Developing and Validating an Automated Test of Spoken MSA

ECOLT at GWU
Washington, DC
8 November 2008

Jared Bernstein
Jian Cheng
Ulrike Pado
Masa Suzuki
Mohamed Al-Saffar
Alistair van Moere

Pearson Knowledge Technologies
Palo Alto, California
Outline

1. How Versant tests work
2. Versant Arabic Test (development)
3. Validation evidence
4. Predictive accuracy
“The train has been delayed by one hour”

How Versant tests work

Versant Database

Test Delivery Server

Scoring
Versant Arabic Test

- DLI purpose
  ~1000 students at DLI need predictive speaking tests

- Requirements
  Accurate test of Arabic listening & speaking
  Convenient to use at DLI and worldwide
  Suitable for repeated formative testing
  High peak capacity for mass screening
Versant Arabic Test

Test Structure

Part A: Reading

Part B: Repeat -1

Part C: Short Answers

Part D: Sentence Builds

Part E: Repeat -2

Part F: Passage Retelling
Schematic test structure

Read aloud → Repeat Sentence 1 → Short Question → Sentence Build → Repeat Sentence 2 → Passage

17 minutes
## Prompt recordings

<table>
<thead>
<tr>
<th>Country</th>
<th>Egypt</th>
<th>Iraq</th>
<th>Jordan</th>
<th>Morocco</th>
<th>Lebanon</th>
<th>Palestine</th>
<th>Syria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voices</td>
<td>F, M</td>
<td>F, M</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F, M</td>
<td>F, M</td>
</tr>
</tbody>
</table>
How Versants are developed
Pilot data sources

### Native Data

<table>
<thead>
<tr>
<th></th>
<th>Egypt</th>
<th>Syria</th>
<th>Iraq</th>
<th>Palestine</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>484</td>
<td>281</td>
<td>179</td>
<td>187</td>
<td>517</td>
<td>1648</td>
</tr>
</tbody>
</table>

### Learner Data

<table>
<thead>
<tr>
<th></th>
<th>DLI</th>
<th>Non-DLI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>1120</td>
<td>552</td>
<td>1672</td>
</tr>
</tbody>
</table>
# Reliability

<table>
<thead>
<tr>
<th>Score</th>
<th>Split-Half Reliability ((N = 134))</th>
<th>Test – Retest Reliability ((N = 100))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Sentence Mastery</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.96</td>
<td>0.94</td>
</tr>
</tbody>
</table>
## Machine – Human Comparison

<table>
<thead>
<tr>
<th>Score</th>
<th>Correlation (N = 134)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.97</td>
</tr>
<tr>
<td>Sentence Mastery</td>
<td>0.97</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.96</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.84</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Native ~ Non-Native Scores
Natives by Countries

- **Egypt**: N=148, mean=85.4
- **Syria**: N=156, mean=87.0
- **Iraq**: N=149, mean=86.2
- **Palestine**: N=128, mean=84.3
- **Saudi Arabia**: N=49, mean=86.0
- **Morocco**: N=41, mean=82.0
- **Yemen**: N=43, mean=85.0

Percent Cumulative Density vs. Arabic Overall Score
Concurrent Validation

1. Versant Machine Scores

2. ILR-OPI Human Interview Scores
# Rater-rater & rater-OPI correlations

<table>
<thead>
<tr>
<th>Interview/Rater</th>
<th>N</th>
<th>Rater-rater Correlation</th>
<th>Rater-OPI Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>85</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>B</td>
<td>84</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>C</td>
<td>92</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>D</td>
<td>76</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>E</td>
<td>49</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>F</td>
<td>53</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>G</td>
<td>33</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>472</td>
<td><strong>0.88</strong></td>
<td><strong>0.90</strong></td>
</tr>
</tbody>
</table>

*In all cases, rating different OPI performances
Rater-based prediction range: 0.84-0.93*
How Versants Compare to OPIs (3)

ILR OPI Score (logits) vs. Versant Arabic Overall Score

N = 118
r = 0.87
Spanish Data: Versant ~ OPI

ILR OPI Score (logits)

Versant Spanish Score

N = 37
r = 0.92
# Concordance Table

<table>
<thead>
<tr>
<th>SST Overall Score</th>
<th>Best Estimate of ILR Score</th>
<th>$\geq$ ILR Score with 80% Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21-35</td>
<td>0+</td>
<td>At least 0+</td>
</tr>
<tr>
<td>36-43</td>
<td>1</td>
<td>At least 0+</td>
</tr>
<tr>
<td>44-49</td>
<td>1+</td>
<td>At least 1</td>
</tr>
<tr>
<td>50-55</td>
<td>2</td>
<td>At least 1+</td>
</tr>
<tr>
<td>56-60</td>
<td>2</td>
<td>At least 2</td>
</tr>
<tr>
<td>61-66</td>
<td>2+</td>
<td>At least 2</td>
</tr>
<tr>
<td>67-71</td>
<td>2+</td>
<td>At least 2+</td>
</tr>
<tr>
<td>72-77</td>
<td>3</td>
<td>At least 2+</td>
</tr>
<tr>
<td>78-80</td>
<td>3</td>
<td>At least 3</td>
</tr>
</tbody>
</table>
ILR triage steps

- Versant scores provide ILR level floors (twice) with 92% accuracy

LowerBound(ILR) = ILR - (t-score)(standard error of the estimate)

For 80% confidence, 36 df:  \( t = 0.85 \) (one tailed)
Construct Comparison

**COMMUNICATIVE COMPETENCE***

- Organization
  - Grammar
    - V
    - M
    - S
    - P
  - Text
    - Coh
    - Rh
  - Illocution
    - Ideat
    - Manip
    - Huer
    - Imag
  - Socioling.
    - Dial
    - Reg
    - Nat
    - Cult

**LANGUAGE FACILITY***

- Grammar
  - V
  - M
  - S
  - P
- Skill
  - Rate
  - Fluency

*Bachman

*Versant

**Taxonomic**

**Metric in time**
Construct Comparison

OPI Construct: Oral Proficiency as manifest in an Oral Proficiency Interview, is compatible with communicative competence as reflected in the functional level and/or complexity of content accurately produced.

Versant Construct: facility in spoken language – the ability to understand spoken language and speak appropriately in response at a conversational pace on everyday topics.
Shared variance

COMMUNICATIVE COMPETENCE*

Organization

Pragmatics

Grammar

Text

Illocution

Socioling.

VERSANT tests contain sufficient material for predictive scoring

VERSANT

0.87 prediction

0.87 prediction

LANGUAGE FACILITY

Grammar

Skill

V M S P

Rate Fluency

V M S P

Coh Rh Ideat Manip Huer Imag Dial Reg Nat Cult

Pearson Knowledge Technologies, Palo Alto, California

ECOLT November 2008
Summary

• Versant Arabic Test (VAT) is in operation
• Based on a large and wide body of transcribed spoken material
• VAT is available on demand
• VAT can triage or screen for OPI tests

• Useful score correspondence may be provided even when constructs differ
النهاية