122nd ASEE Annual Conference & Exposition
June 14 - 17, 2015
Seattle, WA
Making Value for Society
Paper ID #13102

Randomized Exams for Large STEM Courses Spread via Communities of Practice

Prof. Matthew West, University of Illinois, Urbana-Champaign

Matthew West is an Associate Professor in the Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign. Prior to joining Illinois he was on the faculties of the Department of Aeronautics and Astronautics at Stanford University and the Department of Mathematics at the University of California, Davis. Prof. West holds a Ph.D. in Control and Dynamical Systems from the California Institute of Technology and a B.Sc. in Pure and Applied Mathematics from the University of Western Australia. His research is in the field of scientific computing and numerical analysis, where he works on computational algorithms for simulating complex stochastic systems such as atmospheric aerosols and feedback control. Prof. West is the recipient of the NSF CAREER award and is a University of Illinois Distinguished Teacher-Scholar and College of Engineering Education Innovation Fellow.

Mariana Silva , University of Illinois at Urbana-Champaign
Dr. Geoffrey L Herman, University of Illinois, Urbana-Champaign

Dr. Geoffrey L. Herman is a visiting assistant professor with the Illinois Foundry for Innovation in Engineering Education at the University of Illinois at Urbana-Champaign and a research assistant professor with the Department of Curriculum & Instruction. He earned his Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign as a Mavis Future Faculty Fellow and conducted postdoctoral research with Ruth Streveler in the School of Engineering Education at Purdue University. His research interests include creating systems for sustainable improvement in engineering education, promoting intrinsic motivation in the classroom, conceptual change and development in engineering students, and change in faculty beliefs about teaching and learning. He serves as the webmaster for the ASEE Educational Research and Methods Division.

Randomized Exams for Large STEM Courses Spread via Communities of Practice

1. Introduction

This paper describes `randexam`: a new computerized system for the generation, grading, and analysis of randomized multiple-choice paper exams using Scantrons for student responses. In the past three years since this technology was developed at the University of Illinois at Urbana-Champaign, it spread from the original course to a total of 12 courses. This spread is documented and analyzed from two different perspectives: diffusion of innovations [14] (change theory perspective) and communities of practice [8;16;17] (education theory perspective). The key question that this paper addresses is: why did the `randexam` system spread so rapidly at University of Illinois, and what lessons can be learned from this to facilitate the spread of other pedagogical innovations?

Multiple-choice questions have been widely used for student's assessment due to their ease of grading and processing. Research has shown that multiple-choice exams can be a reliable tool for assessment when the problems are carefully constructed [1;12]. For example, Scott et al. [15] investigated the validity and reliability of scores from multiple-choice exams constructed and administered in large introductory physics courses. In one of their studies, they observed that the multiple-choice exams yielded statistically equivalent assessment of student's understanding when compared to free response and oral exams. Unfortunately, the creation of reliable exams requires a great deal of research work and assessment iterations, and inevitably some poor questions make their way into exams. In this paper, we describe how we can detect poor questions and adequately adjust the scores to remove grade discrepancy among students.

Another major concern about the use of multiple-choice questions is the inability to give students partial credit. To address this issue, some authors have investigated the use of partial credit in multiple-choice exams [10;19]. Lin and Singh [10] proposed the creation of multiple-choice questions in which the distractors mimic results obtained when typical mistakes are made during intermediate steps of the problem. These distractors received partial credit points, according to the rubric of the free response question. One group of students received the question in multiple-choice format, while another group received the same question in free response format. The results show that both formats yield statistically equivalent performance.

Among some of the drawbacks of multiple-choice questions is the propensity for cheating, since it is relatively easy to see and copy answers from neighbors. Cheating becomes more difficult through the use of randomized multiple-choice exams [2;5;11]. Bresnock et al. [2] investigates the influence of questions and answer order within the exam in student's performance. Their findings indicate that random arrangement of questions has little influence in student's performance. They also noticed the same trend when the answer order is shuffled randomly. However, asymmetric answer-order distribution (uneven distribution of "A" answers, "B" answers, etc.) might result in substantial discrimination among the students and compromise fairness (for example, they observed that students that receive a large number of "A" responses may have better performance than students that receive a large number of "D" responses, since they were not introduced to the confusion of the distractors).

2. The `randexam` system

The `randexam` system is implemented in an open-source Python script[13], which processes a LaTeX file containing a library of questions, each with one or more variants. Per-student randomized exams are generated by selecting random variants of each question, and randomly permuting the order of questions and the order of answers to each question, with restrictions as specified by the user (for example, only the first set of problems are permuted). The randomized exams are generated as LaTeX files, which are processed into PDF format for printing. Students take the exam using Scantron bubble forms, which are then scanned and the raw data is processed by the `randexam` script to undo the randomization, grade the exams, and produce summary statistics for analysis. This process is illustrated in Figure 1.

Each exam is identified by a unique *exam key*, such as ACAEDD, which is printed on the front of the exam paper. Students copy this code onto the Scantron forms by using each letter in the key as the answer to one question on the Scantron form (typically the final questions). The key generation algorithm (see Appendix A) produces keys which differ in at least 3 letters from all other exam keys (that is, keys have Hamming distance of at least 3 from each other). This means the set of exam keys forms an error-correcting code, which is able to detect and correct single-letter errors and to detect, but not correct, double-letter errors. In the use of `randexam` over several years and thousands of students, we have never observed a case where a student incorrectly transcribed an exam key in a way that was not automatically detected by `randexam`.

The implementation of the `randexam` system has a number of appealing features from the diffusion of innovations perspective, including relative advantage (improved exam security, better statistics output), compatibility with existing systems (still have 2 or 3 midterms and a final, do it on the same schedule (evening or in-class), still have all students coming together, familiar exam format (Scantrons) for students, use existing Scantron infrastructure on campus, exams are already being written in LaTeX), low complexity (for adopters who can run Python and LaTeX), immediately observable impact (security and better output statistics), and ease of reinvention (flexible enough to adapt for use in evening exams, finals, in-class quizzes, etc). See Section 7 for further discussion of the diffusion of this technology.

3. Randomized multiple-choice compared to free-response exams

One way to address the concern that multiple-choice questions do not allow students to receive partial credit (currently we don't have this option in `randexam`) is to construct questions that are short and require the knowledge of one or two basic concepts. By removing questions with several partial steps we remove the need to give partial credit to students. The question that arises

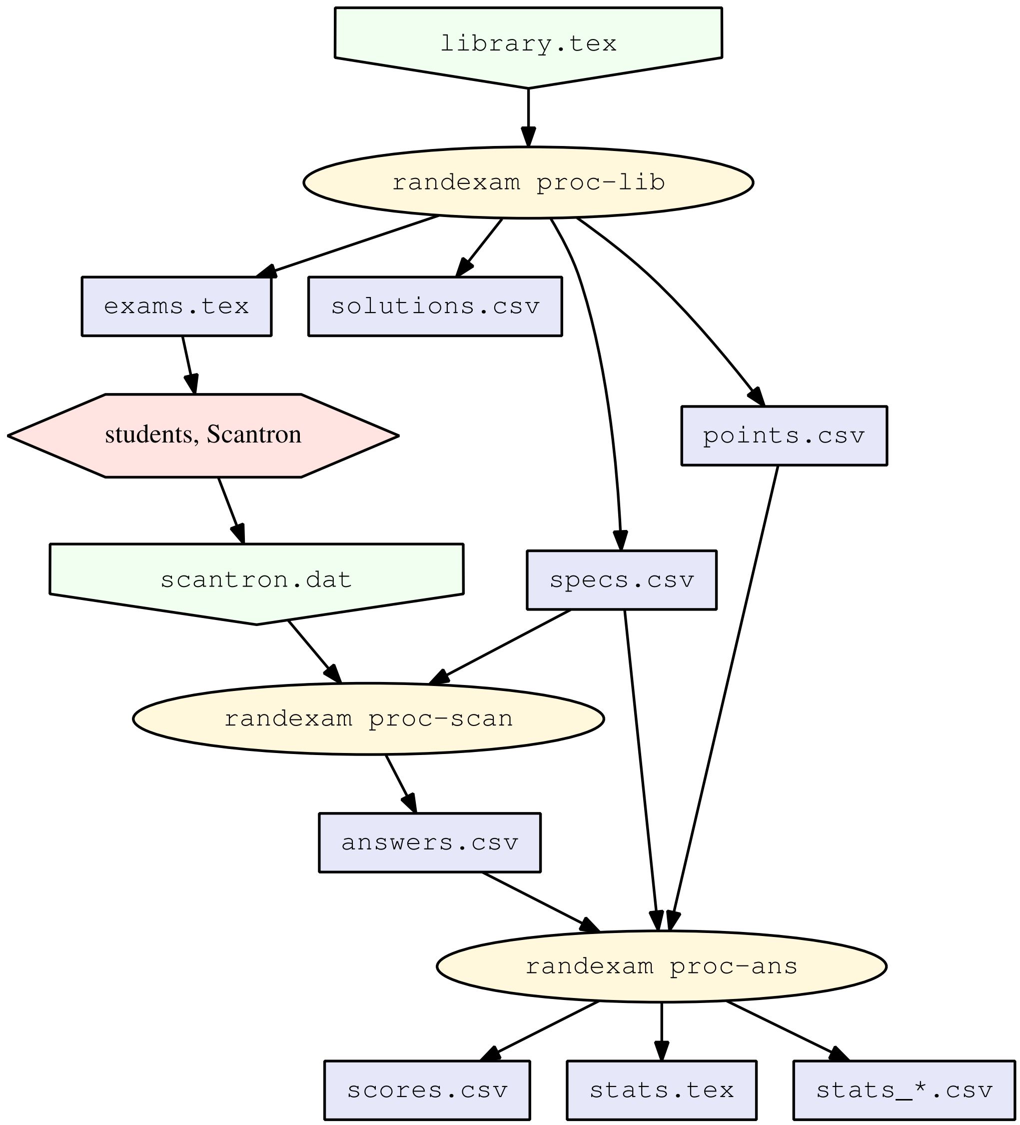


Figure 1: The pipeline for making and grading the randomized exams. The green pentagons are input files that require human effort to generate, the purple boxes are output files, the yellow ellipses are runs of the `randexam` program, and the pink hexagon is the students sitting the exam and the scanning of the Scantron forms.

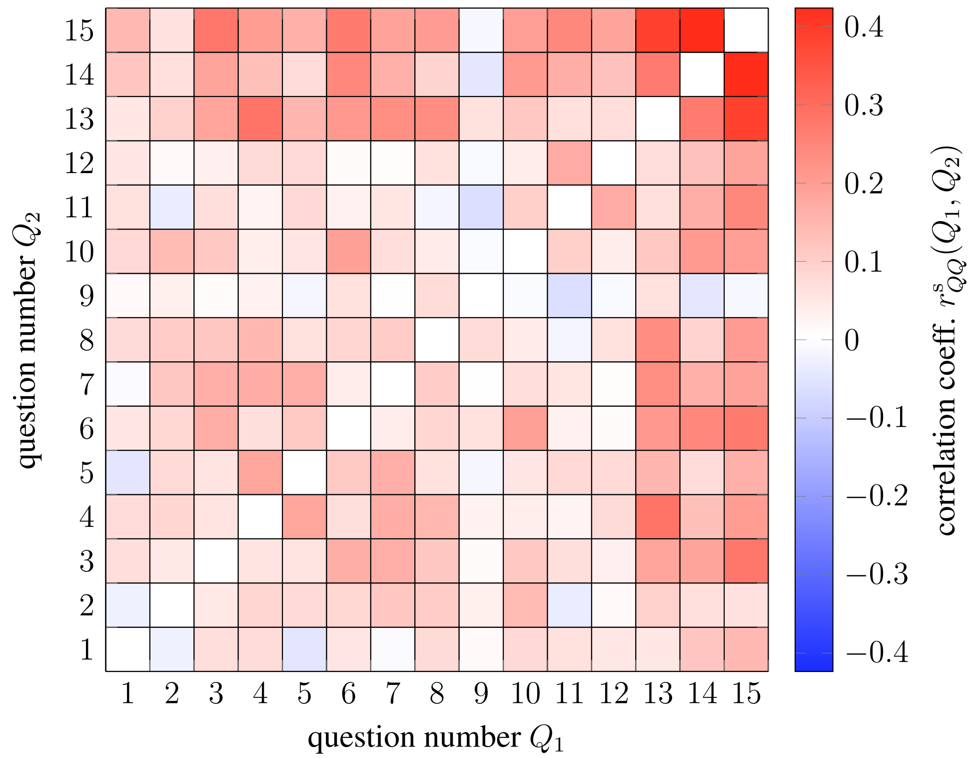


Figure 2: Correlation coefficients $r^{\mathrm{s}}_{\mathrm{QQ}}$ between the student scores on exam questions for Midterm 1. Positive and negative correlations are shown in red and blue, respectively. Self-correlations of a question with itself are always $r = 1$ and are not shown.

is: will students be prepared to address more complicated problems which combine many different concepts, such as the more traditional free-response questions?

In this section, we investigate the performance of 600 students that took the first exam (Midterm 1) in the Statics course in the Fall 2014. The exam consisted of 12 multiple-choice questions (1–12) and 3 free-response questions (13–15).

Figure 2 shows the correlation between exam questions. Free-response question 13 involves the concept of particle equilibrium in 2D. The same concept is assessed in multiple-choice question 4. As expected, questions 4 and 13 have a strong correlation as illustrated in Figure 2. Free-response question 14 asks students to determine the resultant moment of a given set of forces about a specified point in a 3D problem. On the other hand, multiple-choice question 6 asks students to determine the resultant moment of a given set of forces in a 2D problem. Again, we note a strong correlation between these two questions. The last free-response question involves several concepts in Statics, such as unit vectors (question 3), resultant forces and moments (questions 4, 6, 8, 13, 14) and support conditions (question 11). Note that the correlation between these questions corresponds to points of higher correlations in Figure 2. These results are very encouraging and indicate that students that acquire knowledge of basic concepts and therefore are able to perform well in simple multiple-choice questions are likely equipped to perform well in more complex free-response questions.

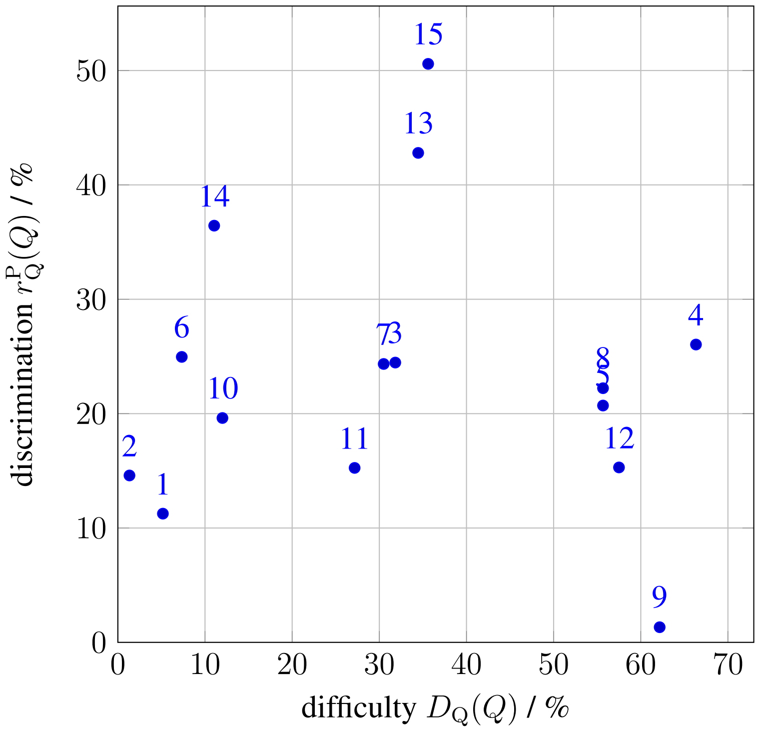


Figure 3: Plot of *discrimination* against the *difficulty* for each question. Questions should ideally be high on this plot (discriminating well), and there should be a mixture of left-to-right (difficulty) values. The difficulty $D_{\mathrm{Q}}(Q)$ is defined to be the fraction of students who get question Q incorrect, while the discrimination $r_{\mathrm{Q}}^{\mathrm{P}}(Q)$ is defined to be the correlation coefficient of scores between question Q and the total exam score.

4. Evaluating quality of multiple-choice questions

A further investigation of Figure 2 reveals very little correlation of question 9 with the rest of the exam. Indeed, as indicated in Figure 3, question 9 was one of the hardest problems of the exam with discrimination nearly zero. This was a true-false question, in which the correct answer was false, marked as answer "A" in Figure 4. Note that only 50% of the students in the top 20 percentile obtained the correct answer for question number 9, which is the same as random guessing. It is clear that question 9 was not well constructed and therefore cannot be considered a reliable measure to assess student's performance in this Statics course. In such cases it is easy to re-grade the exam without this question being included, which allows post-exam quality control to ensure a fair assessment.

Figure 5 shows the relative points awarded for each question variant in this Statics exam. To ensure fairness of these random exams, it is imperative to produce variants that yield similar relative points for each questions. Unfortunately, for this exam, question 11 had one variant that was found by students to be harder than the other one. Figure 6 illustrates both variants: variant 1 asks for the reactions at a pin connection and variant 2 asks for the reactions at a fixed connection. According to Figure 7, about 40% of students that received variant 1 marked "A" as the correct answer. In future exams, the variants should involve the same type of support conditions, but

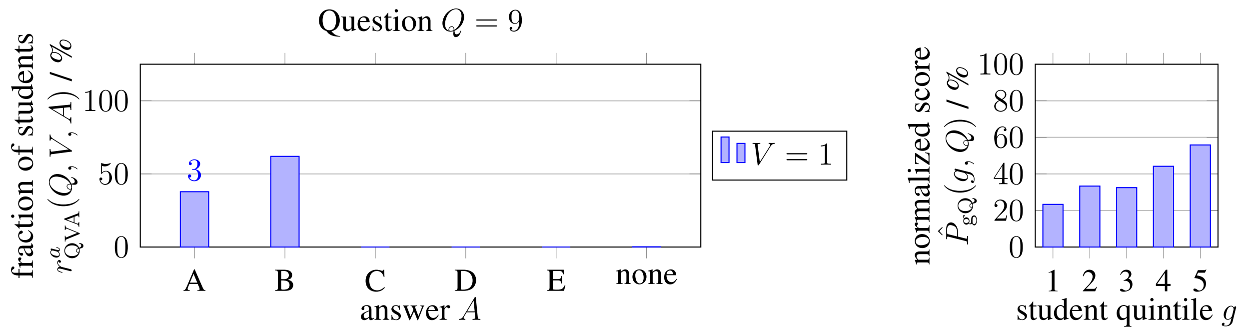


Figure 4: Detailed information for Question 9, showing which answers were given by students (left), and the average scores for each quintile of students (right). See Section 4 for discussion.

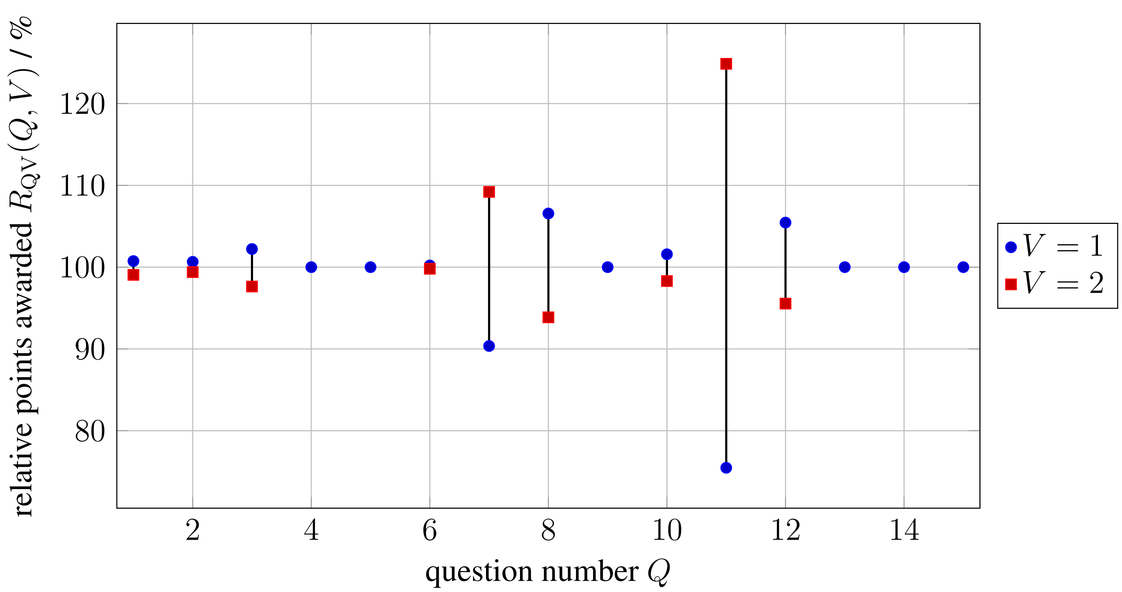


Figure 5: Relative points for the question variants. Variants with $R_{\mathrm{QV}}(Q, V)$ above 100% are easier than average (more points awarded), while values below 100% indicate a harder-than-average variant.

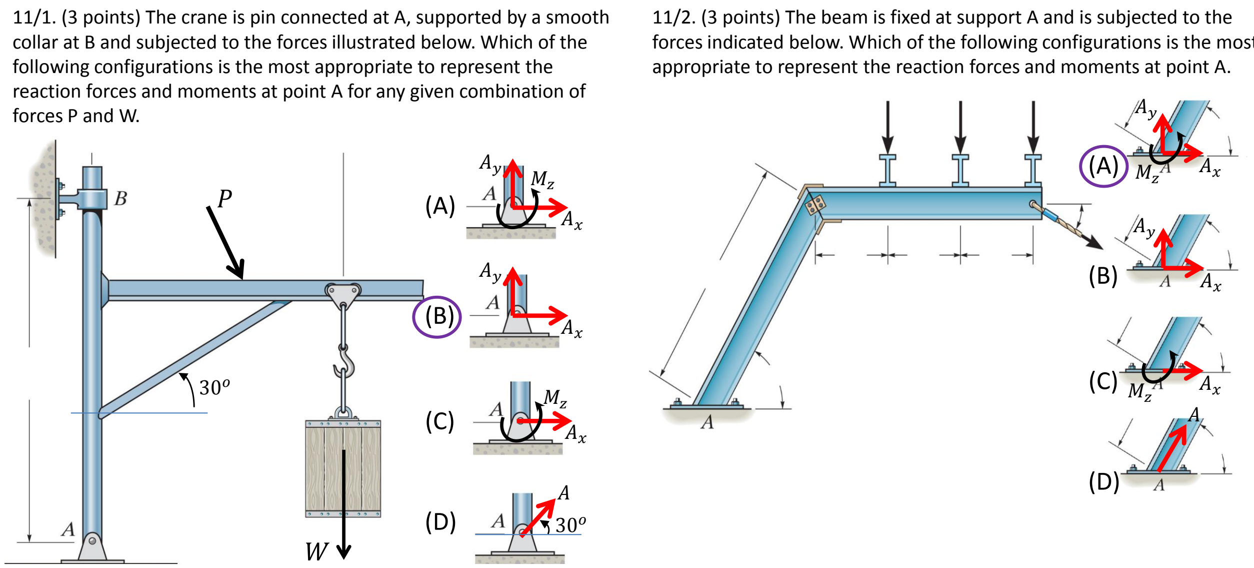


Figure 6: Variants for Question 11, Midterm 1, Fall 2014, Statics. Left: variant $V = 1$. Right: variant $V = 2$.

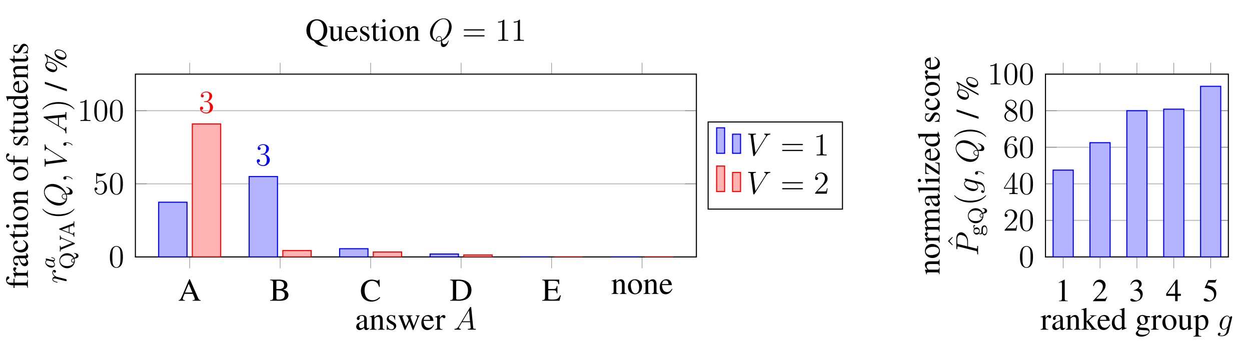


Figure 7: Detailed information for Question 11, showing which answers were given by students (left), and the average scores for each quintile of students (right). See Section 4 for discussion.

using different configurations, to improve consistency and fairness of the randomized multiple-choice exams.

5. Student perceptions survey

A survey addressing student perceptions of the randomized exam format was administered in two courses (*Calc 2 Eng* and *Statics*) in Fall 2014, with results as shown in Figure 8. This shows that students think that it is important to ensure the integrity of exams (71% important versus 10% unimportant), but that they think cheating on regular multiple-choice exams is easy (59% easy versus 12% hard). In contrast, students believe that cheating on randomized multiple-choice exams with the `randexam` system is difficult (91% hard versus 2% easy), and that the randomized format does not significantly alter the fairness of the exam (60% no change, 12% find randomized less fair, 28% find non-randomized less fair). Summarizing their thoughts, students significantly prefer randomized over non-randomized multiple choice exams (48% prefer randomized, 43% have no preference, and just 9% prefer non-randomized).

Correlations between survey items are shown in Figure 9. Notably, items 2 and 3 are positively correlated, indicating that students have a consistent view of the ease of cheating on randomized and non-randomized exams, and items 1 and 5 are negatively correlated, showing that students who think it is important to stop cheating are more in favor of using randomized exams.

6. Mapping the spread of `randexam`

The spread of `randexam` is illustrated in Figure 10.

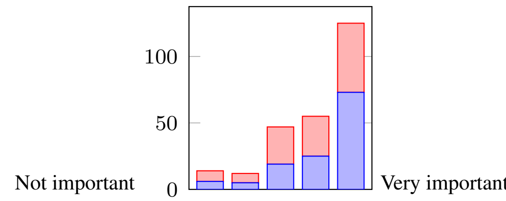
The `randexam` system was created in Fall 2012 in the course *Calc 2 Eng* by one of the authors (MW). At this time, MW was participating as the engineering instructor in a co-taught Math/Engineering section of *Calculus 2*. In the following semester MW returned to Engineering and co-taught TAM *Dynamics* course with another engineering instructor, bringing the new `randexam` system to this course. Both of the *Dynamics* instructors were participating in a CoP within TAM, which focused on the three-course introductory sequence *Statics*, *Dynamics*, and *Solids*. Through the TAM CoP, the `randexam` system was adopted into the other TAM courses.

In a separate development, one of the Math co-instructors of *Calc 2 Eng* took `randexam` to the *Stoch Proc* course that he was teaching, and a separate group of instructors adopted `randexam` in *Calc 2 non-Eng*.

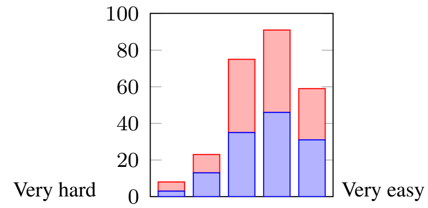
Even as the above transfers were occurring, the instructor of *MatSE Mech* was sitting in as an observer in the TAM CoP. Having seen the use of `randexam` within TAM courses, he then adopted it in the MatSE course, from where it was discussed in the MatSE CoP and used in *Therm & Mech*.

The final pathway by which `randexam` was spread involved the CS CoP. As part of the College of Engineering program that supported the departmental CoPs, MW was acting as an Education Innovation Fellow mentor to the CS CoP and attending their weekly meetings. In this context, the `randexam` system was discussed, and subsequently adopted by CS instructors in four courses.

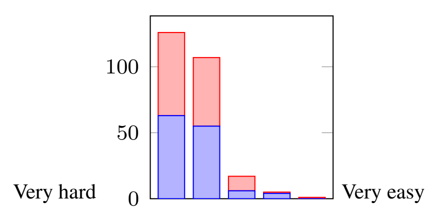
1. How important is it to you that professors and course staff stop other people from cheating on exams?



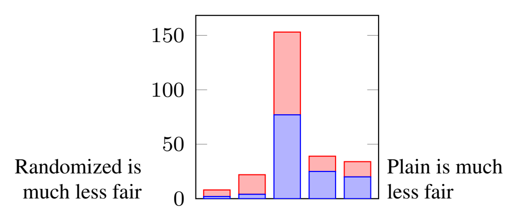
2. On regular multiple-choice exams where everyone gets the same questions in the same order, how easy do you think it is for other people to copy answers from someone sitting near them?



3. [This class] used “randomized multiple-choice exams”. This meant that everyone had a different question order, different answer order, and different selection of question variants (different numbers, etc). How easy do you think it is for other people to copy answers from someone sitting near them on this type of exam?



4. From a difficulty point of view, do you think that randomized multiple-choice exams are more or less fair than plain (non-randomized) multiple-choice exams?



5. Given that [this class] uses multiple-choice exams, would you prefer to them to be randomized or plain (non-randomized)?

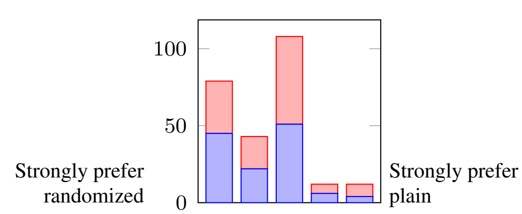


Figure 8: Student survey results from Fall 2014. Students were surveyed from two courses, with 128 responses from *Calc 2 Eng* (shown in blue) and 123 responses from *Statics* (shown in red), for a total of 251 student responses. All questions used 5-level Likert-type scales. See Section 5 for discussion.

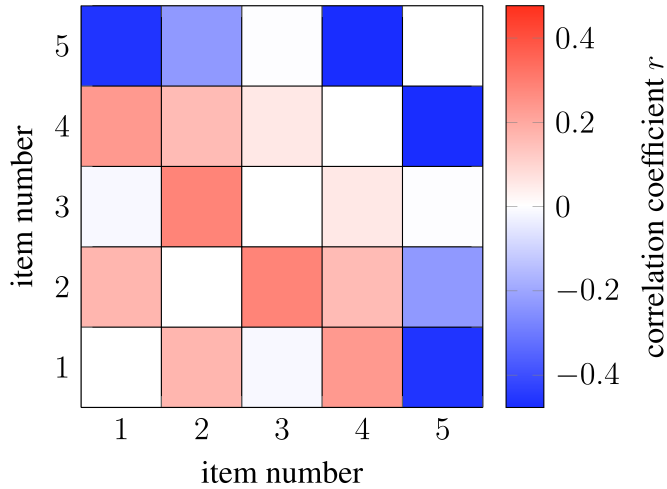


Figure 9: Correlation coefficients between the survey items shown in Figure 8. Positive and negative correlations are shown in red and blue, respectively. Self-correlations of an item with itself are not shown. See Section 5 for discussion.

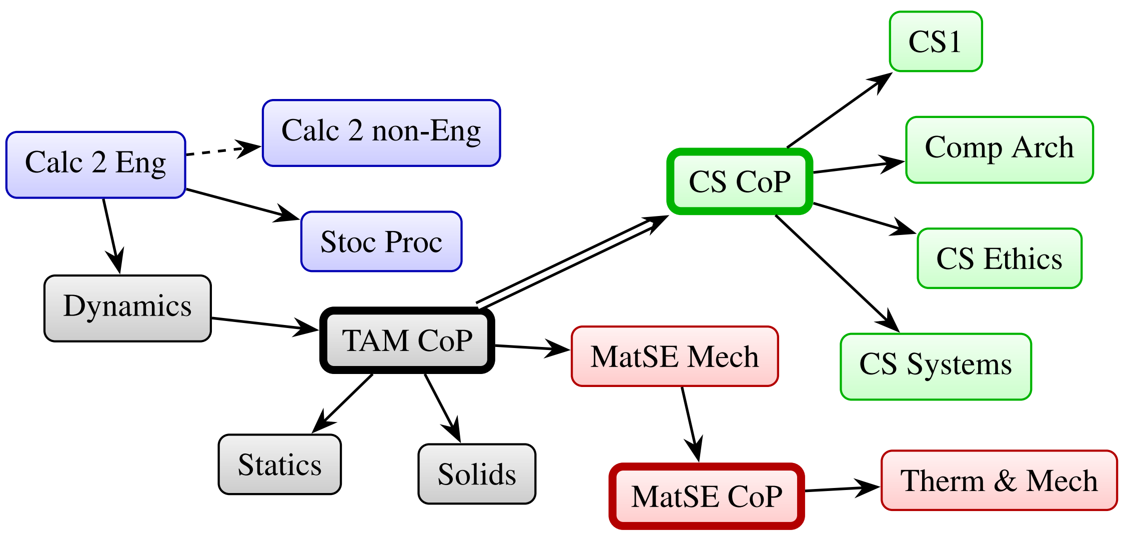


Figure 10: Spread of the `randexam` system from its creation in the *Calc 2 Eng* course in Fall 2012 to reach a total of 12 courses by Fall 2014. Thin borders denote individual courses, while thick borders indicate Communities of Practice (CoP). The arrows between boxes show the direction of spread, where solid lines indicate an embedded or co-teaching arrangement (at least one semester duration), dashed lines indicate word-of-mouth dissemination, and double-solid lines indicate an embedded Education Innovation Fellow (EIF) supported by the College of Engineering (at least one semester duration). See Section 6 for discussion.

7. A diffusion of innovations perspective on `randexam` spread

The diffusion of innovations literature[14] shows that diffusion is enhanced by key characteristics of: (1) the innovation, (2) individual adopters, and (3) the organization.

The six key characteristics[6] of the innovation itself are relative advantage, compatibility, complexity, trialability, observability, and reinvention. The `randexam` system fits well with this model, as it has five out of the six characteristics: it brings immediate relative advantage (improved exam security, better statistics output), is very compatible with the existing exam system (same exam arrangements for scheduling, usage of the same exam rooms, familiar exam format for students; see Section 2), is low complexity (for adopters with the appropriate technical background), has immediately observable impact, and is easy to reinvent (many adopters have customized the system to their particular environment). The one characteristic that `randexam` is lacking is trialability, because running a trial would involve filling out Scantron sheets by hand and having them scanned, which is too much effort for any adopter to have done (we are unaware of anyone having done this).

While classifications of individual adopters have been proposed in the literature, there is little agreement or empirical support for these categories[6]. We do not attempt to analyze features of individual adopters in this paper.

Seven characteristics of the organizational or system context can be identified[6;14] that assist diffusion: network structure, homophily, opinion leaders, harnessing of the opinion leaders' influence, champions, boundary spanners, and formal dissemination programs. In the case of `randexam`, six of these seven characteristics were present. The network structure was essential and greatly enhanced by departmental communities of practice (see Section 8). The early adopters of `randexam` were highly homophilous, having very similar professional, technical, and cultural backgrounds. Opinion leaders were explicitly enabled by the Education Innovation Fellows program within the College of Engineering and were important links both between and within departments, while champions within departments provided important early support. Boundary spanners were especially important in spreading `randexam` between departments, which are the units of organization for teaching, and were enabled by organization programs, including the Math/Engineering Calculus co-teaching project, the Education Innovation Fellows program, and the college-supported Communities of Practice. The one characteristic that was not present for `randexam` was formal dissemination programs.

A key aspect of the spread of `randexam` was the fact that almost every link in Figure 10 had a long-term faculty-member involvement on both ends of the link (at least one semester). That is, it was not the case that a faculty member heard about the innovation at a workshop or other one-time event, but rather that they participated in an extended conversation that allowed familiarity to build over at least one semester. Explicit programs of the College of Engineering were instrumental in producing these long-lasting links, with the co-taught Math/Engineering calculus section providing the initial link, the college-support Communities of Practice acting as hubs (see Section 8), and the Education Innovation Fellows mentor program connecting CoPs.

8. Communities of Practice (CoP) as key elements of `randexam` spread

The theory of Communities of Practice was developed by Lave and Wenger[8] to explain and understand learning in informal settings. In the context of this paper, a CoP is a group of faculty who regularly interact to support each other in the teaching of a set of specific courses. The creation of CoPs has been shown to effectively spread tacit knowledge[4;7], decreasing the learning curve for novices, reducing creation of redundant resources or reenactments of failures, and promoting creativity[9].

The rapid and wide-ranging diffusion of `randexam` was accelerated by Communities of Practice (CoP) within departments, as they acted as central hubs for information dissemination and support, and it is likely that without these CoPs the `randexam` system would only have been adopted within two or three courses at most. These CoPs were explicitly created and supported at the University of Illinois by an education innovation program (SIIP: Strategic Instructional Initiatives Program) within the College of Engineering[3;18].

9. Conclusions

In this paper we have presented the `randexam` system for randomized exams and documented its spread to 12 courses over three years at the University of Illinois at Urbana-Champaign. Furthermore, we have provided a critical examination of this spread from the perspectives of both change theory (diffusion of innovations, Section 7) and education theory (communities of practice, Section 8).

Considering both perspectives, we can identify the two key characteristics that enabled the spread of `randexam` to be:

1. Innovation characteristics: the innovation brought immediate observable benefits while integrating easily into pre-existing exam systems, both in terms of technology (LaTeX), and exam format (same times, same rooms, same arrangements).
2. Organization and system characteristics: the organizational network was especially effective at spreading the innovation, with CoP hubs and long-term links, both of which were explicitly created and supported by programs in the College of Engineering.

Acknowledgments. This work was supported by the College of Engineering and the Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign as part of the Strategic Instructional Initiatives Program (SIIP), as well as by the National Science Foundation (NSF) award DUE-1347722.

A. Key generation

The exam keys are encoded within the final questions on the Scantron form, each of which has N_{A} possible answers. The keys are generated by encoding the exam number e in base-N_{A} using N_{D} digits and then appending two or three checksum digits. The exam-number digits are in little-endian order, with least significant digit first, so for zero-based digits K_i and zero-based indexes i they are

$$K_i = \left\lfloor \frac{e}{(N_\mathrm{A})^i} \right\rfloor \mod N_\mathrm{A}, \text{ for } i = 0, \ldots, (N_\mathrm{D} - 1). \tag{1}$$

The N_C checksum digits are the *parity digit* K_P, a *Fletcher-type* checksum K_F, and possibly a modified Fletcher-type checksum K_M, defined for zero-based digits K_i and zero-based indexes i by

$$K_\mathrm{P} = \sum_{i=0}^{N_\mathrm{D}-1} K_i \mod N_\mathrm{A} \tag{2}$$

$$K_\mathrm{F} = \sum_{i=0}^{N_\mathrm{D}-1} w_{\mathrm{F},i} K_i \mod N_\mathrm{A} \qquad w_{\mathrm{F},i} = \Big(i \mod (N_A - 1)\Big) + 1 \tag{3}$$

$$K_\mathrm{M} = \sum_{i=0}^{N_\mathrm{D}-1} w_{\mathrm{M},i} K_i \mod N_\mathrm{A} \qquad w_{\mathrm{M},i} = \left(\left\lfloor \frac{i}{N_\mathrm{A} - 1} \right\rfloor \mod (N_\mathrm{A} - 1)\right) + 1. \tag{4}$$

Using the first two checksum digits ensures a minimum Hamming distance of 3 between keys for at most $(N_\mathrm{A} - 1)$ non-checksum digits (up to $N_\mathrm{A}^{(N_\mathrm{A}-1)}$ different random exams). If more digits are required to encode the exam numbers then the third checksum digit is appended to the key, ensuring a Hamming distance of at least 3 between keys for up to $(N_\mathrm{A} - 1)^2$ non-checksum digits, allowing up to $N_\mathrm{A}^{(N_\mathrm{A}-1)^2}$ different random exams. In the case of $N_\mathrm{A} = 5$ the key lengths are:

number of exams	exam-number digits N_D	checksum digits	key length $N_\mathrm{K} = N_\mathrm{D} + N_\mathrm{C}$
5	1	2	3
25	2	2	4
125	3	2	5
625	4	2	6
3125	5	3	8
15625	6	3	9

References

[1] G. Aubrecht and J. Aubrecht. Constructing objective tests. *American Journal of Physics*, 51:613, 1983.

[2] A. Bresnock, P. Graves, and N. White. Multiple-choice testing: Question and response position. *Journal of Economic Education*, 20:239–245, 1989.

[3] L. Crowley and G. L. Herman. Using faculty communities to drive sustainable reform: Learning from the Strategic Instructional Initiatives Program. In *ASEE 2014: Proceedings of the American Society for Engineering Education 121st Annual Conference and Exposition*, 2014. Paper ID #9052.

[4] T. H. Davenport and L. Prusak. *Working Knowledge: How Organizations Manage What They Know*. Harvard Business School Press, second edition, 2000.

[5] S. Gohmann and L. Spector. Test scrambling and student performance. *Journal of Economic Education*, 20:235–238, 1989.

[6] T. Greenhalgh, G. Robert, F. Macfarlane, P. Bate, and O. Kyriakidou. Diffusion of innovations in service organizations: Systematic review and recommendations. *The Milbank Quarterly*, 82(4): 581–629, 2004. doi: 10.1111/j.0887-378X.2004.00325.x.

[7] P. Hildreth and C. Kimble. The duality of knowledge. *Information Research*, 8, 2002.

[8] J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, 1991.

[9] L. E. Lesser and J. Storck. Communities of Practice and organizational performance. *IBM Systems Journal*, 40, 2001.

[10] S.-Y. Lin and C. Singh. Can free-response questions be approximated by multiple-choice equivalents? *American Journal of Physics*, 81:624, 2013. doi: http://dx.doi.org/10.1119/1.4804194.

[11] I. McLeod, Y. Zhang, and H. Yu. Multiple choice randomization. *Journal of Statistics Education*, 11 (1), 2003.

[12] M. D. Miller, R. Linn, and N. Gronlund. *Measurement and Assessment in Teaching*. Pearson, 1995.

[13] randexam. `http://lagrange.mechse.illinois.edu/randexam/`, 2015.

[14] E. M. Rogers. *Diffusion of Innovations*. Free Press, 1995.

[15] Michael Scott, Tim Stelzer, and Gary Gladding. Evaluating multiple-choice exams in large introductory physics courses. *Phys. Rev. ST Phys. Educ. Res.*, 2:020102, 2006. doi: 10.1103/PhysRevSTPER.2.020102.

[16] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1998.

[17] E. Wenger, R. McDermott, and W. M. Snyder. *Cultivating Communities of Practice*. Harvard Business Press, 2002.

[18] M. West and G. L. Herman. Sustainable reform of “Introductory Dynamics” driven by a community of practice. In *ASEE 2014: Proceedings of the American Society for Engineering Education 121st Annual Conference and Exposition*, 2014. Paper ID #10519.

[19] B. Wilcox and S. Pollock. Couple multiple-response versus free-response conceptual assessment: an example from upper-division physics. *Physics Education Research*, 10:020124, 2014. doi: http://dx.doi.org/10.1103/PhysRevSTPER.10.020124.