Automated and scalable assessment: present and future

Dr. Edward F. Gehringer, North Carolina State University

Dr. Gehringer is an associate professor in the Departments of Computer Science, and Electrical & Computer Engineering. His research interests include computerized assessment systems, and the use of natural-language processing to improve the quality of reviewing. He teaches courses in the area of programming, computer architecture, object-oriented design, and ethics in computing. He is the lead PI on a multi-institution NSF IUSE grant to construct web services for online peer-review systems.

Automated and Scalable Assessment: Present and Future

Abstract

A perennial problem in teaching is securing enough resources to adequately assess student work. In recent years, tight budgets have constrained the dollars available to hire teaching assistants. Concurrent with this trend, the rise of MOOCs, has raised assessment challenges to a new scale. In MOOCs, it's necessary to get feedback to, and assign grades to, thousands of students who don't bring in any revenue. As MOOCs begin to credential students, accurate assessment will become even more important. These two developments have created an acute need for automated and scalable assessment mechanisms, to assess large numbers of students without a proportionate increase in costs. There are four main approaches to this kind of assessment: autograding, constructed-response analysis, automated essay scoring, and peer review. This paper examines the current status of these approaches, and surveys new research on combinations of these approaches to produce more reliable grading.

Keywords: scalable assessment, autograding, constructed-response analysis, automated essay scoring, peer review, self-review

1. Introduction

Limited resources to adequately assess student work are a major problem in higher education. A 2012 *New York Times* article [1] on automated essay scoring contains the following anecdote:

> For a question asking students to discuss why college costs are so high, Mr. Perelman wrote that the No. 1 reason is excessive pay for greedy teaching assistants.
>
> "The average teaching assistant makes six times as much money as college presidents," he wrote. "In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, starring roles in motion pictures."
>
> E-Rater gave him a [perfect score]. He tossed in a line from Allen Ginsberg's "Howl," just to see if he could get away with it.

This statement both caricatures the magnitude of the problem and illustrates the difficulty of finding an automated solution. Indeed, assessment of student work is a growing problem.

- Per-student instructional budgets have fallen across the country since the Great Recession, and show no signs of turning around anytime soon.
- A growing body of evidence indicates that one of the most important factors in the success of at-risk students is the amount, and promptness, of the formative feedback they receive.

- The rise of MOOCs poses a new challenge, as they can reach their full potential only if certificates of completion can be taken as evidence that students have mastered the work … and this can only be demonstrated if scalable methods of grading are devised.

- Compared to advanced countries, universities in less-developed countries rely on exams, rather than homework, for almost all of their assessment, and this encourages students to cram, which is not an effective way to learn for retention. Even in advanced countries, there is a tendency to increase the weight of exams, in order to avoid incentivizing Internet-based cheating.

But help is on the horizon. Four approaches to automating grading are being deployed in a growing number of courses, and new research promises to increase their effectiveness.

2. Autograding

We will use the term "autograding" to mean matching a student's response to a question with a predetermined "correct" answer or answers. Almost every LMS has a quizzing module that can be used to administer multiple-choice and fill-in-the-blank questions. The questions and answers can usually be randomized so that no two students see the exact same questions and answers. One way of doing this is to have the software select a certain number of questions for each student from a test "bank" that contains a larger number of questions than any student will be asked to answer. Some systems allow numeric parameters to be varied, so that each student needs to do a different calculation. If the question is a multiple-choice question, the answers can be presented in different orders to different students.

The same functionality is available in proprietary publishers' systems like Wiley Plus, and in third-party applications like Webassign.

These systems also collect analytic data that can be used to improve the assignments that are presented to students. For example, data can be collected on the distribution of scores per assignment, or per question. For multiple-choice questions, the system can record the number of times each distractor was chosen. This data offers great benefits to instructors who want to make sure that work is challenging but not too challenging, or to make their multiple-choice distractors correspond to common misconceptions.

Systems can be designed administer tests as "branched surveys" [3]. Based upon how students do on early questions, they can be presented later on with questions of appropriate difficulty.

Autograding systems have also been designed for specialized domains, notably computer programming. A system such as Web-CAT [4] compiles students' code, then runs it

against a suite of automated tests written by the course staff. The student's grade is computed based on how many of the tests pass. The same concept has been applied to automated testing of student-written web applications [5].

3. Constructed-response analysis

Of course, not all material can be adequately tested by autograded questions. Requiring a student to *construct* a short-answer response to a question requires a deeper kind of understanding than answering a multiple-choice question. These answers can be analyzed by an automated system, which through linguistic analysis, groups together responses that seem to be similar. In the AACR (Automated Analysis of Constructed Response) project at Michigan State [6], the data are then utilized to develop rubrics for human scoring. These rubrics are used in formative assessment. Microsoft's Powergrading approach [7, 8] is similar. A similarity metric is used to group student responses into clusters. An instructor can grade a representative item from each cluster, and assign a score to each cluster. Each cluster is also given customized formative feedback.

The clustering approach has proven useful in other domains as well. A recent paper on mathematical language processing uses cluster-based analysis to assign partial credit to mathematical derivations [9]. This is probably more challenging than doing the same with prose answers, because it must be able to recognize different ways of expressing a formula, and determine whether one is simpler than another (an example from the paper is that $\sin^2 x + \cos^2 x + x$ is the same as $1 + x$, but the latter form is the only correct answer if the question asks the student to simplify the expression).

It should be stressed that at the current level of technology, constructed-response techniques are not ready for prime time. They require considerable coding effort on the part of the instructor, and in the case of AACR, at least, the responses must be downloaded and analyzed by a third-party statistical application. Nonetheless, with more experience and tool integration, these techniques offer the promise of substantial speedup in the grading task.

4. Automated essay scoring

Constructed-response analysis is intended to help instructors rate student answers for correctness. In the case of creative writing, aesthetics is the goal instead of correctness. There is a long history of work in automated essay scoring, and it is widely used commercially in high-stakes testing (such as the SAT), and programs that prepare students for taking such tests. Studies have shown [10] that automatic ratings of essays correlated more highly with human raters than the human raters correlated with each other. Commercial products can issue reports on various characteristics of essays, like vocabulary, text complexity, sentence variety, style, and errors in word usage.

AES applications calculate their metrics using several factors, including average word length, number of grammar errors, and similarity to human-scored essays with similar vocabulary. However, as illustrated by the anecdote at the beginning of this paper, AES applications do not "understand" text; they can easily be gamed by a good writer. Whether they can be gamed by an ineffective writer is not so clear.

AES is used extensively on the EdX MOOC platform [11]. EdX also uses self- and peer assessment, but other platforms have developed these techniques more fully.

5. Peer review

Peer review asks students to assess each other's work. Like the other techniques, it can be used both formatively and summatively. For formative review, the student reviewer is asked to fill out a rubric that asks certain questions about the student author's work. For summative review, the reviewer is asked to rate the work numerically based on a set of criteria (organization, clarity, etc.).

Peer review has been widely used in higher education since the 1970s, and online systems have been available for over 20 years. The largest ongoing project in this area is the NSF-funded Calibrated Peer Review (CPR) project [11], which has been used by more than a quarter-million students. While the pedagogical benefits of peer review are well established, students must be trained in how to write an effective review. CPR does this by having students review three artifacts supplied by the instructor: one is a model artifact, and the other two have known defects. How close the student reviewer comes to the instructor's rating of these three artifacts determines the reviewer's *reviewer competency index*, and the RCI is used to weight the reviewer's scores in a calculation of the author's grade. The Coursera MOOC platform employs a takeoff on this strategy: a student is asked to rate one actual student submission, which has also been rated by an instructor. If the student's rating is "close enough," the student is allowed to assess peers' work. Otherwise, the student is given another pre-rated artifact to assess. If the student's score is "close enough," (s)he is allowed to go on to assess peers; otherwise, the process is repeated up until the fifth attempt, after which the student assesses others' work anyway. As with Calibrated Peer Review, a reviewer reputation is computed, and used in combination with other data to derive student grades.

Unlike peer review, peer grading is controversial. The largest studies conducted on peer grading are those from the Coursera platform. Piech et al. [12] studied over 7200 students in a MOOC on Human-Computer Interaction. Grades were assigned by peers, and effectively "spot-checked" by instructors. Even with their best statistical model, at least 26% of the peer-assigned grades were more than 5% away from the "true" grade that instructors would have assigned. In a study of 5876 students in other Coursera MOOCs, Kulkarni et al. [13] found that 40% of assignment grades derived from peer

review were at least one letter grade away from the correct grade. Thus, in a classroom situation, it is recommended that the instructor assign final grades. With peer reviews available, however, grading goes more rapidly, because the peer comments point out characteristics of the work that are relevant to the grade.

Like autograding, peer review is supported by all of the major LMSs, but some of the standalone systems like Peerceptiv and Peer Scholar tend to be more advanced. The Canvas LMS/MOOC platform has an interface that allows peer reviewers to annotate the documents that are submitted by students, which is often an easier and more effective way to provide feedback than writing comments in a prose review.

Before leaving peer review, we should point out that peer review is often combined with self-review. This practice seems to have originated in CPR, where the agreement between self-review and peer review is one of the criteria for assigning a grade. It is also supported by the Coursera, EdX, and Google Course Builder MOOC platforms.

Table 1 summarizes the various types of tools described in this paper.

Table 1. Approaches to Scaling Grading

Approach	**Type of questions**	**Limitations**	**Where to find tools**
Autograding	Multiple choice, checkbox, T/F, matching, numeric, fill-in-the-blank	Difficult to assess deep learning with objective questions	Any LMS, Webassign, Wiley Plus
Constructed-response analysis	Short answer, mathematical derivations	Requires considerable effort to set up, cluster, and analyze	AACR, CRASE
Automated essay scoring	Essay	Can be fooled by work with proper form but inadequate substance	IntelliMetric, e-rater, PEG, EdX
Peer review	Almost anything, esp. where other approaches cannot be used	Requires training of students, often inaccurate for summative assessment	Calibrated Peer Review, Peerceptiv, Peer Scholar, Expertiza, etc.
Self-review	Almost anything	Cannot be used as the only assessment mechanism	Calibrated Peer Review, Course Builder, EdX, etc.

6. The future of scalable grading

Many of the techniques we have described can be used in conjunction with each other. Peer reviews of student work can be combined with metrics derived from the student text to improve the reliability of reviews [14]. Peer reviewers' annotations of the author's

document could be enhanced to give different reviewers an opportunity to respond to each other's annotations on the document [15]. Reviews can be analyzed in various ways using natural-language processing techniques to compute reviewer reputations and provide feedback to the reviewer on how to improve reviews [16]. Peer review can be applied to student-submitted exercises over the course material [17], which may then be used as questions for an autograding system. Latent semantic analysis, which has been used extensively in AES, can be applied to other domains, such as grading computer programs for style as well as correctness [18]. The clustering approach applied to constructed responses can also be applied to improve reliability of peer grading [19]. Work on clustering and summarizing of posts on a class message board [20] could be applied to peer reviews, to present the student author with a summary that was less repetitive than reading the reviews sequentially.

Most of these approaches can benefit from advances in natural-language processing. In the case of constructed-response analysis and automated essay scoring, NLP is applied to the submitted work. In peer and self-review, it is applied to assess the reliability of the review and estimate whether the review is an accurate assessment of the submitted work.

References

[1] Winerip, Michael, "Facing a robo-grader? Just keep obfuscating mellifluously," *New York Times*, April 22, 2012.

[2] Fox, Armando; Canny, John, "Autograding and online ed technology," https://docs.google.com/document/d/11e7HzGGRAvAhTce6L7P33fyQUo67wO_Qbec6cGynrKo/edit#heading=h.vo90ekim8uj0, accessed Feb. 2, 2015

[3] Beitzel, B. D.; Gonyea, N. E., "The rubric interview: a technique for improving the reliability of scoring written products," Proc. 2014 Virginia Tech Conference on Higher Education Pedagogy, p. 242.

[4] Edwards, S.H; Perez-Quiñones, M.A., "Web-CAT: automatically grading programming assignments." In *Proceedings of the 13th annual conference on Innovation and technology in computer science education* (ITiCSE '08). ACM, New York, NY, USA, 328-328, 2008. DOI=10.1145/1384271.1384371 http://doi.acm.org.prox.lib.ncsu.edu/10.1145/1384271.1384371

[5] Warne, K., "Experiences in implementing automated evaluation and grading in a MOOC using a behavior-driven testing framework," master's thesis, University of New Mexico, 2014.

[6] Ha, M.; Nehm, R. H;, Urban-Lurain, M.; Merrill, John E., "Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations," CBE—Life Sciences Education 10:4, pp. 379–393, 2011, doi 10.1187/cbe.11-08-0081

[7] Basu, S.; Jacobs, C.; Vanderwende, L., "Powergrading: a clustering approach to amplify human effort for short-answer grading," *Transactions of the Association for Computational Linguistics*, 2013.

[8] Brooks, M.; Basu, S.; Jacobs, C.; Venderwende, L., "Divide and correct: using clusters to grade short answers at scale," *L@S '14: Proceedings of the First ACM Conference on Learning @ Scale*, Mar. 2014, pp. 89–98, doi 10.1145/2556325.2566243

[9] Lan, A.S.; Vats, D.; Waters, A.E.; Baraniuk, R.G., "Mathematical language processing: automatic grading and feedback for open-response mathematical questions," *L@S '15: Proceedings of the Second ACM Conference on Learning @ Scale*, Mar. 2015.

[10] Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K., "Automated essay scoring: Writing assessment and instruction." In E. Baker, B. McGaw, & N.S. Petersen (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 75–80). Elsevier, 2010.

[11] Balfour, S.P., "Assessing writing in MOOCs: automated essay scoring and Calibrated Peer Review," *Research & Practice in Assessment*, 2013.

[12] Piech C, Huang J, Chen Z, Do C, Ng A, Koller D. Tuned Models of Peer Assessment in MOOCs. *7th Int Conf Educ Data Min*. 2013. Available at: http://www.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf. Accessed September 17, 2013.

[13] Kulkarni C, Wei KP, Le H, et al. Peer and Self Assessment in Massive Online Classes. *ACM Trans Computer-Human Interaction*. 2013;20(6):33:1–33:31. doi:10.1145/2505057.

[14] Vista, A.; Care, E.; Griffin, P., "A new approach towards marking large-scale complex assignments: developing a distributed marking system that uses an automatically scaffolding and rubric-targeted interface for guided peer-review," *Assessing Writing* 14:1–15, April 2015.

[15] Sikka D. NORA: No One Revises Alone. 2013. Master's thesis, MIT. Available at: http://groups.csail.mit.edu/uid/other-pubs/denzil-meng-thesis.pdf. Accessed January 30, 2014.

[16] Ramachandran, Lakshmi, "Automated assessment of reviews," Ph.D. Dissertation, North Carolina State University, May 2013.

[17] Denny, Paul, et al. "PeerWise: students sharing their multiple choice questions." *Proceedings of the fourth international workshop on computing education research*. ACM, 2008.

[18] Srikant, S.; Aggarwal, V., "Automatic grading of computer programs: a machine-learning approach," *12th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2013.

[19] Shah, N.B.; Bradley, J.; Balakrishnan, S.; Parekh, A.; Ramchandran, K; Wainwright, M.J., "Some scaling laws for MOOC assessments," 2014.

[20] Reich, J.; Tingley, D.; Leder-Luis, J.; Roberts, M.E., "Computer assisted reading and discovery for student-generated text," *Journal of Learning Analytics*, to appear.