Journal of Statistics Education

ISSN: (Print) 1069-1898 (Online) Journal homepage: http://www.tandfonline.com/loi/ujse20

Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course

Anna Helga Jonsdottir, Audbjorg Bjornsdottir & Gunnar Stefansson

To cite this article: Anna Helga Jonsdottir, Audbjorg Bjornsdottir & Gunnar Stefansson (2017): Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course, Journal of Statistics Education

To link to this article: http://dx.doi.org/10.1080/10691898.2017.1291289

Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course

Abstract

A repeated crossover experiment comparing learning among students handing in pen-and-paper homework (PPH) with students handing in web-based homework (WBH) has been conducted. The system used in the experiments, the tutor-web, has been used to deliver homework problems to thousands of students in mathematics and statistics over several years. Since 2011 experimental changes have been made regarding how the system allocates items to students, how grading is done and the type of feedback provided. The experiment described here was conducted annually from 2011 to 2014. Approximately 100 students in an introductory statistics course participated each year. The main goals were to determine whether the above mentioned changes had an impact on learning as measured by test scores in addition to comparing learning among students doing PPH with students handing in WBH.

The difference in learning between students doing WBH compared to PPH, measured by test scores, increased significantly from 2011 to 2014 with an effect size of 0.634. This is a strong indication that the changes made in the NAMEOFSYSTEM have a positive impact on learning. Using the data from 2014 a significant difference in learning between WBH and PPH for 2014 was detected with an effect size of 0.416 supporting the use of WBH as a learning tool.

Key Words: Web-based homework (WBH), pen-and-paper homework (PPH), learning environments, repeated crossover experiments, statistics education.

1 Introduction

Enrolment to universities has increased substantially in the past decade in most OECD (Organisation for Economic Co-operation and Development) countries. In COUNTRYOFAUTHORS, the

increase in tertiary level enrolment was 40% between 2000 and 2010 (OECD, 2013). This increase has resulted in larger class sizes at the University of COUNTRYOFAUTHORS, especially in undergraduate courses. As stated in Black and Wiliam (1998), several studies have shown firm evidence that innovations designed to strengthen the frequent feedback that students receive about their learning yield substantial learning gains. Providing students with frequent quality feedback is time consuming and in large classes this can be very costly. It is therefore of importance to investigate whether web-based homework (WBH), that does not require marking by teachers but provides feedback to students, can replace (at least to some extent) traditional pen-and-paper homework (PPH). To investigate this, an experiment has been conducted over a four year period in an introductory course in statistics at the University of COUNTRYOFAUTHORS. About 100 students participated each year. The experiment is a *repeated crossover experiment* so the same students were exposed to both methods, WBH and PPH.

The learning environment *NAMEOFSYSTEM* (`http://NAMEOFSYSTEM.net`) used in the experiments has been under development during the past decade in the University of COUNTRYOFAUTHORS. Two research questions are of particular interest:

1. Have changes made in the NAMEOFSYSTEM had an impact on learning, as measured by test performance?

2. Is there a difference in learning, as measured by test performance, between students doing WBH and PPH after the changes made in the NAMEOFSYSTEM?

In this section, an overview of different learning environments in the context of the functionality of the NAMEOFSYSTEM is given (Section 1.1), focusing on how to allocate exercises (problems) to students. A literature review of studies, conducted to investigate a potential difference in learning between WBH and PPH, is given in Section 1.2 followed by a brief discussion about formative assessment and feedback (Section 1.3). Finally a short description of the NAMEOFSYSTEM is given in Section 1.4.

1.1 Web-based learning environments

A number of web-based learning environments are available on the web, some open and free to use, others commercial products. Several types of systems have emerged, including the *learning management systems* (LMS), *learning content management systems* (LCMS) and *adaptive and intelligent web-based educational systems* (AIWBES). The LMS is designed for planning, delivering and managing learning events, usually adding little value to the learning process nor supporting internal content processes while the primary role of a LCMS is to provide a collaborative authoring environment for creating and maintaining learning content (Ismail, 2001). In AIWBES the focus is on the student. Such systems adapt to the needs of each and every student (Brusilovsky & Peylo, 2003) in contrast to many systems that are merely a network of static hypertext pages (Brusilovsky, 1999).

A number of web-based learning environments use intelligent methods to provide personalized content or navigation such as the one described in Own (2006). However, only few systems use intelligent methods for exercise item allocation (Barla et al., 2010). The use of intelligent item allocation algorithms (IAA) is, however, a common practice in testing. Computerized Adaptive Testing (CAT) (Wainer, 2000) is a form of computer-based tests where the test is tailored to the examinees ability level by means of Item Response Theory (IRT). IRT is the framework used in psychometrics for the design, analysis and grading of computerized tests to measure abilities (Lord, 1980). As Wauters, Desmet, and Van Den Noortgate (2010) argue, IRT is potentially a valuable method for adapting the item sequence to the learners knowledge level. However, the IRT methods are designed for *testing*, not *learning*, and as shown in AUTHREF and AUTHREF the IRT models are not appropriate since they do not take learning into account. New methods for IAA in learning environments are therefore needed.

Several systems can be found that are specifically designed for providing content in the form of exercise items. Examples of systems providing homework exercises are the WeBWork system (Gage, Pizer, & Roth, 2002), ASSiSTments (Razzaq et al., 2005), ActiveMath (Melis et al., 2001), OWL (Hart, Woolf, Day, Botch, & Vining, 1999), LON-CAPA (Kortemeyer, Kashy, Benenson, & Bauer, 2008) and WebAssign (Brunsmann, Homrighausen, Six, & Voss, 1999). None of those

systems use intelligent methods for item allocation, instead a fixed set of items are submitted to the students or drawn randomly from a pool of items.

1.2 Web-based homework vs. pen-and-paper homework

A number of studies have been conducted to investigate a potential difference in learning between WBH and PPH. In most of the studies reviewed, no significant difference was detected (Bonham, Deardorff, & Beichner, 2003; Cole & Todd, 2003; Demirci, 2007; Gok, 2011; Kodippili & Senaratne, 2008; LaRose, 2010; Lenz, 2010; Palocsay & Stevens, 2008; Williams, 2012). In three of the studies reviewed, WBH was found to be more effective than PPH as measured by final exam scores. In the first study, described in Dufresne, Mestre, Hart, and Rath (2002), data was gathered in various offerings of two large introductory physics courses taught by four lecturers over a three year period. The OWL system was used to deliver WBH. The authors found that WBH lead to higher overall exam performance, although the difference in average gain for the five instructor-course combinations was not statistically significant. In the second paper, VanLehn et al. (2005) describe Andes, a physics tutoring system. The performance of students working in the system was compared to students doing PPH homework for four years. Students using the system did significantly better on the final exam than the PPH students. However, the study has one limitation; the two groups were not taught by the same instructors. Finally, Brewer and Becker (2010) describe a study in multiple sections of college algebra. The WBH group used an online homework system developed by the textbook publisher. The authors concluded that the WBH group generally scored higher on the final exam but no significant difference existed between mathematical achievement of the control and treatment groups except in low-skilled students where the WBH group exhibited significantly higher mathematical achievement.

Even though most of the studies performed comparing WBH and PPH show no difference in learning, the fact that students do not do worse than students doing PPH makes WBH a favourable option, especially in large classes where correcting PPH is very time consuming. Also, students' perception towards WBH has been shown to be positive (Demirci, 2007; Hauk & Segalla, 2005; Hodge, Richardson, & York, 2009; LaRose, 2010; Roth, Ivanchenko, & Record, 2008; Smolira,

2008; VanLehn et al., 2005).

All the studies reviewed were conducted using a quasi-experimental design, i.e. students were not randomly assigned to the treatment groups. Either multiple sections of the same course were tested where some sections did PPH while the other(s) did WBH or the two treatments were assigned on different semesters. This could lead to some bias e.g. due to difference in the student groups or lecturers participating in the two treatment arms of the experiments. The experiment described in this paper is a *repeated randomized crossover experiment* so the same students were exposed to both WBH and PPH, resulting in a more accurate estimate of the potential difference between the two methods.

1.3 Assessment and feedback

Assessments are frequently used by teachers to assign grades to students (assessment *of* learning) but a potential use of assessment is to use it as a part of the learning process (assessment *for* learning) (J. Garfield et al., 2011). The term *summative assessment* (SA) is often used for the former and *formative assessment* (FA) for the latter. The concepts of *feedback* and FA overlap strongly and, as stated in Black and Wiliam (1998), the terms do not have a tightly defined and widely accepted meaning. Therefore, some definitions will be given below.

Taras (2005) defines SA as "... a judgement which encapsulates all the evidence up to a given point. This point is seen as a finality at the point of the judgement" (p. 468) and about FA she writes "... FA is the same process as SA. In addition for an assessment to be formative, it requires feedback which indicates the existence of a 'gap' between the actual level of the work being assessed and the required standard" (p. 468). A widely accepted definition of *feedback* is then provided in Ramaprasad (1983): "Feedback is information between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p. 4).

Stobart (2008) suggests making the following distinction between the *complexity* of feedback; *knowledge of results* (KR) only states whether the answer is incorrect or correct, *knowledge of correct response* (KCR) where the correct response is given when the answer is incorrect and *elaborated feedback* (EF) where, for example, an explanation of the correct answer is given.

The terms formative assessment, feedback and the distinction between the different types of feedback will be used here as defined above.

1.4 The NAMEOFSYSTEM

The NAMEOFSYSTEM (`http://NAMEOFSYSTEM.net`) project is an ongoing research project. The functionalities of the system have changed considerable during the past decade. A pilot version, written only in HTML and Perl, is described in AUTHREF. A newer version, implemented in Plone (Nagle, 2010), is described in detail in AUTHREF. The newest version, described in AUTHREF, is a mobile-web site and runs smoothly on tablets and smart phones. Also, users do not need to be connected to the Internet when answering exercises, but only when downloading the item banks.

The NAMEOFSYSTEM is an LCMS including exercise item banks within mathematics and statistics. The system is open and free to use for everyone having access to the web. At the heart of the system is the formative assessment. Intelligent methods are used for item allocation in such a way that the difficulty of the items allocated adapts to the students' ability level. Since the focus of the experiment described here is on the effect of doing exercises (answering items) in the system, only functionalities related to that will be described. A more detailed description of the NAMEOFSYSTEM is given in the above mentioned papers.

1.4.1 Item allocation algorithm

In the systems used for WBH named in Section 1.1 a fixed set of items are allocated to students or drawn randomly, with uniform probability, from a pool of items. This was also the case in the first version of the NAMEOFSYSTEM. A better way might be to implement an IAA so that the difficulty of the items adapts to the students' ability. As discussed in Section 1.1, current IRT methods are not appropriate when the focus is on learning, therefore a new type of IAA has been developed using the following basic criteria:

- Increase the difficulty level as the student learns
- select items so that a student can only complete a session with high grade by completing the

most difficult items

- select items from previous sessions to refresh memory.

Items are grouped into *lectures* in the NAMEOFSYSTEM system where each lecture covers a specific topic. This could be *discrete distributions* in material used in an introductory course in statistics or *limits* in a basic course in calculus. Within a lecture, the difficulty of an item is simply calculated as the ratio of incorrect responses to the total number of responses. The items are then ranked according to their difficulty, from the easiest item to the most difficult one.

The implementation of the first criterion (shown above) has changed over the years. In the first version of the NAMEOFSYSTEM all items within a lecture were assigned uniform probability of being chosen for every student. This was changed in 2012 with the introduction of a *probability mass function* (pmf) that calculates the probability of an item being chosen for a student. The pmf is *exponentially* related to the ranking of the item and also depends on the student's grade:

$$p(r) = \begin{cases} \frac{q^r}{c} \cdot \frac{m-g}{m} + \frac{g}{N \cdot m} & \text{if } g \leq m, \\ \\ \frac{q^{N-r+1}}{c} \cdot \frac{g-m}{1-m} + \frac{1-g}{N \cdot (1-m)} & \text{if } g > m, \end{cases} \tag{1}$$

where q is a constant ($0 \leq q \leq 1$) controlling the steepness of the function, N is the total number of items belonging to the lecture, r is the difficulty rank of the item ($r = 1, 2, ...N$), g is the grade of the student ($0 \leq g \leq 1$) and c is a normalizing constant, $c = \sum_{i=1}^{N} q^i$. Finally, m is a constant ($0 < m < 1$) so that when $g < m$, the pmf is strongly decreasing and the mass is mostly located at the easy items, when $g = m$ the pmf is uniform and when $g > m$ the pmf is strongly increasing with the mass mostly located at the difficult items. This was changed in 2013 in such a way that the mode of the pmf moves to the right with increasing grade which is achieved by using the following pmf based on the *beta* distribution:

$$p(r) = \frac{1}{\sum_{i=1}^{N} \left(\frac{i}{N+1}\right)^{\alpha} \cdot \left(1 - \frac{i}{N+1}\right)^{\beta}} \left(\frac{r}{N+1}\right)^{\alpha} \cdot \left(1 - \frac{r}{N+1}\right)^{\beta}, \tag{2}$$

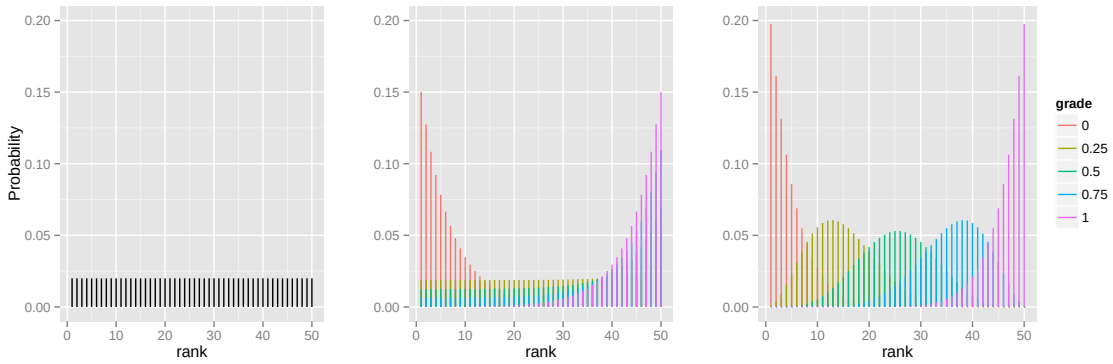

Figure 1: The different probability mass functions used in the item allocation algorithm. Left: uniform. Middle: exponential. Right: beta.

where r is the ranked item difficulty ($r = 1, 2, ..., N$) and α and β are constants controlling the shape of the function. The three different pmfs used over the years (uniform, exponential and beta) are shown in Figure 1. Looking at the last figure, showing the pmf currently used, it can be seen that a beginning student (with a score 0) receives easy items with high probability. As the grade increases the mode of the probability mass functions shifts to the right until the student reaches a top score resulting in a high probability of getting the most difficult items. Using this pmf, the first two of the criteria for the IAA listed above are fulfilled.

The last criterion for the IAA is related to how people forget. Ebbinghaus (1913) was one of the first to research this issue. He proposed the *forgetting curve* and showed in his studies that learning and the recall of learned information depends on the frequency of exposure to the material. It was therefore decided in 2012 to change the IAA in such a way that students are now occasionally allocated items from previous lectures to refresh memory.

1.4.2 Grading

Although the main goal of making the students answer exercises in the NAMEOFSYSTEM is learning there is also a need to evaluate the students' performance. The students are permitted to continue to answer items until they (or the instructor) are satisfied, which makes grading a non-trivial issue. In the first version of the NAMEOFSYSTEM, the last eight answers counted (with equal weight) towards the NAMEOFSYSTEM grade. Students were given one point for a correct answer and mi-

nus half a point for an incorrect one. The idea was that old sins should be forgotten when students are learning. This had some bad side effects with students often quitting answering items after seven correct attempts in a row AUTHREF, which is a perfectly logical result since a student who has a sequence of seven correct and one incorrect will need another eight correct answers in sequence to increase the grade. The NAMEOFSYSTEM grade was also found to be a bad predictor of students' performance on a final exam, the grade being too high AUTHREF. It was therefore decided in 2014 to change the grading scheme (GS) and use $min(max(n/2, 8), 30)$ items after n attempts when calculating the NAMEOFSYSTEM grade. That is, use a minimum of eight answers, then after eight answers use $n/2$, but no more than 30 answers. Using this GS, the weight of each answer is less than before (when $n > 8$), thus eliminating the fear of answering the eighth item incorrectly, simultaneously making it more difficult for students to get a top grade since more answers are used when calculating the grade.

1.4.3 Feedback

The quality of the feedback is a key feature in any procedure for formative assessment (Black & Wiliam, 1998). In the first version of the NAMEOFSYSTEM, only KR/KCR type feedback was provided. Sadler (1989) suggested that KR type feedback is insufficient if the feedback is to facilitate learning so in 2012 an explanation was added to items in the NAMEOFSYSTEM item bank, thus providing students with EF. A question from a lecture covering inferences for proportions is shown in Figure 2. Here the student has answered incorrectly (marked by red). The correct answer is marked with green and an explanation given below.

1.4.4 Summary of changes in the NAMEOFSYSTEM

In the sections above, changes related to the IAA, grading and feedback were reviewed. A summary of the changes discussed is shown in Table 1.

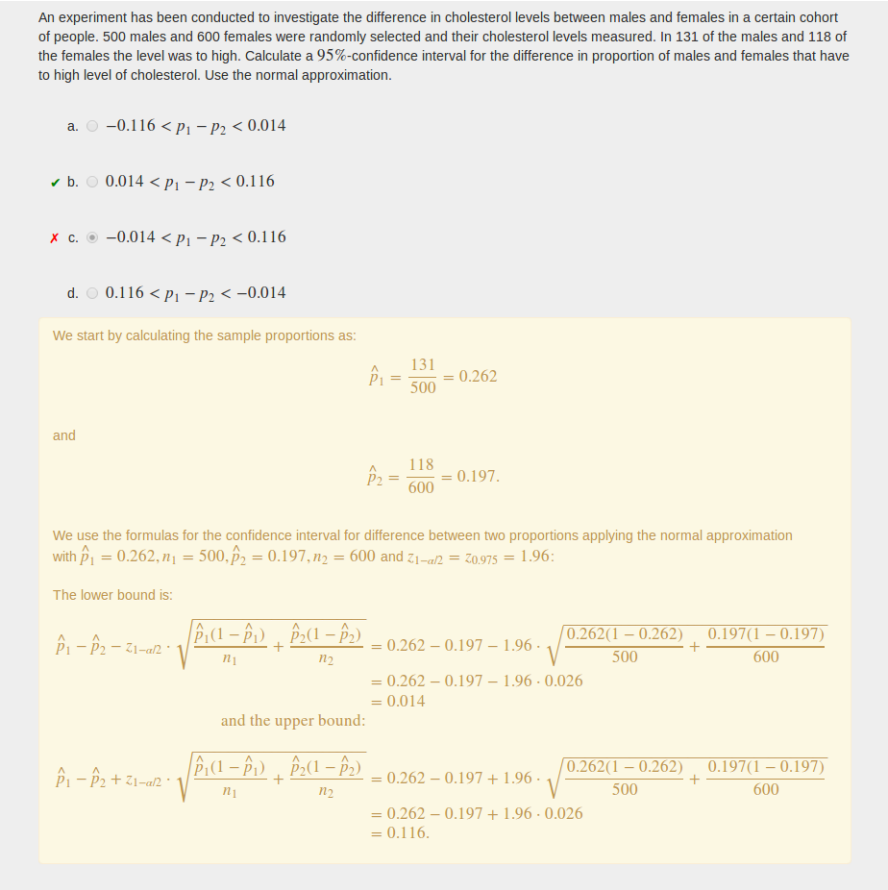

Figure 2: A question from a lecture on inferences for proportions. The students are informed what the correct answer is and shown an explanation of the correct answer.

Table 1: Summary of changes in the NAMEOFSYSTEM.

Year	IAA difficulty	IAA refresh memory	Grading	Feedback	Mobile-web
2011	uniform	no	last 8	KR/KCR	no
2012	exponential	yes	last 8	EF	no
2013	beta	yes	last 8	EF	no
2014	beta	yes	min(max(n/2,8),30)	EF	yes

2 Material and methods

The data used for the analysis was gathered in an introductory course in statistics in the University of COUNTRYOFAUTHORS from 2011-2014. Every year some 200 first year students in chemistry,

biochemistry, geology, pharmacology, food science, nutrition, tourism studies and geography were enrolled in the course. The course was taught by the same instructor over the timespan of the experiment. About 60% of the students had already taken a course in basic calculus the semester before while the rest of the students had much weaker background in mathematics. Around 60% of the students were females and 40% males. The students needed to hand in homework four times during the course. The subjects of the homework were: discrete distributions, continuous distributions, inference about means and inference about proportions. The students were told in the beginning of the course that there would be several in-class tests during the semester but they were not told how many, at what timepoints or from which topics they would be examined in. The final grade in the course consisted of four parts, the final exam (50%), the four homework assignments (10%), in-class tests (15%) and assignments in the statistical software R (25%).

The experiment conducted is a *repeated randomized crossover experiment*. The design of the experiment is shown in Figure 3.

Each year the class was split randomly into two groups. One group was instructed to do exercises in the NAMEOFSYSTEM system in the first homework assignment (WBH) while the other group handed in written homework (PPH). The exercises on the PPH assignment and in the NAMEOFSYSTEM were similar and covered the same topics. Shortly after the students handed in their homework they took a test in class. The groups were crossed before the next homework, that is, the former WBH students handed in PPH and vice versa and again the students were tested. Each year this procedure was repeated and the test scores from the four exams registered. The students were not made aware of the experiment but were told that the groups were made to manage the number PPH homework that needed to be corrected at a time. There were no indications that the students

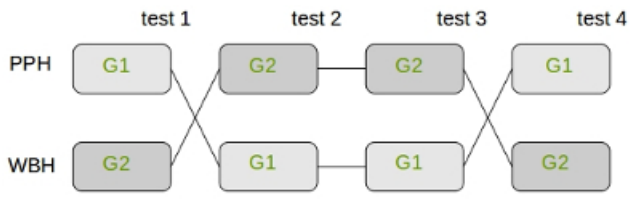

Figure 3: The design of the experiment. The experiment was repeated four times from 2011-2014.

Table 2: Number of students taking the tests.

	Discrete	Continuous	Means	Proportions
2011	91	84	122	115
2012	113	113	100	65
2013	117	123	110	99
2014	129	130	111	110

were aware of the experiment. The number of students taking each exam is shown in Table 2.

To answer the first research question, stated in Section 1, the following linear mixed model is fitted to the data from 2011-2014 and nonsignificant factors removed:

$$g_{mlhyi} = \mu + \alpha_m + \beta_l + \gamma_h + \delta_y + (\alpha\gamma)_{mh} + (\beta\gamma)_{lh} + (\delta\gamma)_{yh} + s_i + \epsilon_{mlhyi} \tag{3}$$

where g is the test grade, α is the *math background* (m = weak, strong), β is the *lecture material* (l = discrete distributions, continuous distributions, inference about means, inference about proportions), γ is the type of *homework* (h = PPH, WBH), δ is the *year* (y = 2011, 2012, 2013, 2014) and s is the random student effect ($s_i \sim N(0, \sigma_s^2)$). The interaction term ($\alpha\gamma$) measures whether the effect of type of homework is different between students with strong and weak math background and ($\beta\gamma$) whether the effect of type of homework is different for the lecture material covered. The interaction term ($\delta\gamma$) is of special interest since it measures the effect of changes made in the NAMEOFSYSTEM system during the four years of experiments.

To answer the second research question, only data gathered in 2014 is used and the following linear mixed model fitted to the data:

$$g_{mlhi} = \mu + \alpha_m + \beta_l + \gamma_h + (\alpha\gamma)_{mh} + (\beta\gamma)_{lh} + s_i + \epsilon_{mlhi} \tag{4}$$

with α, β, γ and s as above. If the interaction terms are found to be nonsignificant, the γ factor is of special interest since it measures the potential difference in learning between students doing WBH and PPH.

In addition to collecting the exam grades, the students answered a survey at the end of each semester. 442 students in total responded to the surveys (121 in 2011, 88 in 2012, 131 in 2013 and 102 in 2014). Two of the questions are related to the use of the NAMEOFSYSTEM and the students' perception of WBH and PPH homework:

1. Do you learn by answering items in the tuto-web? *(yes/no)*
2. What do you prefer for homework? *(PPH/WBH/Mix of PPH and WBH)*

3 Results

3.1 Analysis of exam scores

In order to see which factors relate to exam scores the linear mixed model in Eq. (3) was fitted to the exam score data using R (R Core Team, 2014). The `lmer` function in the `lme4` package, which includes functions to fit linear and generalized linear mixed-effects models (Bates, Maechler, Bolker, & Walker, 2014), was used. The interaction terms (*mh*) and (*lh*) were found to be nonsignificant and therefore removed from the model. This indicates that the effect of homework type does not depend on math background nor lecture material covered. However, the (*yh*) interaction was found to be significant implying that the effect of the type of homework is not the same during the four years. The resulting final model can be written as:

$$g_{mlhyi} = \mu + \alpha_m + \beta_l + \gamma_h + \delta_y + (\delta\gamma)_{yh} + s_i + \epsilon_{mlhyi} \tag{5}$$

The estimates of the parameters and the associated t-values are shown Table 3 along with p-values calculated using the `lmerTest` package (Kuznetsova, Brockhoff, & Christensen, 2013). Estimates of the variance components were $\hat{\sigma}_s^2 = 1.84$ and $\hat{\sigma}^2 = 3.33$. The reference group (included in the *intercept*) are students in the 2011 course with weak math background handing in PPH on discrete distributions. Residual plots revealed no violation of model assumptions, such as non-normal errors or random effects.

Table 3: Parameter estimates for the final model used to answer research question 1. The reference group are students in the 2011 course with weak math background handing in PPH on discrete distributions. Grades were given on the 0 - 10 scale.

Parameter estimates	Estimate	Std. Error	df	t value	Pr($>\|t\|$)
(Intercept)	4.416	0.211	1123.789	20.957	0.000
year2012	0.326	0.244	1039.348	1.336	0.182
year2013	0.785	0.234	1039.243	3.349	0.001
year2014	0.540	0.234	1013.152	2.313	0.021
WBH	−0.228	0.186	1206.998	−1.229	0.219
strongMath	1.680	0.146	580.124	11.515	0.000
test2	1.255	0.126	1236.322	9.924	0.000
test3	0.015	0.128	1250.851	0.117	0.907
test4	1.337	0.133	1268.752	10.057	0.000
year2012:WBH	0.519	0.267	1220.682	1.942	0.052
year2013:WBH	0.201	0.259	1244.169	0.774	0.439
year2014:WBH	**0.634**	0.252	1189.315	2.515	0.012

By looking at the estimate for the *year2014:tw* term it can be noticed that the difference between the WBH and PPH groups is significantly different in 2011 (the reference group) and 2014 (p = 0.012), indicating that the changes made to the NAMEOFSYSTEM had a positive impact on learning. The difference in effect size between WBH and PPH in 2011 and 2014 is 0.634. It should also be noted that the effect size of math background is large (1.680).

In order to answer the second question, the model in Eq. 4 was fitted to the data from 2014. The interaction terms were both nonsignificant and therefore removed from the model. The final model can be written as:

$$g_{mlhi} = \mu + \alpha_m + \beta_l + \gamma_h + s_i + \epsilon_{mlhi} \tag{6}$$

The estimates of the parameters, the associated t- and p-values are shown Table 4. Estimates of the variance components were $\hat{\sigma}_s^2 = 1.48$ and $\hat{\sigma}^2 = 2.84$. The reference group (included in the *intercept*) are students with weak math background handing in PPH on discrete distributions. By looking at the table it can be noted that the difference between the WBH and PPH groups is significant (p = 0.009) and the estimated effect size is 0.416 indicating that the students did better after handing in WBH than PPH. Again, the effect size of math background is large (1.379).

Table 4: Parameter estimates for the final model used to answer research question 2. The reference group (included in the *intercept*) are students with weak math background handing in PPH on discrete distributions. Grades were given on the 0 - 10 scale.

Paramter estimates	Estimate	Std. Error	df	t value	Pr($>\|t\|$)
(Intercept)	5.080	0.239	349.520	21.279	0.000
mathStrong	1.379	0.251	158.556	5.502	0.000
test2	0.137	0.216	347.434	0.633	0.527
test3	1.254	0.228	360.445	5.493	0.000
test4	1.719	0.228	358.667	7.538	0.000
WBH	**0.416**	0.158	336.485	2.640	0.009

3.2 Analysis of student surveys

In general, the students' perception of the NAMEOFSYSTEM system is very positive. In student surveys conducted over the four years over 90% of the students feel they learn using the system. Despite the positive attitude towards the system about 80% of the students prefer a mixture of PPH and WBH over PPH or WBH alone.

It is interesting to look at the difference in perception over the four years shown in Figure 4. As stated above, the GS was changed in 2014 making it more difficult to get a top grade for homework in the system and more difficult than in PPH. This lead to a general frustration in the student group. The fraction of students preferring only handing in PPH, compared to WBH or mix of the two, more than tripled compared to the previous years.

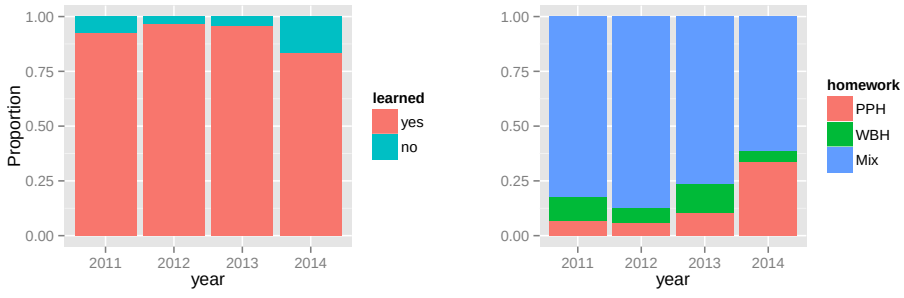

Figure 4: Results from the student survey. Left: "Do you learn from the NAMEOFSYSTEM?". Right: "What is you preference for homework"?

4 Conclusion and future work

The learning environment NAMEOFSYSTEM has been under development during the past decade at the University of COUNTRYOFAUTHORS. An experiment has been conducted to answer the following research questions:

1. Have changes made in the NAMEOFSYSTEM had an impact on learning as measured by test performance?

2. Is there a difference in learning, as measured by test performance, between students doing PPH and WBH after the changes made in the NAMEOFSYSTEM?

The experiment was conducted over four years in an introductory course on statistics. It is a repeated crossover experiment so students were exposed to both methods, WBH and PPH.

The difference between the WBH and PPH groups was found to be significantly different in 2011 and 2014 (p = 0.012), indicating that the changes made to the NAMEOFSYSTEM have made a positive impact on learning as measured by test scores. The difference in effect size between WBH and PPH in 2011 and 2014 is 0.634. Several changes were made in the system between 2011 and 2014 as shown in Table 1. As can be seen in the table the changes are somewhat confounded but moving from uniform probability to the pmf shown in Eq. 2 when allocating items, allocating items from old material to refresh memory, changing the grading scheme so that $min(max(n/2, 8), 30)$ items count in the grade instead of eight, providing EF instead of KR/KCR type feedback and having a mobile version appears to have had a positive impact on learning.

To answer the second research question, only data gathered in 2014 was used. The difference between the WBH and PPH groups was found to be significant (p = 0.009) with effect size 0.416 indicating that the students did better after handing in WBH than PPH. In both models the effect size of math background was large (1.680 and 1.379).

The NAMEOFSYSTEM project is an ongoing research project and the NAMEOFSYSTEM team will continue to work on improvements to the system. Improvements related to the exercise items are *quality of items and feedback*, *the grading scheme* (GS) and *the item allocation algorithm* (IAA).

4.1 Quality of items and feedback

As pointed out in J. B. Garfield (1994), it is important to have items that require student understanding of the concepts, not only test skills in isolation of a problem context. It is therefore important to have items that encourage *deep learning* rather than *surface* learning (Biggs, 1987).

One goal of the NAMEOFSYSTEM team is to collect metadata for each item in the item bank. One classification of the items will reflect how deep an understanding is required using e.g. the *Structure of the Observed Learning Outcomes* (SOLO) taxonomy (Biggs & Collis, 1982). According to SOLO the following three structural levels make up a cycle of learning. "*Unistructural*: The learner focuses on the relevant domain, and picks one aspect to work with. *Multistructural*: The learner picks up more and more relevant or correct features, but does not integrate them. *Relational*: The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning" (p.152).

In addition to the SOLO framework, to reflect difficulty of items in statistics courses, items could also be classified based on cognitive statistical learning outcomes suggested by delMas (2002); J. Garfield and Ben-Zvi (2008); J. Garfield and delMas (2010).These learning outcomes have been defined as (J. Garfield & Franklin, 2011): "*Statistical literacy*, understanding and using the basic language and tools of statistics. *Statistical reasoning*, reasoning with statistical ideas and making sense of statistical information. *Statistical thinking*, recognizing the importance of examining and trying to explain variability and knowing where the data came from, as well as connecting data analysis to the larger context of a statistical investigation" (p.4-5). Items measuring these concepts could be ranked in hierarchical order in terms of difficulty, starting with statistical literacy items as less difficult and ending with most difficult items measuring statistical thinking.

4.2 Grading scheme

The GS used in a learning environment such as the NAMEOFSYSTEM influences the behaviour of the students AUTHREF. The GS used in the NAMEOFSYSTEM was changed in 2014 eliminating some problems but introducing a new one; the students found it unfair. The following criteria will be used to develop the GS further.

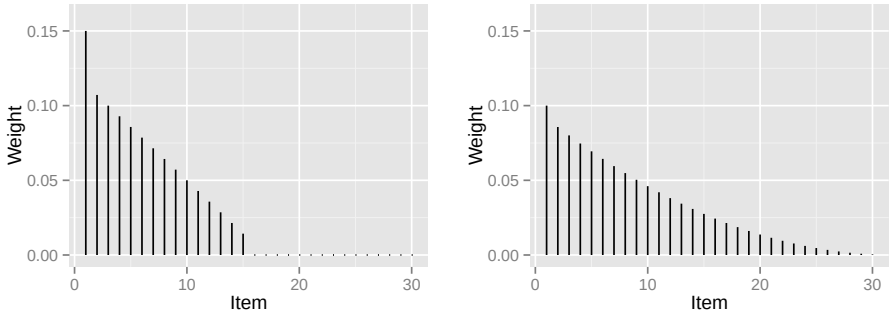

Figure 5: The weight function for a student that has answered 30 items for different values of the parameters. Left: $\alpha = 0.15, s = 1, n_g = 15$. Right: $\alpha = 0.10, s = 2, n_g = 30$.

The GS should:

- Entice students to continue to request items, thus learning more
- reflect current knowledge well
- be fair in students' minds.

Currently a new grading scheme is being implemented. Instead of giving equal weight to items used to calculate the grade, newer items are given more weight using the following formula:

$$w(l) = \begin{cases} \alpha & \text{when } l = 1, \\ (1-\alpha) \cdot \dfrac{\left(1 - \frac{l}{n_g+1}\right)^s}{\sum_{i=2}^{n_g} \left(1 - \frac{i}{n_g+1}\right)^s} & \text{when } 1 < l \leq n_g \\ 0 & \text{when } l > n_g \end{cases} \tag{7}$$

where l is the lagged item number (l = 1 being the most recent item answered), α is the weight given to the most recent answer, n_g is the number of answers included in the grade and s is a parameter controlling the steepness of the function. Some weight functions for a student that has answered 30 items are shown in Figure 5. As can be seen by looking at the figure, the newest answers get the most weight and old (sins) get less.

The students will be informed of their current grade as well as what their grade will be if they

answer the next item correctly to entice them to continue requesting items. Studies investigating the effect of the new GS will be conducted in 2016 - 2017.

4.3 Item allocation algorithm

In the current version of the IAA, the items are ranked according to difficulty level, calculated as the ratio of incorrect responses to the total number of responses. This is, however, not optimal since the ranking places the items with equal distance apart on the difficulty scale. A solution to this problem could be to use directly the ratio of incorrect responses to the total number of responses in the IAA instead of the ranking. Another solution would be to implement a more sophisticated method for estimating the difficulty of the items using IRT but as mentioned earlier those methods are designed for testing not learning. However, it would be interesting to extend the IRT models by including a *learning parameter* which would make the models more suitable in a learning environment. Finally, it is of interest to investigate formally the impact of allocating items from old material to refresh memory.

Acknowledgements

References

Barla, M., Bieliková, M., Ezzeddinne, A., Kramar, T., Simko, M., & Vozár, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computers & Education*, *55*(2), 846–857.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4.
(arXiv:1310.8236)

Biggs, J. B. (1987). *Student approaches to learning and studying. research monograph*. Melbourne: Australian Council for Educational Research Ltd.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The solo taxonomy*. New York: Acedemic Press.

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Bonham, S. W., Deardorff, D. L., & Beichner, R. J. (2003). Comparison of student performance using web and paper-based homework in college-level physics. *Journal of Research in Science Teaching*, *40*(10), 1050–1071.

Brewer, D. S., & Becker, K. (2010). Online homework effectiveness for underprepared and repeating college algebra students. *Journal of Computers in Mathematics and Science Teaching*, *29*(4), 353–371.

Brunsmann, J., Homrighausen, A., Six, H.-W., & Voss, J. (1999). Assignments in a virtual university–the webassign-system. In *Proc. 19th world conference on open learning and distance education.* Vienna, Austria: Citeseer.

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

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

Cole, R. S., & Todd, J. B. (2003). Effects of web-based multimedia homework with immediate rich feedback on student learning in general chemistry. *Journal of Chemical Education*, *80*(11), 1338-1343.

delMas, R. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, *10*(3).

Demirci, N. (2007). University students' perceptions of web-based vs. paper-based homework in a general physics course. *Eurasia Journal of Mathematics, Science & Technology Education*, *3*(1), 29–34.

Dufresne, R., Mestre, J., Hart, D. M., & Rath, K. A. (2002). The effect of web-based homework on test performance in large enrollment introductory physics courses. *Journal of Computers in Mathematics and Science Teaching*, *21*(3), 229–251.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (No. 3). New York: Teachers college, Columbia University.

Gage, M., Pizer, A., & Roth, V. (2002). WeBWorK: Generating, delivering, and checking math homework via the internet. In *Ictm2 international congress for teaching of mathematics at the undergraduate level.* Crete, Greece: John Wiley & Sons Inc.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.

Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2–7.

Garfield, J., & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*, 133–145.

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G. W., Chance, B. L., & Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, *65*(1), 1–10.

Garfield, J. B. (1994). Beyond testing and grading: Using assessment to improve student learning. *Journal of Statistics Education*, *2*(1), 1–11.

Gok, T. (2011). Comparison of student performance using web-and paper-based homework in large enrollment introductory physics courses. *International Journal of Physical Sciences*, *6*(15), 3778–3784.

Hart, D., Woolf, B., Day, R., Botch, B., & Vining, W. (1999). OWL: An integrated web-based learning environment. In *International conference on mathematics/science education and technology* (pp. 106–112). San Antonio.

Hauk, S., & Segalla, A. (2005). Student perceptions of the web-based homework program WeBWorK in moderate enrollment college algebra classes. *Journal of Computers in Mathematics and Science Teaching*, *24*(3), 229–253.

Hodge, A., Richardson, J. C., & York, C. S. (2009). The impact of a web-based homework tool in university algebra courses on student learning and strategies. *Journal of Online Learning and*

Teaching, *5*(4), 616–628.

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

Kodippili, A., & Senaratne, D. (2008). Is computer-generated interactive mathematics homework more effective than traditional instructor-graded homework? *British Journal of Educational Technology*, *39*(5), 928–932.

Kortemeyer, G., Kashy, E., Benenson, W., & Bauer, W. (2008). Experiences using the open-source learning content management and assessment system lon-capa in introductory physics courses. *American Journal of Physics*, *76*, 438–444.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). *R package version*, 2–0. Retrieved from `http://cran.r-project.org/web/packages/lmerTest/`

LaRose, P. G. (2010). The impact of implementing web homework in second-semester calculus. *Primus*, *20*(8), 664–683.

Lenz, L. (2010). The effect of a web-based homework system on student outcomes in a first-year mathematics course. *Journal of Computers in Mathematics and Science Teaching*, *29*(3), 233–246.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: L. Erlbaum Associates.

Melis, E., Andres, E., Budenbender, J., Frischauf, A., Goduadze, G., Libbrecht, P., ... Ullrich, C. (2001). ActiveMath: A generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education*, *12*, 385–407.

Nagle, R. (2010). *A user's guide to plone 4*. Houston, TX: Enfold Systems Inc.

OECD. (2013). *Education at a glance 2013*. Organisation for Economic Co-operation and Development.

Own, Z. (2006). The application of an adaptive web-based learning environment on oxidation–reduction reactions. *International Journal of Science and Mathematics Education*, *4*(1), 73–

96.

Palocsay, S. W., & Stevens, S. P. (2008). A study of the effectiveness of web-based homework in teaching undergraduate business statistics. *Decision Sciences Journal of Innovative Education*, *6*(2), 213–232.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.R-project.org/`

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13.

Razzaq, L. M., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., ... others (2005). The Assistment project: Blending assessment and assisting. In *Proceedings of the 12th annual conference on artificial intelligence in education* (pp. 555–562). Amsterdam.

Roth, V., Ivanchenko, V., & Record, N. (2008). Evaluating student response to webwork, a web-based homework delivery and grading system. *Computers & Education*, *50*(4), 1462–1482.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119–144.

Smolira, J. C. (2008). Student perceptions of online homework in introductory finance courses. *Journal of Education for Business*, *84*(2), 90–95.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. New York: Routledge.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*(3), 147–204.

Wainer, H. (2000). *Computerized adaptive testing*. Hillsdale, NJ: L. Erlbaum Associates.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562.

Williams, A. (2012). Online homework vs. traditional homework: Statistics anxiety and self-efficacy in an educational statistics course. *Technology Innovations in Statistics Education*,