Test Administration and Scoring Technologies

Statewide Assessment Reauthorization
Work Group
June 12, 2012

Eric Zilbert, Administrator
Psychometrics and Assessment Analysis Unit

Primary Considerations

- The purpose for which test results will be used is the key element informing how an assessment is administered and scored.
- Standardized
  - Experience is as uniform as possible to lead to a valid result
  - Non-uniform administration can affect measurement
  - Scorers’ judgment can vary
- Higher stakes
  - Stricter control of administration and scoring
  - Security paramount
  - Concerns about cheating and item exposure
  - Independent scoring with high level of quality control

Test Administration

- Directions for Administration (DFAs)
  - Key element of testing process
  - Some assessments require administrator and proctor training
  - Specify testing conditions, manipulatives, room and equipment preparation, etc.
- Incident reports
  - Testing irregularities and problems are recorded and reported
- Audits
  - Check for standard conditions
  - Detect problems, inform development of DFAs and training requirements

TOM TORLAKSON
State Superintendent
of Public Instruction

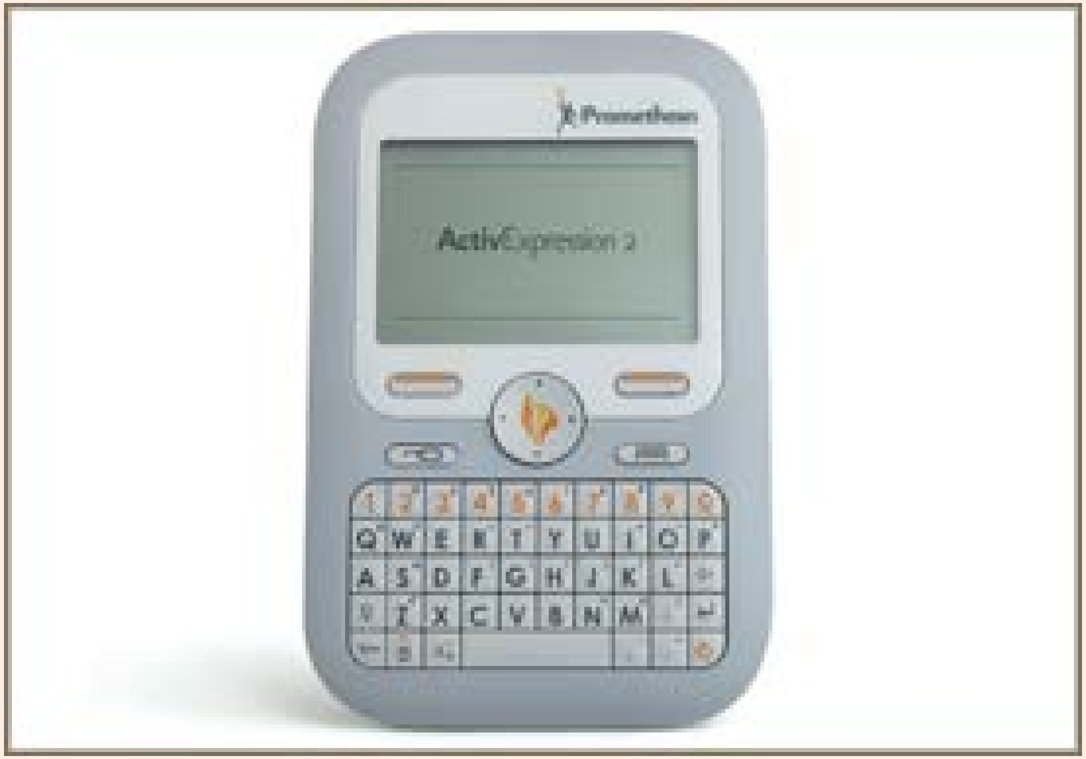
Test Administration Technology

- Bubble form and #2 pencil
  - Scoring technology closely linked to response format
  - Higher complexity than simple multiple-choice possible (e.g., number grids, coordinate graph grids)
- Performance assessment
  - Live performance in front of judges (e.g., speech, high jump, ice skating, oral exam)
  - May be recorded for scoring
- Computer based testing
  - Allows for additional item types
  - Can shorten scoring period
  - Can reduce paper, printing, and shipping costs
- Computer adaptive testing
  - More efficient and more secure
    - Fewer questions are required to accurately determine each student’s achievement level which make the items more secure
  - Based on student responses, the computer program adjusts the difficulty of questions throughout the assessment. (e.g., a student who answers a question correctly will receive a more challenging item, while an incorrect answer generates an easier question)

TOM TORLAKSON
State Superintendent
of Public Instruction

Test Administration Technology

- Clicker or Student Responder



TOM TORLAKSON
State Superintendent
of Public Instruction

Examples of Computer Based Tests Used in California

- GED, CBEST, TOEFL, and GRE all have an option for computerized assessment and are currently administered in California.

- Many benchmark/interim assessment systems use computers. Some include test design and/or data management systems:
  - DATAWISE from Measured Progress
  - Pearson Benchmark
  - Discovery Education Benchmarks

TOM TORLAKSON
State Superintendent
of Public Instruction

General Educational Development Test (GED) – Pearson

- Purpose: High school equivalency test
- Content: Mathematics; language arts, reading; language arts, writing (including essay); science; and social studies
- Format: Computer based, fixed form
  - Paper and pencil version available
  - Three parallel versions per year of each subject
- Length: 7 ½ hours
- Scoring: Essays scored by two scorers

California Basic Educational Skills Test (CBEST) – Pearson

- Purpose: Assess basic skills of prospective teachers
- Content: Reading, mathematics, and writing skills
- Format: Computer based, fixed form
  - Reading, 50 questions; mathematics, 50 questions; writing, 2 essays
- Length: 4 to 5 hours
- Scoring: Essays scored by two scorers

Test of English as a Foreign Language (TOEFL) – ETS

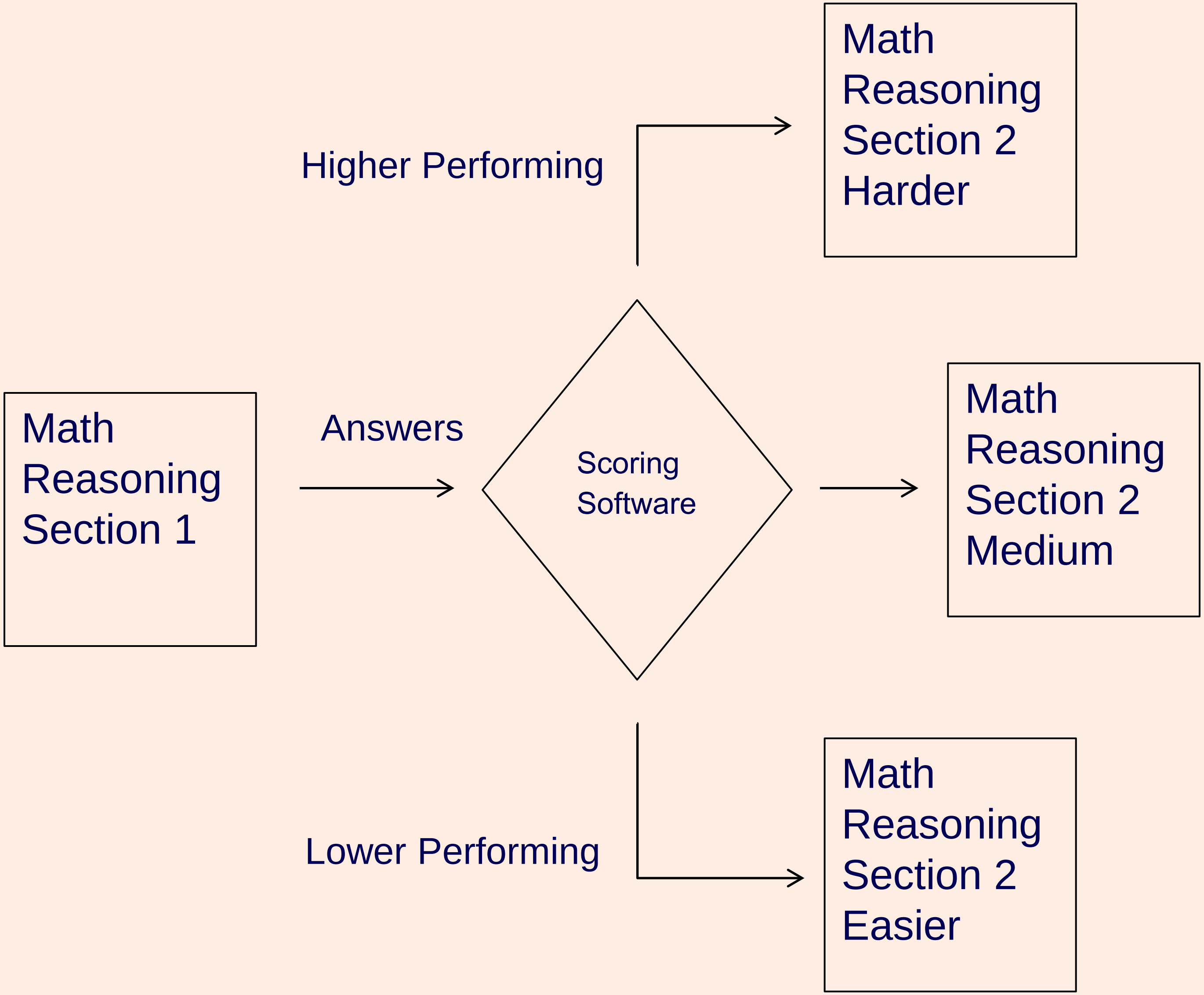
- Purpose: Evaluate the English proficiency of people who are non-native English speakers.
- Content:
  - **Listening:** 30 to 49 questions, with 15-25 minutes to answer the questions; 40-60 minutes to complete entire section.
  - **Structure:** 20-25 questions, with 15-20 minutes to complete the questions.
  - **Reading:** 44-55 questions, with 70-90 minutes to complete the section (includes time spent reading passages and answering questions).
  - **Writing:** One assigned essay topic, with 30 minutes to write the essay. Scored by two scorers.
- Format: Fixed-form computer based test (CBT)
- Length: 170-225 minutes

Graduate Record Exam (GRE) – ETS

- Purpose: Admissions test for graduate school
- Content: Analytical writing, verbal reasoning, quantitative reasoning
- Format: Section level adaptive or P&P (limited)
- Length: 190 minutes
- Scoring: Essays scored by two scorers

Measure	Number of Questions	Allotted Time
Analytical writing (One section with two separately timed tasks)	One "Analyze an Issue" task and one "Analyze an Argument" task	30 minutes per task
Verbal reasoning (Two sections)	Approximately 20 questions per section	30 minutes per section
Quantitative reasoning (Two sections)	Approximately 20 questions per section	35 minutes per section

Section Level CAT



Math Reasoning Section 1

Answers

Scoring Software

Higher Performing

Math Reasoning Section 2 Harder

Math Reasoning Section 2 Medium

Lower Performing

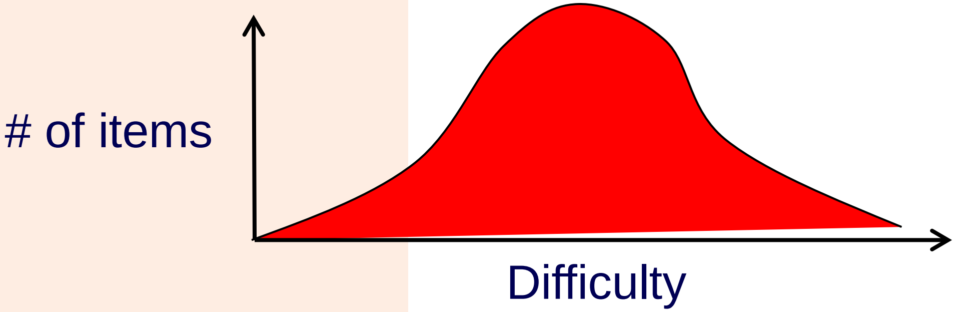
Math Reasoning Section 2 Easier

Section Level CAT (cont.)

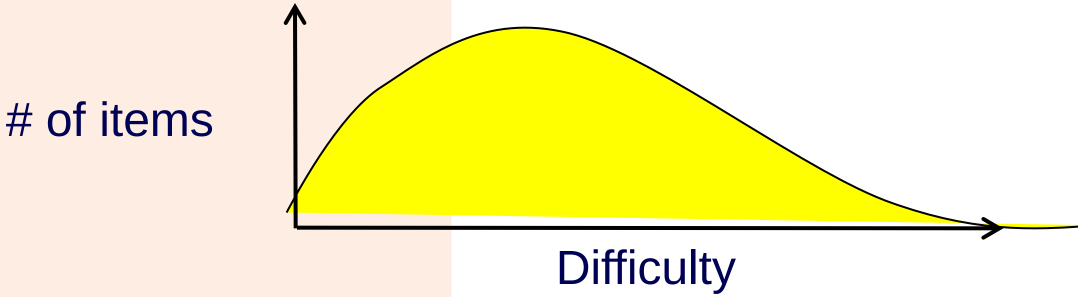
- All three of the Section 2 tests cover the whole range of performance; however, each has a higher density of items in the region of the scale in which the student is expected to score based on Section 1
- Total score is based on Section 1 and Section 2
- All student scores are placed on the same scale
- Test difficulties overlap; scores are compensatory (i.e., composite of scores)

TOM TORLAKSON
State Superintendent
of Public Instruction

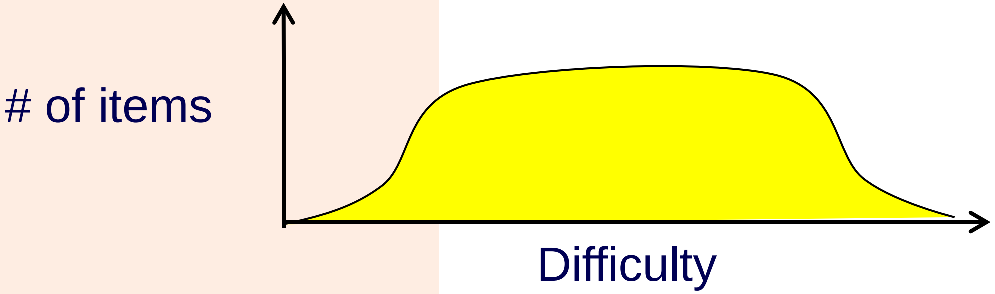
Section 1



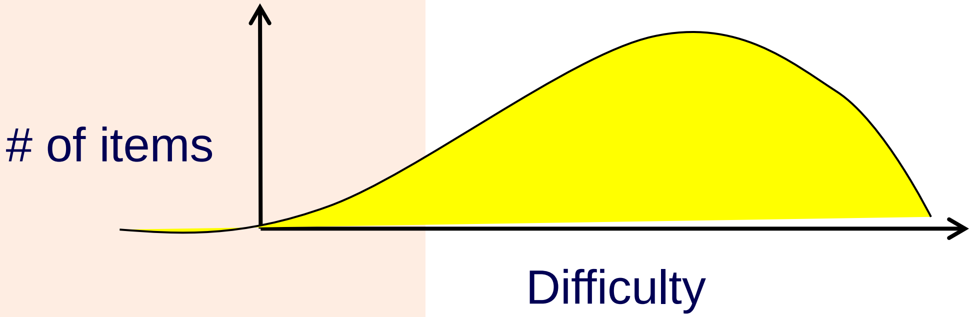
Section 2
Easier



Section 2
Medium



Section 2
Harder



Scoring Technology

- Templates
- Optical Scanning
  - Scantron
  - Electronic image based scoring (e.g., Pearson e-Pen)
  - Scan to Score
- Traditional machine scoring
  - Dichotomous (correct/incorrect) scoring most common
  - Exact word, number, or grid matches
  - No partial credit
- Automated Scoring
  - Allows scoring of short answer and essay questions
  - Requires set of human-scored papers to develop the scoring model
  - Can give partial credit, or multiple point scores

How Automated Scoring Works

- Uses a set of human-scored examples to develop a statistical model used to analyze answers (e.g., latent semantic analysis or natural language processing)
- Generally examines overall form and specific combinations of words
- Has an extensive library of possible meanings for words

What Can Be Scored?

- Written responses
  - Prompt specific essays
  - Prompt independent essays
  - Short answers
  - Summaries
- Spoken language
  - Correctness
  - Fluency
- Responses to simulations
  - Diagnosis of a patient’s illness
  - Landing a plane

TOM TORLAKSON
State Superintendent
of Public Instruction

How Good Is Automated Scoring?

ETS, Pearson, and the College Board, in the recent report *Automated Scoring for the Common Core Standards,* offered the following as a checklist to answer the question “How do you know automated scoring works effectively?”:

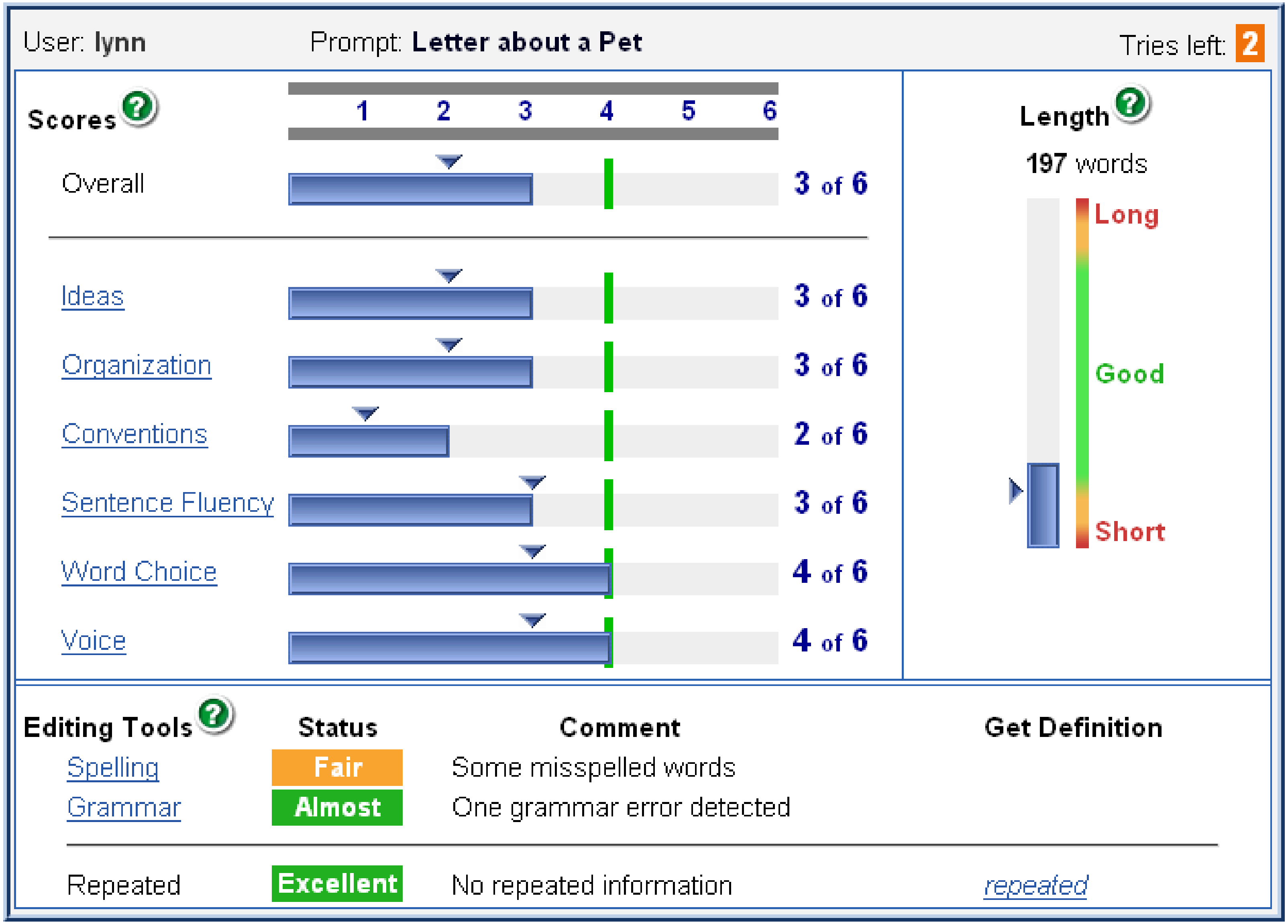
- Automated scores are consistent with the scores from expert human graders.
- The way automated scores are produced is understandable and substantively meaningful.
- Automated scores are fair.
- Automated scores have been validated against external measures in the same way as is done with human scoring.
- The impact of automated scoring on reported scores is understood (i.e., If, item by item, the automated scoring appears to perform well, an evaluation at the test level may reveal notable differences between automated and human scores).

Autoscoring Performance

Response	Assessment Prompt Material	N	Machine-Human Correlation	Human-Human Correlation	Source
Written	81 published essay prompts (grade 6-12)	400	0.89	0.86	Prentice Hall
	18 research-leveled essay prompts (grades 4-12)	635	0.91	0.91	MetaMetrics
	5 synthesizing memos from multiple sources	1239	0.88	0.79	Council for Aid to Education
Spoken	2000 spoken English items	50	0.97	0.98	Balogh & et al. (2005)
	3000 spoken Arabic items	134	0.97	0.99	Bernstein et al. (2009)
	9 Oral Reading Fluency Passage Grades 1-5	248	0.98	0.99	Downey et al. (2011)

Source: Streeter et. al. *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*, Pearson, May 2011.

Example Essay Feedback



User: **lynn** Prompt: **Letter about a Pet** Tries left: **2**

Scores

Overall	3 of 6
Ideas	3 of 6
Organization	3 of 6
Conventions	2 of 6
Sentence Fluency	3 of 6
Word Choice	4 of 6
Voice	4 of 6

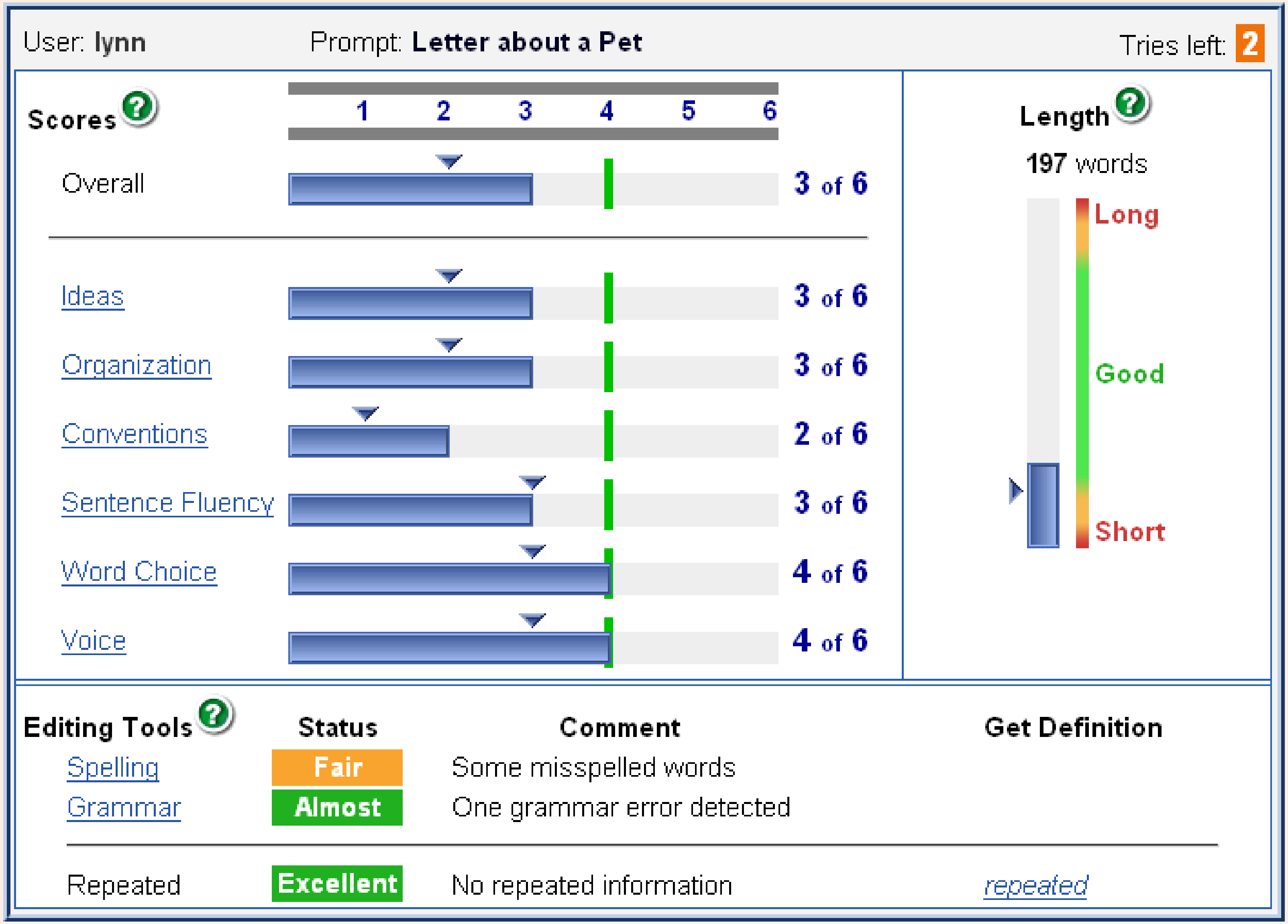
Length

197 words

Editing Tools	Status	Comment	Get Definition
Spelling	Fair	Some misspelled words	
Grammar	Almost	One grammar error detected	
Repeated	Excellent	No repeated information	*repeated*

Source: Streeter et. al. *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*, Pearson May 2011.

Example Essay Feedback



User: **lynn** Prompt: **Letter about a Pet** Tries left: **2**

Scores

Overall	3 of 6
Ideas	3 of 6
Organization	3 of 6
Conventions	2 of 6
Sentence Fluency	3 of 6
Word Choice	4 of 6
Voice	4 of 6

Length

197 words

Editing Tools	Status	Comment	Get Definition
Spelling	Fair	Some misspelled words	
Grammar	Almost	One grammar error detected	
Repeated	Excellent	No repeated information	*repeated*

Source: Streeter et. al. *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*, Pearson May 2011.

Data Requirements for Various Types of Automated Scoring

Item Type	Response Length in Words	Typical Data Requirements for development	Measures Returned
Prompt-Specific Essays	100-500	200-250 double-scored student essays	Overall score, trait scores, grammar & mechanics feedback
Prompt Independent Essays **(general models)**	100-500	Approximately 1000 essays per grade	Overall score, select trait scores, grammar & mechanics feedback
Short Answers	~10-60	500 double-scored student answers	Total or partial-credit content score
Summaries	50-250	Readings to be summarized divided by major sections	Content coverage score for each section; checks copying, length, redundancy and irrelevance.

Source: Streeter et. al. *Pearson's Automated Scoring of Writing, Speaking, and Mathematics*, Pearson, May 2011.

TOM TORLAKSON
State Superintendent
of Public Instruction

Questions?

Contact Information

- Eric Zilbert

  Administrator

  Psychometrics and Assessment Analysis Unit

  Assessment Development and Administration Division

  E-mail: ezilbert@cde.ca.gov

  Phone: 916-445-4902