Measuring the effectiveness of online problem solving for improving academic performance in a probability course

José Antonio González¹ · Mónica Giuliano² · Silvia N. Pérez²

Received: 20 October 2021 / Accepted: 16 December 2021 / Published online: 13 January 2022
© The Author(s) 2022

Abstract

Research on impact in student achievement of online homework systems compared to traditional methods is ambivalent. Methodological issues in the study design, besides of technological diversity, can account for this uncertainty. Hypothesis This study aims to estimate the effect size of homework practice with exercises automatically provided by the ‘e-status’ platform, in students from five Engineering programs. Instead of comparing students using the platform with others not using it, we distributed the subject topics into two blocks, and created nine probability problems for each block. After that, the students were randomly assigned to one block and could solve the related exercises through e-status. Teachers and evaluators were masked to the assignation. Five weeks after the assignment, all students answered a written test with questions regarding all topics. The study outcome was the difference between both blocks’ scores obtained from the test. The two groups comprised 163 and 166 students. Of these, 103 and 107 respectively attended the test, while the remainder were imputed with 0. Those assigned to the first block obtained an average outcome of -1.85 , while the average in the second block was -3.29 (95% confidence interval of difference, -2.46 to -0.43). During the period in which they had access to the platform before the test, the average total time spent solving problems was less than three hours. Our findings provide evidence that a small amount of active online work can positively impact on student performance.

Keywords Adaptive computer learning · Experimental research · Statistics · Higher education · Automated grading

---

✉ José Antonio González
jose.a.gonzalez@upc.edu

Extended author information available on the last page of the article

1 Introduction

1.1 Framework

Computer-assisted assessment, often integrated within Computer-assisted instruction (CAI) systems, is widespread in undergraduate education. It is used in both formative and summative assessments, depending on whether the emphasis is placed on (self-)training, grading, or feedback generation (Chalmers and McAusland, 2002; Massing, Schwinning, Striewe, Hanck, and Goedicke, 2018).

Online formative assessments cultivate the complex cognitive processes that today's students need to succeed, according to McLaughlin and Yan (2017). These authors are in agreement with Clark (2012), who states "The development of self-regulation leads to higher achievement, drive and overall success". Online means should be employed to a) share 'exemplars' of performance; b) post self-grading objective tests; c) disseminate group and individual feedback; and d) elicit feedback from students (Nicol and Milligan, 2006). Clearly, such processes require the appropriate conditions to succeed. For example, digital resources can support students' learning if they are aligned in the course and seem valuable to the students. More specifically, tasks could include applied problems and real-life data sets that students find reasonable and beneficial (Zetterqvist, 2017).

Compared with other CAI cases, automated formative assessment constitutes the most effective use of digital technologies in the classroom, according to Tomasik, Berger and Moser (2018). Nonetheless, using online homework systems has raised concerns about relevant differences from the traditional paper-based model. Trussell and Gumpertz (2020) mention a number of works to emphasize that online homework systems effectiveness in student performance is still unclear. There is a vast literature on teacher and student perceptions toward the modern model (Nardi and Ranieri, 2019), but evidence of online homework enhanced benefits for students is, at best, scattered (Magalhães, Ferreira, Cunha and Rosário, 2020).

1.2 CAI in statistics education

Knowledge of statistics is important to any engineer, since engineering practice involve working with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes (Montgomery and Runger, 2018). Garfield and Ben-Zvi (2009) state that the effective teaching of a subject requires a suitable environment to develop statistical thinking in learners and, further, that information and communications technology plays a central role. In statistics education, many studies have proved positive effects of CAI when comparing them with traditional lecture-based instruction (Kolpashnikova and Bartolic, 2019). Boyle et al. (2014) compiled a narrative review of different systems (based on games, animations and simulations) applied to statistics. A larger and systematic review by Davis, Chen, Hauff and Houben (2018), which was not restricted to statistics, extracted 126 papers that empirically evaluated learning strategies and

identified the most promising as cooperative learning, simulations, gaming, and interactive multimedia.

Larwin and Larwin (2011) present a meta-analysis of 70 studies conducted over 40 years on the use of CAI in statistics education. Such a long period unavoidably reveals profound heterogeneity in the studies analyzed, and they point out that specific features of the course setting may largely influence its impact. For instance, a strictly online format is associated with negative effect sizes, but both face-to-face and blended learning produce positive student achievements. In addition, Sosa, Berger, Saw and Mary (2011) use their meta-analysis to address the benefits of computer-assisted instruction in statistics; one of their findings refers to larger effects reported in studies employing an embedded assessment.

The use of randomized controlled trials to assess the effectiveness of online homework tools for statistical education is scarce, if not absent. Palocsay and Stevens (2008) present a comparison of traditional homework and three other web-based options, oriented to business statistics courses. However, the intervention group was self-selected by the participant. Shotwell and Apigian (2015), in a business statistics setting, compare on-ground students with online students, which use McGraw-Hill Connect and other resources, but once more the study is not randomized. Wood and Bhute (2019) compare pen-and-pencil work with an online tool (WebWork), though they do not employ an experimental framework since their objective is mainly exploratory.

The work presented hereafter aims to contribute to existing knowledge about computer-assisted assessment, specifically to find the effect estimate of probability and statistics students using an online self-assessment tool, e-status.

2 Motivation of the study

2.1 Outline of e-status

The web-based platform used in this study, e-status, was originally developed in 2003 for automated testing in probability and statistics courses (González and Muñoz, 2006). The tool is used to provide the students with a set of problems covering most of the topics in our courses, allowing them to practice with realistic data generated individually by means of the R language. The R programming language is one of the most widely used among statisticians: it ranks 14th in the TIOBE index (2021), which focuses on all types of programming languages.

Typically, an e-status administrator creates *subjects*, assigns teachers to them and registers the students to allow them access to the platform. Following authentication on the platform, students registered in a given subject can access specific *problems* developed by teachers. In essence, a problem is: a probabilistic model implemented with commands written in R language; a set of questions related to the data generated; and a procedure (also in R) for assessing the validity of the student's answer to each question. After solving an instance (or *exercise*), the student can try the same problem with new data. Depending on the intricacy of the model, this new data can take the form not only of changes in numbers and/or charts, but also in the wording

and meaning of the questions. In this way, teachers can encourage students to persist and continue their efforts in order to reinforce their abilities. As an illustration, Fig. 1 displays a screenshot of the platform when a student is solving an exercise.

The student computes an answer by suitable means (e.g., a calculator or statistical software) and enters it in the pertinent text boxes. When required, e-status assesses and corrects the given answers immediately and gives feedback through the R procedure. Moreover, every attempt at answering is saved and can be recovered later, providing both students and teachers with valuable feedback about individual and group performance. Although information about other students is restricted, the platform provides different tools and rankings for students to place themselves among their anonymized colleagues.

When an exercise reaches its time limit before all the questions have been answered, it is saved as an unfinished exercise and does not count in the general statistics on the student. Recovering a specific exercise is possible, but the student cannot resume an unfinished one: he or she must start a completely new instance. One advantage of the mechanism is that the students can glance over a problem to get some idea of it without penalties if they leave it incomplete.

Teachers can incorporate their activities in e-status in a way that closely follows the guidelines of the constructive alignment principle: students construct meaning from the tasks they do and, thus, teachers align the learning activities with assessment in order to facilitate the intended learning goals (Zetterqvist, 2017). The students' efforts produce a phenomenal amount of data: on which students participate more or less; how many times each problem has been solved; which questions were harder to answer; how long an exercise has taken on average; as well as many others.

Calculator Execution: New tires

Normal
t-Student
Chi-square
F of Fisher

User
Welcome pepe
Language: English
Select language
Notify translation error
Change password
Close session

Student
PE Tardor18
History
Select block
Configuration
Select subject

Last passed problem

Time limit: 42 min 58 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 as follows:

tire	1	2	3	4	5	6	7	8	9
life	43.2	37.8	40	40.1	39.7	39.3	35.8	35.7	42.9

1. The null hypothesis is as follows:
 $H_0: \mu = \mu_0$

Which should be the value of μ_0 ?

2. We want to test the claim from the user's point of view (who suspects about the tire's quality). What is the correct direction to take into account this point of view?

1. one-sided (to the left, <)
2. one-sided (to the right, >)
3. two-sided ($\neq$)

3. Which is the sample estimation of the mean life for the new tires?

4. Find the value of the test statistic (three figures).

5. Compute the p-value of the test data (three figures).

6. Find a 90 percent confidence interval for the mean life of the tires tested (two figures of accuracy).

Fig. 1 Screenshot of e-status. The image shows an exercise as seen by a student. White rectangles on the right are text boxes to enter the answers to the questions on the left. A button below (not shown in the image) is intended to assess the responses. The specific problem shown in the image was not one of the employed in the study

Thus, lecturers can use this data to assess the learners' progress virtually and in real time.

The online address of the e-status platform used in this study is <http://estatus.unlam.edu.ar>

2.2 Research objective

The *Universidad Nacional de La Matanza* (UNLaM), under supervision of the *Departamento de Ingeniería e Investigaciones Tecnológicas* (DIIT), offers several engineering programs, all of which have in common the subject Probability and Statistics (P&S). It is worth highlighting that a large fraction of students combine work and study (70% at UNLaM, according to Giuliano, Martínez and García, 2016).

We have currently conducted an experiment at DIIT-UNLaM according to the design described by González, Jover, Cobo and Muñoz (2010), with a version of e-status technologically more advanced, larger sample sizes, an experienced team of various teachers involved instead of only one, and greater variability among the students. Hence, it is of the greatest interest to analyze and answer the following research question:

Does online practice with e-status have a positive impact on probability and statistics learning for a population with high heterogeneity?

3 Study methods

We conducted an experimental study by randomly assigning P&S students to two groups, followed by five weeks of individual training with e-status and, lastly, an assessment through a written test.

The study protocol was presented as an original contribution by Giuliano, Pérez, Falsetti and González (2019) at an international conference on the learning assessment of statistics. The final methodology, described below, remained unchanged from that which was stated in the original protocol.

3.1 Participants

All the students enrolled in the subject P&S were included in the study. The subject is taught by the DIIT in the following five Engineering programs: Informatics, Electronic, Industrial, Civil and Mechanical. Further information was collected from the students through an online survey (via Google form) in the beginning of the course. In view of the significance of new technologies being used to boost student learning, the Dean of the Engineering Department gave consent to implement the study.

Each student was registered to one of six lecture groups (listed as the rows in Table 1). Weekly lectures lasted four hours, with regular breaks in between. Two teachers were present in the theoretical and practical sessions. The teaching

Table 1 Participant data disaggregated by lecture group (six categories), type of degree (two categories), repeating status (yes or no) and intervention group (A or B)

Lecture group (n _A /n _B)	Degree	GA (n=163)		GB (n=166)	
		Repeating	Not repeating	Repeating	Not repeating
Wed. after. (28/29)	Informatics	6	10	6	10
	Other	5	7	5	8
Wed. even. (31/31)	Informatics	11	10	12	10
	Other	3	7	2	7
Thu. morn. (27/29)	Informatics	6	9	6	10
	Other	5	7	6	7
Thu. even. (33/32)	Informatics	12	8	13	8
	Other	8	5	8	3
Fri. morn. (15/15)	Informatics	3	2	2	3
	Other	5	5	5	5
Sat. morn. (29/30)	Informatics	8	5	9	5
	Other	9	7	9	7
N = 329		n=81	n=82	n=83	n=83

after.: afternoon; even.: evening; morn.: morning

materials, schedules, practical work and exams were the same for all lecture groups, which were supervised by a head professor (MG).

The P&S course lasted 16 weeks, with two mid-term tests and one final exam. The first mid-term test corresponded to probability topics and the second one to statistics. The study described here runs up to the first test.

3.2 Interventions and randomization

The teachers of the subject defined two separate sets of topics covering most of the probability course's contents: Contents A (CA) and Contents B (CB). Some of the last topics in the probability part of the course were not included in either CA or CB, since students would be short on time for practicing them properly. We randomly allocated the students into two groups: Group A (GA) and Group B (GB), with allocation ratio 1:1. In order to ensure proper access, students in GA or GB were registered in different "subjects" within e-status, depending on their assignment. A student in any intervention group could access the platform for practice with a number of problems, but both groups differed in the topics covered: GA students could practice with problems related to CA but not with problems related to CB, and GB students with problems related to CB but not with those related to CA. As an example, among the other problems, CA included three on the Bernoulli process while CB included three on the Poisson process. The only common problem available to all the students was one intended to familiarize them with the platform.

A simpler design for comparing one group with access to the platform against another group lacking access was considered, but this option was disregarded to avoid the ethical implications of depriving students of such a resource.

The head P&S professor stratified the students according to: a) the six lecture groups; b) enrolment in the computer science degree program or some other; and c) whether the student was repeating the year or not. Then, she randomly assigned the students into the corresponding e-status group, either GA or GB (see Table 1). The randomization was performed using an Excel spreadsheet. All teachers (but MG) were masked to the intervention group. The students were not, since they could compare the problems and thus find out the assigned intervention.

3.3 The e-status exercises

In our case, we employed 19 problems designed for the e-status platform: nine problems related to CA probability topics for students assigned to GA; nine problems covering CB for students assigned to GB; and one introductory problem that was recommended for becoming familiar with the tool and which did not focus on either CA or CB topics. Table 2 describes the distribution of topics and e-status problems. The answers to questions were numerical, except for the problems that involved identifying variables, which were based on true/false choices. After logging in to e-status with their personal password, students could see ten problems but had no way to practice on the problems of the other group. Since the problems start each time with new data, the students were encouraged to practice with them as often as they felt them helpful in reinforcing their learning. The P&S subject included additional and complementary activities that are common to all students, such as completing work by hand.

Table 2 Main topics and brief description of the deployed e-status problems

	Topics	Description	Number of problems
CA	Probability	Conditional and marginal probability. Bayes' theorem	3
	Bernoulli process	Binomial and geometric distributions	3
	Variable identification	Identifying variables of a Bernoulli process within a problem	3
CB	Probability	Disjoint events. Independence of events	3
	Poisson process	Poisson, exponential and gamma distributions.	3
	Variable identification	Identifying variables of a Poisson process within a problem	3
Common		Probability and combinatorics.	1

3.4 Outcomes

The main outcome was obtained from the mid-term test, which was common to all the students and included questions related to topics from both the CA and CB sets. The student tests were assessed by the teacher responsible for each lecture group, who, as said above, was masked to the intervention assignment. Let QA be the set of questions appearing on the test related to topics in CA (idem with QB and CB). Every question in QA or QB was assessed on a scale from 0 (completely wrong) to 6 (perfect): we denote the score reached by student i in question j , where either $j \in QA$ or $j \in QB$, as y_{ij} . The exam included additional questions related to topics not covered either in CA or CB, because they were relevant for the student's final score and were used to assess the reliability of the scale through Cronbach's alpha.

Finally, the outcome observed for student i (Y_i) was defined as the sum of scores for QA questions minus the sum of scores for QB questions:

$$Y_i = \sum_{j \in QA} y_{ij} - \sum_{j \in QB} y_{ij}. \quad (1)$$

The outcome Y in eq. (1) represents the benefit to students using e-status. If e-status is advantageous in learning, it should on average be a positive number for a student from the GA group and negative for a student from the GB group, and the difference between averages—the effect size—is a measure of that advantage. Conversely, if e-status were neutral, then both groups would have the same expected value in the outcome (not necessarily zero), and thus the effect size would be null. The diagram in Fig. 2 provides a scheme of the design described above, where the shifted curves on the right would represent the distribution of Y in GA centred on positive values and that of Y in GB on negative values, that is, a positive effect of the use of e-status.

Fig. 2 Diagram of the randomized study. The main outcome is the difference between two sets of questions to answer on a written test. The statistical analysis would attempt to find differences between the outcome averages from the two groups. An effect of e-status would express itself as two separate distributions

4 Statistical analysis

All the randomized students taking the subject were included in the main analysis. Students that did not attend the test were imputed with $Y=0$. According to the protocol, the statistical analysis consists of comparing the means of Y from two independent samples —students assigned to GA and students assigned to GB— using Welch’s method (the Welch method is similar to the common t-Student’s test, without the premise that the variances in each group are equal). The method obtains a 95% confidence interval (CI) for the difference of means in the outcome Y between GA and GB, which can be interpreted as an estimate of the effect of e-status on learning a general topic.

Prior to starting the academic year, we performed a series of analyses using simulated data, under the assumption that 200 students would take the test and 100 would not. These simulations aimed to confirm if the expected number of students in the subject would provide enough power for the statistical analysis should the effect be moderate. Those results showed that the (likely) available sample size was sufficiently large for detecting an effect equivalent to 0.5 points on a scale from 0 to 10, with power greater than 80% and two-side risk $\alpha=5\%$. The preliminary simulations proved the analysis to be reasonably robust with regard to Normality assumptions, given the discrete nature of the outcome. Robustness was also checked for the number of questions on the mid-term test and for the granularity of the assessment: two questions in each QA and QB were assessed on a scale from 0 to 5, for which we obtained about 80% power. Ultimately, four QA and four QB questions were used on the test and measured on a 7-level scale, which means less random error and more precision at the end. Data was simulated assuming statistical within-student independence of the scores, which is probably not the case; however, even a moderate, positive correlation between the student’s answers would result in less variability in the outcome and thus provide still more power for detecting the effect.

Secondary analyses were envisaged as either sensitivity analyses or an exploration of interesting relationships among certain subgroups of the sample. We took into account the following information: survey data such as age, gender, whether the student is working or not, hours of work, and other details; academic data like first year in university and number of subjects passed, among others; performance on the platform such as mean score on solved e-status exercises, time spent using the platform, and others. The compared outcomes for given groups were especially relevant in that we focused only on the students that attended the mid-term test, or only on those that used the platform to practice the topics.

5 Results

The study was set up during the first weeks of the 2018 spring term (August to December in Argentina). The anonymized collected data are available in the Zenodo repository (González, Giuliano and Pérez, 2019).

The process of random allocation began in the second week of classes and continued until the fifth week. At the beginning, 67% of the students from the total 329 final participants had already completed registration and could be assigned to either GA or GB. All remaining students were randomly allocated upon completing registration, with particular attention being paid to maintaining balance within the intervention groups in regard to the above-mentioned factors. The mid-term test was taken in the 8th week, precisely on the 6th of October.

5.1 Characteristics of the participants

As expected, randomization generated comparable, well balanced intervention groups. No particular characteristic of the students seems to be unequally distributed between GA and GB, as can be seen in Table 3. Around 81% of students are

Table 3 Characteristics of participants at the beginning

		GA (n=163)	GB (n=166)
Gender – Count (%)	Man	133 (81.6%)	135 (81.3%)
	Woman	29 (17.8%)	30 (18.1%)
	NA	1 (0.6%)	1 (0.6%)
Age – mean (SD)		25.0 (5.2)	25.3 (4.8)
Repeating year – Count (%)	Yes	81 (49.7%)	83 (50%)
	No	82 (50.3%)	83 (50%)
Degree – Count (%)	Informatics	90 (55.2%)	94 (56.6%)
	Industrial	33 (20.2%)	31 (18.7%)
	Civil	19 (11.7%)	19 (11.4%)
	Electronic	16 (9.8%)	20 (12%)
	Mechanical	5 (3.1%)	2 (1.2%)
Lecture group – Count (%)	Wednesday afternoon	28 (17.2%)	29 (17.5%)
	Wednesday evening	31 (19%)	31 (18.7%)
	Thursday morning	27 (16.6%)	29 (17.5%)
	Thursday evening	33 (20.2%)	32 (19.3%)
	Friday morning	15 (9.2%)	15 (9%)
	Saturday morning	29 (17.8%)	30 (18.1%)
Starting year – Count (%)	Until 2005	10 (6.1%)	9 (5.4%)
	2006 — 2010	17 (10.4%)	29 (17.5%)
	2011 — 2015	99 (60.7%)	98 (59%)
	After 2015	37 (22.7%)	30 (18.1%)
Passed subjects – Count (%)	Less than 10	4 (2.4%)	3 (1.8%)
	10 — 20	95 (58.3%)	91 (55.8%)
	21 — 30	48 (28.9%)	55 (33.1%)
	More than 30	16 (9.8%)	17 (10.4%)
Work hours/week – Count (%)	0 — 4	48 (29.4%)	49 (29.5%)
	4 — 12	19 (11.4%)	20 (12.3%)
	12 — 20	11 (6.6%)	12 (7.4%)
	20 — 30	25 (15.3%)	26 (15.7%)
	More than 30	60 (36.8%)	59 (35.5%)

SD: standard deviation

male, and the average age is 25 years, with standard deviation equal to 5 years. Among these students, 36% were working 30 h or more per week and 30% 4 h or less. Half of the students were repeating the previous year: typically, a P&S student delays their studies by about two years in order to work. More than half (56%) of the students are enrolled in the Informatics program. Every lecture group has around 60 students, with the exception of the “Friday morning” group, which has 30 students; and all the lecture groups are split with a maximum imbalance of two on “Thursday morning”. This means that the lecturer, who is likely a relevant covariate in the study, is completely independent of the intervention group.

In order to facilitate balance in the randomization, the degree information was simplified into just two categories: *Informatics* or *Other*. Table 1 explains in detail the distribution of students by the factors “Repeating or Not” and “Informatics or Other”, depending on the lecture group and the intervention group. As stated in the “Interventions and randomization” subsection, random allocation was controlled by those three factors; thus, the final distribution is reasonably satisfactory.

Table 4 Statistics of problems specific to the intervention arms (total=18)

		GA (n=127; NA=36)	GB (n=135; NA=31)
Number of finished exercises	median	11	12
	mean	11.17	12.02
	SD	7.13	6.11
	maximum	50	42
Number of different problems tried	median	9	9
	mean	7.58	7.78
	SD	2.92	2.53
Time spent with exercises (hours)	median	2.23	2.14
	mean	2.35	2.51
	SD	1.73	1.88
	maximum	10.3	9.2
Sum of average scores of problems ^a	median	66.88	60.42
	mean	59.6	55.0
	SD	20.6	19.3
	NA ^b	47	38
Number of logins to the platform	median	7	8
	mean	8.35	8.95
	SD	6.27	7.35
	maximum	48	49
	NA	22	30

SD: standard deviation

NA: not available

^a The sum of average scores from the problems that every student tried is a performance indicator that is more reliable than the mean of average scores, since the number of attempted problems can be anywhere from 1 to 9.

^b The increase in NAs is due to students who had not finished any exercise and thus lacked any score within e-status

Table 5 Statistics of the common problem

		GA (<i>n</i> =127; NA=36)	GB (<i>n</i> =135; NA=31)
Number of finished exercises	<i>median</i>	1	1
	<i>mean</i>	1.63	1.81
	<i>SD</i>	1.23	1.68
Time spent on exercises (minutes)	<i>median</i>	20.5	21.5
	<i>mean</i>	31.1	30.9
	<i>SD</i>	37.5	30.4
	<i>maximum</i>	292	151
Average score	<i>median</i>	6.67	6.67
	<i>mean</i>	6.7	6.1
	<i>SD</i>	2.70	2.97
	<i>NA</i>	43	38

SD: standard deviation

NA: not available

5.2 Usage of e-status

Once the mid-term test was taken, one of the authors (JAG) accessed the data stored on the platform and recovered that which was related to student performance (summarized in Tables 4 and 5). The evaluators had no access to these data, and neither was JAG involved in assessing the test.

Globally, one out of five students never accessed the platform (22% and 19%, respectively, of GA and GB). Table 4 shows some indicators related to the 18 problems linked to the intervention (9 problems each arm). Table 5 summarizes the performance obtained with the 19th problem, shared by both arms. Observed differences between groups are only minor. Overall, the students logged in to e-status 8.65 times on average and up to 49 times in the most extreme case; they practiced with every problem, but with only a low number of repetitions (11.6 on average for the nine intervention problems, with a maximum of 50); and they typically spent a bit less than three hours in total solving exercises on the platform. It is worth mentioning that, as expected, the distributions of these variables are strongly right-skewed.

5.3 Performance on mid-term test

Eight questions were added to the October 6th test: four related to CA topics and four to CB topics. In each group, one question was related to probability and three were related, respectively, to variables within Bernoulli and Poisson processes. Evaluators scored each one with an integer number between 0 (wrong) to 6 (perfect). Therefore, the outcome derived from assessing these questions (*Y*) would range between −24 to 24. The remaining test exercises were not accounted for.

A total of 210 students attended the test: 103 from GA and 107 from GB. Cronbach's alpha was employed to assess the reliability and internal consistency of the measure. The obtained value is 0.80 for the set of eight questions related to the outcome (either in QA or QB). When we extend the set to include two more questions,

Fig. 3 Participant flow diagram of the study

present in the test but unrelated to the outcome, and the mark obtained in the second mid-term ($n=134$), alpha increases slightly up to 0.82.

The non-attending students were imputed with $Y=0$ (see Fig. 3 for details on the participant flow). The average values of the outcome among those attending the test were -2.92 ($SD=5.32$) and -5.10 ($SD=5.27$), respectively, for GA and GB. Considering all the students included in the study, the main analysis finds averages of -1.85 ($SD=4.45$) and -3.29 ($SD=4.89$). The minus sign in these figures reveals that QA questions may be harder, which does not jeopardize the analysis since it rests on the difference of means. We have appended a third analysis

Table 6 Results from intervention group comparison

		GA	GB	estimate 95% CI
All included (main analysis)	mean (SD)	-1.85 (4.45)	-3.29 (4.89)	-1.44
	n	163	166	(-2.46, -0.43)
Only those attending test	mean (SD)	-2.92 (5.32)	-5.10 (5.27)	-2.18
	n	103	107	(-3.62, -0.74)
Only those using e-status	mean (SD)	-2.16 (4.83)	-3.91 (5.09)	-1.75
	n	127	135	(-2.96, -0.55)

SD: standard deviation

on the 262 students that had used the platform to solve at least one exercise. Of them, 60 (30 from each group) did not attend the test, although this particular analysis basically shows a similar in-between effect. These results appear in Table 6, with their respective effect sizes measured using 95% confidence intervals (CI). In all cases, the CI was obtained by means of an unadjusted Welch test for means comparison.

Secondary analyses adjusted by relevant covariates have little impact on the previous results. Table 7 summarizes some models by incorporating the factors used to control the randomization and the weekly working time declared by the participants. None of them explains a significant variation in the outcome (with an exception in the case of the 12-to-20-h group, which could be the result of chance since there are no clear signs of a trend). More remarkably: they do not modify in any relevant way the main estimate of the intervention effect sizes. These models include all the students and the original group assignment. Neither the age nor gender of the students are statistically significant factors.

Additionally, the influence of e-status on the outcome may have an interesting interpretation. Figure 4 shows the outcome Y versus the sum of average scores for the e-status problems, with intervention groups differentiated by two gray shades. The figure suggests that the difference between groups GA and GB increases with more practice in e-status. For groups that perform better on the platform, the

Table 7 Results from linear models after adding one factor to the simple model with only the intervention. Estimates of the effect size are in boldface

	parameters	model estimates	95% CI
With “repeating” factor	intercept	−1.49	(−2.36, −0.61)
	group	−1.44	(−2.45, −0.43)
	repeating (Yes)	−0.72	(−1.73, 0.29)
With “degree” factor (reference is “Civil”)	intercept	−2.18	(−3.76, −0.61)
	group	−1.48	(−2.49, −0.46)
	degree (Electronic)	0.23	(−1.91, 2.36)
	degree (Industrial)	0.62	(−1.27, 2.50)
	degree (Informatics)	0.44	(−1.20, 2.08)
	degree (Mechanical)	−1.82	(−5.61, 1.96)
With “lecture group” factor (reference is “Thursday morning”)	intercept	−2.63	(−3.97, −1.30)
	group	−1.43	(−2.45, −0.42)
	lecture gr. (WeAn)	0.71	(−1.01, 2.44)
	lecture gr. (WeEv)	1.08	(−0.62, 2.77)
	lecture gr. (ThEv)	1.05	(−0.63, 2.72)
	lecture gr. (FrMn)	−0.08	(−2.16, 1.99)
	lecture gr. (SaMn)	1.43	(−0.28, 3.14)
With “work hours/week” factor (reference is “0–4 h”)	Intercept	−1.69	(−2.28, −0.70)
	group	−1.45	(−2.47, −0.44)
	4–12 h	0.07	(−0.93, 1.08)
	12–20 h	−1.47	(−2.80, −0.14)
	20–30 h	0.69	(−0.60, 1.99)
	> 30 h	0.83	(−0.84, 2.49)

Estimates of the effect size appear emphasized in bold

WeAn: Wednesday afternoon; WeEv: Wednesday evening; ThEv: Thursday evening; FrMn: Friday morning; SaMn: Saturday morning.

Fig. 4 Outcome (A-B) comparing both intervention groups' outcomes by performance on e-status problems: very low (less than 30), low (30–45), medium (45–60), satisfactory (60–75) and excellent (more than 75). Maximum is 90 (ten times nine problems). On the right, marginal distribution of outcome for GA and GB

benefit obtained from using e-status is apparent in GB and moderate in GA. Thus, the intervention could be more effective among those students who did more and solved the exercises better.

5.4 Fidelity of implementation

At the time of analysis, we were aware of an error involving ten students who had been inadvertently allocated to both groups: that is, they had access to all 18 problems. All of them had been assigned to GA originally. Two of these students never accessed e-status; seven practiced mostly with the originally intended intervention (of these, one tried only one problem and another tried

two problems from the other group); the tenth student chose exclusively problems from GB. Her outcome was $Y=-7$, which is somewhat suitable to the actual intervention. Anyway, all the analyses account for the original assignment.

6 Discussion

6.1 Findings

The results of this randomized experiment are promising for future students. Evidence is provided that using the platform improves scores on the subject. The difference among averages in the outcome, -1.85 and -3.29 in GA and GB, respectively, is 1.44 . With a pooled standard deviation equal to 4.70 , the observed effect size can be interpreted by computing Cohen's $d=0.307$, which fits somewhere between medium and low magnitudes. In the case of an analysis that excludes students who did not attend the test, Cohen's d statistic rises to 0.412 , which is closer to the medium category.

The present figures are inferior in comparison with its precedent (González, Jover, Cobo and Muñoz, 2010), where the d index found was 0.454 . However, they correspond to those of Sosa, Berger, Saw and Mary (2011) who report a global effect of $d=0.33$ in their meta-analysis on computer-assisted instruction in statistics. Global effect size is larger for Larwin and Larwin (2011): 0.566 , although those studies based on exam scores (total of 104 measures, half of the meta-analyses) give a mean effect size of 0.43 . It is worth considering Bernard, Borokhovski, Schmid and Tamim's (2018) extensive collection of the twenty best-quality meta-analyses on the effectiveness of integrating educational technology, in which, notably, the effect sizes are rarely over 0.4 . However, these systematic reviews include all kinds of learning research that is not necessarily devoted to statistics education. Finally, Magalhães, Ferreira, Cunha and Rosário (2020) do not include a global effect size in their systematic review of 31 studies (of which only 1 was a randomized one), but point out that 9 studies found some benefit for the use of online tools and 15 were found as neutral, because of similar results between online and traditional homework groups.

The systematic review from Shi, Yang, MacLeod, Zhang and Yang (2019), based in 31 high-quality journal articles and college students' cognitive learning as one of the outcomes, notes strong evidence of improvement but also the importance of the social context as a moderator of the effect. One focus of our study was the group of working students: almost 60% of the students declared that they work at least 12 h a week, with most of them reporting more than 30 weekly hours. Consequently, the period of time dedicated to completing one's studies can increase from five to more than ten years (Panaia, 2015). As one would expect, students working less than 4 h a week are the most active group on e-status: only 10% of them solved zero problems, while it was 29% for the other students working between 4 and 30 h, and 35% for those working more than 30 h. However, according to our results, we can see no drop in performance relative to the amount of working time among the active students. A previous study conducted

in the UNLaM by Giuliano, Pérez and García (2016) also found that the students who most greatly appreciated the tool were workers who were pressed for time. Moreover, Berretta, Pérez and Giuliano (2019) have shown that students who used e-status also had a greater tendency to pass the subject than those who did not use it or did so with less intensity. Thus, in our reading, it is worthwhile to give working students even more encouragement to compensate for a loss in face-to-face classes by complementing their studies with an active workload, such as problem-solving in e-status.

We have shown that internal validity is likely not an issue in our setting. The two groups being compared were created by randomizing the students and, after comparing the baseline characteristics and several indicators collected from those using the platform, no relevant difference can be found between them. Teachers were masked, meaning that none of them had access to the platform and therefore could not find out the assignments; thus, it is our opinion that the test assessment was effectively blinded. One cannot exclude, however, that some students may have described their group's problems to the teacher during office hours or a consultation. Similarly, cross-overs between both groups could have happened if students shared information about their problems. However, these situations could only mitigate the real effect size: that is, if the cross-overs had been perfectly prevented, we would have estimated larger effect sizes.

Our students appear to have performed better with questions in QB, which could be explained by factors that are external to e-status (for instance, the difficulty of the Poisson process makes students more likely to seek consultation). A different but interesting interpretation may lie in differing intensities of knowledge transfers between topics. Students practicing CA topics with e-status (e.g., binomial distribution) could benefit more when solving questions related to CB topics (e.g., Poisson distribution) than the opposite. Knowledge transfers derived from practice could result in infra-estimation of the effect size, but they are positive and desirable in the learning process. Even though this study was not designed to measure the effect of these transfers, teachers should not have a concern about them.

6.2 Generalizability and limitations

As often occurs in education research, the validity of the main outcome chosen remains an important limitation; and this is even truer when bearing in mind that the conceptual understanding of mathematical topics is more difficult to measure than procedural understanding (Jones, Bisson, Gilmore and Inglis, 2019). Learning probability and statistics is complex and determined by a student's strengths in different abilities such as recall, understanding, application, and analysis. The instrument employed should correspond to that complexity, and we rely on our experience when devising the tests that determine the level of a student's knowledge. The proposed tests attempt to optimize our ability to evaluate a student's critical thinking skills, as well as to make correction more efficient in the face of having limited time available. In this regard, the questions on the test were developed and arranged

jointly by all the teachers involved in the subject, under the supervision of the P&S head professor, and with the aim of quantifying student achievement of the proposed educational objectives. The Cronbach's alpha obtained from our data (0.82) can be qualified as good, so the used items demonstrate internal consistency.

Our study results could be applicable to other situations, likely with other online platforms, although external validity of this research is compromised due to particular circumstances that may not be possible to replicate. Influential factors that may impede generalization can be type of degree, socioeconomic background, and cultural appreciation for personal work. The knowledge gathered by the authors in recent years provides some advice, for instance: integrate as good as possible your problems with other educational materials on the subject; involve the students; give them easy access to the platform and problems; solve some problems in the classroom to encourage them to try the tool; explain often to the students that it is desirable and even appropriate to adopt an active attitude (e.g., using the platform on a regular basis) for delving into the subject.

The results found are rather conservative, as they are a result of combining students who attended the test and mostly had practiced before, with students who did not attend it and could not be assessed. Thus, the only intervention we are able to evaluate is the advice of working with e-status. However, ancillary analyses (such as those performed on the subset of students attending the test or on the favorable interaction with the platform) suggest very strongly that the benefit to the score would be positively correlated with prior effort. The results were non-conclusive when we included in the statistical model covariates such as gender, degree, work hours and lecture group, which means that they do not likely have a notable effect on the outcome.

7 Conclusion

This work shows that using the e-status platform in P&S was effective in improving the academic performance of engineering students. This affirmation is supported by the thorough study set forth in the previous sections.

Although practice with computer-based systems has been criticized because it mainly involves procedural learning instead of higher orders of cognitive thinking skills, this result of this randomized study proves that active work on an online system can make a difference in student performance. Furthermore, the positive interaction between work and feedback may help learners stay on track with the course's objectives and reduce the dropout rate, which is commonly high among working students.

According to Bernard, Borokhovski, Schmid and Tamim (2018), computers can provide a valuable service to teachers and students. Unquestionably, computer-assisted instruction involves an enormous variety of approaches to learning, and researchers have an obligation to ascertain which of these contribute to effective learning for students. Considering the volume of research on various e-learning methods, it may seem superfluous for us to test one specific approach;

nevertheless, we advocate here that the highest standards of quality be applied to empirical studies on efficacy—as well as to authors’ reported findings.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bernard, R. M., Borokhovski, E., Schmid, R. F., & Tamim, R. M. (2018). Gauging the effectiveness of educational technology integration in education: What the best-quality meta-analyses tell us. *Learning, design, and technology*, 1–25. https://doi.org/10.1007/978-3-319-17727-4_109-2.
- Berretta, G. G., Pérez, S. N., & Giuliano, M. (2019). Análisis y comparación de indicadores de uso de una plataforma de e-learning: e-status. *Revista de Investigación del Departamento de Humanidades y Ciencias Sociales*, (15), 137–153. Available online at: <https://rihumso.unlam.edu.ar/index.php/humanidades/article/view/145/pdf> (accessed 16 May 2019).
- Boyle, E. A., MacArthur, E. W., Connolly, T. M., Hainey, T., Manea, M., Kärki, A., & Van Rosmalen, P. (2014). A narrative literature review of games, animations and simulations to teach research methods and statistics. *Computers & Education*, 74, 1–14. <https://doi.org/10.1016/j.compedu.2014.01.004>
- Chalmers, D. & McAusland, W. D. M. (2002). Computer-assisted assessment. *The handbook for economics lecturers*, 2–12. Available online at: https://www.economicsnetwork.ac.uk/handbook/printable/caa_v5.pdf ().
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24, 205–249. <https://doi.org/10.1007/s10648-011-9191-6>
- Davis, D., Chen, G., Hauff, C., & Houben, G. J. (2018). Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education*, 125, 327–344. <https://doi.org/10.1016/j.compedu.2018.05.019>
- Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics*, 31, 72–77. <https://doi.org/10.1111/j.1467-9639.2009.00363.x>
- Giuliano, M., Pérez, S., & García, M. (2016). Teaching probability and statistics with e-status. In Kaiser, Gabriele (ed.) *Proceedings of the 13th International Congress on Mathematical Education (ICME-13)*. Available online at: http://iase-web.org/documents/papers/icme13/ICME13_S3_Giuliano.pdf (accessed 22 May 2019).
- Giuliano, M., Martínez, M., & García, A. (2016). Clasificación de experiencias de permanencia enunciadas por estudiantes de ingeniería de acuerdo a lógicas de acción del ámbito académico. In *Proceedings of the VI Conferencia Latinoamericana sobre el abandono en la educación superior (CLABES)*. Quito, Ecuador. Available online at: <https://revistas.utp.ac.pa/index.php/clabes/article/view/1396/1899> (accessed 20 June 2019).
- Giuliano, M., Pérez, S., Falsetti, M., & González, J. A. (2019). Diseño experimental para la evaluación de aprendizajes de la estadística con la plataforma e-status. *Actas del Congreso Internacional Virtual de Educación Estadística*. Editores: J. M. Contreras, M. M. Gea, M. M. López-Martín y Elena Molina (Eds.). Available online at: <https://www.ugr.es/~fqm126/pagesCIVEEST/comunicaciones.html> (accessed 22 June 2019).

- González, J. A. & Muñoz, P. (2006). e-status: An automatic web-based problem generator—Applications to statistics. *Computer Applications in Engineering Education*, 14, 151–159. <https://doi.org/10.1002/cae.20071>
- González, J. A., Jover, L., Cobo, E. & Muñoz, P. (2010). A web-based learning tool improves student performance in statistics: a randomized masked trial. *Computers & Education*, 55(2), 704–713. <https://doi.org/10.1016/j.compedu.2010.03.003>
- González, J. A., Giuliano, M. & Pérez, S.N. (2019). Dataset of UNLaM + e-status study. August 3, 2019; Version 1. <https://doi.org/10.5281/zenodo.3359615>
- Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45(3), 662–680 <https://doi.org/10.1002/berj.3519>
- Kolpashnikova, K., & Bartolic, S. (2019). Digital divide in quantitative methods: The effects of computer-assisted instruction and students' attitudes on knowledge acquisition. *Journal of Computer Assisted Learning*, 35, 208–217 <https://doi.org/10.1111/jcal.12322>
- Larwin, K., & Larwin, D. (2011). A meta-analysis examining the impact of computer-assisted instruction on postsecondary statistics education: 40 years of research. *Journal of Research on Technology in Education*, 43(3), 253–278 <https://doi.org/10.1080/15391523.2011.10782572>
- Magalhães, P., Ferreira, D., Cunha, J., & Rosário, P. (2020). Online vs traditional homework: A systematic review on the benefits to students' performance. *Computers & Education*, 152, 103869 <https://doi.org/10.1016/j.compedu.2020.103869>
- McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning*, 33, 562–574 <https://doi.org/10.1111/jcal.12200>
- Massing, T., Schwinning, N., Striwe, M., Hanck, C., & Goedicke, M. (2018). E-assessment using variable-content exercises in mathematical statistics. *Journal of Statistics Education*, 26(3), 174–189 <https://doi.org/10.1080/10691898.2018.1518121>
- Montgomery, DC. & Runger, GC. (2018). Applied statistics and probability for engineers, 7th ed, Sons.
- Nardi, A., & Ranieri, M. (2019). Comparing paper based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology*, 50(3), 1495–1506 <https://doi.org/10.1111/bjjet.12644>
- Nicol, D. & Milligan, C. (2006). Rethinking technology supported assessment in terms of the seven principles of good feedback practice. In C. K. C. Bryan (Ed.), *Innovative assessment in higher education* (pp. 64–77). Routledge.
- 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 <https://doi.org/10.1111/j.1540-4609.2008.00167.x>
- Panaia, M. (2015). Nuevas demandas para las ingenierías tradicionales. Proceedings XI Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires, Buenos Aires. Available online at: <https://www.aacademica.org/000-061/380> (accessed 12 May 2019).
- Shi, Y., Yang, H., MacLeod, J., Zhang, J., & Yang, H. H. (2019). College students' cognitive learning outcomes in technology-enabled active learning environments: A meta-analysis of the empirical literature. *Journal of Educational Computing Research*, vol., 58(4), 791–817 <https://doi.org/10.1177/2F0735633119881477>
- Shotwell, M. & Apigian, C.H. (2015). Student performance and success factors in learning business statistics in online vs. on-ground classes using a web-based assessment platform. *Journal of statistics education* volume 23, number 1(2015). <https://doi.org/10.1080/10691898.2015.11889727>.
- Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research*, 81(1), 97–128 <https://doi.org/10.3102/0034654310378174>
- TIOBE - The Software Quality Company. (2021). Accessed 18 October 2021, from <https://www.tiobe.com/tiobe-index/>
- Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 2245 <https://doi.org/10.3389/fpsyg.2018.02245>

- Trussell, H. J., & Gumpertz, M. L. (2020). Comparison of the effectiveness of online homework with handwritten homework in electrical and computer engineering classes. *IEEE Transactions on Education*, 63(3), 209–215 <https://doi.org/10.1109/TE.2020.2971198>
- Wood, P.M. & Bhute, V. (2019) Exploring student perception toward online homework and comparison with paper homework in an introductory probability course. *Journal of college science teaching*. May/June 2019 (volume 48, issue 5).
- Zetterqvist, L. (2017). Applied problems and use of technology in an aligned way in basic courses in probability and statistics for engineering students—a way to enhance understanding and increase motivation, *Teaching Mathematics and its Applications: An International Journal of the IMA*, Volume 36, Issue 2, June 2017, 108–122, <https://doi.org/10.1093/teamat/hrx004>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

José Antonio González¹ · Mónica Giuliano² · Silvia N. Pérez²

Mónica Giuliano
mgiuliano@unlam.edu.ar

Silvia N. Pérez
sperez@unlam.edu.ar

¹ Universitat Politècnica de Catalunya, UPC Barcelona-Tech, Campus Nord C5, Jordi Girona 1-3, 08034 Barcelona, Spain

² Universidad Nacional de La Matanza, Buenos Aires, Argentina