Can automated assessment technology improve student writing?

Jamey Heit, Nathan Rudemiller, and Robin Donaldson

Introduction

Current automated assessment systems can be split into two categories, either those using topical analysis or those using pattern-based analysis. The former do not analyze content, instead focusing on aspects such as spelling, grammar and word choice. The latter require a large body of essays to train the systems for a particular essay question. In either case, key aspects of quality writing (e.g. organization, development of argumentation, explanation of evidence) are never measured directly. Therefore, these systems cannot deliver high quality feedback to the student and cannot be used effectively as a learning tool.

We have developed a new automated assessment technology called mechanistic assessment that computes accurate percentage (i.e. 0-100) grades for essays, and provides the student with substantive feedback throughout their essay, within 30 seconds. The technology computes thousands of data points throughout an essay, which in turn inform dozens of metrics measuring aspects of quality writing. If the student performs poorly in any of these aspects, they are given feedback on how to improve their writing with respect to that specific aspect.

We have implemented a web app using this technology, which allows professors to set essay assignments and students to submit drafts of their essays to those assignments. The students automatically receive feedback on each draft, which allows them to improve their essay. As our software does not require training, students can get feedback on any essay in the humanities.

Our thesis is that the writing ability of students using our software will improve due to the feedback that they receive. In this paper we present results from pilot projects that use our software, showing that our automated assessment technology helps to improve students' writing ability.

Background

There are clear, established concerns with automated assessment software. We cite Dr. Les Perelman, former Professor of Composition at MIT, who led the academic argument claiming automated assessment could not provide a viable educational resource. We consider two of Dr. Perelman's concerns in this paper.

Performance Concern. Dr. Perelman's first concern is about the performance of automated assessment: "My first and greatest objection to the research is that they did not have any valid statistical test comparing the software directly to human graders."[1]

Mechanism Concern. Dr. Perelman's second concern is about the mechanisms behind automated assessment: "Computers cannot 'read.' They cannot measure the essentials of effective written communication: accuracy, reasoning, adequacy of evidence, good sense, ethical stance, convincing argument, meaningful organization, clarity, and veracity, among others."[2]

As outlined, there are two categories of automated assessment technology, either those using *topical analysis* or those using *pattern-based analysis*.

Systems that use topical analysis grade essays based on topical elements such as spelling, grammar and word choice, which correlate poorly with grade, and of course they are not measures of "effective written communication." Therefore Dr. Perelman's two concerns hold.

Systems that use pattern-based analysis rely on artificial intelligence or machine learning algorithms that must be trained with hundreds of sample essays for a particular essay question. The training process finds elements of the sample essays that correlate with the grades for the sample essays, and these elements are then used to predict the grades of future essays. This process is opaque, and one never knows what elements are used to grade the papers, nor whether to trust these elements. Certainly, they are not "essentials of effective written communication." However, as the algorithms are highly trained, they can perform well in terms of grade prediction, though one can argue that these systems are myopic: "they grade based only on what they've seen before."

In addition to failing Dr. Perelman's two concerns, neither type of system provides substantive feedback because they do not directly measure effective written communication. Furthermore, the more sophisticated of the two systems, pattern-based systems, requires training, which can be prohibitive especially in a learning environment. For example, one system requires 350 essays graded by 2 professors over an extensive rubric for a single essay question. Because of these issues, the current systems are ineffective as learning tools and the area requires a new technology to facilitate learning.

---

[1] Quoted in John Markoff, "Essay-Grading Software Offers Professors a Break," *New York Times* (April 4, 2013).
[2] Ibid.

Results

We have developed a new automated assessment technology called *mechanistic assessment* that addresses Dr. Perelman's two concerns and can be used in an effective learning tool. The technology computes thousands of data points throughout an essay, which in turn inform dozens of metrics measuring aspects of quality writing. The metrics are combined to produce an overall grade for the essay. If the student performs poorly in any of these metrics then feedback is automatically generated on how to improve on that writing aspect in the future. The technology does not require training, so it can be used "out of the box" for any essay question in the humanities.

Our software, which uses mechanistic assessment, (i) computes accurate percentage grades for essays, (ii) directly measures aspects of quality writing, (iii) provides substantive feedback to the student, and (iv) does not require training.

Addressing the Performance Concern. To address the performance concern, we have compared our software's grades with the human grades over a set of essays.

We have collected 2,000 essays from various humanities classes covering eight professors. Each paper was provided as an anonymized electronic file with the percentage grade assigned by the professor *a priori* (i.e. before the software grade was computed). We then randomly selected 500 essays from the collection, and ran those essays through our software. **Figure 1** shows the results. In the overwhelming majority of cases, the algorithmic and human scores converged within a narrow range.

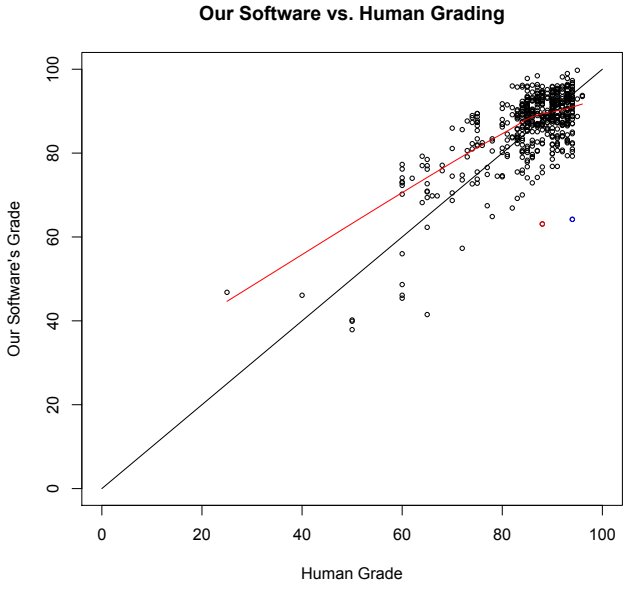

Figure 1. The comparison of the human grades (x-axis) versus our software's grades (y-axis) over 500 essays covering 18 essay questions and eight different professors.

Addressing the Mechanism Concern. Our algorithm analyzes essays using dozens of metrics that map to specific aspects of quality writing that humans look for when evaluating an essay. For example, just as professors will look for a thesis statement and adjust a paper grade based on the quality of that statement, so too does our algorithm. The set of metrics that we use to compute the essay grade and student feedback forms part of our intellectual property.

Effectiveness of Mechanistic Analysis. To demonstrate the effectiveness of mechanistic assessment, we show the performance of our system on a particular essay question: "Write a Love Story." This essay question is unusual because it asks students who are supposed to be practicing essay writing to write a story.

Upon reviewing the paper submissions for this assignment, we noticed that some students wrote stories and some wrote essays. We asked the professor, before grading the papers, to categorize each paper as either an essay or a story. Then we plotted in **Figure 2** the grades of the essays and the grades of the stories and found that our algorithm graded essays far higher than stories (P value < 0.0001). This reflects the mechanistic assessment aspect of our software, in which we truly capture aspects of good argumentation.

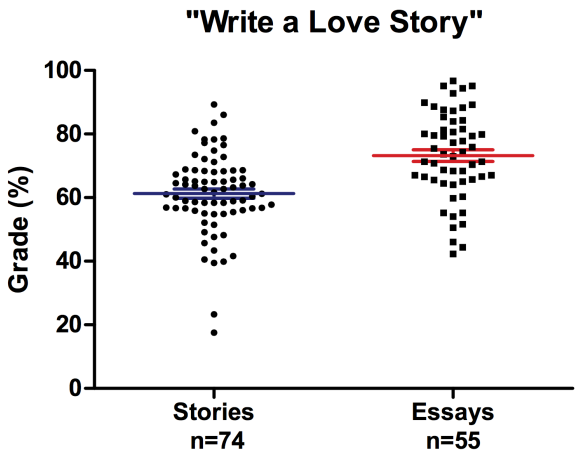

Figure 2. Two types of papers were submitted to the "Write a Love Story" assignment, stories and essays. Our software has graded the papers that were essays significantly higher than those that were stories.

Use as a Learning Tool. No current system offers substantive feedback on student essays. A result of our mechanistic assessment is that when a student performs poorly according to one of the metrics, feedback is automatically generated on how to improve in the future. Also, if a student performs well on a given metric, feedback

is provided to reiterate the positive performance vis. that metric. Given the consensus that immediate feedback on academic work is known to help students learn[3], we have solved a known roadblock: "substantive feedback on written communication cannot be automated." This feedback is a significant technical and pedagogical breakthrough, only possible because of our new approach, mechanistic assessment.

Automated feedback also allows students to submit subsequent drafts of their essay to refine the essential elements of good argumentation. With such a resource, the expectation would be a material improvement in student writing.

Along with the lack of feedback in current systems, all systems using pattern-based analysis (the more sophisticated of the two types of systems) require training. Our automated assessment system takes as input simply the essay question, 5 to 10 key themes (providing extra context to the essay question) and the essay text. The essay is then graded according to predefined metrics that mirror standard essay assessment rubrics, discussed above; hence, there is no training of our system, and we can accurately assess essay on any topic within the humanities.

Effectiveness as a Learning Tool. Our automated assessment system can be used as a learning tool due to the breakthroughs described above. We have used the system as a learning tool in over 20 pilot projects around the world to date. We now present results that show the effectiveness of this tool on student learning. Note that we have anonymized the data resulting from the pilot projects, i.e. classes are identified using letters and assignments are identified using numbers.

The ideal student usage of our system, from a learning point of view, is that the student will *iterate* their essays; the student will submit subsequent drafts of an essay, each time improving their writing using the automated feedback. We refer to the process of submitting multiple drafts as *iteration*.

We have randomly selected several classes from the pilot projects to show the effectiveness of our system. We then select those students who have iterated their essays for assignments in each class, and in **Figure 3** show the grades for the first and final draft of the essays. If students are improving, we expect that the grade for their final draft will be higher than for their first draft.

In all classes, the grades for the final drafts are higher than the grades of the first drafts. We have computed the P values for the significance of this difference. Hence,

---

[3] See, for example, F.M. Van der Kleij, R.W.C. Feskens, and T.J.H.M. Eggen, "Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis," *Review of Educational Research* (January 8, 2015).

a P value of 0.05 (the boundary of *statistical significance*) says that there is a 95% confidence that the difference between the first and final drafts in the class is not by chance. Class E and Class F have a P value that exceeds the threshold for statistical significance (i.e. the P value is less than 0.05). Classes A, B, C & D, which represent four different sections of English Composition, have a P value of 0.059 (around 94% confidence) which is just below the threshold of statistical significance. Finally, Classes G & H, which are honors literature classes, show a P value of 0.076 (around 92% confidence).

If we take the largest P value in this data set (i.e. the least convincing data, Classes G & H), we still see a high degree of confidence that this software helps students improve their writing. Taking account that the classes with this P value are honors classes, we can surmise that the slightly lower confidence level results in part from the overall higher student aptitude. In other words, students who are not in honors classes consistently improve more, which meets the threshold for being statistically significant. The fact that no students in these honors classes show improvements of more than 10 points justifies this claim.

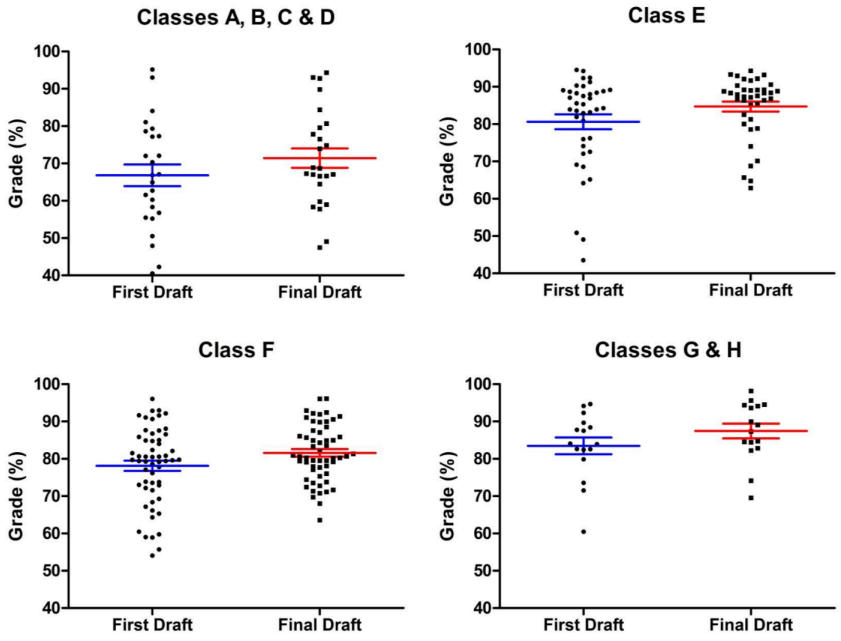

Figure 3. These graphs show the improvement in writing quality between first and final draft of students essays over eight different classes.

For some students, the effect of feedback is striking; there is a clear subset of students who improve by 10 or more points. The four sections of English Composition showed that 28% of students who iterated showed improvements of at least one letter grade.

This data shows the pedagogical value of providing immediate feedback to students on their written work. When students submit drafts and receive substantive feedback as many times as they need before submitting the final draft of their essay, the general result is an increase in the quality of the student's writing.

If we shift from class-wide view of the data and look at individual assignments, in many instances the improvement students realize is even stronger than suggested in the above results. Taking Assignment 136 from Class F in isolation, shown in **Figure 4**, we see that the improvement is conclusive.

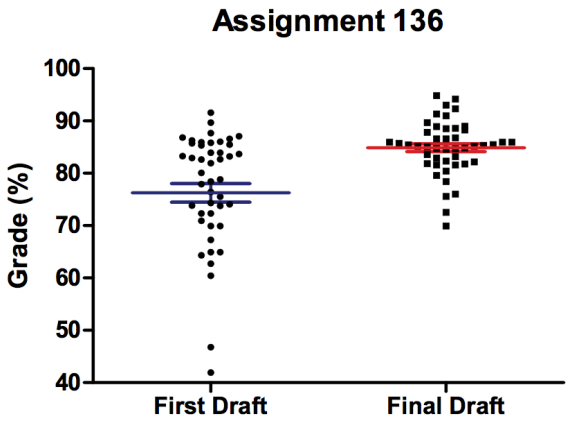

Figure 4. This graph shows the improvement in writing quality between first and final draft of students essays in Assignment 136.

Two important points are captured in this graph. The first is the conclusive improvement for students who iterate in this assignment. The average improvement was around 8.5 points.[4] With a P value less than 0.0001, we can conclude that this improvement is not by chance.

Student Participation. We note that in this specific assignment, as with the class-wide results mentioned in this study, students were not required to use the software. Those who chose to iterate did so without external incentive. Despite the voluntary nature of using the software, 41% of the students in the class iterated on this assignment. This participation rate mirrors the rate of iteration across the different classes in this study. The level of iteration strongly suggests that students understand that the feedback they are receiving is useful to improve their writing.

Beyond the compelling data we have presented in this paper, we see as well in student surveys that our software provides better educational engagement than

[4] The average improvement is in line with broader trends in student improvement; during the Fall 2015 term, the average increase for students who iterated was 7.6 points.

human teachers. Students were asked to rate the impact of our software on their writing. 92% of students responded that the software was as good as or better than the engagement they received from their human teachers.

Conclusion

We have introduced in this paper a breakthrough technology called mechanistic assessment. Mechanistic assessment computes percentage grades for essays by directly measuring various aspects of quality writing. We have shown that the technology is accurate in the grade prediction of 500 essays.

Because mechanistic assessment directly measures aspects of quality writing, substantive feedback can be generated automatically to help the student improve their writing. The system also does not need to be trained. In other words, we have removed the two most significant barriers to using automated assessment technology as a learning tool.

We have shown results in this paper on the effectiveness of our system as a learning tool; students who iterate their essays using our software (i.e. improve their writing based on the software's feedback) improve. We also show that the student participating rate is high, without any external influence, suggesting that the students realize the value of the feedback they receive.

In summary, students are using our system globally in pilot projects and are improving as a result. We can now answer the question, *"Can automated assessment technology improve student writing?"* with a clear *yes*.