Symposium:
Innovations in the automated scoring of spoken responses

Evaluating the constructs and automated scoring performance for speaking tasks in the Versant Tests and PTE Academic

Alistair Van Moere
Knowledge Technologies
Pearson
Automated Tests of Spoken Proficiency

- **Versant Test**
  - Listening-Speaking test
  - Uses: Job recruitment, placement, progress monitoring
  - Available in English, Spanish, Arabic, Dutch, (French, Chinese)

- **PTE Academic**
  - 4-skills language proficiency test
  - Uses: Entrance into English-speaking universities
Assessment argument

(Mislevy 2005)
Assessment argument

(Mislevy 2005)
Versant Tasks and Scoring
Versant Tasks and Scoring

- Pronunciation
- Fluency
- Sentence Mastery
- Vocabulary

Tasks:
- Read Aloud
- Repeat Sentence
- Answer Question
- Sentence Build
- Story Retell
Versant Tasks and Scoring

OVERALL

20% Pronunciation
30% Fluency
30% Sentence Mastery
20% Vocabulary

Read Aloud
Repeat Sentence
Answer Question
Sentence Build
Story Retell
Versant Tasks and Scoring

20% Pronunciation
30% Fluency
30% Sentence Mastery
20% Vocabulary

OVERALL

Read Aloud  Repeat Sentence  Answer Question  Sentence Build  Story Retell

63 responses, 3’30 mins speech
## Versant Test Scoring

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Temporal features of speech predict expert human judgments</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>Spectral properties and segmental aspects predict human judgments</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>i) Rasch-based ability measures from dichotomous-scored vocabulary items; ii) LSA-based measures on constructed responses predict human judgments</td>
</tr>
<tr>
<td>Sentence Mastery</td>
<td>Rasch-based ability measures from word errors on increasingly complex sentences</td>
</tr>
</tbody>
</table>
## Versant Test Scoring

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scoring</th>
<th>Machine-Human, r</th>
<th>Human split-half</th>
<th>Machine split-half</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Temporal features of speech predict expert human judgments</td>
<td>.94</td>
<td>.99</td>
<td>.97</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>Spectral properties and segmental aspects predict human judgments</td>
<td>.88</td>
<td>.99</td>
<td>.97</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>i) Rasch-based ability measures from dichotomous-scored vocabulary items; ii) LSA-based measures on constructed responses predict human judgments</td>
<td>.96</td>
<td>.93</td>
<td>.92</td>
</tr>
<tr>
<td>Sentence Mastery</td>
<td>Rasch-based ability measures from word errors on increasingly complex sentences</td>
<td>.97</td>
<td>.95</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td><strong>Overall</strong></td>
<td><strong>.97</strong></td>
<td><strong>.99</strong></td>
<td><strong>.97</strong></td>
</tr>
</tbody>
</table>

Validation sample, n=143, flat score distribution
Scoring Model for Sentence Mastery

Repeat Sentence:
Scoring Model for Sentence Mastery

Repeat Sentence:

*I’ll catch up with you soon.*

“uh .. I’ll catch up you ... I don’t know”

*Security wouldn’t let him in because he didn’t have a pass.*

“*Security wouldn’t help him pass*”
Scoring Model for Sentence Mastery

Repeat Sentence:

*I’ll catch up with you soon.*

“uh .. I’ll catch up you ... I don’t know” = 2 word errors

*Security wouldn’t let him in because he didn’t have a pass.*

“Security wouldn’t help him pass” = 7 word errors
Scoring Model for Sentence Mastery

Repeat Sentence:

*I’ll catch up with you soon.*

“uh .. I’ll catch up you ... I don’t know” = 2 word errors

*Security wouldn’t let him in because he didn’t have a pass.*

“Security wouldn’t help him pass” = 7 word errors
Versant’s Domain of Use

Hulstijn (2010)
Versant’s Domain of Use

Hulstijn (2010)
Versant’s Domain of Use

Versant test score correlations with communicative tests

<table>
<thead>
<tr>
<th>Communicative test</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test of Spoken English (TSE)</td>
<td>0.88</td>
<td>58</td>
</tr>
<tr>
<td>New TOEFL Speaking</td>
<td>0.84</td>
<td>321</td>
</tr>
<tr>
<td>BEST Plus interview</td>
<td>0.86</td>
<td>151</td>
</tr>
<tr>
<td>IELTS interview test</td>
<td>0.76</td>
<td>130</td>
</tr>
</tbody>
</table>
PTE Academic: Broader construct

- Read Aloud
- Repeat Sentence
- Describe Image
- Retell Lecture
- Answer Short Question
PTE Academic: Broader construct

<table>
<thead>
<tr>
<th></th>
<th>Describe Image</th>
<th>Retell Lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preparation time</strong></td>
<td>25 secs</td>
<td>40 secs</td>
</tr>
<tr>
<td><strong>Response time</strong></td>
<td>40 secs</td>
<td>40 secs</td>
</tr>
</tbody>
</table>
PTE Academic: Broader construct

- **Pronunciation**
- **Fluency**
- **Accuracy**
- **Content**
- **Vocabulary**

<table>
<thead>
<tr>
<th>Activity</th>
<th>Describe Image</th>
<th>Retell Lecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation time</td>
<td>25 secs</td>
<td>40 secs</td>
</tr>
<tr>
<td>Response time</td>
<td>40 secs</td>
<td>40 secs</td>
</tr>
</tbody>
</table>
PTE Academic: Sampling Academic Domain

• **5 tasks:**
  ~ 36 responses
  ~ 8 minutes of speech

• **Input:**
  – Reading texts
  – Listening texts
  – Visual (non-linguistic)

• **Output:**
  – Prepared monologues
  – Short, real-time responses
Content Scoring of Constructed Responses

- Word choice (Latent Semantic Analysis)
- Content relevance
- Lexical measures
- Words in sequence; collocations

Sample response

“the lecture was given about biotic cells prokaryotic cell was first described and eukaryotic cell was secondly ref uh described uh it was said eukaryotic cells are more complicated than prokaryotic cell eukaryotic cell is microorganisms where it is it has one single cell and multi cell organisms are also present in eukaryotic cell this more complicated than prokaryotic cell which is placed in right side of the screen”
PTE Academic: Reliability

Validation sample n=158

<table>
<thead>
<tr>
<th>Scoring</th>
<th>Machine-Human, r</th>
<th>Human split-half</th>
<th>Machine split-half</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>.97</td>
<td>.97</td>
<td>.96</td>
</tr>
</tbody>
</table>
Claim about student

unless Alternate explanations

since so

Data concerning student

Data concerning situation

Warrant concerning scoring

Warrant concerning task design

Student acting in assessment situation
The tasks are valid for assessing spoken language proficiency.
The tasks tap real-time automatic processes, and sample academic language & domain interactions.

The tasks are valid for assessing spoken language proficiency.
The tasks are valid for assessing spoken language proficiency.

The tasks tap real-time automatic processes, and sample academic language & domain interactions.

Some tasks are not authentic; the interactions are too constrained.

**SCORING**

- Claim about student
- Data concerning student
- Data concerning situation
- Warrant concerning scoring
- Warrant concerning task design
- Student acting in assessment situation

**TASKS**

- REBUTTAL
- COUNTER
- BACKING
- CLAIM
The tasks are valid for assessing spoken language proficiency. They tap real-time automatic processes and sample academic language & domain interactions. Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances). Some tasks are not authentic; the interactions are too constrained. The tasks are not always authentic; the interactions are too constrained.
The tasks are valid for assessing spoken language proficiency. The scoring is sufficiently accurate to replace humans.

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances).

Some tasks are not authentic; the interactions are too constrained.

The tasks tap real-time automatic processes, and sample academic language & domain interactions.

The scoring is sufficiently accurate to replace humans.
The tasks are valid for assessing spoken language proficiency.

The scoring is sufficiently accurate to replace humans.

Machine-to-human score correlations ~0.97 (same tasks, same performance instance).

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances).

Some tasks are not authentic; the interactions are too constrained.

The tasks tap real-time automatic processes, and sample academic language & domain interactions.

Scoring

Tasks

Claim

- The tasks are valid for assessing spoken language proficiency

Backing

- Machine-to-human score correlations ~0.97 (same tasks, same performance instance)

Counter

- The tasks tap real-time automatic processes, and sample academic language & domain interactions

Rebuttal

- Some tasks are not authentic; the interactions are too constrained

- Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)
**TASKS**

- **CLAIM**: The tasks are valid for assessing spoken language proficiency.

- **BACKING**: The tasks tap real-time automatic processes, and sample academic language & domain interactions.

- **COUNTER**: Some tasks are not authentic; the interactions are too constrained.

- **REBUTTAL**: Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances).

**SCORING**

- **CLAIM**: The scoring is sufficiently accurate to replace humans.

- **BACKING**: Machine-to-human score correlations ~0.97 (same tasks, same performance instance).

- **COUNTER**: Machines are notoriously error prone; scores may be triple counting poor pronunciation.

- **REBUTTAL**: Some tasks are not authentic; the interactions are too constrained.
The tasks are valid for assessing spoken language proficiency.

The scoring is sufficiently accurate to replace humans.

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances).

Scores are relatively insensitive to simulations of worse recognition; systems should be optimized for score accuracy.

Some tasks are not authentic; the interactions are too constrained.

Machines are notoriously error prone; scores may be double counting poor pronunciation.

The tasks tap real-time automatic processes, and sample academic language & domain interactions.

Machine-to-human score correlations ~0.97 (same tasks, same performance instance).

The scoring is sufficiently accurate to replace humans.
Acknowledgements

Dr Jared Bernstein,
Consulting Scientist, Knowledge Technologies, Pearson

Prof John De Jong,
SVP Global Strategy & Business Development, Pearson