Factor structure of a spoken Chinese test: investigating five subskill scores for diagnosis

LTRC 2013 in Seoul

Masanori Suzuki
Yujie Jia
Pearson Knowledge Technologies
Overview of the Presentation

- Background on Spoken Chinese Test
- Test Structure
- Scoring logic
- Research Questions
  - Construct validation
  - Reliability analysis of subscores
- Analysis and Results
- Summary
Spoken Chinese Test

- Fully automated test of spoken Mandarin Chinese
- Jointly developed between Peking Univ and Pearson
- Development period: June 2010 - August 2012
- Integrated listening-speaking item types
- Delivery options: phone, computer
- Scored automatically by in-house, custom-developed speech processing technologies and computerized scoring systems
Test Design
**Spoken Chinese Test**

- **8 item types**
- **70 items**
- **20 min**
- **Overall score**
- **5 analytic subscores**
- **Score Scale: 20-80**
Facility in spoken Chinese

The ability to understand spoken Chinese on everyday topics and to respond intelligibly in Chinese at a native-like conversational pace.

Adapted from Levelt, 1989
<table>
<thead>
<tr>
<th>Part</th>
<th>Item Type</th>
<th>Response Characteristics</th>
<th>Scored Trait(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Tone Phrases</td>
<td>Word, Phrase</td>
<td>Tone Production</td>
</tr>
<tr>
<td>B</td>
<td>Read Aloud</td>
<td>Sentence</td>
<td>Pronunciation, Tone production</td>
</tr>
<tr>
<td>C</td>
<td>Sentence Repetition</td>
<td>Sentence</td>
<td>Grammar, Pronunciation, Fluency, Tone production</td>
</tr>
<tr>
<td>D</td>
<td>Short Answer Questions</td>
<td>Word, Phrase</td>
<td>Vocabulary</td>
</tr>
<tr>
<td>E</td>
<td>Recognize Tones - Word</td>
<td>Word</td>
<td>Tone reception</td>
</tr>
<tr>
<td>F</td>
<td>Recognize Tone - Sentence</td>
<td>Word</td>
<td>Tone reception</td>
</tr>
<tr>
<td>G</td>
<td>Sentence Builds</td>
<td>Sentence</td>
<td>Grammar, Pronunciation, Fluency, Tone production</td>
</tr>
<tr>
<td>H</td>
<td>Passage Retellings</td>
<td>Passage</td>
<td>Vocabulary, Fluency</td>
</tr>
</tbody>
</table>
2 Test Methods

• Test Method 1: Short Responses
  Tone Phrase
  Read Aloud
  Repeat
  Short Answer Questions
  Sentence Builds
  Recognize Tones

• Test Method 2: Long Responses
  Passage Retelling
Example: Short response

Sentence Repeats

要下雨了。
*It’s going to rain.*

后来他又去了一次。
*Afterwards, he went back again.*

报纸上说明天下午两点开始。
*The newspaper says that it starts at two o’clock tomorrow afternoon.*

*Scored traits: Grammar, Pron, Fluency, Tone Production*
Cellphones are great. No matter where you are, you can always make calls and send text messages. Nowadays cellphones have even more functions. You can use them not only to listen to music but also to surf the Internet.

Scored traits: Vocabulary, Fluency
Score reporting design

Overall

Content
  - Grammar
  - Vocabulary

Manner
  - Fluency
  - Pronunciation
  - Tone

Five analytic subscores
Multi-Trait, Multi-Method

Grammar
Vocabulary
Fluency
Pronunciation
Tone

Tone
Read Aloud
Repeat
Short Answer Question
R Tone Word
R Tone Sent
Sentence Build
Passage Retell

20 minutes

Grammar, Vocabulary, Fluency, Pronunciation, and Tone are interconnected through various activities such as Tone Phrase, Read Aloud, Repeat, Short Answer Question, R Tone Word, R Tone Sent, Sentence Build, and Passage Retell. This diagram illustrates the integration of multi-trait and multi-method approaches in language learning.
Five Analytic Subscores

- Fox and Fraser (2009)
  "What is surprising here is that the test developers have not emphasized the obvious diagnostic properties of the *The Versant Spanish™ Test.* “ (p.319)

- SCT’s test design and scoring logic is similar to that of Versant Spanish
  → Can SCT’s subscores be useful for diagnostic purposes?
Conditions for Being Useful

• SCT should exhibit evidence that the test measures facility in spoken Chinese (construct validation)

• SCT’s subscores should be stable estimates (reliability)
Research Questions

RQ 1: Does the SCT discriminate as expected according to known populations?

RQ 2: What is the factor structure of the SCT test scores?

RQ 3: Are SCT’s subscores reliable enough to be taken as stable indicators of test-takers’ strengths and weaknesses?
Native Data Set

- 1,822 completed tests
- Various dialect groups
- Used to develop scoring systems

<table>
<thead>
<tr>
<th>Dialect</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>57%</td>
</tr>
<tr>
<td>Wu</td>
<td>13%</td>
</tr>
<tr>
<td>Min</td>
<td>9%</td>
</tr>
<tr>
<td>Yue</td>
<td>6%</td>
</tr>
<tr>
<td>Xiang</td>
<td>3%</td>
</tr>
<tr>
<td>Gan</td>
<td>2%</td>
</tr>
<tr>
<td>Hakka</td>
<td>2%</td>
</tr>
<tr>
<td>Jing</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
Learner Data Set 1

- 3,845 completed tests
- Various countries and L1 backgrounds
- Used to develop scoring systems

<table>
<thead>
<tr>
<th>L1</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>17%</td>
</tr>
<tr>
<td>Japanese</td>
<td>13%</td>
</tr>
<tr>
<td>Korean</td>
<td>9%</td>
</tr>
<tr>
<td>Russian</td>
<td>7%</td>
</tr>
<tr>
<td>Spanish</td>
<td>6%</td>
</tr>
<tr>
<td>Arabic</td>
<td>5%</td>
</tr>
<tr>
<td>Thai</td>
<td>5%</td>
</tr>
<tr>
<td>Bengali</td>
<td>5%</td>
</tr>
</tbody>
</table>
Learner Data Set 2

- 166 learners of Chinese
- Various countries and L1 backgrounds
- Set aside for validation

<table>
<thead>
<tr>
<th>L1</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>24%</td>
</tr>
<tr>
<td>Korean</td>
<td>13%</td>
</tr>
<tr>
<td>Cantonese</td>
<td>11%</td>
</tr>
<tr>
<td>Japanese</td>
<td>8%</td>
</tr>
</tbody>
</table>
Research Questions

RQ 1: Does the SCT discriminate according to known populations as expected?

RQ 2: What is the factor structure of the SCT test scores?

RQ 3: Are SCT’s subscores reliable enough to be taken as stable indicators of test-takers’ strengths and weaknesses?
Educated native Chinese speakers of different dialects
Educated native speakers and learners of Chinese
Educated native speakers, learners, heritage-Mandarin speakers, and heritage-Non-Mandarin speakers
## Accuracy of SCT’s Automated Scoring

<table>
<thead>
<tr>
<th>Score</th>
<th>Machine – Human Correlation (n=166)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.98</td>
</tr>
<tr>
<td>Grammar</td>
<td>0.97</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.97</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.93</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.90</td>
</tr>
<tr>
<td>Tone</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Automated Scoring vs. Human Scoring

N=166
r=0.98
Research Questions

RQ 1: Does the SCT discriminate according to known populations as expected?

RQ 2: What is the factor structure of the SCT test scores?

RQ 3: Are SCT’s subscores reliable enough to be taken as stable indicators of test-takers’ strengths and weaknesses?
Eight Variables

1. Grammar
2. ShortVoc
3. LongVoc
4. ShortFlu
5. LongFlu
6. Pronunciation
7. Tone Production
8. Tone Recognition
Data Analysis

• Confirmatory Factor Analysis
  
  Four CFA models were hypothesized and tested in this study

  1) One-factor model (Unidimensional construct)
  2) Correlated two-factor model
  3) Bi-factor model
  4) Correlated trait-uncorrelated method model
One-Factor Model

Oral

Grammar
Pron
LongVoc
ShortVoc
ShortFlu
LongFlu
ToneRec
ToneProd
Correlated Two-Factor Model

Content

Manner

Grammar

Pron

LongVoc

ShortVoc

ShortFlu

LongFlu

ToneRec

TonePro
Correlated-Traits, Uncorrelated-Method

- **Content**
  - Grammar
  - Pron
  - LongVoc
  - ShortVoc
  - ShortFlu
  - LongFlu
  - ToneRec
  - ToneProd

- **Manner**
- **Long**
- **Short**

**Correlated Traits (yellow) and Uncorrelated Method (blue) connecting mechanisms.**
## Results

<table>
<thead>
<tr>
<th>Model Description</th>
<th>One-factor</th>
<th>Correlated two-factor</th>
<th>Bi-factor</th>
<th>Correlated trait-uncorrelated method</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>20</td>
<td>19</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Minimum fit function chi-square</td>
<td>3256.17</td>
<td>2151.25</td>
<td>336.62</td>
<td>194.35</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.273</td>
<td>0.227</td>
<td>0.148</td>
<td>0.074</td>
</tr>
<tr>
<td>CFI</td>
<td>0.836</td>
<td>0.892</td>
<td>0.984</td>
<td>0.991</td>
</tr>
<tr>
<td>NFI</td>
<td>0.835</td>
<td>0.891</td>
<td>0.983</td>
<td>0.990</td>
</tr>
<tr>
<td>NNFI</td>
<td>0.770</td>
<td>0.840</td>
<td>0.971</td>
<td>0.983</td>
</tr>
</tbody>
</table>
## Model Comparison

<table>
<thead>
<tr>
<th>Models compared</th>
<th>df difference</th>
<th>Chi-square difference</th>
<th>Significance (p&lt;.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-factor vs. Correlated two-factor</td>
<td>1</td>
<td>1104.92</td>
<td>Significant</td>
</tr>
<tr>
<td>Correlated two-factor vs. Bi-factor</td>
<td>3</td>
<td>1814.63</td>
<td>Significant</td>
</tr>
<tr>
<td>Bi-factor vs. Correlated trait-uncorrelated method</td>
<td>1</td>
<td>142.27</td>
<td>Significant</td>
</tr>
</tbody>
</table>
Factor Loadings and Correlations

<table>
<thead>
<tr>
<th>Variables</th>
<th>Content</th>
<th>Manner</th>
<th>Long</th>
<th>Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>0.88*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LongVoc</td>
<td>0.73*</td>
<td>0.57*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ShortVoc</td>
<td>0.76*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronunciation</td>
<td></td>
<td>0.83*</td>
<td></td>
<td>0.71*</td>
</tr>
<tr>
<td>ShortFluency</td>
<td>0.80*</td>
<td></td>
<td>0.18*</td>
<td></td>
</tr>
<tr>
<td>LongFluency</td>
<td>0.73*</td>
<td>0.57*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToneReception</td>
<td>0.33*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToneProduction</td>
<td>0.80*</td>
<td></td>
<td></td>
<td>0.43*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Manner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>0.915*</td>
</tr>
</tbody>
</table>

*p<.05
Research Questions

RQ 1: Does the SCT discriminate as expected according to known populations?

RQ 2: What is the factor structure of the SCT test scores?

RQ 3: Are SCT’s subscores reliable enough to be taken as stable indicators of test-takers’ strengths and weaknesses?
# Test Reliability

<table>
<thead>
<tr>
<th>Score</th>
<th>Split-half Method (n=166)</th>
<th>Test – Retest Method (n=158)</th>
<th>Human Scoring (n=166)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td><strong>0.97</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>Grammar</td>
<td>0.92</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.94</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.97</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>0.96</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>Tone</td>
<td>0.93</td>
<td>0.87</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Summary

• SCT appears to measure facility in spoken Chinese, discriminating different test-taker groups (L1 Chinese, Heritage speakers, Non-heritage learners)
• SCT’s MTMM is supported
• SCT’s score reporting logic with content and manner aspects of language performance is supported
• SCT’s reliability estimates are high and stable
Implications

• Subscores do not provide any specific problem areas that could be useful for intervention
e.g., no specific grammar points or phonetic points

• SCT subscores can still be good indicators of *relative* strengths and weaknesses

• Perhaps, conduct a survey with test score users to understand whether and how these subscores are useful for diagnostic purposes
Thank you

masanori.suzuki@pearson.com
yujie.jia@pearson.com