Overview

• Automated scoring
• What’s available
• How to implement it
Why Constructed Response Items?

• Constructed Response – examinee generates a response rather than selecting from presented options

• Challenges
  – Development and administration
  – Scoring: recruitment, training, score quality, multiple raters
  – Score turnaround
  – Information/reliability relative to multiple-choice per unit time

• Demand
  – Construct coverage – address something that is valued and thought to be inadequately covered by MC
  – Face validity – real-world fidelity to naturalistic tasks is valued
Why Automated Scoring?

• Construct Representation
  – The use of constructed-response items

• Efficiency
  – Time
  – Cost
  – Scheduling

• Quality
  – Consistency
  – Tractability
  – Performance Feedback
Automated Scoring at ETS

• Automated scoring capabilities
  – e-rater®: essays
  – c-rater™: correctness of short responses
  – m-rater™: mathematical expressions, figures and plots
  – SpeechRater™: spontaneous speech
e-rater® — ETS’s Automated Scoring of Essays
What is *e-rater*?

- State-of-the-art automated scoring of English language essays
- First automated scoring engine for essays in high stakes assessment (GMAT®, 1999)
- Incorporates more than 10 years of advances in Natural Language Processing technology, providing:
  - Summary scores
  - Performance feedback (e.g. *Criterion*)
  - Customizable filters to identify atypical essays (e.g. off-topic)
  - Multiple options for model building and implementation
- Agreement with human scores similar to and often higher than interrater agreement
How is e-rater used?

• High Stakes Admissions & Licensure
  – GRE
  – TOEFL
• Placement & Practice
  – TOEFL Practice Online
  – SATOnline
• Learning
  – Criterion
How Does e-rater work?

- **Grammar**
  - Pronoun Errors
  - Ill-formed Verbs
  - Number of Discourse Elements

- **Usage**
  - Subject-verb Agreement
  - Garbled Sentences
  - Run-ons
  - Possessive Errors

- **Mechanics**
  - Fragments
  - Spelling
  - Capitalization
  - Punctuation
  - Nonstandard Verbs or word forms

- **Style**
  - Reverberations
  - Fragmented Sentence
  - Missing Question Marks
  - Missing Final Punctuation
  - Missing Commas or Apostrophes

---

**e-rater Construct Decomposition**

- **Organization**
  - Number of Discourse Elements

- **Development**
  - Length of Discourse Elements

- **Lexical Complexity**
  - App. Word Length
  - Sophistication of Word Choice

- **Topospecific Vocabulary Usage**
  - Theme Analysis to Stimulate with Similar Vocabulary
  - Theme Analysis to Stimulate with Tactful Vocabulary

---

Copyright © 2010 by Educational Testing Service. All rights reserved.
Example of **Criterion Feedback**

**Example of Criterion Feedback**

- **Student:** Susan Takazono
- **Grade:** 8th Grade
- **Teacher:** Ms. Alvarez
- **Submitted:** November 10, 2005, 01:32:48 PM EST

**Trait Feedback Analysis Menu**
- **Grammar**
- **Usage**
- **Mechanics**
- **Style**
- **Organization & Development**

**Summary of Style Comments**
- **Repetition of Words**
  - Inappropriate Words or Phrases
  - Sentences Beginning with Coordinating Conjunctions
  - Too Many Short Sentences
  - Too Many Long Sentences
  - Passive Voice

- **Number of Words:** 160
- **Number of Sentences:** 15
- **Average number of words per sentence:** 11.5

**View Score Analysis**

**View Combined Feedback Report...**

**Close Report**

---

**Repitition of Words**

1. **First reason:** A child need to have a job in order to do good in school. First think if a child is doing good in school, they have a job. Second a student can help buy school clothes. Third a child need to have a job, because it will help them in life to be a adult.

2. The first reason is a child need to have a job. If they have good grades, then they be doing in the classes and behavior to.

3. A student need to know that when school work is in important thing. Good in school is a good way of working.

4. The second reason is a child need to pay for the school clothes. They can buy their clothes their own money. They can get it for them self and buy more clothes and name brand clothes. And buy brothers and sister clothes to.

5. Third reason is a job will help.
How Are *e-rater* Scoring Models Developed?

- Assure quality items and scoring rubrics
- Administer item *on computer*, under conditions similar to operational use
- Conduct human scoring, preferably double-scored
- Compute the e-rater scoring features
- Determine optimal weights of features in predicting human scores via linear regression
- Cross-validate against an additional set of human-scored cases
c-rater — ETS’s Automated Scoring of Short Answers
What is c-rater?

• First automated scoring system for correct answers in short computer-entered English language text
• First used operationally in high-stakes statewide assessment in 2003
• Provides multiple scoring options:
  – Summary scores
  – Performance feedback - which concepts were identified in the response
Example Question & Rubric

• Question: “Identify TWO common ways the body maintains homeostasis during exercise.”

• Key elements include any of the following:
  – Sweating (perspiration)
  – Increased breathing rate (respiration)
  – Decreased digestion
  – Increased circulation rate (heart speeds up)
  – Dilation of blood vessels in skin (increased blood flow)

• Rubric:
  – 2 points - two key elements
  – 1 point - one key element
  – 0 point - other
How does c-rater work?

• Linguistic model constructed from rubric and sample of human-scored responses
• Automated text parser identifies phrases with content sufficiently similar to keyed concepts
  – Synonym engine allows for multiple expressive forms of concepts
• c-rater is similar in performance to human raters for appropriate items
m-rater — ETS’s Automated Scoring of Mathematics Items
What is *m-rater*?

- Automated scoring engine for mathematics questions
- First used operationally in high-stakes statewide assessment in 2003
- Meets or exceeds the performance of human grading
- Robust and flexible scoring:
  - Equations, graphs, geometric figures and numeric response
  - Accommodates *any* response that is mathematically equivalent to the key through *computed* scoring
  - Partial credit scoring (e.g. form of the equation)
  - Interdependent multi-part responses (e.g. graph/equation)
  - Summary scores and/or performance feedback
Multiple-Choice Item

A family travels at a constant speed during a road trip. After 3 hours, they stop for 2 hours to eat and rest. They then continue for 4 more hours at the same speed.

Which of these graphs best represents this situation?

- From 2003 HSA Algebra I practice items, available at:
Example Item – Graph

Question 3

A family travels at a constant speed during a road trip. After 3 hours, they stop for 2 hours to eat and rest. They then continue for 4 more hours at the same speed.

On the coordinate system below, draw a graph that could represent this situation.

Distance Traveled (in miles)

Time (in hours)

Copyright © 2010 by Educational Testing Service. All rights reserved.
Multiple-Choice Item

19. In one day, a store sold 300 shirts at 25% off the regular price of $x$ dollars. Which of these expressions represents the total amount, in dollars, that was received for the sale of shirts on that day?

A. $0.25(x - 300)$
B. $25(x - 300)$
C. $300(x - 25)$
D. $300(x - 0.25x)$

• From 2003 HSA Algebra I practice items, available at:
Example Item – Equation

Question 19

In one day, a store sold 300 shirts at 25% off the regular price of \( x \) dollars. Write an expression that represents the total amount, in dollars, that was received for the sale of shirts on that day.
How does *m-rater* work?

- m-rater scoring meets or exceeds the agreement standard set by human grading
- Most commonly used in high-stakes state assessment
- Scoring
  - Equations - scored by computing mathematical equivalence with key
  - Straight line and broken line plots – line segments converted to equations and scored
  - Curvilinear plots and geometric figures – scored on the basis of plotted points
SpeechRater — ETS’s Automated Scoring of Spontaneous Speech
What is SpeechRater?

• The world’s only automated scoring system for *spontaneous* speech
• ETS’s newest automated scoring system, first used in 2006 for the TOEFL Practice Online program
• Predicts the human score that would be issued using a TOEFL-like rubric
  – Pronunciation
  – Grammar
  – Speech Rate
  – Length of speech between pauses
  – Vocabulary diversity
TOEFL Speaking Construct

Delivery
- Fluency
- Intonation
- Rhythm
- Pronunciation

Language Use
- Diversity
- Vocabulary
- Sophistication
- Precision

Topic Development
- Range
- Grammar
- Complexity
- Accuracy

Diversity
- Range
- Vocabulary
- Sophistication
- Precision

Language Use
- Grammar
- Range
- Complexity
- Accuracy

Topic Development
- Coherence
- Idea progression
- Content relevance
How does *SpeechRater* work?

SpeechRater Data/Outputs

- Audio files
- Recognized words and utterances
- Speech features
- Speaking scores

SpeechRater Components

- Speech recognizer
- Feature extraction programs
- Scoring model
Implementing Automated Scoring: Example for High-Stakes Admissions
Implementation of e-rater

- Construct evaluation
- Automated filters (advisories) selection and evaluation
- Selecting a model
- Implementation models
- Adjudication thresholds
- Impact studies
  - Score changes
  - Comparisons with external variables of interest
Evaluation Criteria

• Construct relevance
  – Conceptual link between features used in e-rater model and rubric for task

• Empirical evidence of validity
  – Relationship to human scores
    • Exact / adjacent agreement [no standard]
    • Pearson correlation ≥ 0.70
    • Weighted kappa ≥ 0.70
    • Reduction compared to human-human agreement < 0.10
    • Difference in standardized mean score < 0.15
    • Subgroups (fairness), difference in standardized mean score < 0.10
  – Relationship to external criteria
Available Filters (Advisories)

• *Reuse of language* – repetitive content
• *Key concepts* – underdevelopment of topics
• *Not relevant* – on a topic other than that assigned
• *Restatement* – repeat of prompt without elaboration
• *No resemblance* – atypical content in essay
• *Too brief* – not of sufficient length for an accurate score
• *Excessive length* – too long for an accurate score
• *Unidentifiable organizational elements* – discourse elements not identified
• *Excessive number of problems* – very high number of errors
• *Unexpected topic* – on a topic other than the prompt indicates
• *Non-essay* – text unrelated to the essay topic
**e-rater model types:**
**prompt-specific**

- Each model is trained on responses to a particular prompt
- **Advantages:**
  - Tailored to particular prompt characteristics
  - High agreement with human raters
  - Incorporates content features
- **Disadvantages:**
  - Higher demand for training data
**e-rater model types:**

**generic**

- A single model is trained on responses to a variety of prompts
- Advantages:
  - Smaller data set required for training
  - Scoring standards the same across prompts
- Disadvantages:
  - No content features
  - Differences between particular prompts are not accounted for
  - Agreement with human raters is lower
## Model Selection: Prompt-Specific or Generic?

### Prompt type A

<table>
<thead>
<tr>
<th>Model</th>
<th>Exact agreement</th>
<th>Exact + adj</th>
<th>Correlation</th>
<th>Wtd. Kappa</th>
<th>Degradation</th>
<th>Mean Score Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>59.4</td>
<td>97.7</td>
<td>0.74 (1)</td>
<td>0.74 (1)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Prompt-specific</td>
<td>60.0</td>
<td>98.5</td>
<td>0.80 (0)</td>
<td>0.77 (0)</td>
<td>+0.03 (0)</td>
<td>0.04 (0)</td>
</tr>
<tr>
<td>Generic</td>
<td>59.2</td>
<td>98.2</td>
<td>0.79 (0)</td>
<td>0.76 (0)</td>
<td>+0.02 (0)</td>
<td>0.05 (4)</td>
</tr>
</tbody>
</table>

### Prompt type B

<table>
<thead>
<tr>
<th>Model</th>
<th>Exact agreement</th>
<th>Exact + adj</th>
<th>Correlation</th>
<th>Wtd. Kappa</th>
<th>Degradation</th>
<th>Mean Score Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>58.2</td>
<td>96.7</td>
<td>0.78 (0)</td>
<td>0.78 (0)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Prompt-specific</td>
<td>52.8</td>
<td>96.0</td>
<td>0.78 (0)</td>
<td>0.76 (0)</td>
<td>0.02 (0)</td>
<td>0.07 (6)</td>
</tr>
<tr>
<td>Generic</td>
<td>51.6</td>
<td>94.5</td>
<td>0.76 (2)</td>
<td>0.73 (15)</td>
<td>0.05 (11)</td>
<td>0.11 (43)</td>
</tr>
</tbody>
</table>
Models for Implementation

• Confirmatory scoring (check score)
  – E-rater “checks” human score, serves as a filter to determine whether an additional human grader is needed

• Contributory scoring
  – One automated and one human score, with discrepancies resulting in additional human adjudication

• Automated scoring alone
Impact Analysis

• What is the anticipated impact on reported scores of replacing a human with an automated score?
• Under multiple variations of implementation:
  – Score contribution
  – Agreement threshold
  – Model building method
Comparing Raw Scores: All Human vs. Human-Automated

- Total raw score (2 prompts) with e-rater check score implementation and exact agreement threshold (.5) compared to all human scoring

<table>
<thead>
<tr>
<th>Percent of scores</th>
<th>-0.5</th>
<th>0.0</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13%</td>
<td>81%</td>
<td>6%</td>
</tr>
</tbody>
</table>

- Contrasting with human prompt agreement

<table>
<thead>
<tr>
<th>Percent Agreement Rates</th>
<th>Exact</th>
<th>Adjacent (1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt type A (H1-H2)</td>
<td>59%</td>
<td>98%</td>
</tr>
<tr>
<td>Prompt type A (H1-Auto)</td>
<td>60%</td>
<td>99%</td>
</tr>
<tr>
<td>Prompt type B (H1-H2)</td>
<td>58%</td>
<td>98%</td>
</tr>
<tr>
<td>Prompt type B (H1-Auto)</td>
<td>53%</td>
<td>96%</td>
</tr>
<tr>
<td>Total Raw Score (Auto-Human)</td>
<td>81%</td>
<td>100%</td>
</tr>
</tbody>
</table>

N=132404
## Associations With Other Measures

<table>
<thead>
<tr>
<th>Correlation with Measure</th>
<th>Final Score All Human</th>
<th>Final Score Automated-Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing Score</td>
<td>--</td>
<td>0.96</td>
</tr>
<tr>
<td>Verbal Score</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Quantitative Score</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>GPA - All classes</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>GPA - Classes in major</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>
A Contrast in Implementation

GRE Program
- Issue and Argument
  - Issue – generic model
  - Argument – prompt-specific models
- Check score use
- Agreement threshold of .5

TOEFL Program
- Independent prompt
- Generic model
- Contributory scoring
- Agreement threshold of 1.5
Operational Maintenance

• Periodic need for engine upgrades
  – Component review
    • Construct evaluation
    • Corpus review
  – Integrated engine evaluation
  – Full model rebuild
    • Impact for operational programs
• Periodic re-evaluation
Automated Scoring: 1941

“The International Business Machine (I.B.M.) Scorer (1938) uses a carefully printed sheet … upon which the person marks all his answers with a special pencil. The sheet is printed with small parallel lines showing where the pencil marks should be placed to indicate true items, false items, or multiple-choices. To score this sheet, it is inserted in the machine, a lever is moved, and the total score is read from a dial. The scoring is accomplished by electrical contacts with the pencil marks. … Corrections for guessing can be obtained by setting a dial on the machine. By this method, 300 true–false items can be scored simultaneously. The sheets can be run through the machine as quickly as the operator can insert them and write down the scores. The operator needs little special training beyond that for clerical work.” (Greene, p. 134)
Want to Know More?

Contact:
David M. Williamson
Senior Research Director, Applied Research & Development
Educational Testing Service
Princeton, NJ 08541
Email: dMwilliamson@ets.org
Phone: 609-734-1303