WriterPlacer® Essay Scoring: The How and the Why

How automated scoring works
How it’s being used at Nassau Community College

Karen Lochbaum
Noreen Wade
June 12, 2014
Why Automated Scoring?

- Immediate & Efficient
  - Evaluate responses in seconds
  - Reduce score turnaround time
  - No waiting for decisions
- Accurate
- Consistent, Objective
Automated Scoring Approach

- Train the engine based on the collective wisdom of human scorers
The Process

- Data Collection
- Human Scoring
- Training the Automated Scoring Engine
- Operational Scoring
- Monitoring
Data Collection

- We collect 500-1000 student responses to each essay prompt
- Want a good representation of different writing abilities
Human Scoring

• Each essay is evaluated by at least two trained human scorers
• Holistically
• And for each of 4-6 different dimensions
WritePlacer Holistic Rubric

8: The essay demonstrates **clear and consistent mastery** of on-demand essay writing
7: …**consistent mastery**…
6: …**reasonably consistent mastery**…
5: …**adequate mastery**…
4: …**developing mastery**…
3: …**little mastery**…
2: …**very little mastery**…
1: …**no mastery**…
WritePlacer Dimensions

• **Purpose & Focus**
  – The extent to which the writer presents information in a unified and coherent manner, clearly addressing the issue.

• **Organization & Structure**
  – The extent to which the writer orders and connects ideas.

• **Development & Support**
  – The extent to which the writer develops and supports ideas.

• **Sentence Variety & Style**
  – The extent to which the writer crafts sentences and paragraphs demonstrating control of vocabulary, voice, and structure.

• **Mechanical Conventions**
  – The extent to which the writer expresses ideas using standard English.

• **Critical Thinking**
  – The extent to which the writer communicates a point of view and demonstrates reasoned relationships among ideas.
WritePlacer ESL Holistic Rubric

6: The writing sample exhibits **strong** control of the elements of on-demand essay writing in English.

5: ...**consistent control**...

4: ...**moderate control**...

3: ...**emerging competency**...

2: ...**limited competency**...

1: ...**little competency**...
WritePlacer ESL Dimensions

• **Word Use**
  - The extent to which the writer is able to use a wide range of words and phrases accurately.

• **Sentence Use**
  - The extent to which the writer is able to use a variety of sentence patterns with both independent and dependent clauses.

• **Grammar**
  - The extent to which the writer is able to express ideas using grammatically correct English.

• **Organization and Development**
  - The extent to which the writer is able to focus on the assigned topic and to develop ideas clearly.
Training the Automated Scoring Engine

• Learn to score based on several hundred human scored responses

• Measure the content and quality of responses by determining
  – The features that human scorers evaluate when scoring a response
  – How those features are weighed and combined to produce scores
The Intelligent Essay Assessor (IEA)

- Overall essay coherence
- Topic development
- Sentence-sentence coherence
- Style, organization and development
- Word Maturity
- Word variety
- Confusable words
- LSA essay semantic similarity
- Vector length
- Content
- n-gram features
- Grammatical errors
- Grammar error types
- Grammar
- Mechanics
- Spelling
- Capitalization
- Punctuation
- Not Just Essay Score
Other Features of IEA

- Uses non coachable measures
  - No counts of total words, syllables, characters, etc.
  - No trigger surface features: “thus”, “therefore”
  - Detects larding of big words

- Knows when it doesn’t know
Confidence in Scoring

Some responses will look unusual

- Off-topic or highly creative
- Not a good faith essay
  - A refusal to write
  - Copying and pasting the same text over and over
- Very different skill level
  - Word choice, structure different than what’s expected
- Not a well-formed response
  - E.g. just a list of words
Content Based Scoring

• Use Latent Semantic Analysis (LSA) to capture the “meaning” of language

• LSA knows that
  – Surgery is often performed by a team of doctors.
  – On many occasions, several physicians are involved in an operation.

mean about the same thing even though they share no words.

• Enables evaluating the content of what is written rather than just matching keywords
Why LSA?

Search for “Cars”...
Why LSA?

- Language is diverse!
- LSA operates on the level of deep word (latent) meaning

<table>
<thead>
<tr>
<th>Key Word</th>
<th>LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor–Doctor</td>
<td>1.0</td>
</tr>
<tr>
<td>Doctor–Physician</td>
<td>0.0</td>
</tr>
<tr>
<td>Doctor–Surgeon</td>
<td>0.0</td>
</tr>
</tbody>
</table>
What does that have to do with automated scoring?

• LSA reads lots of text
• Learns what words mean and how they relate to each other
• Result is a “Semantic Space”
• Every word represented as a vector
• Every paragraph represented as a vector

\[ M(\text{Paragraph}) = M(w1) + M(w2) + \ldots \]
• Every essay represented as a vector
• New essays are placed based on the words they contain
• Compare against pre-scored essays
Summary of IEA

• Trained on several hundred scored candidate responses
• Learns to score based on the collective wisdom of many skilled human scorers
• Measures different aspects of the responses as appropriate for the rubric
  – Semantic similarity to human scored responses, ideas, examples, ....
  – Appropriate word choice, word and sentence flow, fluency, coherence, ....
  – Grammar, word usage, punctuation, spelling, ...
• Combines those measures to produce a score
Where we are in The Process

- Data Collection
- Human Scoring
- Training the Automated Scoring Engine
  - Operational Scoring
- Monitoring
Operational Scoring

- Student submits essay
- IEA evaluates whether it can score the essay accurately and if so does
- If not, the essay is routed for human scoring
  - The score report shows a score pending message
Initial Essay Evaluation

• Is it long enough for accurate evaluation?
• Does it looks like English?
• Does it look on-topic?
• Does it look as expected? I.e. like the essays the system was trained on?
## Individual Score Report

**Student Name:** Lisa Eggers  
**Student ID:** eggell1

<table>
<thead>
<tr>
<th>Date of Birth</th>
<th>Major Name</th>
<th>Site Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-01-01</td>
<td>Omitted</td>
<td>Demo Site</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Testing</th>
<th>Administered By</th>
<th>Voucher Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-10-18</td>
<td>Tigrett, Jennifer</td>
<td>NA</td>
</tr>
</tbody>
</table>

## Time Record

<table>
<thead>
<tr>
<th>Test Started</th>
<th>Test Completed</th>
<th>Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-10-18 02:39 PM CDT</td>
<td>2013-10-18 02:44 PM CDT</td>
<td>00Hr:03Min:23Sec</td>
</tr>
</tbody>
</table>

### Test Results

<table>
<thead>
<tr>
<th>Test Name</th>
<th>CSEM</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisa Test Setting</td>
<td>N/A</td>
<td>4</td>
</tr>
</tbody>
</table>

### Holistic Score Description

The essay demonstrates developing mastery of on-demand essay writing.

### Dimension Title | Dimension Description
--- | ---
Purpose and Focus | Your response does not fully communicate purpose, and focus may be inconsistent.
Organization and Structure | Your response demonstrates limited organization of ideas.
Development and Support | Your response has limited support for your ideas.
Sentence Variety and Style | Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Mechanical Conventions | Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking | Your response shows limited clarity and complexity of thought.
24-Hour Turnaround for Human Scoring

- Essays are routed to Pearson’s Performance Scoring Center
- Trained professional scorers working from home
- Score are reported back to students within 24 hours of students’ completing the test
  - Monday – Friday, excluding holidays.
Monitoring Timeliness

- Reports are sent to scorers and scoring management staff every two hours
  - Essays requiring human scoring
  - When essays are due
Monitoring
Human Scorer Quality

• Validity essays
  • Have “true scores”
  • Interspersed with and indistinguishable from other essays
  • Scorers are given feedback if score is missed
• Maintain acceptable agreement
• A scoring director also “reads behind” scorers
  – Provides feedback on scoring performance
Monitoring
The Annual Audit Process

• Goal: Evaluate automated scoring accuracy and consistency with human scorers
• 5% of the essays are scored by human scorers
• Target: 90% or more agreement within one score point
• Expert human scorer for “resolution”
Summary of the Process

- Data Collection
- Human Scoring
- Training the Automated Scoring Engine
- Operational Scoring
- Monitoring

➢ NCC’s experience
WritePlacer and NCC
Helps NCC achieve goal of moving closer to a one day enrollment process

Helps students stay motivated and engaged in enrollment process
Before WritePlacer

- Blue books
- Scheduling / cost of readers
- Security / Storage / Shredding
- Fatigue / How much is too many?
- Availability of readers
- Student wait time
- Manually entering scores and placements
Why WritePlacer

- Automatic download of scores
- Security
- Allows students to test / retest with ease
- Instant Results
Student Choice

- 3 topics will be presented
- Students will choose one topic
- No attempt will be made to influence their choice
- Students choosing WritePlacer topic will be directed by proctor to the WritePlacer site
- Data will be collected for 100 essays
Scoring

- WritePlacer essays will be printed and scored by 2 human readers using our 6 point rubric.

- Writeplacer essays will also be scored by the computer.

- Each essay will be checked to see if they differ by more than 1 point from the WritePlacer score. If so, they will be scored by a 3rd reader.
Results

- Very little disparity found with WritePlacer score of 6, 7 and 8 (College Level Placement) and human readers 4, 5 and 6 (College Level Placement)
- Score of 6 = college level
- Score of 5 = final evaluation by NCC human reader
- Score of 4 and below = developmental level
Where are we today?

- WritePlacer review committee was formed
- Expand retesting
- Multiple measures used for Placement
- English 100