AUTOMATED STUDENT ASSESSMENT PRIZE: Phase 1 & Phase 2

A Case Study to Promote Focused Innovation in Student Writing Assessment

Jaison Morgan
Mark D. Shermis
Lynn Van Deventer
Tom Vander Ark
# Table of Contents

## Executive Summary

## Background

## Introduction

## How Did We

- Who Is Driving Change? ................................................................. 04
- Exhibit: Better Tests Support Better Learning .................................. 05
- Automated Student Assessment Prize ............................................ 06

## Why Prizes Work

- Leverage Funds .............................................................................. 08
- Mobilize Talent ............................................................................. 08
- Innovate .......................................................................................... 09
- Influence ....................................................................................... 09
- Why a Prize in Education? ................................................................. 10

## ASAP Phase One: Long-Form Constructed Response

- Private Vendor Demonstration ....................................................... 12
- Public Competition ......................................................................... 13
- Vendors and Winners ................................................................. 14

## ASAP Phase Two: Short-Form Constructed Response

- The Challenge ............................................................................... 15
- The Teams ..................................................................................... 16
- The Success .................................................................................. 17

## What’s Next?

- Pricing Study ................................................................................ 19
- Business Incubation Prize ............................................................... 19
- Phase Three—Math and Logical Reasoning .................................. 19
- Classroom Trials ........................................................................... 19
- Item Development ......................................................................... 20

## Conclusion

## Appendix A: ASAP Phase One Winners

## Appendix B: ASAP Phase Two Winners

## Appendix C: Participating Organizations

## Appendix D: Resources

## Authors Bios

## Endnotes
The Automated Student Assessment Prize (ASAP) invites data scientists worldwide to take on the challenge of creating new approaches to high-stakes testing for state departments of education. Sponsored by the William and Flora Hewlett Foundation, the first two phases of ASAP were designed to demonstrate the ability of existing summative writing assessment products and to accelerate innovation in machine scoring. Phase one focused on the ability of technology to assess long-form constructed responses (essays) while phase two focused on short-form constructed responses (short answers). ASAP was designed to answer a basic question: Can a computer grade a student-written response on a state-administered test as well as or better than a human grader?

The results of these two studies present unique opportunities to:

- Establish standards for state departments of education to utilize assessment technologies.
- Advance the field of machine scoring in the application of student assessment.
- Introduce new players with different and disruptive approaches to the field.

ASAP is now expanding to other challenges in machine-scoring applications, including the development of systems to support writing instruction in a classroom setting. ASAP remains committed to its role as an open, fair and impartial arbiter of machine scoring and writing assessment capabilities through a series of scientifically rigorous studies and field trials. ASAP does not endorse or promote any specific technology or provider. Instead, ASAP seeks to deliver critical information to administrators, educators, families and students during a time when student assessment is undergoing a critical shift in schools across America.
Race-to-the-Top (RttT) made up about five percent of the education allocation of the American Recovery and Reinvestment Act, the federal stimulus bill passed in 2009, and almost five percent of that was devoted to supporting development of state assessments aligned with Common Core State Standards. In 2010, two consortia, representing 46 states, received $330 million from the U.S. Department of Education to support a new generation of online assessments that would provide better measures of college and career readiness, offer faster feedback and be affordably priced. Having observed the impact of low-quality assessments on instruction in American classrooms, the Hewlett Foundation recognized the opportunity these new tests could provide to better prepare students for college and career.

To fulfill an aggressive mission—to create better tests and to advance the field—the two phases of ASAP were designed to test vendor claims of efficacy and to determine whether new solutions for machine scoring could be found. Technology providers had long claimed that their systems offered effective and cost-efficient alternatives to hand scoring of student-written responses on state-administered tests. The ASAP team invited the eight leading providers of those technologies, along with another open-source solution developed in a university lab, to compete against each other to determine which solutions most closely approximated hand scores of student-written responses. In addition, the ASAP team invited data scientists from around the world to build new software systems capable of achieving the same or similar outcomes and compete against each other for substantial cash rewards. By compelling existing private providers to demonstrate the effectiveness of their products and by challenging new and diverse players to develop alternative approaches, ASAP committed substantial resources to offer the first comprehensive, unbiased and broadly defensible investigation of machine scoring technology.

ASAP mobilized global talent and produced new innovative approaches to the challenge. According to many key opinion leaders, the results of phase one shocked the field as new players introduced competitive alternatives to the slate of existing products. The investigation proved that, in the case of the machine scoring of student essays, the systems could match the ability of expert graders. The results of phase two, in which the focus shifted to machine scoring of short-answer responses, were not as promising. In both cases, though, ASAP either validated the hypothesis that it will be possible to use a range of machine scoring tools to support more constructed-response items on state-administered tests and established a baseline understanding of those capabilities so that further investment could reveal opportunities to advance the field.
INTRODUCTION

Each spring, millions of students across America sharpen their number two pencils, ready their erasers and sit down to take their states’ assessments of student learning. While the majority of states offer some form of online assessment, most states do not mandate their use. Students primarily fill in bubble sheets, answering multiple-choice questions. A few states require written responses, but most do not.

Why does it matter?

Thinking critically. Communicating effectively. Mastering core academic content. Working collaboratively. Learning to learn. These five core learning outcomes comprise what The William and Flora Hewlett Foundation calls “deeper learning.” To be competitive in the 21st century, students must excel in these skills. Today’s bubble-sheet multiple-choice assessments do not adequately measure students’ progress in these areas, nor do they give teachers the timely information they need to shape individualized instruction. Written essays and short-answer response items can be a better measure of student progress, but they are expensive and time-consuming to score. The Hewlett Foundation is making investments to encourage deeper learning to improve student outcomes and to better prepare students for college and career. Improving assessments is an essential part of this effort, since the Hewlett Foundation believes that better learning is supported by better tests. This is especially important during this critical moment, as the implementation of the Common Core State Standards and the shift to online assessments in the 2014–2015 school year presents an unprecedented chance to create opportunities for students to practice deeper learning and find new ways to measure progress toward a wider band of college- and career-readiness outcomes.
What if there were a way for states to have better tests that cost less to score and produce quicker results? By 2010, the introduction of inexpensive laptops and tablets suggested that it would be cost effective to create high-access environments that would support online assessment. There was also growing evidence that automated scoring could support delivery of tests with more constructed-response items.

In partnership with state testing consortia and under the advisement of leading foundations, the Automated Student Assessment Prize (ASAP) set out to evaluate the current state of automated testing and encourage further developments in the field. The goal of the competition was to assess the ability of technology to assist in grading essays and short answers included in standardized tests. If software could grade these types of items quickly, inexpensively and reliably, then states would be able to include more writing on their summative assessments.

After the 1989 National Education Summit, many states developed systems of standards and assessments. By the mid-1990s, year-end standardized testing was common. In 2001, the reauthorization of the Elementary and Secondary Education Act, referred to as No Child Left Behind, mandated that all schools that accept federal funding administer a statewide standardized test in reading and math each year to all students in grades 3-8. States set their own standards and developed their own assessments. As a result, test scores aren’t comparable across the country, and many of them don’t reflect real college- and career-ready expectations.

The standards movement was initially animated by a rich vision of competency-based learning, authentic assessment, performance demonstrations and student portfolios. Unfortunately, this approach rapidly became expensive and unwieldy, resulting in end-of-year exams in which students primarily answer multiple-choice questions. As accountability pressure mounted and funding of No Child Left Behind was slashed to one-third of the budgeted amount, many schools abandoned a broad curriculum and resorted to a focus on test preparation.

Most state departments of education now exclude written responses on state assessments in favor of multiple-choice questions, which are less suited to assess critical reasoning and writing skills. The states that do include written responses face significant cost and slow turnaround in the scoring of those responses, as nearly all of them are scored by hand.

In many classrooms, what is tested is what is taught. The same problem that led to a reliance on multiple-choice testing—the difficulty and time-intensiveness of grading essays—led to a reduction in the amount of writing in American classrooms. Despite vendor claims of efficacy and more than three decades of
use, writing assessment software has been slow to sell, in part because of the relatively high price, some rather clunky interfaces and limited student access to technology.

Steady progress has been made in adaptive reading and math products with increased philanthropic and private investment over the last five years. That was not the case for writing assessment. The grant-funded opportunity to develop new tests highlighted this market gap.

WHO IS DRIVING CHANGE?

As part of Race-to-the-Top (RttT), two state testing consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), received $330 million from the U.S. Department of Education to develop new assessments to test Common Core knowledge and skills on behalf of 46 member states. The consortia set out to develop better tests with more constructed-response tasks than most states could afford at a price point of about $20 per student—half or a third of what some states with the best tests currently pay. Both consortia assumed that the aggressive budget would require that many of the new assessment items be machine scored.

Both consortia have committed to using online assessment to make the next generation of state tests less expensive to administer with faster turnaround. Most importantly, developments in machine scoring have made it possible to include a significant amount of writing on these new tests, as well as to introduce the possibility of new and innovative performance tasks. The new tests will reflect the deeper learning aspirations of the Common Core State Standards. The online tests are intended to be operational in 2014—2015.

Constructed-response tasks—long and short written responses to complex text and challenging questions—are important ways for students to demonstrate critical reasoning.
Education experts say writing teaches students the critical reasoning skills that they will need to compete in the new century. In part, the move towards more writing is both an affirmation of its inherent promise and a counterpoint to the current approach to assessment, which has been consumed with reciting facts and patterns and less concerned with developing critical reasoning skills. In Oceans of Innovation (2012), Sir Michael Barber and colleagues summarize the new demands:

Innovation requires, first of all, people with the right skills and attributes. In the modern world, individuals must be creative, tenacious and passionate, striving for excellence and the pursuit of new ideas. Regurgitation of existing knowledge, the historical focus of education, is no longer sufficient.

In June 2012, the National Research Council released a study, Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century, which poses a new definition for deeper learning: “The process through which a person becomes capable of taking what was learned in one situation and applying it to new situations—in other words, learning for ‘transfer.’”

“The term ‘deeper learning’ may be new, but its basic concepts are not,” said Bob Wise, president of the Alliance for Excellent Education and former governor of West Virginia, in reaction to the NRC report in Straight As newsletter. “Deeper learning is what highly effective educators have always provided: the delivery of rich core content to students in innovative ways that allow them to learn and then apply what they have learned.” The NRC report confirms that this type of education—once available to only a few elite students—is now necessary for all.

Constructed-response tasks—long and short written responses to complex text and challenging questions—are important ways for students to demonstrate critical reasoning. Yet, over a third of high-school seniors report writing essays requiring analysis or interpretation only a few times a year. If we can increase the quantity and quality of student writing, we can teach them critical reasoning and communication skills that better prepare them for college and career.

One reason that teachers may assign essays infrequently is that grading them is both time consuming and increasingly difficult. If a typical teacher is now instructing a single class of 20–30 students, and—in some cases—those same teachers are carrying 2–4 of those classes, a single long-form essay assignment of 1,000 words (4–5 pages) can generate between 100 and 600 pages of written material. Asking teachers to read, grade and provide substantive feedback on each one represents a formidable commitment of time and attention. The burden should not be underestimated. Providing tools that can assist students in the writing process can be a way to help students learn these necessary skills. Just as math teachers have grown accustomed to using calculators when grading quizzes and performing math instruction functions, new technology can help support the instruction of writing. We can realize this potential by pursuing a rigorous and healthy investigation of these technologies.

“Better tests support better learning,” said Barbara Chow, Education Program Director at the Hewlett Foundation. “Rapid and accurate machine essay scoring will encourage states to include more writing in their state assessments. And the more we can use essays to assess what students have learned, the greater the likelihood they’ll master important academic content, critical thinking and effective communication.”
AUTOMATED STUDENT ASSESSMENT PRIZE

ASAP was designed to demonstrate current machine scoring capability and to focus and accelerate innovation in assessment technology. ASAP is aligned with the aspirations of the Common Core State Standards and seeks to accelerate assessment innovation to help more students graduate with the skills necessary to succeed in college and career.

Sponsored by the Hewlett Foundation, the first two phases of ASAP sought to inform key decision makers who are considering adopting or developing automated scoring systems by providing a fair, impartial and open series of trials to test current capabilities and to drive greater awareness when outcomes warrant further consideration. Combining demonstrations of existing products and open challenges, ASAP invited data scientists worldwide to take on the challenge of improving student assessment capabilities.

As states implement the Common Core State Standards, state leaders are making decisions about what kind of next-generation assessments can measure new expectations quickly with fidelity. Innovative software that can faithfully replicate grading and feedback from trained experts offers a new opportunity to conduct better tests at a lower price than is common today.

If automated scoring systems are proven accurate and affordable, states can use them to incorporate more questions that require written responses into new assessments. Making it easier and more cost effective to include those prompts on standardized tests, experts predict, will drive students to spend more time writing and learning to write in the classroom.

ASAP was designed and directed by Jaison Morgan and Tom Vander Ark. Lynn Van Deventer was the project manager. The project’s principal investigator and lead academic advisor is Dr. Mark D. Shermis, a professor in the College of Education at The University of Akron, the author of Classroom Assessment in Action and the co-editor of Automated Essay Scoring: A Cross-Disciplinary Approach. ASAP was hosted on Kaggle, Inc., the leading platform for predictive modeling competitions. Kaggle’s platform helps companies, governments and researchers identify solutions to some of the world’s hardest problems by posting them as competitions to a community of more than 53,000 data scientists located around the world.
Why Prizes Work

By sponsoring ASAP, the Hewlett Foundation appealed to data scientists worldwide to help focus and accelerate innovation in student assessment. But why choose a prize? Well-constructed prizes produce four benefits:

- **Leverage funds.** Prizes motivate participants to invest time and energy in solving a problem they might not otherwise consider. Prizes are usually performance based and only paid out once a viable solution is demonstrated.

- **Mobilize talent.** Prizes spark the interest of diverse groups of professionals and students. Many prizes are won by scientists several degrees of separation from the subject sector. Prizes are an extremely efficient strategy for mobilizing diverse talent impossible to locate using conventional approaches.

- **Innovate.** The cross-pollination of participants from different backgrounds and with different skill sets unleashes creativity, allowing problem solvers to generate fresh ideas. The use of leader boards and discussion tools promotes transparency and competition, but it also inspires collaboration and innovative discovery.

- **Influence.** The results of prize competitions can garner public attention and influence key decision makers. Good prizes result in news-worthy mobilization and breakthrough outcomes that result in press coverage that can be worth more than the prize purse.
LEVERAGE FUNDS

Prizes motivate participants to invest time and energy in a problem they might not otherwise tackle. And because they’re working towards a deadline, participants quickly dive into the work, spending tens of hours if not hundreds of hours of their personal time. Based on competitor reports, we conservatively estimate that the 250 participants in phase one of ASAP invested time worth more than $7.5 million, and the 187 participants in phase two invested $4.7 million to win a combined purse of $200,000—a total of more than $12 million dollars in research and development time within two three-month time periods. The nine organizations participating in the private vendor demonstrations collectively spent an additional $200,000 to demonstrate the current capabilities of their software. There is simply no other philanthropic strategy that can induce the same level of short-term leverage.

MOBILIZE TALENT

The Kaggle, Inc. website notes that “By exposing the problem to a large number of participants trying different techniques, competitions can very quickly advance the frontier of what’s possible using a given dataset.” Prizes create opportunities for smart people to apply their skills and knowledge to complex problems.

A research fellow in glaciology, a teaching assistant in Slovenia and an actuary all competed in the ASAP. Data scientists from around the globe—in Poland, Singapore, Australia, London and elsewhere—registered for the competition, downloaded the data and designed algorithms to score student responses.

HOW IT WORKS

Machine-scoring software is not independently cognitive; it is composed of either generic algorithms designed to address specific tasks or new formulas trained on particular data sets to replicate expert human grading patterns. More training data (in other words, more graders) typically yields higher interrater reliability and yields more accurate machine scoring. This is an important distinction, and it is for these reasons that ASAP has refrained from using such industry terms as “artificial intelligence” or “robo reading.” It is our position that the application of machine scoring for student assessment shows important promise, but the only relevant application is to support assessment and instruction, not to supplant recognized best practices among teachers and other expert instructors. No matter how valid or reliable they may be, because of the constantly shifting standards for what makes good or great writing, these systems cannot work without a productive relationship between teachers or graders and those responsible for building and adapting the systems.
Individuals competed on their own, but more often than not, participants organized teams that met in the discussion forums. For phase one, a British particle physicist, a data analyst for the National Weather Service in Washington, D.C., and a graduate student from Germany comprised the first-place team.

Certainly the cash reward is a motivator, but there’s more to it than just the money. Well-constructed prizes create interesting challenges. The Kaggle leaderboard recognizes status among a global community. ASAP winners said they like to solve puzzles and contribute to the greater good.

INNOVATE

Prizes spur innovation. The different backgrounds, education and skill sets of the competitors enrich the field to which they’re applying their knowledge. One of the earliest prizes, the Longitude Prize of 1714, sought a solution to the problem of measuring longitude at sea. It’s likely that the British Parliament, the sponsors of the prize, assumed that an astronomer or cartographer would come up with the answer. No one expected that John Harrison, a British clockmaker, would create the marine chronometer.8

The important lesson is that most innovation is translational—something that worked in one field may work in another. The question is how to expose cross-discipline and cross-sector innovation? Prizes are an efficient means of promoting translational innovation.

Data scientists and statisticians in the open competitions deployed different scoring strategies from the natural language processing approach typically taken by vendors in the field. Some of the winners from ASAP phase one are already helping some of the big testing companies improve their services. In phase two, short-answer scoring, competitors were required to open source their code and provide instruction manuals with the goal that others will continue to build on their successes.

INFLUENCE

Both phases of ASAP have been developed with the support of and with the intent to benefit PARCC and SBAC, the RttT state testing consortia that will roll out new online tests in 2014–2015. Both consortia aim to offer much higher-quality assessments than are common today but at much lower prices. To accomplish these objectives, the consortia plan to incorporate scoring solutions similar to those used to license doctors (United States Medical Licensing Examination) and admit students into graduate schools.

Prior to ASAP, it was unknown how these systems compared to human graders. In its role as a fair and impartial arbiter of machine scoring systems, ASAP encouraged the participation of major vendors in these competitions and
introduced them to new talent, resulting in significant breakthroughs in scoring technology. The results of both competitions gained the attention of national news organizations, including National Public Radio and the *New York Times*.

Additionally, ASAP has been able to address some of the criticisms leveled at these systems. Common complaints include the ability of students to “game” the software, getting a good score just by writing a longer essay, and the concern that machine scoring can’t address specific content. By designing prizes to examine those questions, ASAP has helped the consortia understand the current capabilities of machine scoring.

Prizes allow organizations an opportunity to highlight both weaknesses and strengths in a particular field and share the results of the work with a broader community.

**WHY A PRIZE IN EDUCATION?**

While mega-prizes have helped to break open targeted industries, including space exploration\(^9\) and other technical fields, they have not been common in education. As has been true in the fields of chemistry, material sciences and data analytics, a sequence of small, targeted prizes appears to be a promising strategy to produce focused innovations in education.

When designing ASAP, our team felt that influencing the decisions of public-sector leaders required careful timing, rigor and relevance. Rather than looking for a singular breakthrough to radically shift the field of education, ASAP phases one and two set out to make the case for specific tools at a specific time. Key decision makers needed relevant information to make a shift that will change the face of student assessment dramatically within a short time frame.

Prizes work best when the problem is well defined, metrics are quantifiable and not in dispute, and there is a market path to take the innovation to scale. For the two ASAP prizes, we knew scoring of assessments was costly and time-consuming. During the prize design, we defined the metric used to measure success: the quadratic weighted kappa. And lastly, pairing competition winners with established vendors blends the knowledge and skills of the winners with organizations who have the infrastructure to service large state departments of education.

Better tests support better learning. Rapid and accurate machine essay and short-answer scoring will encourage states to include more writing in their state assessments. The more essays can be used to assess what students have learned, the greater the likelihood they’ll master important academic content, as well as critical thinking and effective communication.
Phase one of ASAP demonstrated, via a correlation between aggregate scores generated by human graders and those produced by software designed to score student essays, that the software achieved virtually identical levels of accuracy as human graders, and in some cases proved to be more reliable than hand scoring.

Competition participants had access to 22,000 hand-scored essays that varied in length, type and grading protocols; the competitors were challenged to develop software designed to replicate in a construct-relevant way the assessments of a trained grader. The essays came from six participating state departments of education and encompassed writing assessment items from three grade levels: 7, 8 and 10. The items were evenly divided between source-based prompts (that is, essay prompts developed on the basis of provided source material) or those drawn from traditional writing genre (narrative, descriptive, persuasive). Essays varied in length and scoring protocol.

Phase one was conducted in two segments. The first demonstrated the capabilities of eight commercial providers who market software for grading essays to states, districts and schools; one university team that offered an open-source machine scoring application also participated. The second segment was open to the public, with $100,000 in prize money awarded to the top three competitors.

Establishing standards is an important priority for ASAP. In order to deliver a prize competition, it was critical to identify a unifying metric of system capabilities, a statistical measure of the concordance between hand scores and machine scores. ASAP selected quadratic weighted kappa (QWK) for that purpose. While the study published additional measures of distribution and multiple measures of correlation (for instance, Pearson r, Kappa and others), our emphasis is on QWK. During multiple conferences on the subject of those capabilities (including, NCME, NCSA, TILSA and others), industry product-development teams informed us that QWK is quickly becoming the industry standard for measuring summative student assessment in the aggregate. This development cannot and should not be underestimated. ASAP intends to further advance methods of establishing baseline capabilities among states, the consortia and other partners.
PRIVATE VENDOR DEMONSTRATION

For more than 15 years, companies that provide machine essay-scoring software have claimed that their systems can perform as effectively and affordably as—and faster than—other available methods of essay scoring. The ASAP study was the first comprehensive multivendor trial to test those claims. Eight companies, together constituting a high percentage of the current market, and one university shared the capabilities of their systems. Companies approximated established scores by using their commercial software. The university used an open-source solution.

A study of the demonstration results, authored by Dr. Shermis and Ben Hamner, a data scientist at Kaggle, Inc., found that the scoring software was able to replicate the scores of trained graders with the software in some cases proving to be more reliable. The study was released on April 16, 2012, at the annual conference of the National Council on Measurements in Education. The report will also appear as a chapter in the Handbook of Automated Essay Evaluation: Current Applications New Directions, to be published in spring 2013.

PHASE 1: Essay Scoring Leaderboard Progression

The leaderboard demonstrates how more than 20 teams were able to quickly improve their essay scoring accuracy to surpass that of industry leaders in a period of less than 90 days.
PUBLIC COMPETITION

A British particle physicist and sports enthusiast, a data analyst for the National Weather Service in Washington, D.C., and a graduate student from Germany won the $60,000 first prize in the public competition. None of the team members have a background in education. Their collaborative effort brought together the team’s diverse skill set in computer science, physics and language to create the most innovative, effective and applicable testing model. The competition drew more than 2,500 entries and 250 participants and inspired data scientists to develop innovative, accurate ways to improve on the current standard of essay-scoring technology.

Participants competed to develop new software that could score student essays from state standardized tests that had already been individually graded by hand. The winning team came closest to replicating the expert graders’ results. The open competition website included an active leaderboard to document prize rules, provide regularly updated results and host discussion threads between competitors.

The data that was used to construct the ASAP phase one trials are presented in Table 1.

**TABLE 1: Phase One Test Set Characteristics**

<table>
<thead>
<tr>
<th>DATA SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Grade</td>
</tr>
<tr>
<td>Type of Essay</td>
</tr>
<tr>
<td>M # of Words</td>
</tr>
<tr>
<td>SD # of Words</td>
</tr>
<tr>
<td>Type of Rubric</td>
</tr>
<tr>
<td>Range of Rubic</td>
</tr>
<tr>
<td>Range of RS</td>
</tr>
<tr>
<td>MRS</td>
</tr>
<tr>
<td>SDRS</td>
</tr>
<tr>
<td>Quadratic Weighted</td>
</tr>
</tbody>
</table>

RS - Resolved Score  
*composite score based on four of six traits  
+composite score based on six of six traits  

Source: Handbook of Automated Essay Evaluation: Current Applications and New Directions Edited by Mark D. Shermis, Jill Burstein
Another goal of ASAP is to spur innovation. In May of 2012, the ASAP team hosted a reception at which all eight assessment vendors met members of the three winning teams. As a result, Pacific Metrics entered into an agreement to acquire the first-place technology, which will be integrated into the company’s machine-scoring engine, CRASE. Additionally, staff from Measurement Incorporated joined with members of the third-place team to participate in ASAP phase two, which was focused on short-form constructed responses. Their combined team outperformed all other participants in the second competition.

The members of the winning team of ASAP phase one hail from three different countries: England, Germany and the United States. Their collaborative effort brought together the team’s diverse skill set in computer science, physics and language to create the most innovative, effective and applicable testing model among more than 250 participants. The team, Stefan Henß, Jason Tigg, and Momchil Georgiev, says they believe they have just barely scratched the surface of possibilities with software scoring technology.

Jason Tigg, the spokesman for the team, said, “I am thrilled to win this contest because it gave me an opportunity to think creatively about how we can use technology to improve scoring software that instantly and inexpensively predicts how an educator would hand grade an essay. I enjoyed working on a real-life problem that has the potential to revolutionize the way education is delivered.”
ASAP phase two took on the more difficult challenge of scoring, with a high degree of accuracy, constructed responses of less than 150 words. The goal of the competition was to assess the ability of technology to assist in grading short-answer essays in standardized tests. A range of answer types was included to better understand the strengths of specific solutions. The contest revealed that the software performed well although not to the same level of accuracy as human graders.

THE CHALLENGE

Participants in the competition had access to more than 27,000 hand-scored short-answer responses that varied in length, type and grading protocols. The essays came from three participating state departments of education and encompassed assessment items from two grade levels: 8 and 10. On average, each answer was approximately 50 words in length. Some responses were more dependent upon source materials than others, and the answers cover a broad range of disciplines (from English Language Arts to Science). Essays varied in scoring protocol.

The data used to construct the ASAP phase two trials are presented in Table 2.

### TABLE 2: Phase Two Test Set Characteristics

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>558</td>
<td>426</td>
<td>318</td>
<td>250</td>
<td>599</td>
<td>599</td>
<td>601</td>
<td>601</td>
<td>600</td>
<td>548</td>
</tr>
<tr>
<td>Grade</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Type of Answer</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
<td>Non-Source Dependent</td>
<td>Non-Source Dependent</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
<td>Source Dependent</td>
</tr>
<tr>
<td>Subject</td>
<td>Science</td>
<td>Science</td>
<td>English-LA</td>
<td>English-LA</td>
<td>Biology</td>
<td>Biology</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>Science</td>
</tr>
<tr>
<td>M # of Words</td>
<td>48.47</td>
<td>58.36</td>
<td>48.51</td>
<td>38.62</td>
<td>22.62</td>
<td>24.08</td>
<td>40.96</td>
<td>54.18</td>
<td>51.33</td>
<td>41.13</td>
</tr>
<tr>
<td>SD # of Words</td>
<td>21.92</td>
<td>23.71</td>
<td>15.10</td>
<td>16.26</td>
<td>18.14</td>
<td>21.26</td>
<td>24.82</td>
<td>33.53</td>
<td>40.09</td>
<td>27.33</td>
</tr>
<tr>
<td>Type of Rubric</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
<td>holistic</td>
</tr>
<tr>
<td>Range of Rubric</td>
<td>0-3</td>
<td>0-3</td>
<td>0-2</td>
<td>0-2</td>
<td>0-3</td>
<td>0-3</td>
<td>0-2</td>
<td>0-2</td>
<td>0-2</td>
<td>0-2</td>
</tr>
<tr>
<td>Range of Score</td>
<td>0-3</td>
<td>0-3</td>
<td>0-2</td>
<td>0-2</td>
<td>0-3</td>
<td>0-3</td>
<td>0-2</td>
<td>0-2</td>
<td>0-2</td>
<td>0-2</td>
</tr>
<tr>
<td>M Score</td>
<td>1.53</td>
<td>1.71</td>
<td>0.98</td>
<td>0.68</td>
<td>0.31</td>
<td>0.27</td>
<td>0.76</td>
<td>1.16</td>
<td>1.11</td>
<td>1.22</td>
</tr>
<tr>
<td>SD Score</td>
<td>1.01</td>
<td>1.03</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
<td>0.67</td>
<td>0.83</td>
<td>0.85</td>
<td>0.78</td>
<td>0.68</td>
</tr>
<tr>
<td>Quadratic Weighted</td>
<td>0.95</td>
<td>0.93</td>
<td>0.77</td>
<td>0.75</td>
<td>0.95</td>
<td>0.93</td>
<td>0.96</td>
<td>0.86</td>
<td>0.84</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Source: *Handbook of Automated Essay Evaluation: Current Applications and New Directions* Edited by Mark D. Shermis, Jill Burstein
187 participants across 150 teams from around the world tackled the difficult problem of scoring short-answer responses. They were challenged to develop software to replicate the assessments of trained expert graders using multiple rubrics. The systems do not independently assess the merits of a response; instead, they predict how a person would have scored the response under optimal conditions.

The competing teams developed their systems over three months and shared their technical approaches through an active discussion board. The competition drew more than 1,800 entries, including those from three commercial vendors. For this competition, Measurement Incorporated, a company that provides achievement tests and scoring services, partnered with the third-place team from the first competition, allowing them to outperform all other teams. The $100,000 prize was divided among four individuals and one team. Measurement Incorporated was not eligible for prize money as the company did not open source the code for its solution.
THE SUCCESS

Phase two of the ASAP demonstrated that software that scores short-answer responses has great potential. Documentation and code for the winning submissions has been released under an open license to enable others to build on the outcomes and advance the field of machine assessment. Use of these systems today could supplement and support the work of grading experts.

A mechanical engineering student with interest in aerospace, Luis is a newcomer to data science. He participated in Coursera’s massively open online course (MOOC) in machine learning, which was taught by Stanford’s Andrew Ng, and it ignited a passion. His solution placed 13th in the first phase of ASAP, equaling the performance of vendors who have been in the testing business for decades. After doing so well in data competitions, he’s applying to top graduate programs in computer science.

Luis said, “I am excited to win this contest because it gave me an opportunity to create a futuristic program that reads an essay, finds the answers being asked and scores it as a human would do. I’m hopeful that my model will help advance the field of scoring software so that computers can assist teachers, who can then use the results to provide even more individualized instruction to their students.”
ASAP strives to incentivize and demonstrate new tools to support K-12 teachers and schools. Competitions in development aspire to address several needs:

- Studying the cost of machine scoring on behalf of the state testing consortia.
- Incubating new companies through a business planning competition that will result in viable new commercial enterprises, some of which may build upon the open-source machine scoring systems that resulted from ASAP phase two.
- Launching ASAP phase three, which will focus on the machine scoring of math and logical reasoning items.
- Delivering a demonstration of new assessment tools that teachers can use to support more writing with structured feedback in the classroom.
- Designing competitions that address difficult-to-score items and promote the use of innovative tasks that demonstrate college- and career-ready skills by leveraging new technology.

The ASAP sponsors, advisors and collaborators believe that new tools will help teachers promote more writing, better problem-solving skills and deeper learning.
PRICING STUDY

Next up for ASAP is a study of the current costs to implement machine scoring solutions for testing student academic achievement on high-stakes summative exams. Machine scoring solutions vary in cost and type. The ASAP study will develop a Request for Information (RFI) that includes multiple scenarios against which known providers of machine scoring solutions will submit pricing. The responses received from providers will be compiled into a report that will be shared with PARCC and SBAC. Through this study, ASAP hopes to drive competitive pricing and ultimately lower the cost of assessments for state departments of education.

BUSINESS INCUBATION PRIZE

Building on the success of phase two, ASAP is planning a business plan and incubation competition to help launch start-up ventures. Phase two produced several accurate writing assessment engines that have potential for further development. The competition will likely include an informational campaign, development of a collaborative online community, some incubation support and a prize, as well as the potential for follow-on investment.

PHASE THREE—MATH AND LOGICAL REASONING

In a continuation of the summative assessment work, ASAP phase three will be a competition to improve machine scoring of symbolic mathematical and logic reasoning items. This prize focuses on accelerating development of scoring engines for multistep problem solving, as well as use and interpretation of charts and graphs. Symbolic mathematical representations, including charts and graphs, are the most challenging type of data for machine scoring solutions to address, but this also represents a promising field of applications for future state testing.

CLASSROOM TRIALS

Technologies exist today that are designed to give teachers tools to assess student writing at each phase of the writing process, offer critical and timely feedback and provide the data teachers need to inform their instruction. Such formative assessment is key to improved teaching and learning. However, these technologies have not yet been tested in a fair, uniform and transparent way to inform schools, teachers, parents and students of their capabilities. The ASAP classroom trials are designed to test the capabilities of software systems that claim to support writing instruction in the classroom.
First, we will deliver an open call to commercial providers of available technologies to submit descriptions (or demonstrations) of system attributes, so that we may select only those applications that are most appropriate for a classroom setting. Then, we will place the selected systems in schools and deliver the mechanisms that will allow us to compare them both among each other and across control groups within the same settings.

ITEM DEVELOPMENT

As PARCC and SBAC work towards delivering online assessments aligned to the Common Core State Standards, it’s also time for the consortia to begin thinking about including new approaches to items. ASAP is considering development of a series of challenges to encourage teachers to create great testing items, both to source innovative alternatives and to give teachers a voice in the state assessment process. Using the currently established framework of test item templates and other rules, teachers can offer many more options than are already under consideration. Likewise, we can expand the range of possibilities for test items by pushing further into the capabilities of online assessments and inviting nontraditional contributors to drive even more innovative thinking.

The 2015 tests are just the beginning of what should be an age of innovation in formative and summative assessment. With rapid changes in touch computing and simulation, there is a world of possibility when it comes to innovative assessment items and tasks. Increasing access to touch-enabled tablet computers creates potential for game-based items, graphical manipulations and simulations. As access and computing power increase, the consortia should introduce new testing versions at least every other year.

ASAP can play a role in identifying and developing innovative item types that explore the use of animation and manipulation. Performance tasks—scenario-based application of skills to complex problems that may take 1–2 hours to complete—can be challenging and time-consuming to create. In one example, students are given a list of background links on nuclear power and asked to assume the role of a congressional chief of staff. They are required to conduct research, develop a list of pros and cons and create a press release in support of a position. The rubric-scored task considers purpose, focus and organization as well as use of evidence and elaboration.

For more information on the national shift to the next generation of online assessments, check out "Getting Ready for Online Assessment" and "The Countdown to 2014" from Digital Learning Now!
A series of prizes could be used to develop a bank of high-quality performance tasks within a very short window of time.

The Hewlett Foundation sponsored the ASAP to address the need for high-quality standardized tests to replace many of the current ones, which test rote skills. The goal is to shift testing away from standardized bubble tests to tests that evaluate critical thinking, problem solving and other 21st-century skills. To accomplish this, it’s necessary to develop more sophisticated tests that evaluate these skills and to reduce their cost so they can be adopted widely. Machine scoring can play an important role in achieving this goal.

ASAP has already set a standard measurement (QWK) and, through these first trials, will inform critical decision makers when and how to adopt standards. Here, we offer the outcomes of the first two phases (green) and show how machine scores compare to hand scoring (blue):

ASAP phase one demonstrated that essay-grading software performed extremely well when scoring long-form constructed responses. This information paves the way for states to include more writing on standardized tests, which will lead to more writing in the classroom. And if students are writing more in the classroom, they’re acquiring the 21st-century skills that the global economy demands.
Can a computer grade a student-written response on a state-administered test as well as or better than a human grader?

ASAP phase two determined that software that scores short-form constructed responses did not do as well as expert graders, yet shows great potential. Since the advent of ASAP nearly a year and a half ago, the competition has inspired participants to develop innovative and accurate ways to improve on currently available scoring technologies. We’ll continue to see advancement in the field of machine scoring as existing vendors team up with prize winners and others and as new players continue to enter the field.

Well-constructed prizes mobilize global talent and spur innovation. A college student from Ecuador competed against some of the best testing companies in the world after ten weeks of work. Young men (yes, it still is mostly young men) from Slovenia to Singapore, from Pittsburgh to Poland poured hundreds of hours a week into the competition hoping to see their names creep up a leaderboard. Winners from phase one teamed up with a testing company for phase two to edge out the rest of the competition. Prizes work.
Members of the winning team include:

Jason Tigg – A resident of London, Tigg applies his scientific expertise to predicting financial movements at work and programming games in his free time. Armed with a PhD in particle physics from Oxford University, Tigg is also a champion runner and rower—he won last year’s Barnes and Mortlake Regatta held on the Thames River.

Stefan Henß – The team’s talented rookie, Henß brought his expertise in language and semantics analysis to help guide the team to success. Henß is currently pursuing a Master’s degree in computer science from the Darmstadt University of Technology in Hesse, Germany.

Momchil Georgiev – Georgiev, the son of two teachers, applied his deep appreciation for learning and a passion for data analysis to the team’s winning entry. Born in Bulgaria, Georgiev earned a Master’s degree in computer science from Johns Hopkins University and now works as an engineer for the National Weather Service in Washington, D.C.

The runner-up team stretches across the globe, with members from the United States, Canada and Australia. Members include:

Phil Brierley, based in Melbourne, Australia, has a PhD in engineering and artificial intelligence. Brierley created a popular data-mining system called Tiberius that is a leading competitor in the Heritage Health Prize, a two-year competition to improve health care.

Eu Jin also lives in Melbourne and brought his data-mining and fraud investigation experience to the team.

William Cukierski is a PhD candidate in biomedical engineering at Rutgers University. Cukierski has used his data expertise to predict everything from stock movements to grocery shopping trends.

Christopher Hefele, who currently works for AT&T as a systems engineer, is not new to scoring high in Kaggle competitions—he was part of the team that took home second place in Netflix’s million-dollar competition to improve its movie recommendation engine.

Bo Yang lives and works as a software engineer in British Columbia, Canada, and previously finished first in Kaggle’s photo quality prediction contest.

The third place team is an American duo of data experts. Members include:

Vik Paruchuri is a data modeling and predictive modeling consultant expert who is currently writing a book about statistical programming. Paruchuri served overseas as a U.S. Foreign Service Officer for the State Department. Unlike most data experts, Paruchuri got his degree in American History.

Justin Fister’s educational background in psychology and computer science fueled his interest in the ASAP competition. Fister has worked in the software industry for more than ten years.
APPENDIX B: ASAP PHASE TWO WINNERS

Winners include:

Luis Tandalla, 1st place – Originally from Quito, Ecuador, Luis is currently a college student at the University of New Orleans, Louisiana, majoring in mechanical engineering. A newcomer to data science, Luis’s first experience was one year ago when he took an online machine learning course from Dr. Andrew Ng. Luis also participated as part of a team in phase one, placing 13th.

Jure Zbontar, 2nd place – Jure lives and works in Ljubljana, Slovenia, where he is a teaching assistant at the Faculty of Computer and Information Science. He’s pursuing a PhD in computer science in the field of machine learning. Besides spending time behind his computer, he also enjoys rock climbing and curling.

Xavier Conort, 3rd place – A French-born actuary, Xavier runs a consultancy in Singapore. Before becoming a data science enthusiast, Xavier held different roles (actuary, CFO, risk manager) in the insurance industry in France, Brazil and China. Xavier holds two masters’ degrees and is a Chartered Enterprise Risk Analyst.

James Jesensky, 4th place – With more than 20 years’ experience as a software developer, James currently works in the field of financial services near Pittsburgh, Pennsylvania. He enjoys data competitions because they allow him to combine his computer science expertise with his life-long love of recreational mathematics.

The fifth place team was an international duo of data experts. Members include:

Jonathan Peters, 5th place – Based in the United Kingdom, Jonathan works for the National Health Service as a public health analyst. He spends most of his time modeling death and disease; Kaggle competitions offer some light relief.

Paweł Jankiewicz, 5th place – Paweł lives in Poland and works as a banking reporting specialist. His machine learning experience began when he attended Dr. Andrew Ng’s online machine learning class in 2011. Apart from Kaggle, he enjoys English audiobooks, especially the Wheel of Time series.
APPENDIX C: PARTICIPATING ORGANIZATIONS

American Institutes for Research (www.air.org)
Carnegie Mellon University (http://url.ie/f16y)
CTB/McGraw-Hill (www.ctb.com)
Educational Testing Service (www.ets.org)
Measurement Incorporated (www.measinc.com)
MetaMetrics (www.metametricsinc.com)
Pacific Metrics Corporation (www.pacificmetrics.com)
Pearson Education, Inc. (www.pearson.com)
Vantage Learning (www.vantagelearning.com)
RESOURCES

Automated Student Assessment Prize (www.scoreright.org)
Common Core State Standards (www.corestandards.org)
Elementary and Secondary Education Act (No Child Left Behind) (www.ed.gov/esea)
Getting Smart (www.gettingsmart.com)
Kaggle (www.kaggle.com)
National Council of Teachers of English (www.ncte.org)
National Writing Project (www.nwp.org)
Partnership for Assessment of Readiness for College and Careers (www.parcconline.org)
Race to the Top Fund (www.ed.gov/programs/racetothetop/index.html)
Smarter Balanced Assessment Consortium (www.smarterbalanced.org)
The Common Pool (www.commonpool.org)
The William and Flora Hewlett Foundation (www.hewlett.org)
AUTHORS BIOS:

JAISON MORGAN
Managing Principal of The Common Pool, LLC
Jaison Morgan has been hailed by the BBC as the world’s leading expert in designing innovative prizes. He led the ASAP team during phases one and two. As the Managing Principal of The Common Pool, LLC (TCP), he has worked with such clients as the government of the United Arab Emirates, the First Minister of Scotland, the U.S. Department of Commerce, the Bill and Melinda Gates Foundation and private commercial partners. Mr. Morgan completed graduate studies at the University of Chicago and continues to lecture on the subject of incentive engineering.

DR. MARK D. SHERMIS
Professor, College of Education, University of Akron
Dr. Mark D. Shermis is the principal investigator and academic advisor for ASAP. Dr. Shermis is a frequently cited expert on machine scoring and co-author of Classroom Assessment in Action. He is presently a professor in the Department of Educational Foundations and Leadership and the Department of Psychology at The University of Akron, where he previously served as Dean of the College of Education. Dr. Shermis has also held faculty positions at the University of Florida, Florida International, IUPUI and the University of Texas. Shermis earned a B.A. in developmental psychology at the University of Kansas and Doctorate and Master’s degrees in educational psychology at the University of Michigan.

LYNN VAN DEVENTER
Project Manager, ASAP
Lynn Van Deventer is the project manager for ASAP. Prior to this competition, Lynn led strategy and community engagement projects for Seattle Public Schools. Prior to her work in education, Lynn spent 15 years as a program manager in the software and data management industry. Lynn earned a BA in English from the University of Iowa.

TOM VANDER ARK
Executive Editor, Getting Smart
Tom is the author of Getting Smart: How Digital Learning is Changing the World and the executive editor of GettingSmart.com. He is also a partner in Learn Capital, a venture capital firm that invests in learning content, platforms and services with the goal of transforming educational engagement, access and effectiveness. Previously, he served as president of the X PRIZE Foundation and was the executive director of education for the Bill and Melinda Gates Foundation. Tom was also the first business executive to serve as a public school superintendent in Washington State. Tom is a director of the International Association for K-12 Online Learning (iNACOL) and several other nonprofits.
ENDNOTES:

7. Assuming $125 an hour, 20 hours per week over 12 weeks for phase one and 10 weeks for phase two.