Predicting the difficulty of automatic item generators on exams from their difficulty on homeworks

Binglin Chen, Matthew West, Craig Zilles
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{chen386, mwest, zilles}@illinois.edu

ABSTRACT

To design good assessments, it is useful to have an estimate of the difficulty of a novel exam question before running an exam. In this paper, we study a collection of a few hundred automatic item generators (short computer programs that generate a variety of unique item instances) and show that their exam difficulty can be roughly predicted from student performance on the same generator during pre-exam practice. Specifically, we show that the rate that students correctly respond to a generator on an exam is on average within 5% of the correct rate for those students on their last practice attempt. This study is conducted with data from introductory undergraduate Computer Science and Mechanical Engineering courses.

INTRODUCTION

A significant component of learning most STEM disciplines is procedural knowledge, where specific rules, skills, actions, and sequences of actions are performed to achieve a particular goal. This procedural knowledge is typically assessed by providing the student with a description of a situation and asking the student, explicitly or implicitly, to apply a learned procedure.

Many assessment items of this type have two useful characteristics. First, their answers can be objectively scored, permitting the items to be machine scored even if rich user input (e.g., numeric input or code) is collected from the student. Second, there are many different situations that can be used to test a given procedure that are of similar difficulty. Together these characteristics enable the straightforward development of automatic item generators [3], which are short computer programs that can generate a range of distinct item instances with similar difficulty. In most STEM contexts, these generators will vary the numbers in an item and, perhaps, alter the configuration of the item slightly.

Item generators are extremely useful in education because they significantly reduce the recurring cost of generating homework and exam questions. In fact, with a large enough body of item generators, one can use the same generators for homework, practice exams, and exams, semester after semester, as long as students won't know exactly which generators they'll see on the exam. In this paper, we are particularly concerned with asynchronous computerized exams, where previous work has shown that it is important to randomize which generators each student is given to prevent collaborative cheating [1]. To do this fairly, it is important to create pools of item generators that have roughly the same difficulty.

It can be difficult, however, to predict the difficulty of a generator before its administration on an exam. In high-stakes testing (e.g., ACT, SAT), it is commonplace to calibrate items by including them in one or more rounds of exams for no credit before including them in score computations. This paper explores a cheaper, but less accurate means of calibration that facilitates more rapid adoption of novel content in a higher education context. Specifically, we show that by administering the same item generators as part of homework and/or practice exams during the same semester that they are offered on the exam, we can, on average, estimate an item generator's exam correctness rate to within 5% (averaged over all generators in a course semester).

SOURCE OF THE DATA

The data was collected at a large R1 university in the U.S. from Spring 2015 to Spring 2018. The two courses studied are drawn from introductory sequences in Computer Science and Mechanical Engineering. These courses assigned significant a portion of the homework/practice material via the PrairieLearn system [7] and managed a significant portion of the exams via the PrairieLearn system in the Computer-Based Testing Facility (CBTF) [9].

PrairieLearn

PrairieLearn is an online problem posing system that permits the specification of ***automatic item generators***, each of which is capable of generating a range of randomly parameterized item instances [7]. It is a common practice that the set of parameterized item instances generated by a item generator assess the same set of knowledge/skills. Upon students' submission to an item instance, PrairieLearn will automatically grade the submitted answer and display the result along with optional feedback.

PrairieLearn has been extensively used by courses as a platform to distribute homework and practice material. In the

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

case of homework, students are typically asked to answer item instances generated by a set of item generators to earn credits. Typically a new item instance will be generated each time upon answer submission regardless of the correctness of the submitted answer. The homework itself can be seen as a source of practice material since it is usually made accessible even after the homework deadline. In addition, some courses offer practice exams that mimic the exam format. In such case, PrairieLearn will first randomly select a set of item generators from an instructor-defined pool of item generators and then generate item instances accordingly. These item instances will no longer be replaced upon submission and the number of solution attempts is limited. Practice exams are also typically made accessible throughout the course and students can attempt as many exam instances for practice as they want.

Computer-Based Testing Facility

While PrairieLearn serves well as a homework and practice platform, it can also be used for exams. The mechanism of exam instance generation is essentially the same as in the practice exam case. The only differences are that students only have one exam instance to attempt and they have to take the exam in the Computer-Based Testing Facility (CBTF) [9]. The CBTF is a proctored computer lab with 89 seats for students. Each of the computers is outfitted with a privacy screen that prevents test takers from reading the screens of neighboring computers and the networking and file system are strictly controlled. Students are not permitted to take written notes, photos or other records into or out of the exam room. During the period studied, the facility was open and proctored 10–12 hours a day, 7 days a week. At students' self-scheduled exam time, students have their identity checked by a proctor and are randomly assigned to a computer (to deter coordinated cheating).

Exams within the CBTF are typically administered as follows [8]: Courses assign a 3–5 day period for the students to take an exam depending on the number of students; longer exam periods are used during finals week. Students are free to reserve any time slot during the exam period, provided that there are seats available at that time.

THE DATA

The two courses the data covers are introductory undergraduate engineering subjects in Computer Science and Mechanical Engineering. For each course and semester, we have obtained the information of all the homework/practice exams and exams in the form of **(course id, assessment id, assessment type, accessible period)**. The course id is a unique identifier to differentiate between combinations of courses and semesters. The assessment id is a unique identifier to differentiate between different homework/practice exams and exams. The assessment type indicates whether the assessment is homework, practice exam, or exam. The accessible period indicates the period of time when students can access the assessment[1].

[1]Homework and practice exams are usually made accessible throughout the semester once released. Exams are usually made accessible only during the exam period in the CBTF.

Course and semester	Number of students	Number of questions	Number of submissions	NPAR p-values H_{01}	NPAR p-values H_{02}
Class A1	484	77	50,966	2.97×10^{-10}	3.00×10^{-2}
Class A2	239	43	12,678	4.10×10^{-2}	5.05×10^{-5}
Class A3	456	53	29,816	2.55×10^{-2}	9.34×10^{-7}
Class A4	464	50	8,970	2.17×10^{-4}	5.05×10^{-4}
Class A5	441	88	31,174	1.34×10^{-2}	6.20×10^{-10}
Class A6	403	127	35,104	4.49×10^{-2}	3.59×10^{-19}
Class B1	341	84	9,708	1.44×10^{-9}	3.32×10^{-8}
Class B2	371	80	11,264	1.58×10^{-5}	8.58×10^{-12}
Class B3	202	206	21,662	5.80×10^{-3}	7.52×10^{-28}
Class B4	383	257	44,826	1.80×10^{-1}	2.39×10^{-38}

Table 1. Information for each course and semester. Each course is indicated by a letter (A–B) and a number for the semester. The same number means the same semester for both courses and smaller numbers indicate earlier semesters.

With IRB approval, we obtained the information of all the student answer submissions in the form **(assessment id, item id, student id, submission date, score)**. The assessment id is the same as defined above. The student id is a unique identifier for a student regardless of course. The item id is a unique identifier for item generators. The submission date is a timestamp of when the submission is made. The score is a boolean indicating the correctness of the submitted answer.

Given the raw data, we filtered out records from course instructors and students who did not take any exams. Information for each course and semester is shown in Table 1.

ANALYSIS

In the remainder of the paper, we define the ***correct rate*** for an item generator to be the proportion of students who correctly answered an item instance generated by the item generator. We will show that as a group, students' performance during practice is comparable to their performance during exams, by comparing the correct rate of the last attempt during practice to the correct rate of the first attempt during the exam on the same item generator by the same students.

A particular generator can appear on multiple exams (e.g., a mid-term exam and the final), and, when this happens, we consider these independently. Specifically, we treat items with the same item id but used in exams with different assessment ids as different ***questions***. Every unique **(assessment id, item id)** pair defines a unique ***question*** for the purpose of our analysis. For each question, we computed (1) the number of students who have practiced the item generator outside the exam and were given an item instance of the generator during the exam, (2) their average correct rate on the last attempt during practice before they take the exam, (3) their average correct rate on the first attempt during the exam. The third and forth columns of Table 1 show the number of questions and the number of relevant submissions (last practice attempt and first exam attempt) from each course and semester.

We plotted the first exam attempt correct rate against the last practice attempt correct rate in Figure 1. Each subplot corresponds to a course in a particular semester. Each circle in the figure corresponds to a question whose area is proportional to the number of students. The vertical and horizontal error bars of each circle correspond to the 95% confidence intervals of the first exam attempt correct rate and the last practice attempt

correct rate, respectively. We also plotted an additional circle at the bottom right of each subplot to serve as a reference for the number of students. As the figure shows, most questions are spread around the diagonal and there is a trend where more questions lie further below the diagonal in later semesters as compared to earlier semesters. The diagonal phenomenon seem to suggest that there is no significant systematic difference between the last practice attempt correct rate and the first exam attempt correct rate on the same item generator.

To understand how robust the phenomenon is, we conducted a two one-sided test of equivalence for paired samples. Unlike the normal paired sample test where the null hypothesis states that the mean of the differences between two paired samples is zero, a two one-sided test of equivalence for paired samples assumes that the mean of the differences is outside some equivalence interval $(-\delta, \delta)$ which can be asymmetric. By rejecting the null hypotheses $H_{01} : \mu_1 - \mu_2 \geq \delta$ and $H_{02} : \mu_1 - \mu_2 \leq -\delta$, one can infer that the mean of the differences is within the equivalence interval and, therefore, conclude that there is no significant systematic difference between the two variables. The specific method we used in our analysis is the nonparametric two one-sided test of equivalence for paired samples (NPAR) [6] due to non-normality observed in our data. We reported p-values of the NPAR test where μ_1, μ_2 correspond to overall last practice attempt correct rate and first exam attempt correct rate, respectively, for $\delta = 5\%$ in Table 1. As the table shows, the null hypotheses are rejected at the $p < 0.05$ level for all the cases except Class B4 where H_{01} is not rejected, meaning that overall the last practice attempt correct rate might be 5 percentage points or more better than first exam attempt correct rate.

DISCUSSION AND CONCLUSION

We studied student performance on practice automatic item generators (from homeworks or practice exams) compared to the performance of the same student on the same item generator on a subsequent exam. In each case, the student was answering a particular item instance that was randomly generated by the item generator by varying random parameters. The courses studied were introductory Computer Science and Mechanical Engineering courses at an R1 university in the U.S. and a total of 256,168 student submission records were analyzed from 3,784 student enrollments.

We found that, averaged over all item generators in a course for a semester, the correct rate of student answers on practice item generators was within five percentage points of their correct rate on the same item generators on an exam. This was true ($p < 0.05$) for 9 out of 10 classes studied, and for the one remaining class, students on average scored lower on item generators during exams than during practice.

We find the strength of this observation somewhat surprising because there are various factors that could affect exam performance on the same items relative to the last practice. One obvious negative factor is forgetting; as Ebbinghaus demonstrated in the forgetting curve [2], we forget materials surprisingly fast if we do not restudy. Besides forgetting, test anxiety is another negative factor. According to Hembree's meta-analysis of 562 studies, test anxiety negatively affects test performance of high-test-anxious students relative to their non-test performance [4].

On the other hand, a positive factor is that students are likely to treat exams more seriously than practice, thus would perform better in the exam situation. Another positive factor is that Kornell et al. found an unsuccessful retrieval attempt with feedback can enhance subsequent retrieval [5], which suggests that a student has an increased probability of answering an item correctly on exams even if the last attempt during practice is incorrect. Last but not the least, a final positive factor is that students can practice other questions and study the course material between their last practice and exams, and thus improve their skills and possibly perform better on exams. Despite all the factors that can affect exam performance positively or negatively after the last practice attempt, our results suggest that their effects mostly balance out in the courses and student populations studied here.

REFERENCES

1. Binglin Chen, Matthew West, and Craig Zilles. 2018. How much randomization is needed to deter collaborative cheating on asynchronous exams?. In *Learning at Scale*.
2. Hermann Ebbinghaus. 1913. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms.
3. M.J. Gierl and T.M. Haladyna. 2013. *Automatic Item Generation: Theory and practice*. Routledge.
4. Ray Hembree. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of educational research* 58, 1 (1988), 47–77.
5. Nate Kornell, Matthew Jensen Hays, and Robert A Bjork. 2009. Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 4 (2009), 989.
6. Constance A Mara and Robert A Cribbie. 2012. Paired-samples tests of equivalence. *Communications in Statistics-Simulation and Computation* 41, 10 (2012), 1928–1943.
7. Matthew West, Geoffrey L. Herman, and Craig Zilles. 2015. PrairieLearn: Mastery-based Online Problem Solving with Adaptive Scoring and Recommendations Driven by Machine Learning. In *2015 ASEE Annual Conference & Exposition*. ASEE Conferences, Seattle, Washington. https://peer.asee.org/24575.
8. Craig Zilles, Matthew West, and David Mussulman. 2016. Student Behavior in Selecting an Exam Time in a Computer-Based Testing Facility. In *2016 ASEE Annual Conference & Exposition*. ASEE Conferences, New Orleans, Louisiana. https://peer.asee.org/25896.
9. Craig Zilles, Matthew West, David Mussulman, and Timothy Bretl. 2018. Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility. In *2018 IEEE Frontiers in Education (FIE) Conference*. San Jose, California.

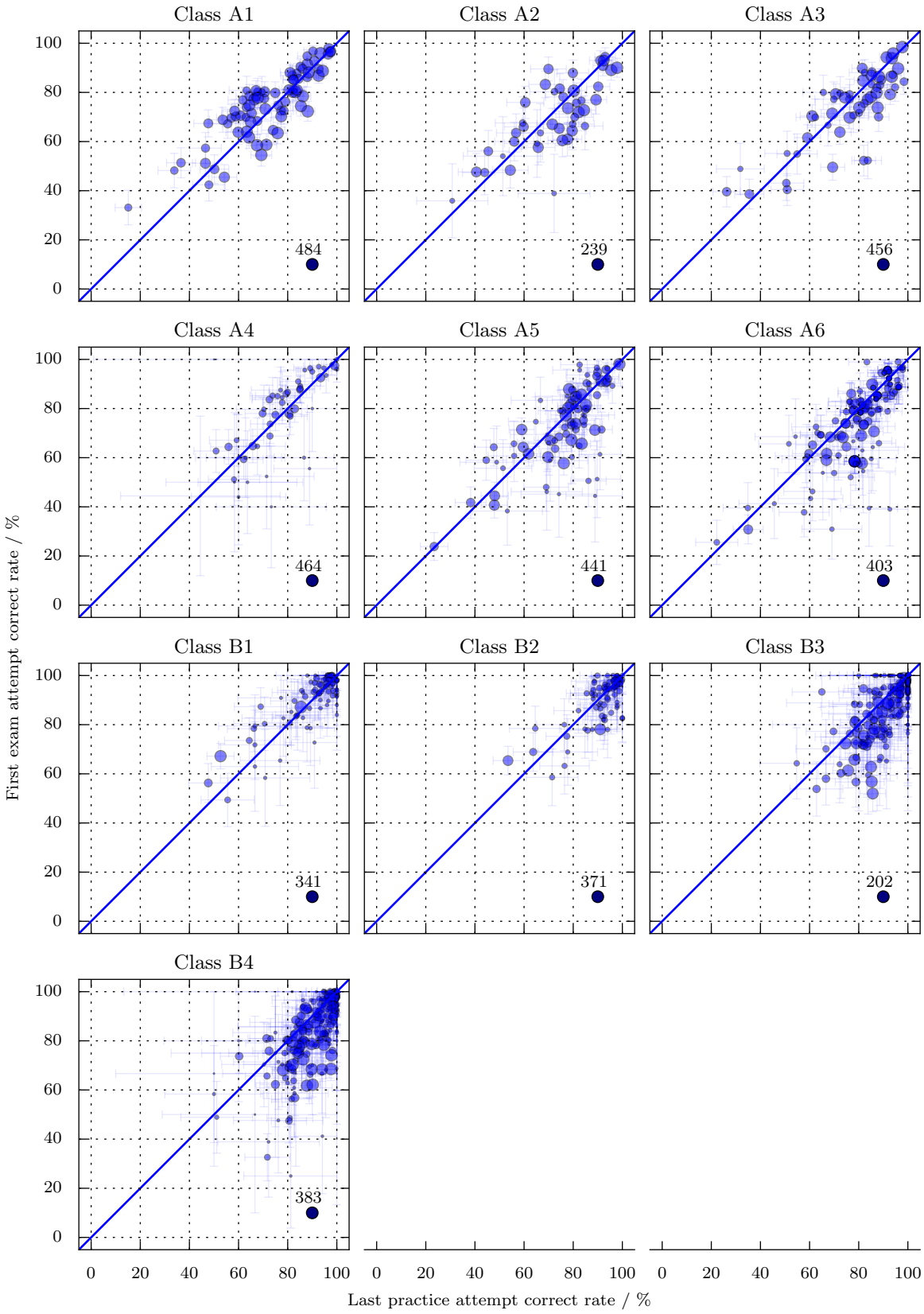

Figure 1. The first exam attempt correct rate versus last practice attempt correct rate for questions in each course and semester. Each circle represents one question, namely one (assessment id, item id) pair. The area of the circles represents the number of students and the dark circle in the bottom right corresponds to the total number of students in the course for the specific semester. The vertical and horizontal error bars of each circle correspond to the 95% confidence intervals of the first exam attempt correct rate and the last practice attempt correct rate, respectively.