1. Introduction

Computer-assisted assessment (CAA)—that is, using digital systems for automated assessment—is nowadays widely used in many university courses from various fields. It can be applied in numerous contexts, such as providing automated feedback for (self-)training as a 24/7 service, or providing automated grading for final exams to relieve teachers from the burden of time-consuming manual grading. As such, it can be used for both summative or formative assessment (Conole and Warburton 2005). There is evidence that using CAA to offer massive (self-)training opportunities helps students to better understand the material of a course and thus improves performance in final exams (Chalmers and McAusland 2002). This article contributes to this literature by reporting on the experiences with the web-based CAA system JACK to support a lecture on mathematical statistics at the University of Duisburg-Essen in Germany during the summer term 2014. JACK is a framework for delivering and grading complex exercises of various kinds. It was originally created to check programming exercises in Java (Striewe, Balz, and Goedicke 2009) and thus named JACK as an acronym for “Java Checker,” but has seen several steps of development in recent years (Striewe, Zurmaar and Goedicke 2015; Striewe 2016). (Selected exercises are available at https://www.jack-demo.s3.uni-due.de/. We invite readers interested in using JACK to contact the authors.)

While the lecture was offered in a traditional manner without CAA support in other years, it made use of JACK both for formative assessment in the form of exercises and summative assessment in the form of tests and exams in 2014 (see Section 3.3 for later years).

The aim of this article is to analyze the effectiveness of JACK for helping students learn statistics. Our results strongly indicate that using such a CAA system is effective in achieving this goal. Our evidence is as follows: first, there is a substantial positive effect of training. Points obtained in training exercises during the term are good predictors for the grades achieved in the final exams (Sections 3.4 to 3.6). Second, students in our term achieved better results in summative assessment compared to other terms without CAA usage (Section 3.3). Third, students reported to invest more time in our course (Section 3.1) and, fourth, students praised the opportunity to use a CAA in the evaluation (Section 3.8).

JACK allows generation of appropriate random numbers to be used in calculations. Hence, we offer “variable content”; students working on the same exercise multiple times as well as different students working on the same exercise at the same time work with different numbers. This may be one of the most useful benefits of CAA systems in quantitative subjects (Milne, Honeychurch, and Barr 2013; Sangwin 2013): it is plausible that it motivates students to tackle exercises multiple times and to discuss strategies to solve exercises instead of exchanging solutions (Bacon 2011). Both features are expected to have a positive effect on students’ performance. Moreover, JACK allows us to offer several increasingly informative hints and feedback when students cannot solve an exercise directly.

The article is organized as follows: Section 2 presents the structure and concept of the course discussed here. Section 3.1 offers an initial quantification of student learning habits using a CAA. The remainder of Section 3 presents our statistical analysis of the effectiveness of JACK. Related work can be found in Section 4, while Section 5 offers concluding remarks.

2. The Setup of Lectures, Tests, and Exams

2.1. Overview of the Course Structure

We now discuss the setup of the course “Induktive Statistik” held in the summer term 2014 by one of the authors at the University
of Duisburg-Essen, a fairly large public German university. This 15-weeks class gives an introduction to mathematical statistics, discussing probability theory, estimation theory, and testing theory. The full schedule can be found in Appendix C. The lecture is the second course in statistics within the curriculum, building on an earlier course in descriptive statistics. It is compulsory for several programs in business and economics. The class met twice per week. Next to classical lectures, 75 exercises were jointly solved and discussed at appropriate moments in class. The course started with about 800 students, of whom about 500 students took an exam at the end of the course. These numbers are close to those of other years. At the beginning of the term, the lecture was visited by about 600 students and at the end, by about 300 students. As attendance in German universities is not compulsory, such drop-out numbers are all but uncommon and, if anything, smaller than those in other years. Those who did not attend the lecture on a regular basis could download the lecture slides in the learning management system Moodle. The material of the course is based on the textbook by Assenmacher (2013).

In other years, homework exercises such as the 75 mentioned above were discussed in tutorial classes. In 2014, we did not offer such tutorials anymore, and instead offered homework and tests online using JACK. This decision was motivated by several reasons. First, students would often only copy the homework solutions presented by tutors in the tutorial classes, rather than attempting to find own solutions prior to the meetings. Lack of manpower implies that it is infeasible to manually correct and grade homework of several hundred students. Second, it is useful that JACK provides students with immediate individual feedback after submitting a solution. (A "submission" is one attempt to solve one specific online exercise.) This may be expected to be more motivating for students than to wait for the next weekly meeting of a tutorial group to see if their solution was correct. Third, JACK offers multilevel hints. Fourth, we are able to provide far more exercises than before, as time constraints in the traditional tutorial classes limited us to a few exercises per week. Fifth, we can offer five online tests during the term in which students could gain bonus points for the final grade (see Section 2.3 for additional details). Again, it would have been impossible to manually grade such a number of tests. The goal of these tests was to motivate students to continuously study during the term and not only shortly prior to the exam. Sixth, JACK has an interface to the statistics software R (R Core Team 2017). R is used to draw appropriate random numbers for the exercises, as well as to automatically and immediately evaluate exercises that are as challenging as those used in nonelectronic exams and exercises in other years. In the future, we moreover plan to create JACK exercises teaching to use R (see Section 5). Schwinning et al. (2017) discussed the technical details of JACK, especially providing information on the connection of JACK and R.

Such a setup seems beneficial both for students and teachers. The operating costs of such a system, once established, imply little if any additional running workload for the teaching staff while improving the quality offered to students.

At the end of the term, two traditional handwritten exams were offered (one in July and one in September). Students can only retake an exam if they failed or did not take the previous ones (so that students can pass at most once), but can fail several times. (Students obtain 6 “malus points” for each failed exam. They may collect at most 180 malus points during their whole bachelor program. When exceeding this threshold, students cannot complete their degree. Students may hence fail each specific exam in the program several times, provided they do not need many attempts in other courses.) In this term, we additionally offered two electronic exams (see Table 1). Students were free to choose which exam to take. According to the German scale, we graded the exam with “100” for very good, “200” for good, “300” for satisfactory, and “400” for modest; “500” stands for failed exams. Thus, while exams were taken at the end of the term, online tests were taken during the term to gain bonus points (see Section 2.3) for the final grade. Table 2 summarizes the different assessment types.

Despite the automated hints and feedback messages to online submissions, many students still value personal contact. We therefore offered additional contact hours in a PC pool (200 student PCs) for individual questions and help with the electronic system. Furthermore, we provide an online forum in Moodle where students may ask questions and discuss their solutions with fellow students and tutors. For additional information, we refer to the (translated) syllabus and to the schedule in Appendices B and C.

### 2.2. Homework Exercises

In total, we created 151 different homework exercises which we made available at several dates during the term (see Appendix C for the schedule). Once an exercise was available, students were allowed to submit solutions to it as often as desired. All homework exercises were voluntary and had no deadline. Table 4 shows the topics covered in the exercises and the lecture.

Figure 1 shows a simple “fill-in” exercise as an example. The student is faced with a certain task and submits a solution in the input field. The present task is to apply the probability mass function of the binomial distribution. Appendix A contains further examples of exercises used in the course, which are available for testing at https://jack-demo.s3.uni-due.de/jack2/demo? id=50302. Exercises were translated into English for this article while the original exercises are in German.

### 2.3. Tests

In five electronic tests (see Table 3), students could gain up to 10 bonus points, to be added to a maximal possible score of 60 points in the final exam. These are only added to the exam score if a student passes the exam. We adopted this policy as students were allowed to take the online tests at home, which enabled them to work collaboratively (see Table 2). We believe that joint work is useful, but still want to make sure that only students

| Table 1. Exams. |
|-----------------|
| Date | No. of students | No. of tasks |
|----------------|
| 1st electronic exam | 07/12/14 | 221 | 6+1 handwritten |
| 1st handwritten exam | 07/28/14 | 218 | 3 |
| 2nd electronic exam | 09/05/14 | 131 | 6+1 handwritten |
| 2nd handwritten exam | 09/22/14 | 124 | 3 |

NOTE: Dates, number of participants, and number of tasks in the exams.
Table 2. Assessment types.

|                | JACK homework | Online tests | Exams        |
|----------------|---------------|--------------|--------------|
| Formative or summative? | Formative     | Summative    | Summative    |
| Graded?         | No            | Only bonus points | Yes        |
| Mandatory to pass? | No            | No           | Yes          |
| Where?          | Anywhere      | Anywhere     | PC/lecture hall |
| When?           | Anytime       | 5 fixed dates during lecture period | 4 dates after lecture period |
| Electronic?     | Yes           |             | First and third yes, second and fourth no |
| Duration?       | No limit      | 40 minutes  | 60+10 minutes |
| Multiple submissions? | Yes          | Only during test | Only during exam |
| Collaborative?  | Possibly      | Only during test | No |
| Overview of statistics | Table 4      | Online test points | Table 5 |
| Measure of success | Score / number of submissions | From 0% to 100% (used for bonus points ranging 0 to 10) | Grade |
| Scale           | From 0 to 15,100 (100 - 151) / in counts |                | “500” (fail), “400” (modest), “300” (satisfactory), “200” (good), “100” (very good) |

NOTE: Summary of the different assessment types in the summer term 2014. Line 1 reports if assessment type is formative or summative. Line 2 reports whether the assessment is graded. Line 3 reports if students have to pass this assessment to pass the course. Line 4 reports the place and line 5 the time when the assessment took place. Line 6 reports if the assessment was electronic. Line 7 compares the duration of the assessment and line 8 if students could try exercises multiple times. Line 9 reports if students could work collaboratively. Line 10 refers to an overview of descriptive statistics. Line 11 and 12 discuss how success is measured.

Figure 1. Fill-in exercise with hints. An exercise dealing with binomially distributed random variables using the R function “dbinom.” The figure shows both the exercise and three hints that have been requested by the student.

Exercise "(Exercise36) Distributions of discrete random variables 2"

Assume that recent demographic studies have shown that the probability of a baby girl being born equals 0.497. What is the probability that 0 boys are part of a family with a total number of 3 kids?

(If necessary, round your result to the fourth decimal and use a point as decimal mark.)

Hint

Think about which distribution the r.v. X follows. (Another hint is available.)

The r.v. X (number of births of boys) is binomially distributed with parameters n = 3 and p = 1 − 0.497. We seek the probability P(X = 0). (Another hint is available.)

P(X = x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}. Now just fill in the corresponding values!

2.4. Exams

As discussed in Section 2.1, we offered two electronic exams. We award partial credit for solutions which we consider as partially correct. For example, when the unbiased sample variance is to be computed, it seems appropriate to award partial credit when a student computes \( \frac{1}{n} \sum (x_i - \bar{x})^2 \) rather than \( \frac{1}{n-1} \sum (x_i - \bar{x})^2 \).

Students only submit final solutions and we, therefore, do not see how they arrived at that solution. This is an important

Table 3. Online test dates.

|       | Date    | No. of students | No. of solutions | No. of tasks |
|-------|---------|-----------------|------------------|--------------|
| 1st T | 05/06/14| 731             | 2794             | 4            |
| 2nd T | 05/20/14| 600             | 1731             | 3            |
| 3rd T | 06/03/14| 472             | 2024             | 5            |
| 4th T | 06/17/14| 382             | 1393             | 4            |
| 5th T | 07/01/14| 330             | 1255             | 4            |
| Add T | 07/08/14| 98              | 305              | 4            |

NOTE: Dates and number of participants of the online tests. The additional online test (08.07.2014) was a make up test.

with a sufficient individual understanding of the material pass the course. At any rate, as we also used variable content in the tests, each student still got to see different numbers for his or her exercises. Hence, students working on the tests together would still need to share strategies to solve a question rather than just numerical answers.

The tests’ questions ranged from easy ones with no context (e.g., “Compute the probability P(A ∪ B).”) to more difficult ones. Within the time limit of 40 min students may resubmit their answers as often as desired as they do not get any feedback on their submissions while still working on the test.
difference between an e-exam and a handwritten exam. In the latter, the grader has a chance to see the mistake a student makes and award partial credit. Automated partial credit is therefore important. Of course, such a list of anticipated mistakes in the e-exam will never be exhaustive, but we try to cover as many likely and typical mistakes as possible. When discussing the e-exam with students this was one of their main concerns, causing us to treat it carefully.

Additionally, we assigned partial credit to follow-up mistakes. (This is to minimize the effect of, e.g., erroneous uses of students’ calculators.)

Figure 20 in Appendix A presents an additional approach to control for the effect of such minor mistakes. Instead of only asking for the correct numeric solution we also ask for the students’ solution approach in a drop-down menu. This menu lists many (in total 108) possible approaches, so that students are unlikely to pick the correct approach by chance. We acknowledge that not all fairness concerns can be fully addressed in this manner. At the same time, believe that traditional grading is not free of such concerns either, for example, due to different standards of graders or fatigue. Finally, not every type of exercise is appropriate for an e-exam. Since analytical exercises like derivations or proofs typically do not ask for a single number we did not cover these types of exercises in the electronic part. Instead, the e-exams also contained a small handwritten part (worth 15% of the points).

3. Evaluation

3.1. JACK Data

We now discuss the data on learning habits in the formative assessments (see Section 2.2), the data from the online tests (see Section 2.3), and exams (see Section 2.4). From a student’s submission, we observe the name of the exercise, the students’ ID, the date when the exercise was started, and the points at the end of the exercise (from 0 to 100). We received 131,426 submissions in total. We focus on the 118,423 submissions from the 493 students who participated in an exam. The other submissions were from students who did not show up or did not register for any exam. This information is summarized in Table 4. For each of the covered topics, it presents the number of different exercises, the number of submissions and the number of correct submissions. Column 4 presents the average number of points achieved per category. Figure 2 shows a bar plot of this column. For more advanced topics, the average number of points is lower than for introductory exercises due to their increasing degree of difficulty. The second-to-last column reports the number of students who submitted exercises in that category. While the total number of students is 493, not every student tried every exercise. This is, for example, because we announced that some of the last parts of the lecture would not be covered in the early exams (see Appendix C). Figure 3 visualizes the last column of Table 4. It presents the average number of submissions for each student per exercise, that is, the number of submissions divided by the number of exercises in each category and the number of different students submitting to this category. In that column, the numbers in parenthesis use 493 (= total number of students) instead. Comparing Figure 3 with Figure 2 shows that difficult categories with a low number of average points (e.g., confidence intervals) do not necessarily lead the students to work more in this category. On the other hand, not every category with a high number of submissions has a high number of average points. It therefore seems that there are other reasons for a high number of submissions than success, for example, personal interest or students’ opinion of what is most important for the final exam.

We now present an analysis when students learn. Figure 4 shows the number of submissions per day against the date. Evidently, the submissions peak at a few days during the term. There are at least some submissions (minimum = 4) each day, with an average of 700 daily submissions. We see large values around the days of the online tests and even higher peaks.

| Task                  | No. of exercises | No. of submissions | No. of correct submissions | Average points | No. of different students | No. of submissions / (no. of students - no. of exercises) |
|-----------------------|------------------|--------------------|----------------------------|----------------|--------------------------|---------------------------------------------------------|
| Basics                | 5                | 5937               | 5398                       | 87.96          | 479                      | 2.48 (2.41)                                             |
| Probability theory    | 8                | 11,510             | 7905                       | 64.22          | 470                      | 3.06 (2.92)                                             |
| Conditional probability | 10              | 8265               | 5543                       | 60.33          | 454                      | 1.82 (1.68)                                             |
| Combinatorics         | 8                | 11,456             | 7338                       | 59.11          | 456                      | 3.14 (2.90)                                             |
| One-dimensional r.v.s  | 15               | 16,690             | 9280                       | 51.25          | 448                      | 2.48 (2.26)                                             |
| Discrete r.v.s        | 20               | 16,619             | 10,664                     | 59.36          | 441                      | 1.88 (1.69)                                             |
| Continuous r.v.s      | 16               | 13,728             | 8426                       | 54.16          | 416                      | 2.06 (1.74)                                             |
| Convergence           | 4                | 5106               | 2445                       | 44.82          | 418                      | 3.05 (2.59)                                             |
| Estimation theory     | 5                | 3419               | 1777                       | 42.04          | 403                      | 1.70 (1.39)                                             |
| Confidence intervals  | 8                | 3650               | 1627                       | 34.23          | 274                      | 1.67 (0.93)                                             |
| Hypothesis testing    | 11               | 4014               | 2310                       | 49.23          | 268                      | 1.36 (0.74)                                             |
| Two-dimensional r.v.s | 8                | 1739               | 591                        | 34.03          | 173                      | 1.25 (0.44)                                             |
| Online test exercises | 24               | 11,437             | 5122                       | 37.97          | 338                      | 1.41 (0.97)                                             |
| Exam exercises        | 9                | 4853               | 1613                       | 29.45          | 349                      | 1.55 (1.09)                                             |

Table 4. Topics and submissions.

NOTE: Covered topics from the lecture with number of exercises, number of submissions, number of correct submissions (more than 50% of the points), average points, the number of students who submitted at least one exercise from each category, and the number of submissions per student and exercise. In the last column, we use first the numbers of different students in the denominator (we divide by 493 = number of total students to obtain the number given in parentheses). As a maximum, one student tried one specific exercise 46 times. The number of different students strongly decreases in the last sections since these topics were not relevant for the first exams. Note that the total number of students in the table is 492 although we have 493 students. This is because one student did not do homework on JACK at all. After the online tests/exam dates, their exercises were published as additional homework (third last and second last line).
Figure 2. Average points per exercise. The average number of points attained with averaging over each submission per category. For more advanced topics the average number of points is lower than for introductory exercises due to their increasing degree of difficulty. After the online tests/exam dates their exercises were published as additional homework (third last and second last bar).

Figure 3. Average number of submissions. The average number of submissions for each student per exercise. It is the number of submissions divided by the number of exercises in each category and the number of different students submitting to this category. After the online tests/exam dates, their exercises were published as additional homework (third last and second last bar). Comparing this figure with Figure 2 shows that difficult categories with just a little number of average points (e.g., confidence intervals) does not necessarily lead the students to work more in this category. On the other hand, not every category with a high number of submissions has a high number of average points. This leads us to the conclusion that there are other reasons for a high number of submissions than success, for example, personal interest or students’ opinion of what is most important for the final exam.

before the exams. We do not observe many submissions on the dates of publishing new exercises (see Appendix C). Hence, students apparently concentrate their efforts on the days prior to a test or an exam. This is somewhat expected, as many instructors report such behavior for their classes. For example, Striewe and Goedicke (2011) observed similar numbers in an earlier computer science lecture on programming that was also supported by JACK. Hence, we arguably increased students’
Figure 4. Time series of submissions. Daily submissions of homework exercises on JACK. Special events are indicated with the different dots and the corresponding date on the time axis. There are peaks around the days of the online tests and even higher peaks before the exam while otherwise there is less students activity.

Figure 5. Student learning hours. Self-assessed weekly student learning time in hours for the years 2012–2016. In 2014, more students invested more time than before and after while fewer students invested less time.

learning efforts during the lecture period (04/07/2014 to 07/18/2014) as there would presumably have been less learning activity in the absence of online tests.

It is therefore plausible that such students’ efforts would have been substantially smaller in other, traditional, editions of the course, where lack of manpower implies total absence of graded weekly homework. To provide additional evidence regarding this claim, we consider the students’ self-assessed number of hours of self-study for the course, which students are asked to estimate in the course evaluation (see Section 3.8). Figure 5 compares the estimated learning time in the summer term 2014 with the two previous and following editions of the course (offered by other lecturers) not using a CAA but running traditional exercise classes. Obviously, students generally invest more time in our edition. In particular, there are more students who invest at least 5 hr per week while there are fewer with
only a few hours. Thus, learning efforts appear to have increased thanks to JACK. (Of course, students’ self-assessment of the amount of self-study may be biased, more likely upwards than downward. That said, we do not believe that the assessment is more or less biased in our edition of the course.) Unfortunately, we cannot include this assessment in the following models (see Sections 3.2, 3.5, and 3.6) as the evaluations are, of course, anonymous.

3.2. Exam Performance

As mentioned in Section 2, we offered four exams in total, alternating e-exams and traditional handwritten exams (see Table 1). All exams lasted 60 min (plus 10 min of introductory reading time) and took place on campus and were proctored to ensure individual efforts. Students could achieve up to 60 points (plus up to 10 bonus points from the online tests, see Section 2.3 and the Appendix). At least 25 points were needed to pass the exam. If so, bonus points from the online tests were added to students’ final grades. We award the worst passing grade “400” for 25 to 30.5 points, a satisfactory “300” for 31 to 39.5, “200” for 40 to 47.5, and “100” for more than 48 points. (The grades contain several intermediate sub-grades (370, 330, etc.), which are not discussed here.) The JACK users’ data allow us to evaluate the learning progress of students and relate it to their exam grades.

Table 5 reports aggregate results for all four exams. The grades of the first e-exam and the first handwritten exam are similar. In the next two exams, the grades deteriorated, likely because many good students passed one of the first two exams. Overall, there is no obvious difference between the two kinds of examinations, which is encouraging as the exams should be comparable. In total, 66.73% of the students passed the course. In comparison to other editions of the course and other introductory courses of the program, this passing rate is rather satisfactory (see in particular Section 3.3).

Figure 6 compares kernel density estimates for the exam points for each of the four exams. (Positive density for negative points is caused by the properties of the kernel density estimator and many “zero points” exams.) The densities for the exams do not differ substantially, which supports the claim that students are comparably successful in the e-exams and the handwritten exams. All estimates are unimodal and nearly Gaussian. Comparing the first e-exam with the first handwritten exam shows that the estimated density is higher in the tails for the handwritten exam, but lower close to the mode than for the e-exam. Possibly, motivated students who attach a higher value to

![Figure 6. Density estimate without bonus points. Comparison of the kernel density estimates of the exam points for each of the four exams. Students could gain at most 60 points and at least 0 and they have passed with at least 25 points (vertical dashed line). Note that positive density for negative points is caused by the properties of the kernel density estimator and many “zero points” exams. The densities for the exams do not differ substantially, which supports the claim that students are comparably successful in the e-exams and the handwritten exams. All estimates are unimodal and nearly Gaussian. The estimated density is higher in the tails for the first handwritten exam but lower close to the mode than for the e-exam. Possibly, very good students tended to be risk averse and did not participate in the new electronic assessment. Furthermore, poor students who rarely did JACK homework avoided the e-exam as well but, due to their weak preparation, gained few points in the traditional exam.](image-url)
good grades are risk averse and did not participate in the new electronic assessment. Furthermore, poorly prepared students who rarely did JACK homework may have avoided the e-exam as well. Due to their weak preparation, they may have scored no or few points in the traditional exam. Figure 7 shows the kernel density estimates of total points, including the bonus points from the online tests for each of the four exams. This causes the density estimates to be bimodal (except for the second handwritten exam), with one mode for failing students and one for passing students. The first e-exam has a higher rightmost mode. This may be due to students with a positive experience of JACK homework and tests tending to take the first e-exam.

Three hundred and eighty out of a total of 493 students participated at only one of the four exams (231 of them passed), 130 participated in two exams (81 eventually passed), 30 at three exams (16 passed), and 5 at all exams (1 passed). Of the 165 multiple exam participants only 29 exclusively took handwritten exams (14 passed) and 13 exclusively electronic exams (10 passed). The 123 remaining students switched types, which suggests that both types of exams were accepted by the majority of students.

### 3.3. Comparison to Other Terms

It is interesting to compare the exam results to those of earlier and later terms with more traditional setups to investigate whether the e-learning approach is beneficial for students’ success. We hence compare our term with those in 2013, 2015, and 2016. Being offered by other lecturers, no CAA was used those terms and only handwritten exams were offered. Although the exams largely cover the same material, they may differ in their level of difficulty. Table 6 gives an overview of final grade counts. The grades in 2014 are at least as good as those in editions of the course without JACK (see the “average” grade). In addition, the total number of students is much higher in our edition. This supports the claim that the CAA attracts students (who are free to choose the term in which they take this course).

However, it is not easy to identify any differences as causal effects of the CAA, as the lecture in 2014 was held by the two lecturers for the first (and up to now only) time. Grading and teaching styles vary between professors. For example, possible "grading-to-the-curve" implies that a grade distribution may be a rather poor reflection of students’ learning outcomes. To reduce such effects, Figure 8 compares the average exam points in percentage points.

In 2014 and 2015, students did better in earlier exams. This is reasonable as the good students who passed the first exam do not
participate at the next, such that, on average, weaker students take later exams. Hence, the average number of points decreases. Interestingly, in 2013 and 2016 the average number of points is lower in the first exam. We cannot offer a conclusive explanation as we did not offer the courses in these years. (There is, of course, quite some noise in exam grades, for example, because students may have found certain questions to be much harder than anticipated by the author of the exam. Hence, a comparison of only 10 exams may be affected by more factors.)

The first and second e-exam of 2014 score better than the first handwritten and second handwritten exam of any other term. Moreover, the difference in points of the first e-exam to any other first exam is significant at the level 5% (Welch’s t-test). For the second e-exam, the difference is significant only for the second handwritten exam in 2016. Besides that, the average for our first handwritten exam is still higher than in any other handwritten exam in other terms, even though many good students have already passed the first e-exam. Only the second handwritten exam of our edition compares poorly to the other terms. Again, this is very likely due to many good students having already passed in earlier exams.

To conclude, there is evidence that students perform better in summative assessments when they have access to a CAA. However, we note that there are some potential confounding variables (e.g., different lecturers). Hence, we caution against interpreting the usage of JACK as the (only) causal effect for the better results.

### 3.4. Learning Progress

We focus on data from the summer term 2014 in what follows and analyze the relationship between JACK usage for the formative (“homework”) exercises and the final grade. Section 3.5 will discuss the relationship between JACK usage and the online tests.

We split the analysis into two parts, investigating how the effort (in terms of the number of homework submissions in JACK) is related to the grade and how the success (in terms of points in JACK) is related to it. Of course, both are highly correlated since successful students tend to persevere, while students with low or no success may become frustrated and stop investing effort (Weiner 1972).

To understand the impact of effort, Figure 9 shows the average number of submissions per student. It is separated into five time periods. The x-axis reports the time period and the y-axis gives the average number of submissions for all (black), for passing (green), and for failing students (red). We divide the number of submissions in this period by the size of the group. Consider the bars corresponding to the lecture period. The black bar reports the average number of submissions of the 493 exam-writing students in that period. The green bar shows the average number of submissions of the 329 students who passed one of the exams. The red bar is for the 164 students who did not pass any exam. While the first period represents the average submissions of all exam-writing students, the second block focuses on the 221 students who took the first e-exam. The third time period shows the average number of submissions before the other three exams. The numbers of participants, fails, and passes can be found in Table 5. We only compare values within time periods and not between them as the reference groups change. We report five segments because during the lecture period almost every student uses JACK (or should have, recall Table 4), whereas in the other time periods only those students who took the next exam are relevant. The other students did not use JACK during this period, likely busy preparing for other exams. The different segments of Figure 9 show that, over
Figure 9. Average number of submissions per student. The average number of submissions per student in the corresponding time period. The three bars of the first time period during the semester (from left to right) consider all 493 students (black), the 329 students who passed one exam (green) and the 164 students who failed any of the four exams (red). The second segment considers the 221 students who participated in the first electronic exam (black), the 122 who passed (green) and the 99 who failed (red). The third segment considers the 218 students who participated in the first handwritten exam. The fourth segment considers the 135 students who participated at the second electronic exam. The fifth segment considers the 124 students who participated at the second handwritten exam. For all bars the average number of submissions of the passing students is higher than the average number of submissions of all students, which in turn is higher than the average number of submissions of the failing students.
all time periods, the average number of submissions of passing students is higher than that of failing students. As expected, higher effort correlates with a higher probability of passing the exam. Students taking an e-exam (second and fourth period) used JACK more actively compared to students taking a handwritten exam (third and fifth period). That is likely because the latter students believe that the JACK exercises are less relevant for handwritten exams.

Furthermore, the average number of daily submissions is 1.42 when considering all days and the whole term. Split up according to the five grades, average daily submissions are 1.01 for grade "500" (fail), 1.41 for "400," 1.5 for "300," 1.73 for "200," and 2.21 for "100," showing that students with better grades did more homework in JACK.

We now analyze if students who achieved better homework results perform better in the exam than those with fewer points. The final sum of points of homework exercises could easily be increased by many correct submissions of the same introductory or easy exercise. We therefore do not look at the total number of submissions or correct submissions, but estimate the learning progress for every student. We do that by tracking every student’s "score" of points, defined as the sum of points of the last submissions to every exercise. To give an example, if someone got 100 points in one specific exercise at the first attempt but only 50 points at the second, his score decreases by 50 points from the first to the second try. This is to reflect that if a student initially correctly solved an exercise but could not do so anymore later, the "score" should negatively reflect his/her learning progress. The individual scores vary from 0 to the total number of exercises times 100 (in this term, 15,100). Figure 10 shows the average progression of the student score for each grade group. Additionally, the solid black line describes the average progression for all students. We observe that the better the final grades, the better the students have developed according to their score. In fact, this claim almost always holds true during the entire term. The curves for students with a moderate exam ("400") and those who failed ("500") are similar at the beginning of the term. In July the "moderate" students, however, progress substantially while the failing students barely improve. Hence, both groups are moderately successful in the beginning, but the passing students improve shortly prior to the exams.

To conclude, both effort and learning progress have a strong positive effect on the final grade. Section 3.6 elaborates on this discussion. We, however, discuss the relationship between exercises and online tests first.

3.5. Test Performance

Table 7 shows results of the online tests (see Section 2.3) in percentages along with the number of participants. The number of participants in the tests decreases, mostly due to students who quit the course. (We recall that the completion of an early compulsory course is generally not required for attending other courses of the program in German universities. Hence, students often perceive the costs of postponing completion an introductory course as being low. Needless to say, we do not think that this strategy is useful.) The last two columns are based on the overall number of students who participated in at least one test and the number of students who participated in at least one exam. Again, we focus on these groups because it is not clear how seriously quitting students worked for the course. For instance, there are 40 students who took at least one online test but never submitted any exercise nor registered for any exam. We relate (with standard OLS) the percentage success in the online tests to the final score in the exercises (see Section 3.4). Figure 11 shows the plot of the data with the corresponding regression line. Unambiguously, doing better in the exercises has a significant positive relationship with success.
Table 7. Overview of online tests.

|       | 1st T | 2nd T | 3rd T | 4th T | 5th T | Add T | ∑   | Exam |
|-------|-------|-------|-------|-------|-------|-------|------|------|
| Mean  | 31.65 | 46.04 | 29.54 | 46.80 | 28.21 | 29.72 | 24.58| 31.8 |
| SD    | 17.78 | 23.08 | 21.53 | 24.44 | 19.22 | 23.12 | 19.94| 20.59|
| Max   | 100.00| 100.00| 100.00| 95.00 | 80.00 | 100.00| 86.54| 86.54|
| Upper | 43.75 | 65.00 | 44.40 | 62.50 | 42.50 | 43.75 | 39.71| 47.66|
| Median| 31.25 | 42.50 | 25.60 | 50.00 | 25.00 | 25.00 | 19.07| 29.76|
| Lower | 18.75 | 30.00 | 14.00 | 33.25 | 12.50 | 12.50 | 7.43 | 14.3 |
| Min   | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.00 | 0.00 |
| Participants | 731 | 600 | 472 | 382 | 330 | 98 | 783 | 493 |

NOTE: Descriptive statistics of the result in the online tests. All values (except for the number of participants) are in percentage points of the maximum number of attainable points. The column “∑” gives the statistics of all test participants. In the “Exam” column, we consider only the students who took at least one exam.

Figure 11. Regression online test points on score. Result of the regression of the online tests (in percentage points) versus the final score. The different grades are highlighted with different colors. The colored crosses represent the bivariate means of the score and the online tests’ points with respect to the grades. The dashed lines indicate the marginal mean. The black line is the OLS regression line Online test points $= 12.6 + 3.79 \cdot$ Score, with p-values $8.43 \cdot 10^{-13}$ (Intercept) and $< 2 \cdot 10^{-16}$ (Score).

in the online tests at any conventional level. We estimate that an additional score of 100 points (i.e., one additional correct exercise) increases the online test result by 0.379 percentage points. A student with zero score is expected to obtain (the estimated intercept) 12.6% of the maximum number of bonus points (i.e., min(0.126 \cdot 12,10) \approx 1.51 bonus points). A student with full score is expected to obtain 69.83 percentage points in the online tests (i.e., min(0.6983 \cdot 12,10) \approx 8.38 bonus points). The $R^2$ is 0.2469.

We color the points according to students’ final grades and observe that most green circle crosses (“100”) and blue diamonds (“200”) are in the upper-right corner of the plot: good and very good students already did well in the online tests and the exercises, while the red circles (“500”) center near the origin, indicating low learning activity. We again distinguish different patterns for moderate and failing students. The former are more dedicated in the exercises shortly before the exams (see Section 3.4), whereas the online test points are equally low for both groups.

We may, therefore, conclude that a higher score in the exercises predicts success in the online tests and both jointly a better grade in the exam.

3.6. Prediction

In this subsection, we regress the final grades on the variables from the two previous subsections. We first regress the final grade on the number of submissions using an ordered logistic regression. Since the endogenous variable is an ordered factor variable, we follow the method described in Section 15.9 of Cameron and Trivedi (2005) to fit the probabilities $P(y_i = y)$, where $y_i$ is the grade of student $i$ and $y_i \in \{100, \ldots, 500\}$. Figure 12 (lhs) plots the predicted probabilities to obtain a
Figure 12. Logit regression grade on submissions. The two figures show the predicted probabilities to obtain each grade (predicted with an ordinal logistic regression) against the number of the submissions (lhs) and against the number of correct submissions (rhs). For example, the red long-dashed curve on the lhs is the estimated probability to fail the exam. The probability is about 55% if a student has no submissions at all and it decreases monotonically in the number of submissions. On the other hand, the probability for the best grade is increasing and the modes of the three other curves are ordered as the grades. Due to a high correlation between the regressors of the two models ($\rho = 0.9106$), the two plots look very similar. Both predictors are strongly significant ($p < 10^{-13}$, Wald test).

Figure 13. Logit regression grade on score. The fitted probability for each grade regressed on the score with an ordinal logistic regression. The predicted probability to fail is monotonically decreasing, while the probability for the best grade is increasing. The modes of the three other curves are ordered as the grades. Overall, it is worth it both to invest time in many submissions and attain more points in these. The model is strongly significant ($p < 10^{-16}$, Wald test).

We next regress the grades on the final score of a student to predict the probabilities for each grade. Figure 13 slightly differs from the previous ones, but the conclusion is qualitatively similar: a higher score forecasts a better grade. Again, the predicted probability to fail is monotonically decreasing, while the probability for the best grade is increasing. The modes of the three other curves decrease in the grades. Overall, it is worth both to invest time in many submissions and to attain more points in these. In future terms we shall perform out-of-sample assessments of these relationships. Figure 21 in Appendix A shows qualitatively similar the estimated probabilities from certain grade against the regressor “number of submissions.” For example, the red long-dashed curve is the probability to fail the exam. The probability is about 55% for a student without any submissions and it decreases monotonically in the number of submissions. As expected, the probability for the best grade increases and the modes of the three other curves decrease in the grades. Thus, completing more exercises appears to be effective. Due to the high correlation between the total number of submissions and the number of correct submissions, it is no surprise that the plot is very similar for the regression of the grade on the number of correct submissions (Figure 12 (rhs)).
the regression of the final grades on the results of the online tests.

We also analyzed the JACK data of users who quit the course. It turns out that students who were not going to attend the exams can be identified with high accuracy based on their homework efforts after 8 weeks (end of May), while the exams take place after the 15-week course (no earlier than mid-July). We plan to use this information in the future as an early warning for lazy students. A detailed discussion is available upon request.

3.7. The Role of Previous Unsuccessful Attempts

This subsection briefly considers a rough proxy for mathematical ability (or its absence). To do that, we consider the number of unsuccessful attempts in previous exams. For this purpose, we divide students into four groups: students who took their first exam in mathematical statistics in the summer term 2014, students who failed the course once in a previous term, students who failed twice before and students who failed more than twice. Boxplots of the total number of submissions versus the number of exams in previous terms (Figure 14 (lhs)) show that the first three groups (i.e., the students with zero to two fails) do not differ substantially, while students in the last group work perceptibly less. The results are qualitatively similar for the number of correct submissions versus the number of previous unsuccessful attempts (Figure 14 (rhs)). Students with a high number of previous unsuccessful attempts do not only work less but also have fewer correct submissions. This may be evidence that students with no success in previous terms will not become successful in further exams. (We included the number of unsuccessful attempts as an additional regressor in the logit regressions of Section 3.6 as a rough proxy for intelligence. There are however only 16 students in the group of more than two trials. Therefore, we cannot expect reliable results for checking the hypothesis of worse grades for students with many failures than for others with the same number of submissions.) Another explanation for this behavior may be that these students feel that they need less practice after having taken the course multiple times.

3.8. The Student View

Students generally react very positively to JACK, as evidenced by students’ evaluations of the course. All in all, students grade JACK with 4.1 in a 5-stars-rating. However, we do not have numeric feedback on specific features of JACK such as homework, online tests, and electronic exams as the standard feedback form at our university does not allow for course-specific questions. In the future, we plan to collect feedback before and after the exam dates to study possible differences.

We may however report on positive (Figure 15) and negative (Figure 16) students’ opinions from the free text fields of the feedback forms. Generally, feedback on JACK was positive, with more than 30% of all free text feedback being praise of JACK. Some students argued that the hints and solutions in the exercises are too short and sometimes not helpful. Since then, we invested a lot of time to improve these. Next, many students perceive the online tests as being too difficult. We, however, believe that tests should be challenging because the online tests are open book and should not lead students to falsely expect an easy exam.

4. Related Work

Crompton (2010), Lenz (2010), and Nguyen and Kulm (2005) all found computer supported practice exercises to have a positive effect on final grades in mathematics courses. As we could not perform a controlled experiment with a group not using a CAA system, we cannot add detailed results here (see Section 3.3, though). Nevertheless, our findings support the hypothesis that existing results for school-level mathematics also apply to quantitative training at universities.

Darrah, Fuller, and Miller (2010) argued that CAA systems that are able to give partial credit are rare. It is suggested to use bonus points from extra quizzes to compensate for points lost due to missing partial credit. Path-based exercises and anticipated mistakes as used in JACK to some extent allow for partial credit (see Section 2). Thus, the effectiveness of our approach may be higher than that of other systems.
Jordan, Jordan, and Jordan (2011) studied the role of variable content with respect to fairness. The concerns raised there also apply to our setting. We, however, do not see any evidence that random parameters led to unfair situations in our exercises, tests, or exams.

Tyroller (2005) and Krause, Stark, and Mandl (2009) analyzed e-learning in statistics for the environment Koralle. While these studies can conduct carefully designed experiments, they cannot relate their findings to actual classroom achievements in terms of final grades, as we could in our analysis. Indeed, to the extent that JACK is useful to students, legal concerns prevent us from conducting experiments in which only a part of our students is allowed to use the system.

5. Conclusions and Future Work

The goal of this article was to discuss how the computer-assisted assessment software JACK can support teaching in introductory statistics courses. Our evaluation reveals a clear positive connection between learning effort, learning success, and final grades. The exam grades and points compare good to other terms. Furthermore, students generally accept the e-learning system, with more students participating at the electronic than at the handwritten exams. Moreover students report to invest more time using the CAA than in other terms without CAA. We conclude that the effort in developing JACK was justified, as we observe the beneficial results we had anticipated.

Future research will investigate which empirical strategy relating study effort and success to final grades provides the best out-of-sample predictions when comparing data from upcoming terms with the predictions.

Acknowledgments

The authors thank all colleagues who contributed to the course “Induktive Statistik” in the summer term 2014, especially Robert Czudaj, Natalie Reckmann, Stephan Prell, and Martin Schmelzer. The authors thank Martin Arnold, Alexander Blasberg, Thomas Fuchs, Alexander Gerber, and Kim J. Hermann for excellent research assistance. The authors gratefully acknowledge the valuable advice by the editor, Soma Roy, an associate editor, and two anonymous reviewers who helped to substantially improve this article. Part of the work on this project was funded by the German Federal Ministry of Education and Research under grant numbers 01PL11075 and 01 JA 1610.

References

Assenmacher, W. (2013), Induktive Statistik (Springer-Lehrbuch, 2nd ed.), Berlin, Heidelberg: Springer.
Bacon, D. (2011), “Effective Feedback with Adaptive Questions,” in Proceedings of the 14th CAA Conference.

Cameron, A. C., and Trivedi, P. K. (2005), Microeconometrics: Methods and Applications, Cambridge: Cambridge University Press.

Chalmers, D., and McAusland, W. D. M. (2002), “Computer-Assisted Assessment,” Technical Report, Glasgow Caledonian University. Available at: https://www.economicsnetwork.ac.uk/handbook/printable/caa_v5.pdf.

Conole, G., and Warburton, B. (2005), “A Review of Computer-Assisted Assessment,” Research in Learning Technology, 13(1), 17–31.

Crompton, P. R. (2010), ”Does MapleTA Produce Constructively Aligned Maths assessment?,” in Proceedings of the 13th CAA Conference.

Darrah, M., Fuller, E., and Miller, D. (2010), “A Comparative Study of Partial Credit Assessment and Computer-Based Testing for Mathematics,” Journal of Computers in Mathematics and Science Teaching, 29(4), 373–398.

Jordan, S., Jordan, H., and Jordan, R. (2011), “Same But Different, But Is It Fair? An Analysis of the Use of Variants of Interactive Computer-Marked Questions,” in Proceedings of the 14th CAA Conference.

Krause, U.-M., Stark, R., and Mandl, H. (2009), “The Effects of Cooperative Learning and Feedback on E-learning in Statistics,” Learning and Instruction, 19(2), 158–170.

Lenz, L. (2010), ”The Effect of a Web-Based Homework System on Student Outcomes in a First-Year Mathematics Course,” Journal of Computers in Mathematics and Science Teaching, 29(3), 233–246.

Milne, S., Honeychurch, S., and Barr, N. (2013), ”Module Facelift: Engaging Student Technology Teachers with Math,” in CAA 2013 International Conference.

Nguyen, D. M., and Kulm, G. (2005), ”Using Web-Based Practice to Enhance Mathematics Learning and Achievement,” Journal of Interactive Online Learning, 3(3).

R Core Team (2017), R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing.

Sangwin, C. (2013), Computer Aided Assessment of Mathematics, Oxford: Oxford University Press.

Schwinning, N., Striewe, M., Massing, T., Hanck, C., and Goedicke, M. (2017), ”Towards Digitalisation of Summative and Formative Assessments in Academic Teaching of Statistics,” in Proceedings of the 5th International Conference on Learning and Teaching in Computing and Engineering, LaTICE 2017. Available at: https://arxiv.org/abs/1811.02391.

Striewe, M. (2016), ”An Architecture for Modular Grading and Feedback Generation for Complex Exercises,” Science of Computer Programming, 129, 35–47.

Striewe, M., Balz, M., and Goedicke, M. (2009), ”A Flexible and Modular Software Architecture for Computer Aided Assessments and Automated Marking,” in Proceedings of the First International Conference on Computer Supported Education (CSEDU) (Vol. 2), INSTICC, pp. 54–61.

Striewe, M., and Goedicke, M. (2011), ”Studentische Interaktion mit Automatischen Prüfungssystemen,” in DeLFI 2011 - Die 9. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e. V. (Vol. 188, LNI), pp. 209–220.

Striewe, M., Zurmaar, B., and Goedicke, M. (2015), ”Evolution of the E-Assessment Framework J ACK,” in Gemeinsamer Tagungsband der Workshops der Tagung Software Engineering 2015, Dresden, Germany, pp. 118–120.

Tyroller, M. (2005), ”Effekte Metakognitiver Prompts beim Computerbasierten Statistiklernen,” Ph.D. dissertation, Ludwig-Maximilians-Universität, München.

Weiner, B. (1972), ”Attribution Theory, Achievement Motivation, and the Educational Process,” Review of Educational Research, 42(2), 203–215.