indicators of exam performance. The extended tail described above could be one aspect of this.

Another way of "reducing the tutor-web grade" is to include timeout features. Such a timeout could be a function of grade, i.e. a student can only get into a certain grade range by answering questions correctly within certain time limits.

A different reason for considering a timeout option is the general worry that many students of calculus appear not to have elementary algebra at their fingertips. Thus there is a need to investigate whether "expertise" is a concept which can be developed using the tutor-web system by enforcing some sort of time limit on such elementary issues.

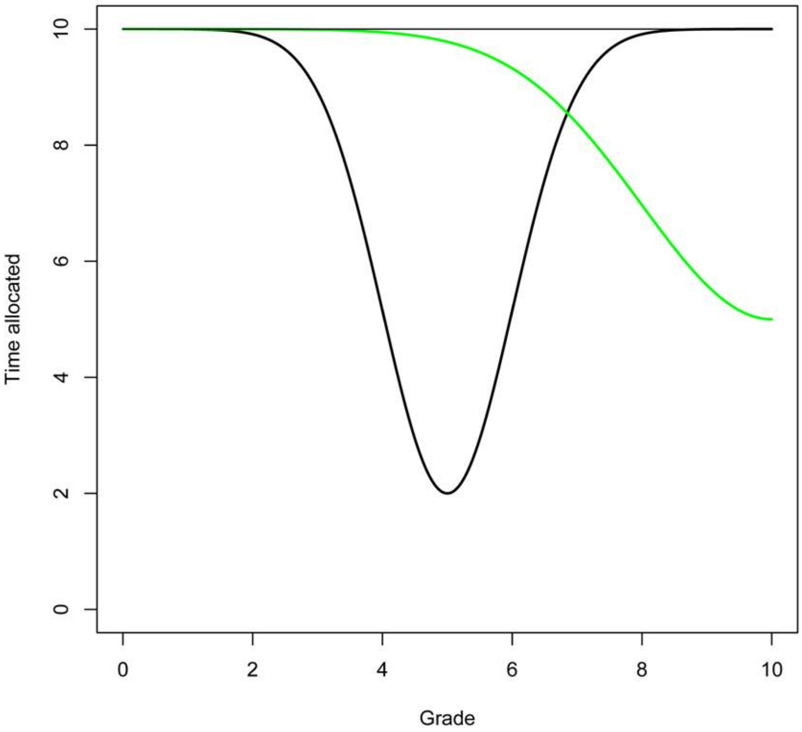


Figure 5. Possible curves to define time allocated to items, as a function of grade. Top solid line: Fixed timeout. Central inverse dome: Timeout set to enforce a threshold of expertise before continuing to higher grade and more difficult items. Green curve: Timeout set to ensure that a high grade is only achieved if the most difficult items are completed quickly.

This will almost certainly keep students working longer within grade intervals with a timeout and this could be used e.g. to ensure expertise within easier items before continuing. With the exception of the very best students, this approach should also increase the number of attempts needed, since this will make simply it harder to obtain a higher grade.

To quantify the effect of the timeout, a future experiment could focus on a single parameter in a formula such as

$$t = a\left[1-\left(1-\frac{b}{a}\right)e^{-\frac{(g-g^*)^2}{2s^2}}\right]$$

which will give an upside-down bell-curve with an upper bound of $t = a$ and a minimum of $t = b$ at $g = g^*$. Given that the median time in the course discussed here is about 2 minutes, one could take e.g.

$a = 10$, $b = 2$, $g^* = 5$ and $s = 1$ as initial values (central inverted dome in Fig. 5) and set up a formal experimental design by selecting either b or g^* at random from within some intervals for each student within each lecture. Performance can be evaluated statistically either by how the number of attempts within a lecture changes as a function of b or by how the performance on an algebra item in an exam varies as a function of b. This particular choice of parameter values enforces a bottleneck where the students have to obtain a certain level of expertise before getting above a certain grade, upon which the timeout parameter is no longer limiting.

Given the complex relationship described in this paper, between time spent on each item and subsequent performance, it is not trivial to predict the full effect of any timeout parameter settings. Other issues also come into account, such as how item difficulty is assessed and why an item is difficult.

For example, one might consider a different approach, using a higher g^* and s, to set a similar limit access to the higher grades (green curve in Fig. 5). This may backfire since the most difficult items will often simply take longer and thus not be doable within a short time implied by the timeout option. Further, item difficulty in the present system is simply assessed based on the proportion of incorrect answers. If the inverse bell has a minimum at an intermediate value for the grade (where medium-difficulty items are answered), one would predict that items which can not be answered within the minimum time will (eventually) be ranked more difficult and moved to a difficulty level where the timeout does not apply. However if the minimum is at the highest grade (and most difficult items), then these items stay in place and can not be answered correctly by anyone.

Finally, since the Poor students (in the GLUP classification) are the poorest performers by all measures, one needs to consider methods to move these students into the otherwise Unclassified group, who spend more time on each item. When students have answered a question the system provides a detailed explanation of how the answer is obtained (most items have such explanations). A possible method to slow these students down is therefore to use pop-ups, such as a warning when a student has answered incorrectly and clearly asks for the next item without first reading the explanation. The net effect of this can easily be tested by randomly assigning such stop-signs to half the P-students and evaluating whether there is a statistical difference in how they move out of the P group.

Acknowledgements

The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 613571 - MareFrame.

The tutor-web project has also been supported by the University of Iceland, the United Nations University, the Marine Research Institute in Reykjavik, Iceland and the Ministry of Education in Iceland. Numerous authors have contributed educational material to the system. The system is based on Plone[4] and educational material is mainly written in LATEX. Examples and plots are mostly driven by R [6]. The Plone version of the tutor-web used here was implemented by Audbjorg Jakobsdottir, but many computer programmers have contributed pieces of code during the lifetime of the project.

The calculus material content used as a basis for this research has been developed by several faculty and students at the University of Iceland.

[4] http://www.plone.org

This paper is based on analyses first presented at JSM 2013 and earlier drafts have been available through arxiv.

The University of Iceland Research Ethics Committee has been informed of the study. In accordance with the Code of Research Ethics of the University of Iceland (in Icelandic: http://bit.ly/1f1jajK), article 2.4.4, informed consent is not required for studies of this type.

References

P. Brusilovsky. Adaptive and intelligent technologies for web-based education. *Kunstliche Intelligenz*, 4:19–25, 1999.

P. Brusilovsky and C. Peylo. Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, 13(2-4):159–172, 2003.

J. Ismail. The design of an e-learning system:: Beyond the hype. *The internet and higher education*, 4(3-4):329–336, 2001.

A.H. Jonsdottir and G. Stefansson. Enhanced learning with web-assisted education. In *JSM Proceedings, Section on Statistical Education: American Statistical Association*, 2011.

G. Piccoli, R. Ahmad, and B. Ives. Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic it skills training. *Mis Quarterly*, pages 401–426, 2001.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

G. Stefansson. The tutor-web: An educational system for classroom presentation, evaluation and self-study. *Computers & Education*, 43(4):315–343, 2004.

G. Stefansson and A. J. Sigurdardottir. Interactive quizzes for continuous learning and evaluation. In *JSM Proceedings, Statistical Education Section. Alexandria, VA: American Statistical Association.*, pages 4577–4591, 2009.