Computers & Education xxx (2010) 1–10

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

A web-based learning tool improves student performance in statistics: A randomized masked trial

José A. González [a,*], Lluís Jover [b], Erik Cobo [a], Pilar Muñoz [a]

[a] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (UPC), Barcelona, C/ Jordi Girona 1-5, Campus Nord, Edif. C5, 08034 Barcelona, Spain
[b] Department of Public Health. Universitat de Barcelona (UB), Barcelona, Spain

ARTICLE INFO

Article history:
Received 28 December 2008
Received in revised form
19 February 2010
Accepted 4 March 2010

Keywords:
Applications in subject areas
Evaluation methodologies
Interactive learning environments
Probability and statistics

ABSTRACT

Background: e-status is a web-based tool able to generate different statistical exercises and to provide immediate feedback to students' answers. Although the use of Information and Communication Technologies (ICTs) is becoming widespread in undergraduate education, there are few experimental studies evaluating its effects on learning.

Method: All of the students (121) from an introductory course for statistics in dentistry were randomly assigned to use the tool with one of two 6-problem sets, known as types A and B. The primary endpoint was the grade difference obtained in the final exam, composed of two blocks of questions related to types A and B. The exam evaluator was masked to the intervention group.

Results: We found that the effect of e-status on the student grade was an improvement of 0.48 points (95% CI: 0.10−0.86) on a ten-point scale. Among the 94 students who actually employed e-status, the effect size was 0.63 (95% CI: 0.17−1.10).

Conclusions: It is feasible to formally assess the learning effect of an innovative tool. Providing e-status exercises to students has a direct effect on learning numerical operations related to statistics. Further effects on higher cognitive levels still have to be explored.

1. Introduction

We present this paper with the aim of demonstrating the effectiveness of a web tool (e-status, González & Muñoz, 2006) for undergraduate students of statistics, based on an experimental approach.

Experimental evaluation has long been established in non-academic practice. Bradford Hill encouraged randomized experimental evaluation of medical interventions over sixty years ago (as cited by Horton, 2000) and, thirty years later, Fletcher and Sackett (1979) proposed "evidence-based medicine". The difficulties in implementing experimental evaluation were well known, but now the medical profession can select interventions with the best efficacy.

In contrast, education is rather lacking in high level experimental evaluation of treatments (Alsop & Tompsett, 2007; Margolis, Nussbaum, Rodríguez, & Rosas, 2006), or recognition of their role in accelerating improvements (Simons, 2003). Evidence-based practice has been considered as inappropriate because it dismisses qualitative approaches to research (Oliver & Conole, 2003).

There are certain inherent problems that accompany statistical instruction at the university level. Some of these problems are related to attitude: students usually consider statistics to be a difficult, demanding subject, irrelevant to their career goals (Carnell, 2008). Moreover, according to Wilks (2006), it is introduced to students relatively late in their academic careers, and teachers have to make an effort to motivate them with the subject. Finally, many experts point out that the root of the problem is related to people's unwillingness to discard misconceptions, intuitions and prejudices (Hagtvedt, Jones, & Jones, 2008), in favor of appropriate reasoning for dealing with uncertainty.

Technological tools appear to help students by providing them with means to represent data, or allowing easy manipulation and exploration of statistical models. For instance, Garfield and Ben-Zvi (2007) point out that well-designed simulations can play a significant role in enhancing students' abilities. Nowadays we can find tens of thousands of web sites containing topics related specifically to statistics

* Corresponding author. Tel.: +34 934015867; fax: +34 934015855.
E-mail addresses: jose.a.gonzalez@upc.edu (J.A. González), lluis.jover@ub.edu (L. Jover), erik.cobo@upc.edu (E. Cobo), pilar.munyoz@upc.edu (P. Muñoz).

0360-1315/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2010.03.003

and learning technologies. They range from pages which perform simple calculations up to complete web-based courses presented as electronic textbooks (Symanzik & Vukasinovic, 2003).

Information and Communication Technologies (ICTs) have been present in education for more than forty years. Several paradigms have appeared successively (*Computer Assisted Learning*, *Intelligent Tutoring Systems*, *Interactive Learning Environments*, *Computer Supported Collaborative Learning*, …), supported by different instructional theories. The rise of the internet in recent years has led web-based education to the center of the discussion, as it offers promising features: resource management is easy, as only a single version of the documents needs to be maintained; collaboration is promoted, since the internet's communication services allow the student the access with other people to share with them experiences and knowledge; learning is no longer bound by space and time (Franklin & Peat, 2001).

However, characteristics like these are rather related to efficiency instead of efficacy of learning. Advanced technologies applied to instruction have been generally assumed to be effective, although the statement cannot be applied without severely constraining the educational context: how is the tool? What type of students? How is the course structure, and how is the tool related to other instructional elements? Large-scale studies were undertaken in the 90s to determine the effectiveness of ICTs (BECTa, 2002), and could not find a conclusion, mainly owing to the high degree of variability in both school practice and impact on learning (from the review of the Impact study, by Wood, Underwood, & Avis, 1999).

Additionally, some authors emphasize the lack of a pedagogical framework for web-based education. Alonso, López, Manrique, and Viñes (2005, p. 218) state: "There are no guidelines for analysing, designing, developing, supplying, and managing e-learning materials pedagogically". They propose an instructional model composed of seven phases: analysis, design, development, implementation, execution, evaluation, and review. One could point out that this evaluation phase is not an actual experimental analysis, so it cannot have a confirmatory value of the methodology efficacy.

In spite of the large body of work that emphasizes the educational advantages of ICTs, related to teaching introductory statistics or other subjects (Basturk, 2005; Larreamendy-Joerns, Leinhardt, & Corredor, 2005), there are few experimental assessments of its effects on improving learning: see, as an example, Krause, Stark, and Mandl (2009). Most work refers to observational or quasi-experimental studies, which leaves a risk of bias in effect estimation, for instance, when students are free to choose the intervention level (Chumley-Jones, Dobbie, & Alford, 2002). Some papers refer to random assignation when they mean in fact that each group (e.g. from two different schools) is assigned to the interventions at random (Dinov, Sánchez, & Christou, 2008; Grubišić, Stankov, Rosić, & Žitko, 2009). Only some authors are aware of these limitations in their studies (Evans et al., 2007).

Statistical education seems to be on its home ground for developing evidence-based methodologies. In this paper, our objective is to evaluate how e-status improves the statistical abilities of students relative to the classical approach where problems are solved on paper, far from the teacher's sight. Specifically, we hypothesize that the grade from an exam at the year end would be higher after exercising with e-status.

The manuscript has been drafted following closely the recommendations offered by the CONSORT statement (Altman et al., 2001; Moher, Schulz, & Altman, 2001). The CONSORT Statement comprises a 22-item checklist and a flow diagram intended to facilitate the reporting of randomized controlled trials, mainly in health research, though it has inspired also a guideline for education research (Newman & Elbourne, 2005). The following section briefly introduces a theoretical background and an outline of e-status. In Section 3, we explain the proposed experimental evaluation. Section 4 shows the results from this trial. Finally, the conclusion discusses the findings and emphasizes some implications of the design.

2. Instructional framework

2.1. Background

Most of universities in developed countries, if not all, use ICTs in their courses to a greater or lesser extent. For instance, our institutions, Universitat Politècnica de Catalunya (UPC) and Universitat de Barcelona (UB), are oriented to classroom teaching and have adopted the e-learning software Moodle, a popular platform because of its pedagogical approach to education, according to constructivist and social constructionist learning theories (Moodle, 2009). However, despite their vast potential as a pedagogical tool, applications like Moodle are frequently used merely as a repository for teaching materials, thus failing to take advantage of their full power.

The courses where e-status is used are generally based on classroom lectures, so that this combination of face to face and online instruction can fit into what is termed "blended learning" or *b-learning*. e-status also shares some of the features from the so-called Integrated Learning Systems (ILS), as Wood et al. (1999) itemized:

- Curriculum content (e-status does not contemplate it, but it is able to include links in the problems at specific addresses);
- A pupil record system;
- A management system.

The roots of ILS are in the 1960s, from the work of Patrick Suppes at Stanford University, and they were modernized through new technologies (internet and multimedia formats) in the 1990s. They are based on a neo-behaviorist model of learning, characterized by automatic task selection, guided practice and individualized feedback (Wood et al., 1999). However, instruction with e-status is characterized also by elements identified with constructivism: learning is centered in the student's activity; the teacher plays the role of a catalyst in a process where knowledge is built, providing a collection of problems at the students' disposal and monitoring their progress; though the activity is individual, the students are not isolated, as they have the means to visualize several performance indicators, ranking themselves among their colleagues.

e-status is relatively simple compared to some other platforms, mainly because it focuses on the topic of numerical solution of problems, and especially in basic statistics for higher education. A specialized tool means that the composition of problems is facilitated to fit the educational goals of the course. The classification of learning domains made by Bloom (1956) is useful for characterizing the kind of problems we pose. Among the well-known categories of cognitive learning (knowledge, comprehension, application, analysis, synthesis and

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

evaluation), we think that e-status is suitable for exercising at least the four more basic ones (since synthesis and evaluation are related to actions like 'argue', 'develop' or 'judge' that need human intervention to be adequately appraised), and the teachers have to strongly consider them when creating interesting problems for the students.

Though firmly convinced that the learning of statistics involves a *genuine gain* in what is known as "statistical thinking" (Chance, 2002), the gain is only possible if lower-order thinking skills are soundly achieved. Therefore, we encourage our students to use e-status as a means to reach the main objectives of the course. The learners have as advantages: open access to the material (any time, any place), immediate assessment of the exercise, and feedback at a glance, from a summary of their activity to full detail of each exercise. This kind of information from the system is in fact what is known as *external representation*, which serves as a focus for learners' attention and may encourage them to work on specific areas (Ertl, Kopp, & Mandl, 2008).

Before undergraduate students face the real and complex problems of their professional lives, they must master the basic steps present in different statistical techniques, and they have to learn them "by doing". Students gain an advantage also when repeating a problem (obviously with new data): well-designed problems can be challenging to the students in the sense that, when presented with different data, they maintain their interest in solving the problem. Repetition has been often associated to memorization and seen as a deficient learning method. Our focus is not so much on *ad nauseam* repetition as in creating a system which instills active learning and raises progress awareness, over other ways of practice (e.g., classical exercises with pencil and paper).

2.2. Outline of e-status

e-status is a web-based application that the student can access from any place, any time, at the URL http://ka.upc.es/. Teachers interested in testing it for academic purposes are invited to contact the authors.

According to Morris (2001), e-status would be a powerful tool oriented to both b-learning and e-learning, as it aims to:

- offer a powerful but simple tool for problem editing and solving,
- give immediate feedback to students, and
- provide the teacher with reliable data about students' performance.

The tool uses the R environment (2008) internally for both computations and graphics management. Teachers edit the problems, make them available to the students and look into the data collected from the exercises solved. Students solve the problems and look up their history records to get feedback and check if they are fulfilling the course objectives. Both teachers and students have to authenticate themselves to log into the system. In order to display the rate of success, as well as the average grade and other student performance indicators, reliable identification of the user is necessary.

The main strength of e-status is its ability to instantly process answers for given questions, even if the problem has variable data. The teacher has to write an analytical model of the problem by means of R instructions (any R procedure can be used in e-status, including graphical capabilities), so it is able to assess the correctness of an answer. The teacher could include additional instructions to deal with predictable errors in the answer and return valuable information to the student, information which is related to the probable source of the mistake.

Fig. 1 shows how e-status looks when it displays an exercise being solved. The problem is provided, followed by questions and spaces for the answers. The student marks the answer to each question and is given an overall grade at the end (not shown in the figure). Every question is also assessed. In this example, the program has generated a random sample of measures from nine fiberglass tires (the sample size is also random). Four questions have been correctly responded to, and it warns that the fifth question is still blank. If a student wants to solve it again, e-status would prepare another version with new data.

Each execution of a problem is registered so both teacher and student can have a record of the completed work. Teachers know the frequency of use for each problem, the grade and even each response for every exercise solved, all of them indicators of possible weaknesses in the students' learning process. Students can schedule their activities and view their performance compared with other students.

3. Participants and methods

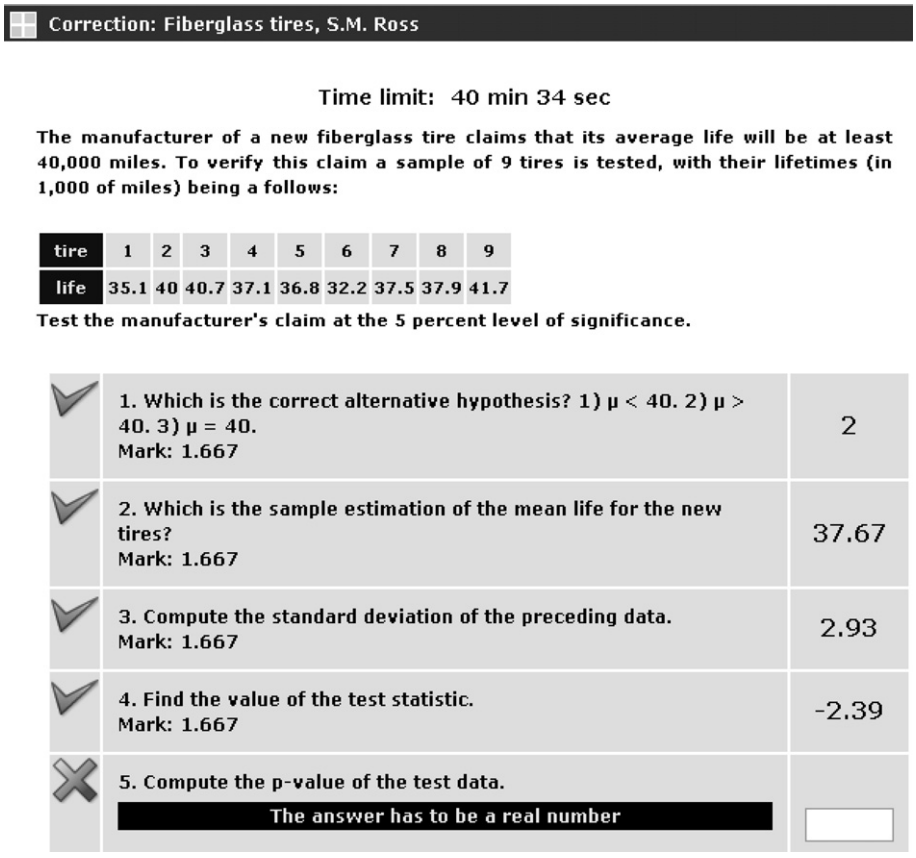
This study compared two groups of students randomly allocated. We decided to give each student the chance to use the tool, so both groups were different only by the topics covered. A simpler design, where a group with access to e-status would be compared to another group without access, was rejected to avoid the ethical implications of having half of the students deprived of such a resource. Randomization was able to be implemented without interfering with participant rights and regular instruction. Masking of any participant in a randomized trial is desirable: the exam evaluator was masked, preventing him from knowing the group for any student but, in our case, students could have been able to compare amongst themselves and being aware of the intervention they received.

3.1. Study design

We planned an experimental approach to evaluate the intervention, i.e. the e-status effect size, in the Biostatistics course 2005–06 at the dentistry school of the UB. The course had 35 h of classroom teaching: 15 one-hour sessions to introduce theoretical subjects, and 10 two-hour practical sessions in the lab. Topics covered were: descriptive statistics, elemental inferential procedures, an introduction to probability and sampling, observer agreement with qualitative measurements, interval estimation, basics of sample size determination and parameter comparison between two populations. Only one teacher was involved in the lecture sessions, ensuring that all the students received the same contents.

Overall content of the course was divided into two non-consecutive parts, A and B, with an attempt to balance the length, difficulty and complexity of statistical procedures to be learned. Thus, we built two sets of six e-status problems each, hereafter E_A and E_B, covering statistical procedures included in course parts A and B respectively.

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003



Correction: Fiberglass tires, S.M. Ross

Time limit: 40 min 34 sec

The manufacturer of a new fiberglass tire claims that its average life will be at least 40,000 miles. To verify this claim a sample of 9 tires is tested, with their lifetimes (in 1,000 of miles) being a follows:

tire	1	2	3	4	5	6	7	8	9
life	35.1	40	40.7	37.1	36.8	32.2	37.5	37.9	41.7

Test the manufacturer's claim at the 5 percent level of significance.

1. Which is the correct alternative hypothesis? 1) μ < 40. 2) μ > 40. 3) μ = 40. Mark: 1.667	2
2. Which is the sample estimation of the mean life for the new tires? Mark: 1.667	37.67
3. Compute the standard deviation of the preceding data. Mark: 1.667	2.93
4. Find the value of the test statistic. Mark: 1.667	-2.39
5. Compute the p-value of the test data. The answer has to be a real number	

Fig. 1. Example of an e-status result. This problem has six equivalent questions (1.667 points), so a perfectly answered exercise scores 10 points.

The official curriculum states that the final score of the biostatistics course is calculated by weighting different tests. One of them is a written exercise to evaluate practical skills: every student, regardless of the allocation group, has to complete three practical questions corresponding to part A and three more to part B. These two sets in the exam will be hereafter referred to as subsets S_A and S_B, both contributing with equal weight to the final examination grade (Table 1). In our experiment, the questions were carefully designed to give to sets S_A and S_B similar levels of difficulty and to require a similar amount of time for their solution.

3.2. Participants

All students enrolled in the course were included in the study ($n = 121$). The number is enough to provide the analysis with a high probability of detecting a significant effect size, if any (see Appendix A). Most of them were new students (101), and 20 students were repeat students. The Academic Advisor of the school gave his approval to this evaluation and the students were informed that the efficacy of this web activity was under evaluation and any comments and suggestions would be appreciated.

3.3. Outcome

The two outcome variables Y_A and Y_B correspond to the grades obtained by a student in parts S_A and S_B, respectively. The main outcome is $D = Y_A - Y_B$, the difference between the grades in part A and part B.

Table 1 shows the maximum score of each question and the sum, corresponding to the top value reachable by Y_A and Y_B. Grades are given in Spain as a number between 0 and 10, often to one decimal place. Typically, the students need a grade greater than five to pass the course.

In the case of e-status effectiveness, a better performance was expected on the part that each student had exercised with the tool, compared to how they would have done without using e-status. That is: students using e-status to solve exercises covering part A (E_A) are likely to obtain higher grades in the exam subset corresponding also to this part (S_A), i.e. $Y_A > Y_B$. A similar effect should be observed in the other group, with higher scores of the exam part S_B, i.e. $Y_A < Y_B$.

In order to remove inter-observer variability, the final examination was graded by the second author only. Aside from this, he was masked to the allocated group to avoid evaluation bias, since the e-status administrator prevented him from having access to the subject.

The secondary outcomes considered are the number of solved exercises and grades obtained on the e-status problems.

3.4. Random allocation

Participants were randomly assigned to one of two groups A and B, by means of random numbers generated by an independent researcher (the third author) using Excel. Randomization was stratified by laboratory group and a previous failing grade in the subject,

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

Table 1
Weights of each practical question in the final exam.

Set	Written exam						Evaluated outcome
	Question number						
	1	2	3	4	5	6	
S_A	2.5			5.0	2.5		$Y_A \leq 10$
S_B		5.0	2.5			2.5	$Y_B \leq 10$

which means six plus one strata, as depicted in Fig. 2. Students who were new to the course (first time students) were randomized within their laboratory group, and the imbalance achieved was not greater than one. Repeat students from the six lab groups were randomized together within a new stratum (number 7), giving ten students for each one of A and B.

3.5. Intervention

One group of students ($n = 61$) had to solve the problems included in E_A and the other group ($n = 60$) those included in E_B. No one had access to both sets of problems.

The students received e-status documentation and training in the second week of November. Students had access to e-status during the last two weeks of November and all of December. The exam date was January 12th. This planning was justified because this period encompassed four weeks of clinical practice with dental patients and without classroom activities followed by Christmas holidays, thus minimizing student interaction (and maintaining independence among units, a basic assumption needed to employ standard statistical methods) while they were solving problems with e-status.

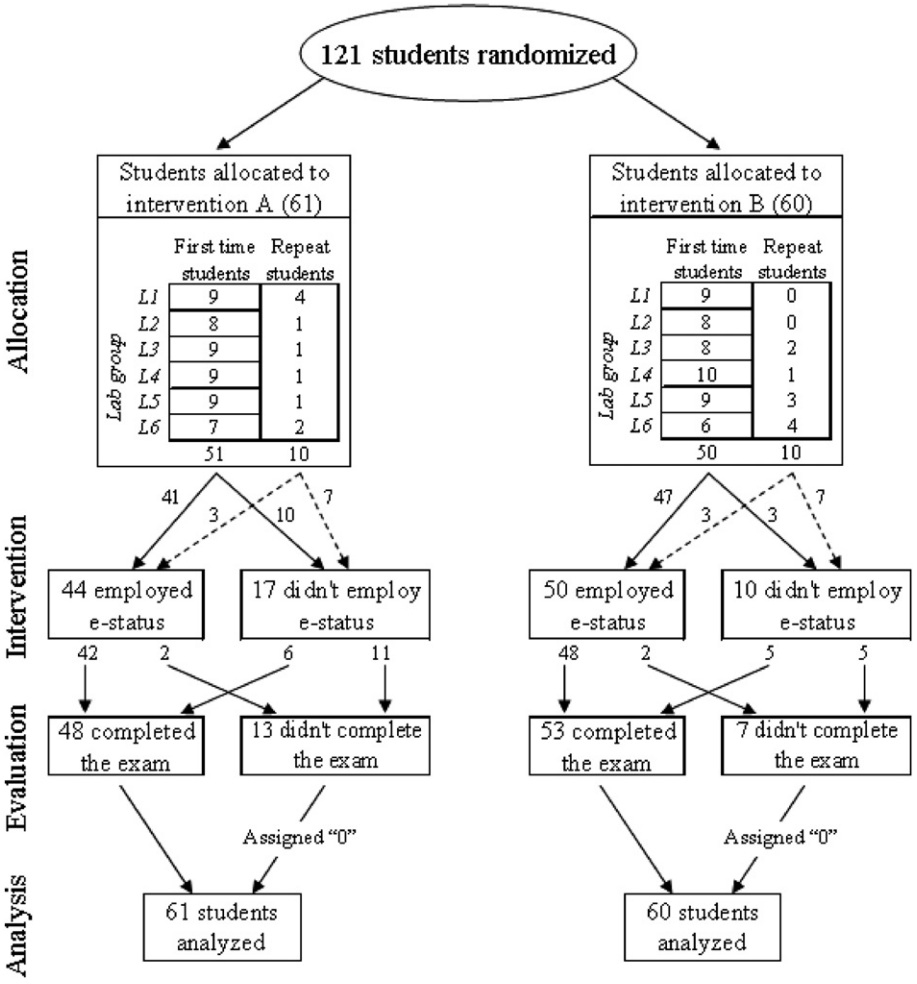


Fig. 2. Groups A and B relate to two non-overlapping sets of topics covered in the course; students could practise the topics they had access to with e-status. Random allocation was balanced over seven strata: six laboratory groups, L1–L6, for the first time students, and one further stratum merging all the repeat students from all the laboratory groups. The final exam was identical for both arms: it included three specific questions focused on A and three on B. The study considered two separate outcomes for every student: Y_{E_A,S_A}, Y_{E_A,S_B}, on the left arm, and Y_{E_B,S_A}, Y_{E_B,S_B}, on the right arm. The exam grade was the mean of both parts, but the main outcome for the study was $Y_{E_x,S_A} - Y_{E_x,S_B}$.

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

Table 2
Mean grades of the problems solved with e-status (grades range from 0 to 10).

	e-status problems of set E_A					
	A_1	A_2	A_3	A_4	A_5	A_6
N. students	44	44	44	42	41	40
Tot. N. sol.	114	107	104	154	96	90
Grade *M*	8.01	7.77	7.62	7.51	8.09	8.48
Grade *SD*	1.38	2.07	2.08	1.97	1.70	1.67
Set average	7.906					
	e-status problems of set E_B					
	B_1	B_2	B_3	B_4	B_5	B_6
N. students	50	50	50	49	49	47
Tot. N. sol.	177	137	96	104	142	126
Grade *M*	7.61	7.92	8.42	8.77	7.38	7.38
Grade *SD*	2.23	1.33	1.76	1.40	1.86	2.90
Set average	7.918					

Tot. N. sol.: total number of exercises solved by problem.

3.6. Statistical analysis

The primary analysis (ITT, for *intention-to-treat*) is a *t* test on the comparison of means in independent samples, given the difference of grades for parts A and B for each student (see Appendix A for details about the underlying statistical model). ITT analysis includes all the randomized students, even if they did not use the tool at all or were absent. Those not taking the practical exam are considered *missing* values, and they have a null (0) grade imputed in order to be included in the primary analysis, in the sense of *no effect*. Our previous experience convinced us that the main outcome had an approximately normal distribution.

As the fulfilment of the students tasks was unequal, several sensitivity analyses are reported for the main outcome to check the soundness of the primary analysis: a) including only students who accessed e-status (PP, for *per-protocol*); b) including only those who had taken their exam (EVAL, for *evaluated*); and c) considering the stratification variable in the model. The latter can also be performed with a dichotomized stratum, considering only whether or not he/she is a repeat student.

PP and EVAL analyses use the *t* test as well, as only one dichotomous factor (the group: A or B) is considered. In the third analysis, secondary variables are included via an ANOVA model with group, student type (repeat or first time) and their interaction.

4. Results

4.1. Performance on the training stage

On January first, access to e-status was closed. Of the 121 randomized students, 44 students from A and 50 from B had employed the tool; 17 from A and 10 from B had not used it. They solved 1447 exercises, mostly during the holiday period, with an average grade of 7.9 ($SD = 1.94$) over 10.

Table 2 shows additional detail for each problem in sets E_A and E_B, used in the training phase with e-status. Table 3 adds more information related to the considered strata. Overall means for a given group do not coincide in the tables (e.g., 7.9 in Table 2 and 7.6 in Table 3, for group A) because in the first case focus is put on the exercise (mean for an exercise), and in the second case we focus on the student (mean for a student). In general, there are only slight differences within the same stratum, both in grade or in number of solved exercises. The most relevant is the group of repeat students (stratum 7), whose grades are far from any other stratum: a decrease of 1.25 points (95% CI 0.57–1.92) with respect to a first time student. However, the difference in number of solved exercises between repeat and first time students is not statistically significant (95% CI 2.1–8.1, for expected decrease).

Results in Tables 2 and 3 are presented to show performance on those intermediate outcomes. From a descriptive point of view, the observed differences cannot be considered as evidence against the comparability between both sets, but they suggest taking into

Table 3
Grades of the problems solved with e-status by stratum and group. n_A/n_B are the size groups for each stratum. Numbers inside the table are *mean (SD)*, both for the grade and for the number of exercises solved.

Stratum (n_A/n_B)	Set E_A		Set E_B	
	Grade	N. solved	Grade	N. solved
1 (7/7)	7.9 (2.1)	21.6 (19.8)	7.5 (1.4)	15.6 (3.6)
2 (6/8)	7.8 (2.6)	13.5 (12.7)	8.6 (1.0)	13.7 (5.0)
3 (8/8)	7.8 (1.1)	12.6 (3.7)	7.8 (1.4)	12.1 (4.2)
4 (6/9)	7.4 (2.1)	10.2 (4.4)	7.2 (2.7)	15.9 (13.9)
5 (7/9)	8.3 (1.1)	10.4 (2.8)	8.4 (1.1)	19.4 (13.2)
6 (7/6)	7.5 (1.1)	24.6 (27.8)	7.8 (0.6)	16.8 (12.8)
7 (3/3)	5.6 (3.1)	8.0 (4.4)	5.1 (0.9)	11.3 (6.7)
All (44/50)	7.6 (1.8)	15.1 (14.8)	7.7 (1.7)	15.4 (9.6)

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

Table 4
Statistical analysis of grades in the written exam parts, and their difference $Y_A - Y_B$. e-status efficiency is related to positive values in the mean of $Y_A - Y_B$ for group A, and negative values for group B.

Analysis	Group	n	M Y_A	M Y_B	M $Y_A - Y_B$	SD $Y_A - Y_B$
ITT	A	61	5.519	4.679	0.840	2.176
	B	60	5.095	5.214	−0.119	2.046
		effect size: 0.96; 95% CI: 0.20–1.72				
PP	A	44	7.238	5.969	1.270	2.309
	B	50	5.764	5.766	−0.002	2.164
		effect size: 1.27; 95% CI: 0.35–2.19				
EVAL	A	48	7.014	5.946	1.068	2.407
	B	53	5.768	5.903	−0.134	2.179
		effect size: 1.20; 95% CI: 0.30–2.11				

consideration whether a student is repeating the subject or not. In our analysis, strata 1–6 are to be merged into one, since we cannot expect to find relevant differences among them.

4.2. Performance on the final exam

48 students from group A and 53 from the other group completed the exam. 13 and 7 students, respectively, did not submit their exams, and a grade of zero was imputed. Few of them had used e-status previously: only 2 of 13 in group A, and 2 of 7 in group B. On the other hand, students having completed the exam had mostly used the tool: 42 of 48 from A, and 48 of 53 from B (see Fig. 2). Six repeat students used e-status, and all of them completed the exam, but only five of 14 who did not use the tool took it. Broadly speaking, there are two clear profiles: a) repeat students were not interested in using e-status (70%) and did not complete the exam (55%); b) first time students did use the tool (87%) and did take the exam (90%). The group allocated was not relevant.

4.3. Main results

Table 4 contains a summary of the outcomes obtained from the written exam. As mentioned above, 61 and 60 students were assigned to A and B respectively; 44 and 50, respectively, solved at least one problem, and 48 and 53, respectively, completed the exam. These numbers are relevant in the analyses provided. The ITT analysis could be interpreted as: "The effectiveness of our advice for using this tool", and the PP analysis as: "What the effect will be on students using this tool," assuming that random balance is not lost. The EVAL analysis is the same as ITT but excludes the missing students, and thus is based on actual grades. The data showed that the normality assumption was reasonable.

All the students are included in the main analysis. The difference $Y_A - Y_B$ obtained in both groups is described in Table 4 as well. The students in group A realized a mean difference of 0.84 points ($SD = 2.18$), whereas the mean difference in students from group B was −0.12 points ($SD = 2.05$), both favoring the intervention. The difference between both means, 0.96, is an estimator of the expected effect of the intervention[1], and it is statistically significant ($t(119) = 2.50$, $p = .014$), giving a 95% CI from 0.20 to 1.72 points. Assuming that the effect size in each group is the same, these figures have to be divided by two, to conclude that in our study a student could expect an increase of 0.48 points—that is, 4.8%—in his/her grade associated to e-status.

4.4. Sensitivity analyses

The results are confirmed by the other analyses, with larger effects: PP, 1.27 points (95% CI 0.35–2.19), as expected since active students are likely to show an interest in the subject; and EVAL: 1.20 points (95% CI 0.30–2.11), since the grade imputed to missing students is implying no effect. Applying the same correction, we obtain an increase of 0.63 or 0.60 points (6.3% or 6.0%), according to PP and EVAL analyses, respectively.

Repeat students were pointed out as a special type, as their outcomes with e-status were substantially lower than outcomes for first time students (see Table 3). Therefore, we fitted a model for the difference D related to group and student type (repeat or first time), in an attempt to explore the difference between repeat students and first time students. Moreover, the model can be useful to elucidate whether or not the problems from E_A are associated with higher performance in the exam, since data suggest that the intervention could be more effective in group A than it could be in group B. The analysis is intention-to-treat as well.

Entering the student type into the model increases its statistical significance and gives an insight into the data. Results from the two-way ANOVA model are: $F(4, 117) = 3.96$, $p = .0048$, the main effect of the group is statistically significant ($t(117) = 3.778$, $p < .001$), as is the main effect of the student type ($t(117) = -2.207$, $p = .029$), but not the interaction ($t(117) = 0.924$, $p = .357$). Fig. 3 shows (95%) confidence intervals for the expected mean difference depending on the group, either for first time or repeat students. For group B, D has been defined as part B minus part A, so that the right side of the axis is related to the benefit of the tool. The confidence intervals for repeat students show lack of precision due to the reduced sample size (10, for each group). Regarding first time students, intervention in group B could not be proved as effective, though we must be cautious when it comes to analyzing the probable reasons. On the other hand, intervention in first time students of group A demonstrates a high positive response: 1.10 points, 95% CI 0.53–1.67, on a scale of 0–10—as percentages, 11%, 5.3%–16.7%.

[1] The mathematical expectation of the effect is derived in the Appendix, equation (A.2).

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

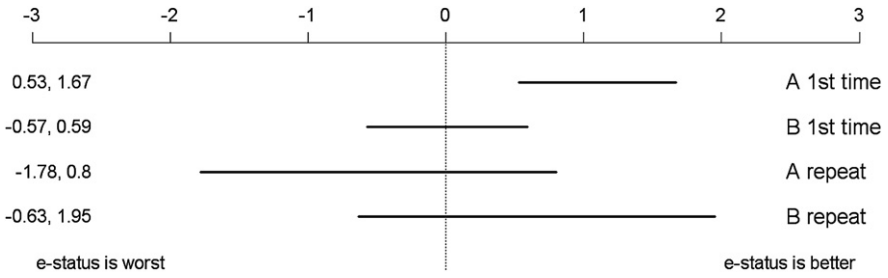


Fig. 3. Effect of intervention estimated through a two-way ANOVA model. Horizontal bars are 95% CI for the mean difference of grades between the part strengthened with e-status and the other part.

5. Discussion and conclusions

In a recent analysis, the U.S. Department of Education (2009) conducted a systematic review of research literature from 1996 to 2008, identifying more than a thousand empirical studies of online learning. After screening to select only those studies considered of sufficient scientific rigor, 51 effects from 46 studies were selected. Key findings include: a) an average effect size of 0.24 ($p < .01$) favoring online conditions compared to traditional face to face instruction; b) a larger advantage (0.35, $p < .001$) for blended learning versus purely face to face settings. The meta-analysis considers only studies with random-assignment or controlled quasi-experimental designs, and examined effects only for objective measures of student learning.

Experimental methods are difficult to apply in an educational environment because they involve many factors, correlations among individuals are common, and conclusions must be drawn carefully (Holt & Oliver, 2002). But only the rigorous evaluation of new learning methods can satisfy both: 1) the evidence-based scientific method currently applied in most disciplines, and 2) the need to select the most efficient methods—while showing the way to improve them.

We have found statistical evidence regarding the effects of using e-status as a learning tool. Our results indicate that an individual student allocated to *half* intervention (so to speak) has improved her/his grade in the practical exam by an overall 5% absolute point improvement, although the intervention, or the outcome measure, has not been symmetrical since it could be proved that group A has had even greater benefit, specifically 11% for first time students. In the group of repeat students one could not be conclusive but, in our opinion, the study is highlighting the need to involve repeat students in educational measures that may be especially useful for overcoming their trouble with the subject.

In the design presented, each student is exposed to *half* intervention since they can only use e-status with problems related to half the contents. Therefore, a student from the study can expect to obtain an average increase of half the true effect. This conclusion arises also from the mathematical model appearing in Appendix A, equation (A.2). The relevant question, however, is: What would be the advantage for students who will have full access to e-status?

For design reasons, one cannot measure the direct, combined effect of the intervention (expressed in Appendix as $\omega_{E_A,S_A} + \omega_{E_B,S_B}$), because it is reduced by some unknown amount (expressed as $\omega_{E_A,S_B} + \omega_{E_B,S_A}$), which may be related to knowledge transfers from one part to the other. This kind of *contamination* implies underestimation of the main effect. However, in the case of a real intervention with full access, its existence would mean that the student is making connections between different topics, a desirable goal in learning: according to constructivist paradigms, the kind of learning related to collateral abilities, relevant to building persistent structures in intellectual thinking. In any case, it is the teacher's duty to include these bridges between topics in the problems, and to foster high quality learning.

One remarkable feature of the presented design is the reduction in the response variance, because the variability between individuals cancels out when the student difference (D) is used as the main outcome. A decrease in the variance of the outcome implies an increase in the test's statistical power and greater accuracy in the estimation of the effect size. On the other hand, underestimation of the direct effect of the intervention—inseparable from other indirect effects—represents a clear limitation for our design. An unbiased estimation could be obtained by comparing one group with full access to another with no access, the simple "treatment vs. control" design. However, this proposal would not obtain any support from administrators or faculty because of its evident inequality among the students. Alternative designs could be worth considering (for instance, one full-access group, AB, versus two half-access groups, A and B…), in order to obtain better estimators of the effect size.

The results achieved are sound with data obtained from students over the next three years in the same course. A significant positive correlation can be observed between the final exam grade and performance with e-status (solving the same twelve problems as the students from 2005 to 06). Although, external validity of the conclusion will rely heavily on the subject's characteristics and on the suitability of e-status exercises. The effect was shown for a specific course with a specific type of problems. Generalization can be hypothesized only with caution. The grade effect in other subjects would be largely dependent on the way the subject organizes the e-status problems proposed to their students, as well as contents, teaching methodology and assessment instruments.

Two points more ought to be made about the external validity of our results. First, the effects of e-status should also be estimated on students with profiles that are different from those of dentistry students at UB. Second, and more important, the desired learning outcome is long-term statistical reasoning. The performance in an exam is a direct indicator of subject success, but just a surrogate of the true long-term effect. Our results are made relevant only by the assumption that analytical thinking—with much consideration in typical exams—promotes statistical reasoning.

We would like to finish by highlighting that we reached our two desired objectives: We showed both that e-status has a *measurable* and positive effect on student performance and that it is feasible to evaluate learning interventions with formal experiments. Given that proper randomization to one of two comparison groups eliminates selection bias, and masking prevents evaluation bias, we claim that any study aiming to measure an effect of some educational intervention should consider including these measures, protecting itself against the risk of biased estimations.

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

Acknowledgements

The authors wish to acknowledge the Agency for Management of University and Research Grants (AGAUR), Catalan Government for partially supporting this research under grant 2005 ECTS 00017. We also wish to thank the Institute of Education Sciences (ICE, UPC) for their financial support. Thanks to prof. Josep Anton Sánchez for his invaluable help at any time. Comments from Mike Campbell and Margaret MacDougall at the Burwalls 2007 meeting have greatly contributed to improve this paper. The work was partially developed when the author P. M. was visiting the Department of Statistics at NCSU.

Appendix A. Statistical model

Let $Y_{t,j,i}$ represent the grade of student i assigned to group t in exam part j; μ the overall mean in the exam; τ_t the fixed overall effect of e-status group t (practice with either E_A or E_B); π_j the fixed effect of the exam's problem set j (completing S_A or S_B), representing the set hardness; $\omega_{t,j}$ the fixed effect of group t on exam part j (that can be seen as an interaction between both fixed factors); ϕ_i the random effect of student i, assuming variance σ^2_ϕ, and $\varepsilon_{i,j}$ the error term of student i in exam subset j, with variance σ^2_ε. Then:

$$Y_{t,j,i} = \mu + \tau_t + \pi_j + \omega_{t,j} + \phi_i + \varepsilon_{i,j} \tag{A.1}$$

The main effect of the intervention is represented by ω_{E_A,S_A} and ω_{E_B,S_B}, whereas ω_{E_A,S_B} and ω_{E_B,S_A} are related to indirect effects (type A intervention affects part B, or viceversa).

Let $D_{E_A,i}$ be the difference of grades between both problem exam sets, for student i receiving intervention E_A : $D_{E_A,i} = Y_{E_A,S_A,i} - Y_{E_A,S_B,i}$. Correspondingly, $D_{E_B,i} = Y_{E_B,S_A,i} - Y_{E_B,S_B,i}$. It could be easily shown that if both groups are of equal size n, and assuming error independence:

$$E\left(\overline{D}_{E_A} - \overline{D}_{E_B}\right) = \left(\omega_{E_A,S_A} + \omega_{E_B,S_B}\right) - \left(\omega_{E_A,S_B} + \omega_{E_B,S_A}\right) \tag{A.2}$$

with variance:

$$V\left(\overline{D}_{E_A} - \overline{D}_{E_B}\right) = \frac{4}{n}\sigma^2_\varepsilon \tag{A.3}$$

That means that 51 students per group provide 80% power (one sided, $\alpha = 5\%$) to highlight a specific effect equal to 0.7 times the intrasubject standard deviation σ_ε. Therefore, the number of students enrolled in the course—121—seems enough to detect a relevant effect, preserving power against deviations from these assumptions.

References

Alonso, F., López, G., Manrique, D., & Viñes, J. M. (2005). An instructional model for web-based e-learning education with a blended learning process approach. *British Journal of Educational Technology, 36*(2), 217–235.
Alsop, G., & Tompsett, C. (2007). From effect to effectiveness: the missing research questions. *Educational Technology & Society, 10*, 28–39.
Altman, D. G., Schulz, K. F., Moher, D., Egger, M. D. F., Elbourne, D., Gøtzsche, P. C., et al. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine, 134*(8), 663–694.
Basturk, R. (2005). The effectiveness of computer-assisted instruction in teaching introductory statistics,. *Educational Technology & Society, 8*, 170–178.
Bloom, B. S. (1956). *Taxonomy of educational objectives: Handbook 1: The cognitive domain*. London: Longmans.
British Educational Communications and Technology Agency (BECTa). (2002). *The impact of information and communication technologies on pupil learning and attainment*. Coventry: BECTa.
Carnell, L. J. (2008). The effect of a student-designed data collection project on attitudes toward statistics. *Journal of Statistics Education, 16*(1). Retrieved from http://www.amstat.org/publications/jse/v16n1/carnell.html.
Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). Retrieved from http://www.amstat.org/publications/jse/v10n3/chance.html.
Chumley-Jones, H., Dobbie, A., & Alford, C. (2002). Web-based learning: sound educational method or hype? A review of the evaluation literature,. *Academic Medicine, 77*(10), S86–S93.
Dinov, I. D., Sánchez, J., & Christou, N. (2008). Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses. *Computers & Education, 50*, 284–300.
Ertl, B., Kopp, B., & Mandl, H. (2008). Supporting learning using external representations. *Computers & Education, 51*, 1599–1608.
Evans, S. R., Wang, R., Yeh, T., Anderson, J., Haija, R., McBratney-Owen, P., et al. (2007). Evaluation of distance learning in an 'introduction to biostatistics' class: a case study. *Statistics Education Research Journal, 6*(2), 59–77.
Fletcher, S., & Sackett, D. (1979). The periodic health examination: Canadian task force on the periodic health examination. *Canadian Medical Association Journal, 121*, 1193–1254.
Franklin, S., & Peat, M. (2001). Managing change: the use of mixed delivery modes to increase learning opportunities. *Australian Journal of Education Technology, 17*(1), 37–49.
Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: a current review of research on teaching and learning statistics,. *International Statistical Review, 75*(3), 372–396. doi:10.1111/j.1751-5823.2007.00029.x.
González, J. A., & Muñoz, P. (2006). e-status, an automatic web-based problem generator: applications to statistics. *Computer Applications in Engineering Education, 14*(2), 151–159.
Grubišić, A., Stankov, S., Rosić, M., & Žitko, B. (2009). Controlled experiment replication in evaluation of e-learning system's educational influence. *Computers & Education, 53*, 591–602.
Hagtvedt, R., Jones, G. T., & Jones, K. (2008). Teaching confidence intervals using simulation. *Teaching Statistics, 30*(2), 53–56.
Holt, R. D., & Oliver, M. (2002). Evaluating web-based learning modules during an MSc programme in dental public health: a case study. *British Dental Journal, 193*(5), 283–286.
Horton, R. (2000). Common sense and figures: the rhetoric of validity in medicine. *Statistics in Medicine, 19*, 3149–3164.
Krause, U.-M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction, 19*, 158–170.
Larreamendy-Joerns, J., Leinhardt, G., & Corredor, J. (2005). Six online statistics courses: examination and review. *The American Statistician, 59*(3), 240–251.
Margolis, J.-L., Nussbaum, M., Rodríguez, P., & Rosas, R. (2006). Methodology for evaluating a novel education technology: a case study of handheld video games in Chile. *Computers & Education, 46*, 174–191.
Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials,. *JAMA, 285*(15), 1987–1991.
Moodle. (2009). *Philosophy*. Retrieved September 6, 2009 from. http://docs.moodle.org/en/Philosophy.
Morris, E. J. (2001). The design and evaluation of link: a computer-based learning system for correlation,. *British Journal of Educational Technology, 2*, 39–52.
Newman, M., & Elbourne, D. (2005). Improving the usability of educational research: guidelines for the REPOrting of primary empirical research studies in education (The REPOSE Guidelines). *Evaluation and Research in Education, 18*(4), 201–212.

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003

Oliver, M., & Conole, G. (2003). Evidence-based practice and e-learning in higher education: can we and should we? *Research Papers in Education, 18*(4), 385–397.
R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0. http://www.R-project.org.
Simons, H. (2003). Evidence-based practice: panacea or over promise? *Research Papers in Education, 18*, 303–311.
Symanzik, J., & Vukasinovic, N. (2003). Teaching experiences with a course on web-based statistics. *The American Statistician, 57*(1), 46–50.
U.S. Department of Education, Office of Planning, Evaluation, and Policy Development. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Washington, D.C: U.S. Department of Education.
Wilks, S. S. (2006). Undergraduate statistical education. *The American Statistician, 60*(1), 39–45. doi:10.1198/000313006X91773.
Wood, D., Underwood, J., & Avis, P. (1999). Integrated learning systems in the classroom. *Computers & Education, 33*, 91–108.

Please cite this article in press as: González et al., A web-based learning tool improves student performance in statistics: A randomized masked trial, Computers & Education (2010), doi:10.1016/j.compedu.2010.03.003