Relating Automatic Spoken Spanish Test Scores to the ILR Scale

29 October 2004

East Coast Organization of Language Testers (ECOLT)
George Washington University

Jennifer Balogh, Jared Bernstein, Isabella Barbier, Elizabeth Rosenfeld

Ordinate Corporation
Menlo Park, California
Presentation

- Spoken Spanish Test (SST) Description
- Relating SST to ILR scale
  - Concurrent validity using ILR scale
  - Predicting ILR scores
Description of SST

• Computerized Spoken Spanish Test
  • Taken over the telephone
    • 15 minutes to complete
    • Landline phone
  • Automated administration and scoring
    • Uses speech recognition technology
    • Scores available on secure web site
SST Construct

- Measures facility in spoken Spanish
  - Ease and immediacy in understanding and producing appropriate conversational Spanish.

Adapted from Levelt, 1989
## SST Design

<table>
<thead>
<tr>
<th>Test Part</th>
<th>Task Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part A</td>
<td>Read Aloud</td>
<td>Julio había recibido de regalo una hermosa bicicleta último modelo. Julio was given the latest model of a beautiful bicycle as a gift.</td>
</tr>
<tr>
<td>Part B</td>
<td>Repeat Sentences</td>
<td>El joven camina por la calle. The man walks along the street.</td>
</tr>
<tr>
<td>Part C</td>
<td>Say the Opposite</td>
<td>alto&lt;br&gt;high</td>
</tr>
<tr>
<td>Part D</td>
<td>Answer Short Questions</td>
<td>¿Cuántas patas tiene un perro? How many legs does a dog have?</td>
</tr>
<tr>
<td>Part E</td>
<td>Build Sentences</td>
<td>te / María / ama&lt;br&gt;you / Maria / loves</td>
</tr>
<tr>
<td>Part F</td>
<td>Answer Open Questions</td>
<td>¿Prefiere usted vivir en la ciudad o en el campo? Por favor explique su selección. Do you prefer to live in the city or the countryside? Please explain your choice.</td>
</tr>
<tr>
<td>Part G</td>
<td>Retell Stories</td>
<td>Tres niñas caminaban a la orilla de un arroyo cuando vieron a un pajarito con las patitas enterradas en el barro...</td>
</tr>
</tbody>
</table>
SST Design and Scoring Logic

**SST** = (30% Sent.M, 20% Vocab, 30% Fluency, 20% Pron)
Presentation

- Spoken Spanish Test (SST) Description
- Relating SST to ILR scale
  - Concurrent validity using ILR scale
  - Predicting ILR scores
Validity Framework

- State argument
- Assemble evidence
- Evaluate most problematic assumptions
- Restate argument (repeat cycle)

ARGUMENT:
SST scores will be highly correlated with human ratings (ILR scale)
Concurrent Validity Evidence

- SST Machine Scores
- ILR-SPT Human Interview Scores
- ILR-SPT Estimates (2 human raters per)
SPT OPI (SPT Interviews)

SPT OPI ~ ILR Estimate-SPT

Same Two Raters
Different Material
\[ r = 0.94 \]

SPT OPI ~ SST

Two Raters ~ Machine
Different Material
\[ r = 0.92 \]
SST ~ ILR Estimate-SPT

Machine ~ Two Raters
Different Material

$ r = 0.89 $
Validity Framework

• State argument
• Assemble evidence
• Evaluate most problematic assumptions
  • Why are correlations so high when constructs are different?
• Restate argument (repeat cycle)
Theory of Language Proficiency: Automaticity

Counsel, persuade, advise

Limited understanding and ability to respond

Better listening and speaking

Language model

resources

Fluent understanding and ability to respond
Presentation

• Description of Spoken Spanish Test

• Relating SST to ILR scale
  • Concurrent validity using ILR scale
  • Predicting ILR scores
Argument

SST scores will accurately predict ILR lower bound scores for military use

1. Methodology
2. Evidence
Predicting ILR Scores from SST Scores

1. Express ILR scores in logits
   Mapping based on IRT analysis of ILR estimates
   Double scoring of 6 responses (same 2 raters)

2. Generate regression equation
Predicting ILR Scores from SST Scores

\[ \text{logit(ILR)} = 0.19(\text{SST}) - 12.69 \]
Predicting ILR Scores from SST Scores

1. Express ILR scores in logits
   Mapping based on IRT analysis of ILR estimates
   Double scoring of 6 responses (same 2 raters)

2. Generate regression equation
   \[ \text{logit(ILR)} = 0.19(\text{SST}) - 12.69 \]

3. Convert logits to ILR scale
   Use thresholds from FACETS analysis
Predicting ILR Scores from SST Scores

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

For 80% confidence, 36 df: \( t = 0.85 \) (one tailed)
## 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>
Validity Evidence

Validate lower bound prediction

- 92% of observed ILR SPT interview scores ≥ lower bound
- 92% of observed ILR SPT estimates ≥ lower bound

What about data not used to generate scores?

DLI OPI data
Validity Evidence: DLI OPIs

Only 6% below lower bound

N=63
r=0.70
Conclusions

• SST scores are highly correlated with human ratings on the ILR scale

  Automaticity theory explains why correlations are high even though constructs are different

• SST scores accurately predict ILR lower bound scores for military use

  Lower bound cut-off scores at 80% confidence account for 92% of observed scores