# SAGrader™ Benefits Can Outweigh Costs in First Semester

Edward Brent, Theodore Carnahan, and Jeff McCully

At what point do the benefits of using SAGrader outweigh the costs to develop automatically-graded writing assignments? This analysis identifies various costs and benefits associated with developing and testing of writing assignments in SAGrader and compares them to those for grading such assignments by hand. It also considers some of the less tangible benefits and costs that must be considered. Examining a range of possible values for the parameters of the cost model discloses that even with relatively pessimistic cost estimates, *SAGrader should generally pay for itself within the first semester when only single submissions are permitted, and should provide time and cost savings over hand grading during the first semester when multiple submissions are permitted.* If the same assignment is used in multiple courses or multiple sections of the same course, even greater cost savings can occur within the first semester. Expected costs for subsequent semesters or when pre-constructed assignments are used should be substantially lower than the costs for comparable hand grading.

For this cost-benefits analysis we compare the costs and benefits of SAGrader with the costs and benefits of traditional "hand-grading" from the perspective of the instructor who would do the grading. For each, we discuss the basic cost components and assumptions, then estimate costs for a realistic range of parameters for essays of varying complexity and length. Finally, we consider costs and benefits for students and administrators.

## Cost Models

### A Cost Model for Traditional Hand-Grading
Our cost model estimates the time required to hand-grade a question ($T_{HG}$) is the product of the number of submissions ($N_{SUB}$) times the sum of time spent determining the score ($T_{SC}$), writing comments or feedback to the student ($T_{FB}$), and totaling and recording the grade ($T_{REC}$)[1].

$$T_{HG} = N_{SUB} * (T_{SC} + T_{FB} + T_{REC})$$

Clearly, traditional hand grading is very sensitive to the number of essays submitted for grading.

### A Cost Model for SAGrader
The costs of SAGrader are very different from those of traditional hand grading. For SAGrader, most of the cost is in implementing questions in the computer program and much less time is required for the automated grading itself. Our cost model for SAGrader assumes the time required to grade ($T_{SAG}$) is the sum of the time required to create a concept map for the concepts ($T_{MAP}$), the time required to identify key phrases that, when found in various combinations in student essays, express those concepts ($T_{KEY}$), the time required to review answers and revise the program until the program is trained well enough to be used ($T_{REV}$), and the product of the number of submissions and the time per submission required for monitoring program performance.

$$T_{SAG} = T_{MAP} + T_{KEY} + T_{TRN} + (N_{SUB} * T_{MON})$$

---

[1] We will ignore the time required to identify a question and write the prompt, since that same amount of time would be required for hand-grading. That time also varies a great deal and is hard to predict.

Creating the concept map is usually the easiest and least time-consuming step in the process. For very simple essay questions this may take a minute or less. For very complex questions it might require careful thought and a few revisions taking an hour or more.

Identifying various key phrases takes longer. When there are important concepts with technical names students should know, this can be a matter of seconds in which one or two acceptable variations are identified. For definitions, complex relationships among concepts, examples, and concepts students are permitted to express in many ways it may take as much as a half hour or even several hours to identify a reasonable number of appropriate phrases for all of the concepts in an assignment.

Training the program and monitoring its performance usually require the most time. By <u>training</u> we mean all those efforts to test the program and improve its performance before beginning to use it in the class. <u>Monitoring</u> includes all those efforts to review student assignments, respond to any student challenges, and correct or improve performance of the SAGrader program as it is used by students. Monitoring the program requires time that generally increases as a function of the number of essays submitted for grading. This monitoring is facilitated by the ability of students to challenge their grade. Students, it turns out, are quite willing to carefully examine their grade and raise concerns if they believe it was unfair. By focusing on challenges and only quickly scanning selected unchallenged essays to assess their validity, monitoring can end up requiring much less time than it would take to hand-grade, provide detailed feedback to students, and record their grades. In one recent course students submitted a challenge for 17% of their submissions. We believe this is a higher rate than usual because many of the challenges were actually off target and more "communications" to instructors instead of pointing out specific areas where the program made a grading error. Based on our analysis we believe a 10% challenge rate is more likely.

Note that only one of these terms is a variable related to the number of students graded.[2] In general, SAGrader can be developed and trained in a fixed period of time to the point where it can run on its own and achieve an acceptable level of accuracy. Only monitoring is related to the number of essays submitted. This means that the costs of grading each with SAGrader are usually mostly determined by the question development and are less affected than hand grading by the number of essays submitted for grading.

**Cost Comparisons for Three Sample Questions**

Actual times required for grading are likely to vary substantially depending on the complexity of questions and the length of the essay. Here we consider three different possibilities encompassing the range of questions from very simple and short essays to moderately complex essays of several pages.

**Case 1: A Simple Essay**
First, let's consider a very simple essay question where only a few words or a phrase or a single sentence provides a good answer. A very optimistic estimate for hand-grading is that it might take as little as 5 seconds to grade, 10 seconds for comments, and 3 seconds per question to total scores and record a grade. This is 18 seconds per student. For SAGrader, such a question might reasonably require 2 minutes to create the semantic map, five minutes to generate key terms, and perhaps a

---

[2] The actual machine-grading itself is related to the number of essays. However, machine grading is totally automated, very fast, and will not be considered in this model.

total of 60 minutes to check and revise the program. Reviewing only challenges and selected essays might take an average of 18 seconds for every 10<sup>th</sup> essay or 2 seconds each.

## Case 2: A 2-Page Essay
For a moderately complex essay of 1 or 2 pages with say five broad questions assessed by the grading rubric, this might require 8 minutes to read and grade, 5 minutes for comments, and 30 seconds per essay to total and record grades. For SAGrader, such an essay might reasonably require 15 minutes to create the semantic maps, 60 minutes to identify key terms, and 2 hours to review challenges and modify the program. Reviewing only challenges and selected essays might take an average of 13.5 minutes for every 10th essay or 1.35 minutes each.

## Case 3: A half-page Essay
For example, a graduate student relatively experienced at using SAGrader timed how long it took for him to create a moderately complex essay question about the social psychological concept, "groupthink." For this assignment, he created a total of 19 codes including "groupthink," its definition, eight "symptoms of groupthink," and a definition for each of those symptoms. The prompt for the assignment looked something like the following:

> *What is groupthink? Define the concept, identify some of the symptoms of groupthink, and explain each of those symptoms.*

Creating the knowledge base (creating the codes and appropriately linking the codes) took 5 minutes. Adding synonyms for the 8 feature codes took 30 minutes. Allowing one hour to review any challenges related to this essay and scan unchallenged essays for problems related to it, this requires a total of 1.58 hours to develop the essay, examine student answers for problems, and refine the program until it grades satisfactorily. Reviewing only challenges and selected essays might take an average of 5 minutes for every 10th essay or 0.5 minutes each.

An optimistic cost analysis for grading the same essays by hand estimates it would take 1 minute to determine a numeric score for the essay, three minutes to write comments to the student, and 1 minute to record the grade and enter it into a grade book.

## Break-Even Analysis
Both hand-grading and grading with SAGrader have fixed times required to set up the essays and variable times required to grade each essay. We can represent the times required for grading by the two methods as

$$Th = Fh + (N*Vh)$$
$$Ts = Fs + (N*Vs)$$

Where T is the total time to grade (Th = by hand, Ts = with SAGrader), Fh and Fs are the fixed times for developing, Vh and Vs are the times required per essay, and N is the number of essays submitted. In traditional hand-grading, the time required to create a question and rubric are usually very small relative to the time required to grade each essay. In contrast, with SAGrader, the fixed setup time is likely to be large compared to the effort required to monitor each essay, as in our examples above.

In order for SAGrader to break even with hand grading, the time required to grade with the program would have to equal the time required for traditional hand-grading. That is,

$$Th = Ts$$

Substituting from the above equations, this becomes

$$Fh + (N*Vh) = Fs + (N*Vs)$$

This is equivalent to

$$N*(Vh - Vs) = Fs - Fh$$

We can replace Vh-Vs with ΔV, the difference in per-essay time required for grading; and we can replace Fs-Fh with ΔF, the difference in fixed time to set up for grading. Solving for the number of essay submissions required for SAGrader to break even with hand-grading, the equation then becomes

$$N*\Delta V = \Delta F \quad \text{or} \quad N = \Delta F/\Delta V$$

For example, if the average time required to grade an essay by hand is 10 minutes (Vh = 10), the average time per essay with SAGrader is 1 minute (Vs = 1), and the added fixed setup costs for SAGrader are 200 minutes (ΔF = 200), then

$$N = \Delta F/\Delta V = 200/(10\text{-}1) = 200/9 = 22.22$$

For 23 or more submissions the SAGrader program would require less time than grading by hand. For 22 or less, SAGrader would require more time.

We can now use this approach to estimate the breakeven numbers for our three sample essay assignments discussed above. The table below includes these estimates for each case and computes the number of students required for the SAGrader program to roughly equal the cost of hand grading.

| SAGrader time estimate | Case 1: Simple Essay | Case 2: 2-page Essay | Case 3: ½-Page Essay |
|---|---|---|---|
| concept map | 2 | 15 | 5 |
| synonyms | 5 | 60 | 30 |
| training | 60 | 120 | 60 |
| Monitoring (minutes per submisison) | 0.03/sub | 1.35/sub | 0.5/sub |
| TOTAL | 67 mins + 0.03/sub | 195 + 1.35/sub | 95 + 0.5/sub |
| | | | |
| Hand-grading time estimate | | | |
| score | 1/12 | 8 | 1 |
| comment | 1/6 | 5 | 3 |
| summarize | 1/20 | 0.5 | 1 |
| TOTAL | 0.3 mins/sub | 13.5/sub | 5/sub |
| | | | |
| Number of submissions required to break even, N = ΔF/ΔV | 67/0.27 = **249** | 195/12.15 = **17** | 95/4.5 = **22** |
| Time saved by SAGrader for Classes of Size 40 | -56 mins | 6 hrs, 51 mins | 1 hr, 25 mins |
| Time saved by SAGrader for Classes of Size 250 | 0 mins | 47 hrs | 17 hrs |

These example cases help clarify the circumstances in which SAGrader can be cost-effective.

**Complexity and length**. For the simplest and briefest essay, the point where costs of SAGrader break even with costs of hand grading is a relatively large 249. However, for both the ½-page and 2-page essays, the breakeven point is 22 submissions and 17 submissions. This analysis suggests that the more complex the essay question and the longer the essay, the sooner SAGrader will become less costly than hand grading. For both the ½ and 2-page essays, SAGrader should become less expensive than hand grading the first time it is used, even in classes of 25 or fewer students.

**Class size.** Obviously, since SAGrader costs are primarily fixed costs unrelated to the number of students, the larger the class size the greater will be the benefits of using SAGrader. For a class of size 40, for example, with these assumptions and Case 2, the SAGrader program would save 6 hours and 51 minutes. For a class of 250 students, SAGrader would save over 47 hours.

We can use the formula, $N = \Delta F/\Delta V$, to estimate the breakeven point for SAGrader under a range of different values for $\Delta F$ and $\Delta V$. The figure below shows the results for $\Delta F$ ranging from 10 to 100 and $\Delta V$ ranging from 1 to 32. Here $\Delta F$ is plotted along the X-axis, and $\Delta V$ is represented by the different lines.
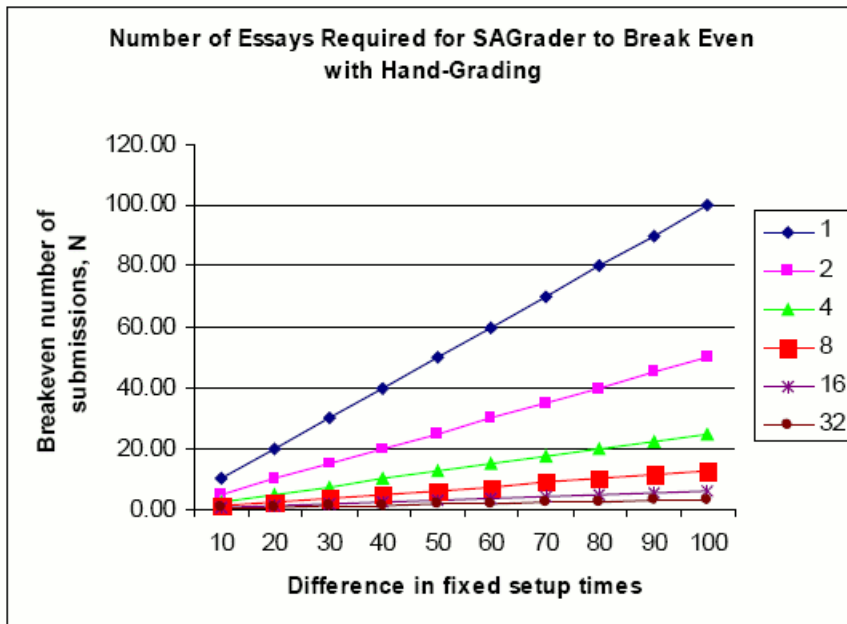


This equation and graph help us examine the circumstances in which SAGrader will be more efficient than hand grading. The fixed time required to set up SAGrader is almost always greater than the fixed time to compose an essay question for hand grading. So SAGrader begins with a disadvantage relative to hand grading due to its higher setup costs. SAGrader can make up this deficit and surpass hand grading in efficiency only if the per-essay times required for grading with SAGrader are lower than those for hand-grading. The larger the difference in initial setup times, $\Delta F$, the greater will be the number of essays required to break even. The larger the difference in times required to grade each essay in the two methods, $\Delta V$, the smaller will be the number of essays required to break even. For example, when $\Delta V$ is 1 the breakeven number of submissions is equal to $\Delta F$. When $\Delta V$ is 2 the breakeven number of submissions is half $\Delta F$. Because SAGrader requires more time to setup, SAGrader can <u>only</u> be cost-effective if it reduces the instructor time required to grade each essay.

**Strategies for Reducing Costs and Increasing Benefits**

This analysis suggests a number of strategies that can reduce the costs and increase benefits of using SAGrader.

INCREASING BENEFITS

1. **Increase the number of essays graded**. Clearly, if SAGrader makes it less expensive per essay, the time and cost savings are increased as the number of essays graded increases. Since the costs of developing automatically graded assignments with SA-Grader are generally onetime costs and do not increase as more students use those assignments, the cost-savings of SAGrader can be enhanced by reusing assignments. This can be accomplished in many ways. SAGrader can be most cost-effective by targeting first high-enrollment classes, classes with multiple sections, or classes where it can be used in subsequent semesters. A particularly effective strategy is to use coursepacks that have already been developed and tested for a course. Another way to maximize the payoff from developing assignments can occur when instructors develop their own coursepacks and make them available to other instructors for a royalty. If SAGrader is not cost-effective for a single semester of a small course, then any of these strategies can be used to increase the benefits with little increase in cost.

2. **Permit multiple revisions**. The earlier cost-benefit analysis assumes students submit only one draft of their essay. With traditional hand grading multiple revisions are often not feasible because of the added costs to grade them. However, once SAGrader has been developed and tested for a specific assignment, the added cost to grade multiple submissions is miniscule. If students are permitted to submit multiple drafts, as SAGrader permits, the break-even point and the cost-benefit advantages of SAGrader relative to hand grading are even greater. For example, in one recent course using SAGrader, students were permitted to submit as many drafts of each essay as they wished. They had an average of 2.4 submissions per essay. The table below shows how the breakeven number of <u>students</u> may be much lower than the number of <u>submissions</u>.

### BREAKEVEN NUMBERS FOR SUBMISSIONS VERSUS STUDENTS

|  | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| **Number of submissions required to break even** | 249 | 17 | 22 |
| **Number of students required to break even (with 2.4 submissions per student)** | 104 | 8 | 10 |

*When multiple revisions are permitted SAGrader becomes cost-effective in the first semester even for small classes*. Even more importantly, an analysis of student performance when multiple revisions are permitted shows that students improve their grades dramatically from a C- average to an A- average.

3. **Re-purpose questions for use in different assignments from one semester or course to another**. SAGrader permits assignments to be constructed from different combinations of questions. So questions designed for one assignment initially can be incorporated into other assignments. For example, a question asking students to describe a historical event could be an entire assignment, or could be part of an assignment asking students to compare and contrast two historical events, or could be used as part of an assignment asking students to identify and describe important events in the life of one of the major participants in that event. Concept maps, and the key terms identified for particular concepts can also be re-used for other questions and assignments.

Once questions have been tested and refined in one assignment, they can generally be re-used in other assignments with little or no additional testing and refinement. There should be few problems and few student challenges requiring additional instructor attention. Using this

strategy, as few as 20 questions covering major issues in a chapter, for example, could be recombined in various ways to generate literally hundreds of possible assignments with little or no need for further testing and refinement. Re-purposed questions can be used to make fresh assignments in subsequent semesters to cut down on plagiarism, can become parts of assignments for other courses, and can be revised to take advantage of recent events.

4. **Reduce training and monitoring costs**. Two of the most time-consuming aspects of developing and testing essays for use with SAGrader are training the program initially and monitoring the program as students submit their work. The most common problem is when SAGrader does not correctly detect important features in student essays. For some features there are literally hundreds or thousands of legitimate expressions. This problem is best resolved by taking maximum advantage of the program's fuzzy logic through the use of key terms or short phrases rather than long ones, adjusting weights to minimize false negatives and false positives, and using negative weights to rule out similar but incorrect phrases.

   **There is a tradeoff between training and monitoring**. Where training focuses on logically possible combinations of terms that express concepts, monitoring focuses on actual phrases used by students in their submissions. With more training, there should be less need for maintenance, the program should perform better from the start, and there should be fewer student challenges. However, in our experience no amount of training can completely assure that all legitimate student answers will be recognized. There must be at least some monitoring of SAGrader as it is used. Less training shifts effort to monitoring, and focuses on real student responses rather than hypothetical training responses, but increases student challenges and possibly student frustration.

   **Monitor efficiently**. Reviewing every student essay by hand will drastically reduce the cost effectiveness of the program. It is much more efficient to selectively review assignments rather than reviewing them all, reviewing more at first until you are confident the program is working well. The most effective way to make monitoring more efficient is to focus on student challenges. Students are very good and generally very willing to point out when they believe the program did not recognize a correct answer. Challenged essays are far more likely to demonstrate a problem needing correction than unchallenged essays.

   **Revise rather than override**. When student challenges identify real problems it is more effective to revise the program to correctly identify the feature rather than overriding the program's grade. When the program is revised it automatically re-grades every student's submission for the assignment reducing the chances of further challenges for the same problem and thereby reducing the time required to monitor. Overriding the program's score does not solve the problem for other students and locks the complaining student into that score, preventing them from benefiting from further improvements in the program. Overriding the program should be reserved for situations where it would be too time-consuming to revise the program and the problem is likely to be rare.

   **Not all student challenges identify a problem with the program**. In our experience, roughly half the time the program is right and the student is wrong. Often student challenges identify problems already identified by other students. So quick responses correcting the issue can help prevent further similar challenges.


**Other costs and benefits for instructors**

Time is not the only way to measure costs and benefits of SAGrader for instructors. Another cost that should be considered is the computer expertise required to create and test assignments and the time required to become familiar with the program. These costs are much smaller for faculty who are comfortable working with computers and can be greatly diminished by using existing coursepacks in your discipline.

There are other advantages offered by SAGrader. SAGrader helps make the logistics of collecting papers, grading them, recording grades, and giving feedback to students much more manageable. Because the program tracks submissions, students can no longer claim they turned in a paper that the instructor lost. SAGrader is an environmentally friendly paperless system. Instructors can easily monitor class progress and identify students needing more help, identify which learning objectives are met and which are not and then do remedial work as necessary. SAGrader also provides a wealth of data that can be used by instructors to study how their course is going or to assess the impact of changes in pedagogical strategies.

## Costs and Benefits for Other Stakeholders

The cost-benefit analysis focused on costs and benefits from the perspective of faculty since they typically make the decision whether or not to use SAGrader. However, costs and benefits should also be considered for other important stakeholders, including students and administrators.

### Students
Students using SAGrader like the opportunity to revise and resubmit their essays, they like the immediate, detailed, personalized feedback SAGrader provides while their answer is still fresh in their minds. There are no longer any complaints about waiting weeks to get an essay graded only to have minimal comments scribbled in the margins. They appreciate that their grades are objective, unbiased, and consistent. They like submitting essays any time of day or night from anywhere via the Internet. SAGrader permits more writing in larger classes and for more assignments. Writing assignments provide more authentic learning experiences than traditional multiple choice tests. SAGrader makes logistics more manageable for students, reduces paper consumption, and lets students easily submit assignments, view feedback, revise, check on their grades, and communicate with their instructor.

Student costs include the fee they pay to use the program and more importantly, the time they spend writing various drafts. The fee is analogous to the costs students face for supplementary texts, CDs, or student responses systems. We doubt most students will complain of the cost when they find that SAGrader can help them dramatically improve their grade in the course. Our analysis shows that when students are permitted to revise and resubmit their papers they improve their grades by as much as 20%. SAGrader also levels the playing field, permitting disadvantaged students to overcome their initial deficits to perform roughly the same as other students.

### Administrators
Administrators, including department chairs, deans, and institutional research staff, can also benefit from SAGrader. The costs of using SAGrader from the perspective of administrators are small. Students can pay the fee or it can be incorporated into standard fees. Costs of training faculty or staff to use the program and setting up assignments are small and quickly made up in cost-savings from more efficient grading. Of course administrators benefit if students learn more and are pleased with the program. SAGrader permits greater emphasis on writing as a more authentic learning experience that should improve the quality of instruction and provide a good selling point for recruiting students. The objectivity of SAGrader should reduce any claims by students that an instructor is biased or unfair in their grading. That objectivity and the detailed data and reports it provides offer an

opportunity to monitor multiple sections of the same course to assure that students are receiving comparable learning experiences. They can identify which learning objectives are being met and which sections are performing below the norm and intervene to make improvements. This can be particularly important for teacher training and quality assurance when graduate students or adjunct instructors are teaching some of their first courses and need extra monitoring and assistance. They can be used to monitor courses over time to identify trends. SAGrader can be integrated with course management systems to pass information to grade book or e-portfolio systems.

**Discussion**

This cost-benefit analysis uses estimates of parameters for both SAGrader and hand grading that may be higher or lower than found in a particular case. These are only estimates and may be wrong. However, the three cases help us understand the sensitivity of the models and the general analysis identifies trends and relationships that should be correct even if particular estimates prove to be inaccurate. By exercising some care it should be possible to use SAGrader in a wide range of courses to provide a win-win-win situation where faculty, students, and administrators all see net benefits.

Students get the opportunity to improve their grades and learn more in return for a modest additional expense and some investment of their time. They experience a learning environment in which they can play a more active role, drafting papers and revising them based on timely and detailed feedback. They may see more timely intervention by their instructor as the information gathered from SAGrader flags students having difficulties. Their grades are more objective and fair. Students can maximize their benefits from SAGrader by taking advantage of the opportunity to revise and resubmit their work to continue until they achieve an acceptable grade (which is precisely what students do, by the way). The benefits are greatest for students who are willing to work hard, interested in learning more, and students who come from disadvantaged backgrounds. Student benefits are less when the program is used only a little, revisions are not permitted, the students are unwilling to invest additional effort to improve their grade, or instructors fail to monitor the program and revise as necessary. Instructors can maximize the benefits of SAGrader for students by using the program extensively in their classes, permitting students to revise and resubmit their work as often as they like, making SAGrader assignments a large part of the class grade, and responding to student challenges in a timely manner.

Faculty can use SAGrader to find ways to improve learning while reducing time spent grading. Instructor benefits of the program are less if it is used on inappropriate problems that are intractable or difficult to implement, if it is used in small classes and only single submissions are permitted, and if totally new questions must be constructed and tested each time without benefiting from prior work. Faculty can get the most out of SAGrader by using it in courses where it can produce the greatest benefits, such as high-enrollment courses or courses with multiple sections. They can produce the greatest learning by permitting students to revise and resubmit their work. They can take advantage of the data and reports provided by SAGrader to monitor student progress, identify students having problems, and intervene.

Administrators can see significant benefits from the use of SAGrader for quality control and training. Administrators will find it less useful if it is used for small portions of courses and counts little, if each class is different, and if development costs are not spread over multiple courses or sections. Administrators can see greatest benefits from SAGrader if it is used in large courses to improve the educational experience, if it is used with the same or comparable assignments for multiple sections of courses to provide for quality assurance, and if instructors share their materials and maximize the benefits of their investment.

Appropriately used, SAGrader should be able to significantly improve the educational experience and enhance student learning, often paying for itself in its first semester of use.