National Assessment Governing Board
Assessment Development Committee

May 17-18, 2012
AGENDA

Thursday, May 17, 2012		
Noon – 4:15 pm	**Closed Session: Noon – 4:15 pm** **ACTION:** Review of Secure NAEP Technology and Engineering Literacy (TEL) Tasks *Lonnie Smith, ETS* *Committee Discussion*	Secure material provided under separate cover
Friday, May 18, 2012 Closed Session 10:00 am – 12:00 pm		
10:00 – 11:15 am	Welcome, Introductions, and Agenda Overview *Alan Friedman, ADC Chair* **ACTION:** Continued Review of Secure NAEP Technology and Engineering Literacy (TEL) Tasks *Committee Discussion*	Secure material provided under separate cover
11:15 – 12:00 pm	Briefing on NAEP Mathematics Special Studies • Math Computer-based Study (MCBS) • Knowledge and Skills Appropriate Study (KaSA) *Gloria Dion, ETS*	Attachment A.1 Attachment A.2
12:00 – 12:30 pm	**Open Session** Discussion of Expert Panel Report on NAEP Background Variables *Mary Crovo, Governing Board Staff*	See Reporting and Dissemination Committee Tab, Attachment D
Information Item	Phase 1 Results: Hewlett Foundation Automated Student Essay Scoring Prize	Attachment B
Information Item	NAEP Item Review Schedule	Attachment C

Mathematics Computer Based Study (MCBS)

Introduction and Goals

In 2011, the National Center for Education Statistics (NCES) conducted a special study called the Mathematics Computer Based Study (MCBS) to start assessing the benefits of adaptive testing for the National Assessment of Educational Progress (NAEP), and to develop knowledge and experience about implementing an operational adaptive testing in the context of a group-score assessment. Adaptive testing in this context means that performance during earlier parts of an administration (e.g., stage 1) is used to determine what items a student receives during later parts (e.g., stage 2). The goal is to match the difficulty level of the test as closely as possible to the performance level of the student. A student who does not answer many questions correctly on an initial set of items subsequently receives a less difficult set of items, while a student who gets many questions correct on the initial set receives a more difficult set of items. The psychometric models used in NAEP, specifically Item Response Theory (IRT) models, make it statistically possible to adjust properly for the varying difficulty of the different sets of items. Therefore, results obtained under adaptive testing should be equally valid and comparable to those obtained under the current matrix sample design.

The following research questions were pursued in this study:

1. How do the results (group averages and achievement level percentages) obtained under one approach to adaptive testing (i.e., multi-stage testing, MST) compare to those obtained under the current design from the main assessment?
2. To what degree did the MST approach increase the precision (i.e., reduce the standard error) of the group-level results reported by NAEP? How successful was the MST adaptive approach used in this study at routing students to the optimal item set in terms of their performance level?

In addition to these questions, the study collected some information about engagement to assess whether a test targeted towards a student's ability level increases the level of engagement with the assessment.

Research Design

The study was conducted with the nationally representative sample of 8th graders using standard NAEP sampling procedures. The total instrument consisted of five blocks of mathematics items from the 2011 operational and pilot assessments... Only items that could be translated directly from a paper- to a computer-based format were selected. The relatively small percentage (i.e., 23%) of items in NAEP that require drawing, producing complex equations, and using auxiliary materials (e.g., protractor) were not included in the study. The five blocks included two routing blocks and three targeted blocks. The assessment itself was designed as a two-stage test, where one of the two routing blocks was administered in stage 1 and one of the targeted blocks in stage 2 for a total of two blocks per student – the same number of blocks given to each student in the main NAEP assessment. The routing blocks were in terms of the distribution of difficulties similar to operational blocks. The targeted blocks were Easy, Medium, and Hard, targeting low, medium, and high performers, respectively.

Despite the aforementioned restrictions on the item pool, the instrument does reflect the content distribution targets (i.e., the proportion of items in each subcontent area) described in the framework. However, the item pool for the MCBS did have a lower proportion of constructed-response items than does the full NAEP item pool. In particular, the first-stage (or routing) blocks consisted entirely of multiple-choice items to facilitate immediate scoring without the need for automated scoring engines.

Percentage of items distributed across content areas by block and across blocks, including the framework targets for the assessment

Block	Numbers & Operations	Measurement	Geometry	Data Analysis, Statistics, & Probability	Algebra
Routing Block A	18	18	18	12	35
Routing Block B	24	18	18	12	29
Easy Block	19	13	19	13	38
Medium Block	19	13	19	19	31
Hard Block	19	13	19	19	31
Total	**20**	**15**	**18**	**15**	**33**
Framework Target	**20**	**15**	**20**	**15**	**30**

A total of 8,400 students participated in this study, which used an experimental design. About 40% of the students were randomly selected in the experimental sample and, therefore, 60% in the control sample. This distribution was by design to ensure a target sample of 1,500 students per item in the control group. The adaptive design used was a Two-Stage Test, containing two

distinct stages and a single decision point. In the experimental group, students received one of the two routing blocks during the first stage and, based on their performance, either the Easy, Medium, or Hard block was presented during the second stage. In the control group, the second stage block was randomly assigned--not based on performance. Figure 1 graphically represents the design of the study in terms of routing and routing decisions. The delivery system captured all student-computer interactions, including time stamps.

Status and Schedule

The core analyses for this study have been completed and a summary will be provided during a closed session of the committee. These results will include:

- Basic performance differences between conditions
- Routing accuracy and routing percentages by student group
- Differences in measurement error at the student and group levels

Extended analyses are currently being completed, which include the use of response time to detect engagement as well as the analysis of response patterns, independently and in relation to performance. In addition, a research memorandum is under development that provides, in addition to the core results, details about scaling methodologies and considerations for student group estimation.

Plans for Future Research

At this point, no specific plans for future research have been finalized. In terms of item development, the focus is changing towards computer based assessments, particularly in terms of taking advantage of technology, and meeting the statistical requirements associated with developing effective multi-stage tests. In terms of design, some simulation work is ensuing around determining optimal designs and using effective measures to evaluate different designs. In addition, some further work is required that focuses on effectively maintaining trends under an adaptive approach.

Routing sequence and design of the study

Experimental Group

Control Group
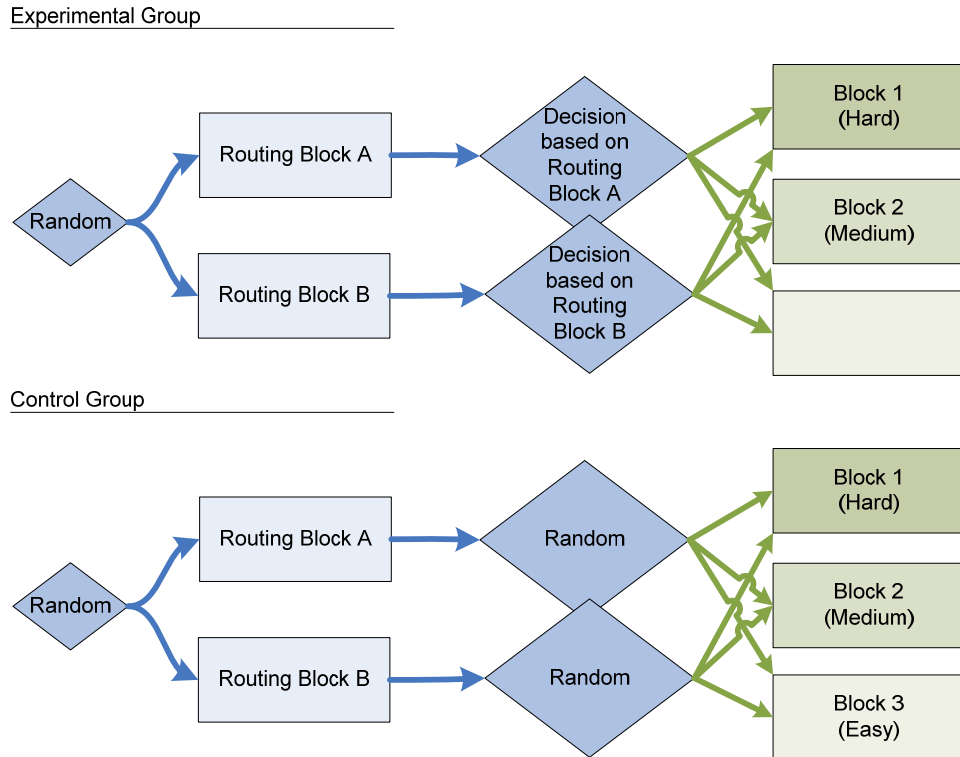
Knowledge and Skills Accessible Study (KaSA)

Introduction and Goals

The National Assessment of Educational Progress (NAEP) is often characterized as an assessment program of broadly defined constructs that is focused on measuring a wide range of performance levels. Over the last decade, this range has expanded considerably with the introduction of the Trial Urban District Assessment as well as the NAEP Mathematics assessment of Puerto Rico in grades 4 and 8. Since 2003, various Puerto Rico assessments have been conducted and they have not been without challenges. While several procedures were modified to address some of the challenges (e.g., different translations, additional assessment time), the core challenge is that a typical NAEP mathematics assessment measures those student groups very well that have average abilities at the middle and upper ends of the NAEP scale, but it is not geared toward the lower end. Combined with the generally low performance levels observed in Puerto Rico public schools, the result is below chance-level performance and high non-response. This, in turn, has yielded unstable, implausible average scores, particularly when looking at trends. To address this misalignment of the NAEP mathematics instrument for student groups with abilities near the lower end of the ability scale, NCES developed the Knowledge and Skills Accessible (KaSA) study with the goals of measuring low performing groups with reasonable accuracy and reporting results from Puerto Rico on the NAEP scale. Note that the desire to better measure low performing groups is a more general goal, beyond Puerto Rico.

As part of the study, KaSA items were developed to address a targeted subset of the NAEP mathematics framework, representing subtopics and objectives in appropriate proportions. While KaSA items are written to address framework objectives, the pool of items does not span the breadth of the framework. In terms of item types, the number of multiple choice items is relatively large in the KaSA item pool; and approximately 70% of the items are of low mathematical complexity, as defined in the framework, and the remainder are of moderate complexity. In comparison, operational assessments have a target of 25% low complexity. For each grade, 60 KaSA items were developed and placed in four 15-item KaSA blocks. The KaSA items were translated to Puerto Rican Spanish for administration in Puerto Rico.

Research Design

In 2011, the KaSA blocks were administered to a representative sample of public school students in Puerto Rico. The goal of the investigation was to report average scores for Puerto Rico on the main NAEP Mathematics scale. Three booklet types were developed: a pair of KaSA blocks, a KaSA block paired with an operational block, and a pair of operational blocks. The scale was developed based on operational items only; then the KaSA items were placed on the scale. To further strengthen the desired link between KaSA and the main assessment, as well as to investigate other potential uses of KaSA items outside of Puerto Rico, a special national U.S. sample also received KaSA books along with books that paired KaSA blocks with main assessment blocks. Below is a table that clarifies the various components and a figure that provides a visual schematic of the components. Sample sizes for the Puerto Rico components were approximately 4,400, while the national components yielded 6,800 and 4,600 for grade 4 and 8, respectively.

Instrument and sample components of the 2011 KaSA study

Sample	Instrument	Contents	Number of Books	Percentage of Students Assessed
	KaSA	Two KaSA blocks	12	41%
Puerto Rico	Mixed	One KaSA, one operational	16	41%
	Main	Two operational blocks	10	18%
	KaSA	Two KaSA blocks	12	45%
National	Mixed	One KaSA, one operational	16	55%
	Main	Main Assessment	50	N/A (150k+)

Schematic of components of the 2011 KaSA study
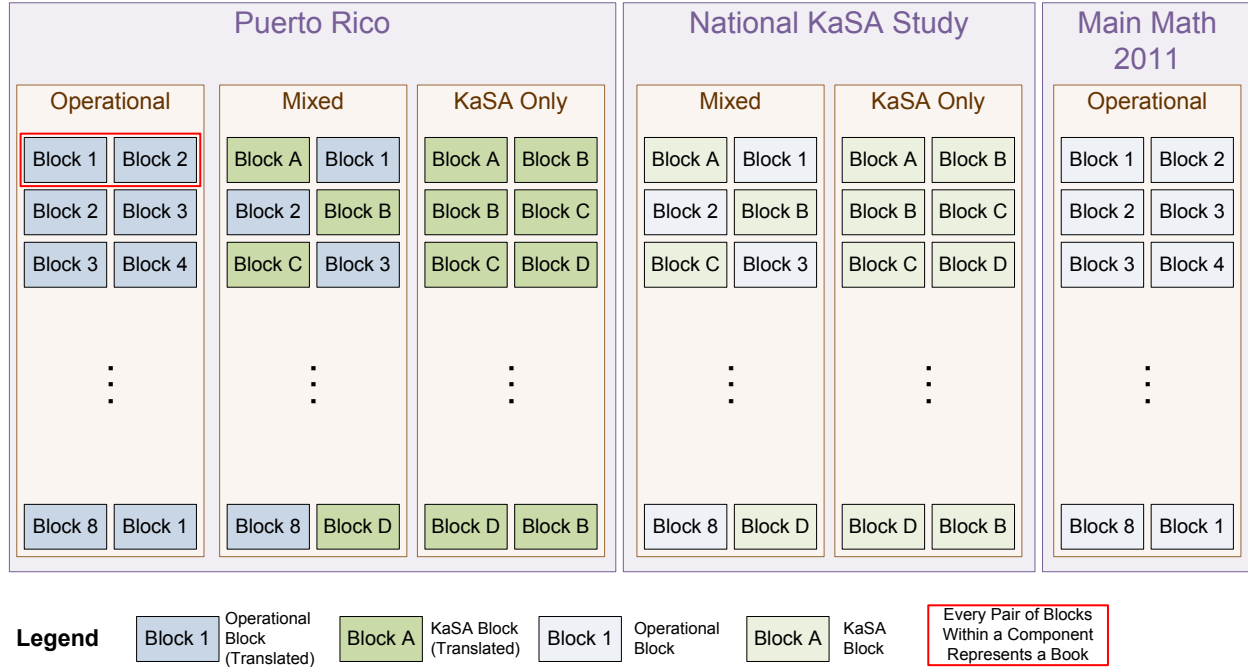
Current Status and Results

The analysis has been completed and results have been discussed with Puerto Rico representatives. The results themselves are under embargo, but the following findings in relation to the measurement aspects of the study can be shared:

- The KaSA item pool yielded lower omit rates and a larger above-chance level student performance compared to the operational items in the Puerto Rico sample.
- The KaSA item pool provided more precise measurement of the performance levels typically found in Puerto Rico compared to the operational item pool.
- Better model-data fit could be obtained in Puerto Rico using the KaSA items than was found for the operational items.
- It appears that Puerto Rico results can be placed on the NAEP scale through the KaSA items and the links established through the national sample. However, given the history of performance on NAEP by students in Puerto Rico, it will be necessary to evaluate the stability of these findings across years to verify the stability of the estimates.

These points will be discussed in more detail during the presentation.

Future Plans

In terms of next steps for Puerto Rico, it is critical to evaluate the success of KaSA in terms of trend as indicated above. Therefore, a replication of the study, using the same KaSA blocks and instrument and sample design, is planned for 2013. Outside of Puerto Rico, these blocks could be used to include more students with an Individualized Education Plan and/or designated as English Language Learner. A special study was conducted in 2011 and it was shown that increased participation could be obtained if KaSA items were available. Finally, development of KaSA and similar efforts serve the goal of enabling NAEP instruments to measure a wider range of abilities accurately and to provide exemplars of what students typically know and can do at various levels located at the lower end of the performance scale. For example, KaSA blocks could serve well as targeted later-stage content for a multi-stage testing approach to NAEP. That is, students exhibiting relatively low performance during a first-stage block could be routed to a KaSA block during the second stage.

Hewlett Foundation Essay Scoring Competition

Note to ADC: This material is being presented in May as an information item only. We are planning a briefing and discussion in August on activities related to this essay scoring competition.

Recent newspaper articles on the Hewlett Foundation competition

New York Times
http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html?pagewanted=all

USA Today
http://www.usatoday.com/news/education/story/2012-04-23/essay-scoring-computer-software/54493662/1

Earlier Press Releases

Hewlett Foundation Sponsors Prize to Improve Automated Scoring of Student Essays
Prize to Drive Better Tests, Deeper Learning
January 9, 2012

MENLO PARK, Calif. – The William and Flora Hewlett Foundation will award a $100,000 prize to the designers of software that can reliably automate the grading of essays for state tests, Foundation education officials announced today.

The software competition is intended to begin to solve the problem of the high cost and the slow turnaround resulting from the time consuming and expensive task of hand scoring thousands of essays for standardized tests. These obstacles typically mean that many school systems exclude essays in favor of multiple-choice questions, which are less able to assess students' critical reasoning and writing skills. The problem is that critical reasoning is one of a suite of skills that experts believe students must be taught to succeed in the new century. The Hewlett Foundation makes grants to educators and nonprofit organizations in support of what it calls "deeper learning," which embraces the mastery of core academic content, critical reasoning and problem solving, working collaboratively, communicating effectively, and learning how to learn independently.

"Better tests support better learning," says Barbara Chow, Education Program Director at the Hewlett Foundation. "Rapid and accurate automated 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."

The competition will determine if current software scoring programs are as effective as expert human scoring and seeks to accelerate innovation for faster and more accurate scoring of student work. If the programs can be shown to be as reliable as human scoring it will increase their acceptance and reduce the need to rely exclusively on costly and time-consuming human scoring.

The competition will be conducted with the support of the two state testing consortia: the Partnership for Assessment of Readiness for College and Careers and Smarter Balanced Assessment Consortium, which together work with forty-four state departments of education. The two testing consortia recently received $365 million from the U.S. Department of Education to develop new assessments.

The competition will be conducted in two phases. The first will demonstrate the capabilities of existing vendors who create and market software for grading essays. The second phase will be open to the public and will award prize money to competitors who demonstrate software that can score essays as well as human graders.

Open Education Solutions, a blended learning service provider that helps educators combine the best of online and classroom work, and The Common Pool, a consulting business that specializes in developing effective incentive models for solving problems, designed and will manage the competition. Tom Vander Ark, CEO of OpenEd, says, "Prizes are a proven strategy for mobilizing talent and resources to solve problems." "We're excited about the potential of emerging assessment capabilities," says Jaison Morgan of The Common Pool, "and confident that focused incentives will accelerate innovation."

Dr. Mark Shermis, Dean of the University of Akron College of Education, author of Classroom Assessment in Action, and noted expert on automated scoring, will chair the Academic Advisory Board.

The competition will be hosted on Kaggle, a platform for predictive modeling competitions that 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 25,000 PhD-level data scientists located around the world. "Kaggle has solved problems for NASA, insurance industry leaders, and HIV researchers," says Anthony Goldbloom, founder and chief executive officer of Kaggle. "The ASAP competition is our most ambitious yet, having the potential to touch more Americans than any other project we've run so far."

The vendor demonstration will be completed in January in time for the results to be incorporated into spring test development. The open competition will run through April to allow competitors time to develop new scoring algorithms. A public leader board will monitor progress.

The Hewlett Foundation: Automated Essay Scoring

Develop an automated scoring algorithm for student-written essays.

The William and Flora Hewlett Foundation (Hewlett) is sponsoring the Automated Student Assessment Prize (ASAP). Hewlett is appealing to data scientists and machine learning specialists to help solve an important social problem. We need fast, effective and affordable solutions for automated grading of student-written essays.

Hewlett is sponsoring the following prizes:

- $60,000: 1st place
- $30,000: 2nd place
- $10,000: 3rd place

You are provided access to hand scored essays, so that you can build, train and test scoring engines against a wide field of competitors. Your success depends upon how closely you can deliver scores to those of human expert graders. While we believe that these financial incentives are important, we also intend to introduce top performers both to leading vendors in the industry and/or an established base of interested buyers. Hewlett is opening the field of automated student assessment to you. We want to induce a breakthrough that is both personally satisfying and game-changing for improving public education.

Today, state departments of education are developing new forms of testing and grading methods, to assess the new common core standards. In this environment the need for more sophisticated and affordable options is vital. For example, we know that essays are an important expression of academic achievement, but they are expensive and time consuming for states to grade them by hand. So, we are frequently limited to multiple-choice standardized tests. We believe that automated scoring systems can yield fast, effective and affordable solutions that would allow states to introduce essays and other sophisticated testing tools. We believe that you can help us pave the way towards a breakthrough. ASAP is designed to achieve the following goals:

- Challenge developers of automated student assessment systems to demonstrate their current capabilities.
- Compare the efficacy and cost of automated scoring to that of human graders.
- Reveal product capabilities to state departments of education and other key decision makers interested in adopting them.

The graded essays are selected according to specific data characteristics. On average, each essay is approximately 150 to 550 words in length. Some are more dependent upon source materials than others. This range of essay type is provided so that we can better understand the strengths of your solution. It is our intent to showcase quality and reliability, based on how well you can match expert human graders for each essay.

You will be provided with training data for each essay prompt. The number of training essays does vary. For example, the lowest amount of training data is 1,190 essays, randomly selected from a total of 1,982. The data will contain ASCII formatted text for each essay followed by one or more human scores, and (where necessary) a final resolved human score. Where it is relevant, you are provided with more than one human score, so that you may evaluate the reliability of the human scorers, but - keep in mind - that you will be predicting to the resolved score. Also, please note that most essays are scored using a holistic scoring rubric. However, one data set uses a trait scoring rubric. The variability is intended to test the limits of your scoring engine's capabilities.

Following a period of 3 months to build and/or train your engine, you will be provided with test data that will contain new essays, randomly selected for blind evaluation. However, you will notice that the rater and resolved score columns will be blank. You will be asked to supply, based on your engine's predictions for each essay, your score in the resolved score column and then submit your new data set on this site.

As part of the file that you will submit with your predictive scores, you will be asked to submit additional information. We would like to understand both the time and capital that you've spent developing your engine, the profile of your team (or you as an individual if you are working alone) and the projected cost to implement your solution on a larger scale, along with any known limitations. Basically, you will have the opportunity to present your case for who you are, why your model is commercially viable and to what extent you can use your model to satisfy the interests of potential buyers. This other information will not be used to determine any prize rewards, and it is optional. But, if you provide it, it will be used to evaluate whether or not your model should be presented to state departments of education and others who stand to benefit from your work.

Also, please note that it is our intention to stage other follow-on ASAP phases in the months ahead. We are starting with graded essays and will follow with new data:

- **Phase 1:** Demonstration for long-form constructed response (essays);
- **Phase 2:** Demonstration for short-form constructed response (short answers);
- **Phase 3:** Demonstration for symbolic mathematical/logic reasoning (charts/graphs).

In every instance, we seek to drive innovation for new solutions to automated student assessment. We hope that you will enjoy this process. May the best model win!

Source: www.kaggle.com

The Hewlett Foundation: Automated Essay Scoring

Develop an automated scoring algorithm for student-written essays.

The William and Flora Hewlett Foundation (Hewlett) is sponsoring the Automated Student Assessment Prize (ASAP). Hewlett is appealing to data scientists and machine learning specialists to help solve an important social problem. We need fast, effective and affordable solutions for automated grading of student-written essays.

Hewlett is sponsoring the following prizes:

- $60,000: 1st place
- $30,000: 2nd place
- $10,000: 3rd place

You are provided access to hand scored essays, so that you can build, train and test scoring engines against a wide field of competitors. Your success depends upon how closely you can deliver scores to those of human expert graders. While we believe that these financial incentives are important, we also intend to introduce top performers both to leading vendors in the industry and/or an established base of interested buyers. Hewlett is opening the field of automated student assessment to you. We want to induce a breakthrough that is both personally satisfying and game-changing for improving public education.

Today, state departments of education are developing new forms of testing and grading methods, to assess the new common core standards. In this environment the need for more sophisticated and affordable options is vital. For example, we know that essays are an important expression of academic achievement, but they are expensive and time consuming for states to grade them by hand. So, we are frequently limited to multiple-choice standardized tests. We believe that automated scoring systems can yield fast, effective and affordable solutions that would allow states to introduce essays and other sophisticated testing tools. We believe that you can help us pave the way towards a breakthrough. ASAP is designed to achieve the following goals:

- Challenge developers of automated student assessment systems to demonstrate their current capabilities.
- Compare the efficacy and cost of automated scoring to that of human graders.
- Reveal product capabilities to state departments of education and other key decision makers interested in adopting them.

The graded essays are selected according to specific data characteristics. On average, each essay is approximately 150 to 550 words in length. Some are more dependent upon source materials than others. This range of essay type is provided so that we can better understand the strengths of your solution. It is our intent to showcase quality and reliability, based on how well you can match expert human graders for each essay.

You will be provided with training data for each essay prompt. The number of training essays does vary. For example, the lowest amount of training data is 1,190 essays, randomly selected from a total of 1,982. The data will contain ASCII formatted text for each essay followed by one or more human scores, and (where necessary) a final resolved human score. Where it is relevant, you are provided with more than one human score, so that you may evaluate the reliability of the human scorers, but - keep in mind - that you will be predicting to the resolved score. Also, please note that most essays are scored using a holistic scoring rubric. However, one data set uses a trait scoring rubric. The variability is intended to test the limits of your scoring engine's capabilities.

Following a period of 3 months to build and/or train your engine, you will be provided with test data that will contain new essays, randomly selected for blind evaluation. However, you will notice that the rater and resolved score columns will be blank. You will be asked to supply, based on your engine's predictions for each essay, your score in the resolved score column and then submit your new data set on this site.

As part of the file that you will submit with your predictive scores, you will be asked to submit additional information. We would like to understand both the time and capital that you've spent developing your engine, the profile of your team (or you as an individual if you are working alone) and the projected cost to implement your solution on a larger scale, along with any known limitations. Basically, you will have the opportunity to present your case for who you are, why your model is commercially viable and to what extent you can use your model to satisfy the interests of potential buyers. This other information will not be used to determine any prize rewards, and it is optional. But, if you provide it, it will be used to evaluate whether or not your model should be presented to state departments of education and others who stand to benefit from your work.

Also, please note that it is our intention to stage other follow-on ASAP phases in the months ahead. We are starting with graded essays and will follow with new data:

- **Phase 1:** Demonstration for long-form constructed response (essays);
- **Phase 2:** Demonstration for short-form constructed response (short answers);
- **Phase 3:** Demonstration for symbolic mathematical/logic reasoning (charts/graphs).

In every instance, we seek to drive innovation for new solutions to automated student assessment. We hope that you will enjoy this process. May the best model win!

Source: www.kaggle.com

**Assessment Development Committee
Item Review Schedule
April 2012 – August 2012**
(Updated 5/3/12)

Review Package to Board	Board Comments to NCES	Background/ Cognitive	Review Task	Approx Number Items	Status
April 12	April 25	Background	2013 Operational Reading & Mathematics (12)	215 items (4 blocks)	✓
April 25	May 8	Background	2014 Technology & Engineering Literacy (TEL) (8)	<60 items (275 with all subitems)	✓
May 3	May 24	Cognitive	2014 Technology & Engineering Literacy (TEL) (8)	21 tasks (pre-clearance)	Review at May Board Meeting
May 29	June 19	Cognitive	2015 Pilot Mathematics (4, 8)	200 items (12 blocks)	
July 5	July 25	Cognitive	2015 Pilot Reading (4, 8)	250 items (12 blocks)	
July 19	August 8	Cognitive	2013 Operational Mathematics (12)	54 items (4 blocks)	
July 19	August 8	Cognitive	2013 Operational Reading (12)	30 items (3 blocks)	
July 19	August 9	Cognitive	2014 Technology and Engineering Literacy (TEL) (8)	21 Tasks, 175 items	