An Automated Spoken Chinese Test: Test Description, Comparison to other Tests and Classroom Application

Xiaoqiu Xu, Masa Suzuki
Pearson Knowledge Technologies
& Xiaoqi Li
Peking University
A Quick Survey

• When each new student enters your program, do you know about his/her proficiency level?
• How often do you assess students’ oral proficiency (as a summative assessment)?
• What are the challenges of assessing students’ oral proficiency?
Imagine there is a test...

- A Spoken Chinese Test
- Can be easily administered by phone or computer
- Can be scored within minutes
- Accurate and reliable
- Detailed score report
- Validated by experts
- A variety of item types
Research Questions

• How to develop a fully automated Spoken Chinese Test with such characteristics (refer to previous slide)?

• What validation evidence demonstrates it is a good, accurate, consistent test?
A bit of history on types of speaking tests

- **Direct testing** (‘50s)
  - Face-to-face interview
  - Paired or group testing
  
- **Semi-direct testing** (‘80s)
  - Computer-mediated
  - Tape-mediated interview

- **Automated testing** (’90s)
  - Computer-mediated
  - Speech processing

Admin: Human  
Score: Human

Admin: Technology  
Score: Human

Admin: Technology  
Score: Technology
## SCT: Test Structure

<table>
<thead>
<tr>
<th>Part (部分)</th>
<th>Item Type (试题)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Tone Phrases (声调词语)</td>
</tr>
<tr>
<td>B</td>
<td>Read Aloud (朗读)</td>
</tr>
<tr>
<td>C</td>
<td>Repeats (重复)</td>
</tr>
<tr>
<td>D</td>
<td>Short Answer Questions (问答)</td>
</tr>
<tr>
<td>E</td>
<td>Recognize Tones – Word (声调识别 – 词)</td>
</tr>
<tr>
<td>F</td>
<td>Recognize Tone – Sentence (声调识别 – 句子)</td>
</tr>
<tr>
<td>G</td>
<td>Sentence Builds (组句)</td>
</tr>
<tr>
<td>H</td>
<td>Passage Retellings (短文重述)</td>
</tr>
</tbody>
</table>
SCT: Score Report

**Content**
- Vocabulary
- Grammar

**Manner**
- Pronunciation
- Fluency
- Tone

Score: 58
SCT: Test Flow

Candidate accesses the testing system by telephone or computer → Candidate responds to the test questions or prompts → Responses are sent to the scoring system → Testing System scores the test and posts scores to the web reporting tool

SCT takes 20 minutes; results available within 10 minutes
SCT: Test Development

PLAN
- Define Test
- Construct
- Design Test
- Create Item Specification

DEVELOP
- Write Items
- Review Items
- Record Items
- Create Data Collection System

SCORE
- Collect Data
- Train Speech Recognizer
- Create Models
Item Writing & Review

- **Vocabulary** (5,186 frequent Chinese words)

- **Item Writing**: adapted from naturally occurring utterances (e.g. TV Programs)
  
  - Original: 后来这段车祸对自己的人生经历有没有什么影响？是否更加珍惜生命？
  - Revised: 这段经历对他的人生有没有什么影响？

  - Original: 哎，我觉得是他可能就是说，回学校这环境可能稍微好一点。
  - Revised: 我觉得那里的环境可能稍微好一点。

- **Item Review**
  
  - 检查语言形式是否与受过教育的中国人使用普通话进行日常对话的形式相吻合；
  - 有没有任何在日常对话中不使用的字词或语句？
  - 有没有一些地域性很强的字词？
  - 有没有一些文化性很强的字词？
Item Recording and Review

- Item Recording - use voice talents from different dialect backgrounds:
  - Beijing natives 北京, Jianghuai 江淮, Jilu 吉鲁, Dongbei 东北, jiaoliao 胶辽, Wu 吴, jin 晋, Xiang 湘, Min 闽, and Taiwanese 台湾, etc.

- Recording Review
  - Count number of pronunciation errors
  - Mark the level of accent
  - Items with errors or strong accents were all removed

good recording                       removed recording
Sampling for Field Test

Purpose of Field Testing:
• Validate the operation of the test items
• Calibrate the difficulty of each item
• Train the automatic speech processing system
• Develop automatic scoring models.

938 native speakers -> 1,969 tests
2,459 non-native speakers -> 4,142 tests
SCT: Human Rating

Human Rating

Fluency
Pronunciation
Tone
Content

>180,000+ ratings
Validation

• How reliable is the test? (reliability)
• What is the relationship between the spoken Chinese test and human rating? (internal validation)
• Is the test able to distinguish native and non-native speakers? (test construct)
• What is the relationship between the spoken Chinese test and other speaking tests? (external or concurrent validation)
### SCT Test Reliability

Estimated through different methods

<table>
<thead>
<tr>
<th>Score</th>
<th>Split-half Reliability (N=166)</th>
<th>Test – Retest Reliability (N=158)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Grammar</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>Tone</td>
<td>0.92</td>
<td>0.87</td>
</tr>
</tbody>
</table>
SCT Machine – Human Correlation

How well do machine scores line up with human scores?

<table>
<thead>
<tr>
<th>Score</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.98</td>
</tr>
<tr>
<td>Grammar</td>
<td>0.96</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.96</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.96</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.95</td>
</tr>
<tr>
<td>Tone</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Native vs. Non-Native

![Graph showing comparison between Native and Non-Native scores](image)
Native, Non-Native, & Heritage

Graph showing percent cumulative density against spoken Chinese test overall score with different groups:
- N=1822, mean=84.2, All Native
- N=3554, mean=46.8, All Non-Native
- N=291, mean=62.3, All Heritage
## SCT vs. Other Speaking Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Correlation</th>
<th>Number of test-takers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSK Oral Test – Intermediate Level vs. SCT</td>
<td>0.86</td>
<td>148</td>
</tr>
<tr>
<td>ILR-OPI</td>
<td>0.87</td>
<td>169</td>
</tr>
</tbody>
</table>
SCT Expert Validation

“北京大学汉语口语自动化考试”是北大和全球领先的教育和评估公司 Pearson 合作开发的自动化评估系统。该系统用于评估和标准化普通话口语能力。


北京大学汉语口语自动化考试

2021 年 8 月 17 日
How is the SCT Used?

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proficiency Test</strong></td>
<td>• Determine test takers' proficiency levels in a standardized manner</td>
</tr>
<tr>
<td><strong>Placement Test</strong></td>
<td>• Accurate placement for targeted levels</td>
</tr>
<tr>
<td></td>
<td>• Verifying Chinese for study abroad programs</td>
</tr>
<tr>
<td></td>
<td>• Graduate and move up one level in program</td>
</tr>
<tr>
<td><strong>Monitor Progress</strong></td>
<td>• Monitor student language progress over a certain period of time</td>
</tr>
<tr>
<td></td>
<td>• Ability to share results and utilize results as learning tool</td>
</tr>
</tbody>
</table>
Questions?