A Common Testing Framework for Computerized Spoken Language Tests in Multiple Languages

19 May 2005

Jennifer Balogh
Ordinate Corporation, Menlo Park, California
Oral Proficiency Interviews

Interviewer 1
Interviewer 2
Candidate

Here is a question
Oral Proficiency Interviews

Interviewer 1
Interviewer 2

Here is an answer

Candidate
Oral Proficiency Interviews

Here is an answer

Interviewer 1
Interviewer 2
Candidate
Oral Proficiency Interviews

Interviewer 1
Interviewer 2
Candidate

Here is an answer
Oral Proficiency Interviews

Interviewer 1
Interviewer 2
Candidate

Here is an answer
Goals

- Create reliable & valid spoken language tests
  - Automatically administered
  - Automatically scored

- Use a common framework for all tests
  - Multiple languages
  - Quick development
Automatic Administration

Call: 1-800-444-7277 or 1-650-328-0336

Test Identification Number
1234 5678

Introduction:
Thank you for calling the Ordinate testing system. Please enter your Test Identification Number on the telephone keypad. Now, please say the city and country you are calling from. Now, please follow the instructions for Parts A through G.

Part A: Reading. Please read the sentences as you are instructed.
1. Julio recibió de regalo una nueva bicicleta último modelo.
2. Iba a la universidad y al trabajo en su nueva bicicleta todos los días.
3. Un día a la salida del trabajo se dio cuenta de que se la habían robado.
4. Se quedó muy triste pensando que no le había durado mucho tiempo.
5. Habían quedado en encontrarse en el parque, al mediodía.
6. Mientras pasaban largo le contó sus planes y le preguntó si quería compartirlos.
7. Ella se quedó pensando un rato, luego le dijo que sí, que le acompañaría.
8. Felices, se abrazaron y decidieron ir a comer juntos al restaurante de la esquina.

Part B: Repeat. Please repeat each sentence that you hear.
Example: a voice says, "Le gustaba cantar canciones románticas." and you say, "Le gustaba cantar canciones románticas."

Part C: Opposites. Now, when you hear a word, just say the opposite.
Example: a voice says, "alto" and you say "bajo".

Part D: Questions. Now, please just give a simple answer to the questions.
Example: a voice says, "Si estuviera enfermo, ¿a quién iría a ver: a un médico o a un vendedor?" and you say, "un médico" or "a un médico."

Part E: Sentence Builds. Now, please rearrange the word groups into a sentence.
Example: a voice says, "y lo vimos" ... "nos asomamos" ... "marcharse" and you say, "Nos asomamos y lo vimos marcharse."
Process flow

1. Proctor downloads test from Ordinate web site and prints out

2. Candidate reviews test sheet & instructions (5min)

3. Candidate calls Ordinate – toll free number on test sheet

4. Candidate keys in unique (8-digit) Test Identification Number

5. Candidate takes test (~10min)

6. Candidate/Proctor retrieves score from Ordinate web site (within minutes after taking the test)
Spoken Spanish Test

Administration Date: Jun 30, 2004 6:38:32 PM (PDT)
Test Identification Number: 68073395

Overall (32)
The Overall Score of the test represents the ability to understand spoken Spanish and speak it intelligibly at a native conversational pace on everyday topics. Scores are based on a weighted combination of four diagnostic subscores. Scores are reported in the range from 20 to 80.

- Test taker can manage slow, short, isolated utterances, or spoken formulas, but has difficulty following any native-paced conversation and will frequently need to ask for repetition; test taker often pauses to search for words and may often be difficult to understand.

Sentence Mastery (38)
Sentence Mastery reflects the ability to understand, recall and produce Spanish phrases and clauses in complete sentences. Performance

Ordinate Corporation  ALTE, Berlin, Germany  May 2005 10
Common Testing Framework

Test Architecture  Testing System  Development and Validation Process

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Task A</th>
<th>Task B</th>
<th>Task C</th>
<th>Task D</th>
<th>Task E</th>
<th>Task F</th>
<th>Task G</th>
</tr>
</thead>
</table>

Phone
Network

![Graph](image)
Common Testing Framework

Test Architecture

<table>
<thead>
<tr>
<th>Task Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A</td>
</tr>
<tr>
<td>Task B</td>
</tr>
<tr>
<td>Task C</td>
</tr>
<tr>
<td>Task D</td>
</tr>
<tr>
<td>Task E</td>
</tr>
<tr>
<td>Task F</td>
</tr>
<tr>
<td>Task G</td>
</tr>
</tbody>
</table>

Testing System

Development and Validation Process
Test Construct

Measures *facility in spoken language*

Adapted from Levelt, 1989
Theory of Language Proficiency: Automaticity

resources

Limited understanding and ability to respond
Theory of Language Proficiency: Automaticity

Better understanding and ability to respond
Theory of Language Proficiency: Automaticity

resources

Counsel, persuade, advise

Language model

Fluent listening and speaking
# Test 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.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The man walks along the street.</td>
</tr>
<tr>
<td>Part C</td>
<td>Say the Opposite</td>
<td>“alto”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>high</td>
</tr>
<tr>
<td>Part D</td>
<td>Answer Short Questions</td>
<td>“¿Cuántas patas tiene un perro?”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How many legs does a dog have?</td>
</tr>
<tr>
<td>Part E</td>
<td>Build Sentences</td>
<td>“te / María / ama”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>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 elección.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>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>
**Introduction:**

*Thank you for calling the Ordinate testing system.*

*Please enter your Test Identification Number on the telephone keypad.*

*Now, please say the city and country you are calling from.*

*Now, please follow the instructions for Parts A through G.*

<table>
<thead>
<tr>
<th>Part A: Reading. Please read the sentences as you are instructed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Julio recibió de regalo una nueva bicicleta último modelo.</td>
</tr>
<tr>
<td>2. Isa a la universidad y al trabajo en su nueva bicicleta todos los días.</td>
</tr>
<tr>
<td>3. Un día a la salida del trabajo se dio cuenta de que se la habían robado.</td>
</tr>
<tr>
<td>4. Se quedó muy triste pensando que no le había durado mucho tiempo.</td>
</tr>
<tr>
<td>5. Habían quedado en encontrarse en el parque, al mediodía.</td>
</tr>
<tr>
<td>6. Mientras pasaban Jorge le contó sus planes y le preguntó si quería compartirlos.</td>
</tr>
<tr>
<td>7. Ella se quedó pensando un rato, luego le dijo que sí, que la acompañaría.</td>
</tr>
<tr>
<td>8. Felices, se abrazaron y decidieron ir a comer juntos al restaurante de la esquina.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B: Repeat. Please repeat each sentence that you hear.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example: a voice says, &quot;Le gustaba cantar canciones románticas,&quot; and you say, &quot;Le gustaba cantar canciones románticas.&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part C: Opposites. Now, when you hear a word, just say the opposite.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example: a voice says, &quot;alto&quot; and you say &quot;bajo&quot;.</td>
</tr>
<tr>
<td>Example: a voice says, &quot;abierta&quot; and you say &quot;cerrado&quot;.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part D: Questions. Now, please just give a simple answer to the questions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example: a voice says, &quot;¿Si estuvieras enfermo, ¿a quién irías a ver: a un médico o a un vendedor?&quot; and you say, &quot;un médico&quot; or &quot;a un médico&quot;.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part E: Sentence Builds. Now, please rearrange the word groups into a sentence.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example: a voice says, &quot;y lo vimos...&quot; and you say, &quot;Nos sorprenderon&quot; &quot;marcharse&quot;</td>
</tr>
<tr>
<td>and you say, &quot;Nos sorprenderon y lo vimos marcharse.&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part F: Open Questions. You will have 30 seconds to answer each of two questions. The questions will be about family life or personal choices. Each question will be spoken twice, followed by a beep. When you hear the beep, you will have 30 seconds to answer the question. At the end of the 30 seconds, another beep will signal the end of the time you have to answer.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Part G: Story Retelling. You will hear two brief stories. After each story, you will have 30 seconds to retell it in Spanish as best you can. Try to retell as much of the story as you can in Spanish, including the situation, characters, actions and ending.</th>
</tr>
</thead>
</table>
Scoring

Content
- Vocabulary
- Sentence Mastery

Manner
- Pronunciation
- Fluency
Scoring Logic

SST = (30% Sent. M, 20% Vocab, 30% Fluency, 20% Pron)
Common Testing Framework

Test Architecture  |  Testing System  |  Development and Validation Process

Task Type
Task A
Task B
Task C
Task D
Task E
Task F
Task G
Public Switched Telephone Network (PSTN)

internet

Virtual Private Network (VPN)

Test Delivery Server (TDS)

servers

Database
Common Testing Framework

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Test Architecture</th>
<th>Testing System</th>
<th>Development and Validation Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Development

**PLAN**

1. Define Test
2. Construct
3. Design Test
4. Create Item Specification
Development

PLAN  DEVELOP

Define Test Construct

Design Test

Create Item Specification

Write Items

Review Items

Record Items

Create Data Collection System
Development

**PLAN** → **DEVELOP** → **SCORE**

- **Define Test**
  - Construct
- **Write Items**
  - Review Items
  - Record Items
- **Create Data Collection System**
- **Collect Data**
  - Train Speech Recognizer
  - Create Models
  - Generate Scores

- **Design Test**
- **Create Item Specification**
- **Train Speech Recognizer**
- **Record Items**
- **Create Models**
- **Generate Scores**
Validation

- Reliability

- Validity
  - Separation of Natives and Non-Natives
  - Correlation with Human Ratings
# Reliability

<table>
<thead>
<tr>
<th>Score Types</th>
<th>SET-10</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>Sentence Mastery</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.97</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Separation of Natives and Non-Natives
# Machine and Human Ratings

### Machine Scores

<table>
<thead>
<tr>
<th>Read</th>
<th>Repeat Sentence</th>
<th>Opposite</th>
<th>Short Question</th>
<th>Build S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Human Interview Scores

Here is a question

Interviewer 1

Interviewer 2

Candidate

### Estimates

2 human raters per
## SET-10 Concurrent Validity

<table>
<thead>
<tr>
<th>Instrument</th>
<th>r</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>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>Common European Framework, 1st experiment</td>
<td>0.84</td>
<td>121</td>
</tr>
<tr>
<td>Common European Framework, 2nd experiment</td>
<td>0.94</td>
<td>150</td>
</tr>
<tr>
<td>Common European Framework, 3rd experiment</td>
<td>0.88</td>
<td>303</td>
</tr>
</tbody>
</table>
SST ~ ACTFL OPI

Two Raters ~ Machine
Different Material
\( r = 0.86 \)
SST ~ SPT OPI

Two Raters ~ Machine
Different Material

$r = 0.92$
SET-10 ~ CEF Estimates

PhonePass
SET-10
Overall
Score

Non-Native
speakers
(n=268)

Native
speakers
(n=35)

Council of Europe Language Levels

r=0.88
SST ~ ILR Estimate-SPT

Two Raters ~ Machine
Different Material
\[ r = 0.89 \]
Common Testing Framework

Test Architecture | Testing System | Development and Validation Process

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Task A</th>
<th>Task B</th>
<th>Task C</th>
<th>Task D</th>
<th>Task E</th>
<th>Task F</th>
<th>Task G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>Phone</td>
<td>Network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phone Network

SST QPI (ECPI Logits)

SST (with mapped logits)

N=37

p<0.02
Conclusions

• Common Testing Framework
  Can be used for any language
  Allows for quick development

• Spoken Language Tests
  Automatically administered and scored
  Reliable
  Valid