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

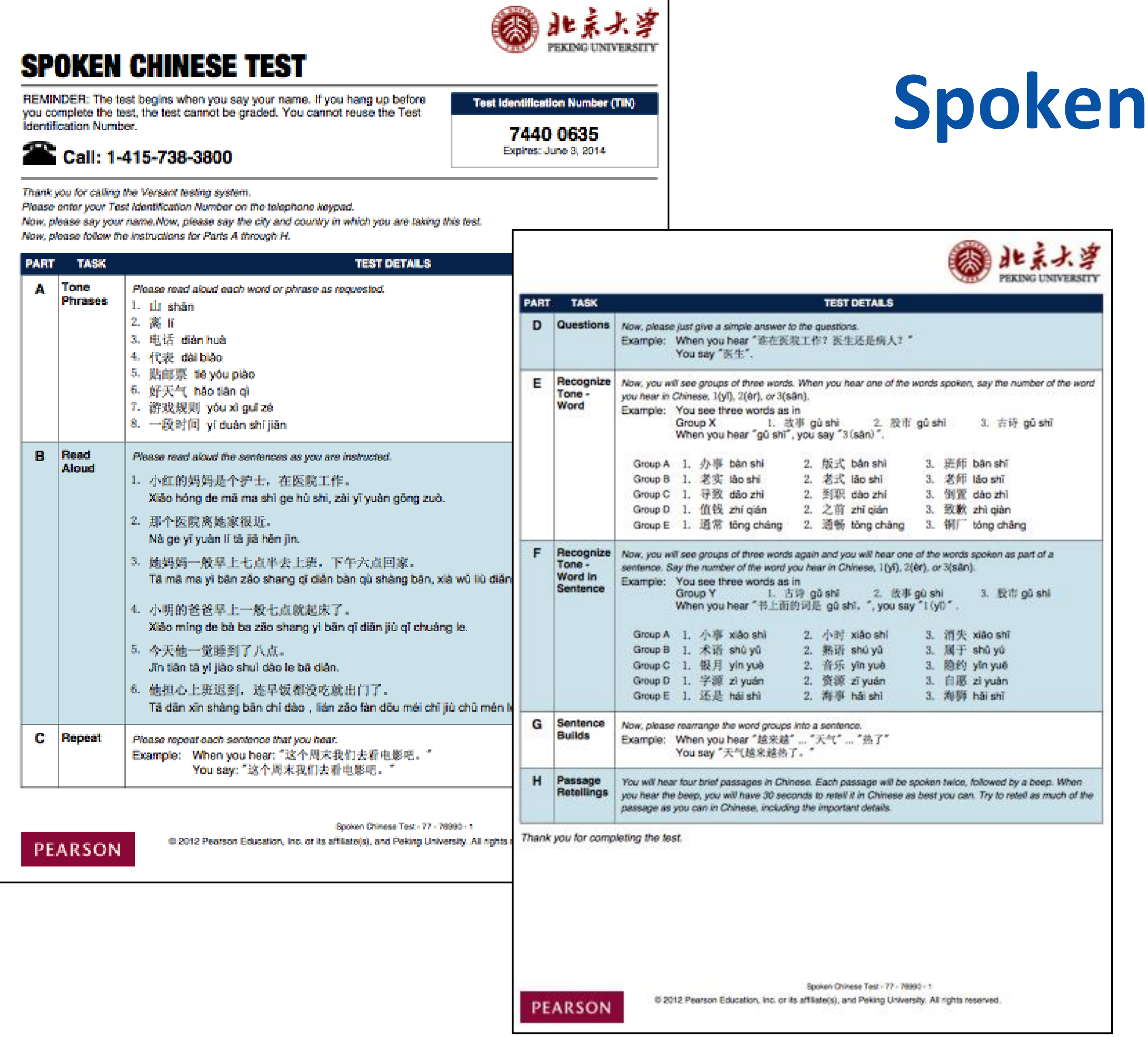
- Background on Spoken Chinese Test
- Test Structure
- Scoring logic
- Research Questions
  - Construct validation
  - Reliability analysis of subscores
- Analysis and Results
- Summary

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



SPOKEN CHINESE TEST

北京大学
PEKING UNIVERSITY

REMINDER: The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.

☎ Call: 1-415-738-3800

Test Identification Number (TIN)
7440 0635 Expires: June 3, 2014

Thank you for calling the Versant testing system.
Please enter your Test Identification Number on the telephone keypad.
Now, please say your name.Now, please say the city and country in which you are taking this test.
Now, please follow the instructions for Parts A through H.

PART	TASK	TEST DETAILS
A	Tone Phrases	*Please read aloud each word or phrase as requested.* 1. 山 shān 2. 离 lí 3. 电话 diàn huà 4. 代表 dài biǎo 5. 贴邮票 tiē yóu piào 6. 好天气 hǎo tiān qì 7. 游戏规则 yóu xì guī zé 8. 一段时间 yí duàn shí jiān
B	Read Aloud	*Please read aloud the sentences as you are instructed.* 1. 小红的妈妈是个护士，在医院工作。 Xiǎo hóng de mā ma shì ge hù shi, zài yī yuàn gōng zuò. 2. 那个医院离她家很近。 Nà ge yī yuàn lí tā jiā hěn jìn. 3. 她妈妈一般早上七点半去上班，下午六点回家。 Tā mā ma yì bān zǎo shang qī diǎn bàn qù shàng bān, xià wǔ liù diǎn 4. 小明的爸爸早上一般七点就起床了。 Xiǎo míng de bà ba zǎo shang yì bān qī diǎn jiù qǐ chuáng le. 5. 今天他一觉睡到了八点。 Jīn tiān tā yí jiào shuì dào le bā diǎn. 6. 他担心上班迟到，连早饭都没吃就出门了。 Tā dān xīn shàng bān chí dào , lián zǎo fàn dōu méi chī jiù chū mén l
C	Repeat	*Please repeat each sentence that you hear.* Example: When you hear: "这个周末我们去看电影吧。" You say: "这个周末我们去看电影吧。"

PEARSON

Spoken Chinese Test - 77 - 78990 - 1

© 2012 Pearson Education, Inc. or its affiliate(s), and Peking University. All rights

PART	TASK	TEST DETAILS
D	Questions	*Now, please just give a simple answer to the questions.* Example: When you hear "您在医院工作？医生还是病人？" You say "医生".
E	Recognize Tone - Word	*Now, you will see groups of three words. When you hear one of the words spoken, say the number of the word you hear in Chinese, 1(yī), 2(èr), or 3(sān).* Example: You see three words as in Group X 1. 故事 gù shi 2. 股市 gǔ shì 3. 古诗 gǔ shī When you hear "gǔ shī", you say "3(sān)". Group A 1. 办事 bàn shì 2. 版式 bǎn shì 3. 班师 bān shī Group B 1. 老实 lǎo shi 2. 老式 lǎo shì 3. 老师 lǎo shī Group C 1. 导致 dǎo zhì 2. 到职 dào zhí 3. 倒置 dào zhì Group D 1. 值钱 zhí qián 2. 之前 zhī qián 3. 致歉 zhì qiàn Group E 1. 通常 tōng cháng 2. 通畅 tōng chàng 3. 铜厂 tóng chǎng
F	Recognize Tone - Word in Sentence	*Now, you will see groups of three words again and you will hear one of the words spoken as part of a sentence. Say the number of the word you hear in Chinese, 1(yī), 2(èr), or 3(sān).* Example: You see three words as in Group Y 1. 古诗 gǔ shī 2. 故事 gù shi 3. 股市 gǔ shì When you hear "书上面的词是 gǔ shī。", you say "1(yī)". Group A 1. 小事 xiǎo shì 2. 小时 xiǎo shí 3. 消失 xiāo shī Group B 1. 术语 shù yǔ 2. 熟语 shú yǔ 3. 属于 shǔ yú Group C 1. 银月 yín yuè 2. 音乐 yīn yuè 3. 隐约 yǐn yuē Group D 1. 字源 zì yuán 2. 资源 zī yuán 3. 自愿 zì yuàn Group E 1. 还是 hái shì 2. 海事 hǎi shì 3. 海狮 hǎi shī
G	Sentence Builds	*Now, please rearrange the word groups into a sentence.* Example: When you hear "越来越" ... "天气" ... "热了" You say "天气越来越热了。"
H	Passage Retellings	*You will hear four brief passages in Chinese. Each passage will be spoken twice, followed by a beep. When you hear the beep, you will have 30 seconds to retell it in Chinese as best you can. Try to retell as much of the passage as you can in Chinese, including the important details.*

Thank you for completing the test.

PEARSON

Spoken Chinese Test - 77 - 78990 - 1

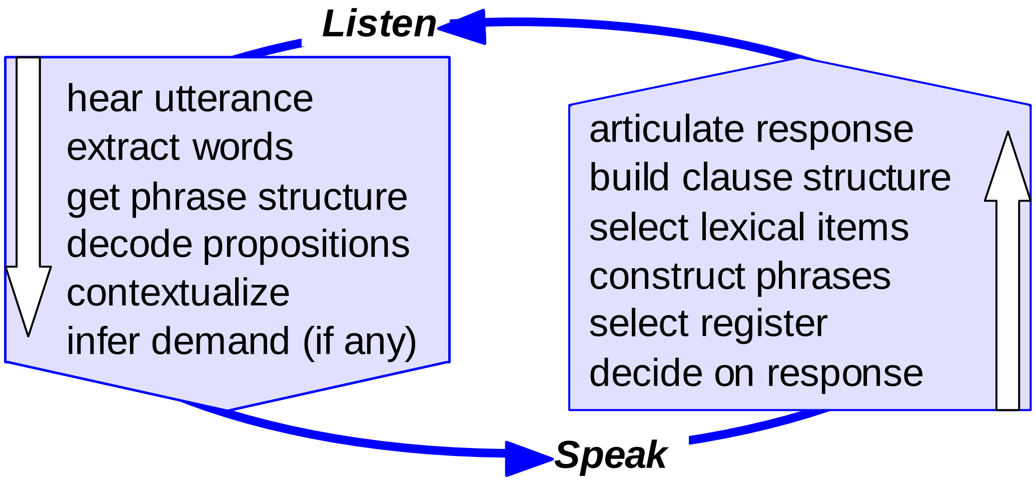
© 2012 Pearson Education, Inc. or its affiliate(s), and Peking University. All rights reserved.

- 8 item types
- 70 items
- 20 min
- Overall score
- 5 analytic subscores
- Score Scale: 20-80

Test Construct

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

Test Structure

Part	Item Type	Response Characteristics	Scored Trait(s)
A	Tone Phrases	Word, Phrase	Tone Production
B	Read Aloud	Sentence	Pronunciation, Tone production
C	Sentence Repetition	Sentence	Grammar, Pronunciation, Fluency, Tone production
D	Short Answer Questions	Word, Phrase	Vocabulary
E	Recognize Tones - Word	Word	Tone reception
F	Recognize Tone - Sentence	Word	Tone reception
G	Sentence Builds	Sentence	Grammar, Pronunciation, Fluency, Tone production
H	Passage Retellings	Passage	Vocabulary, Fluency

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

Example: Long Response

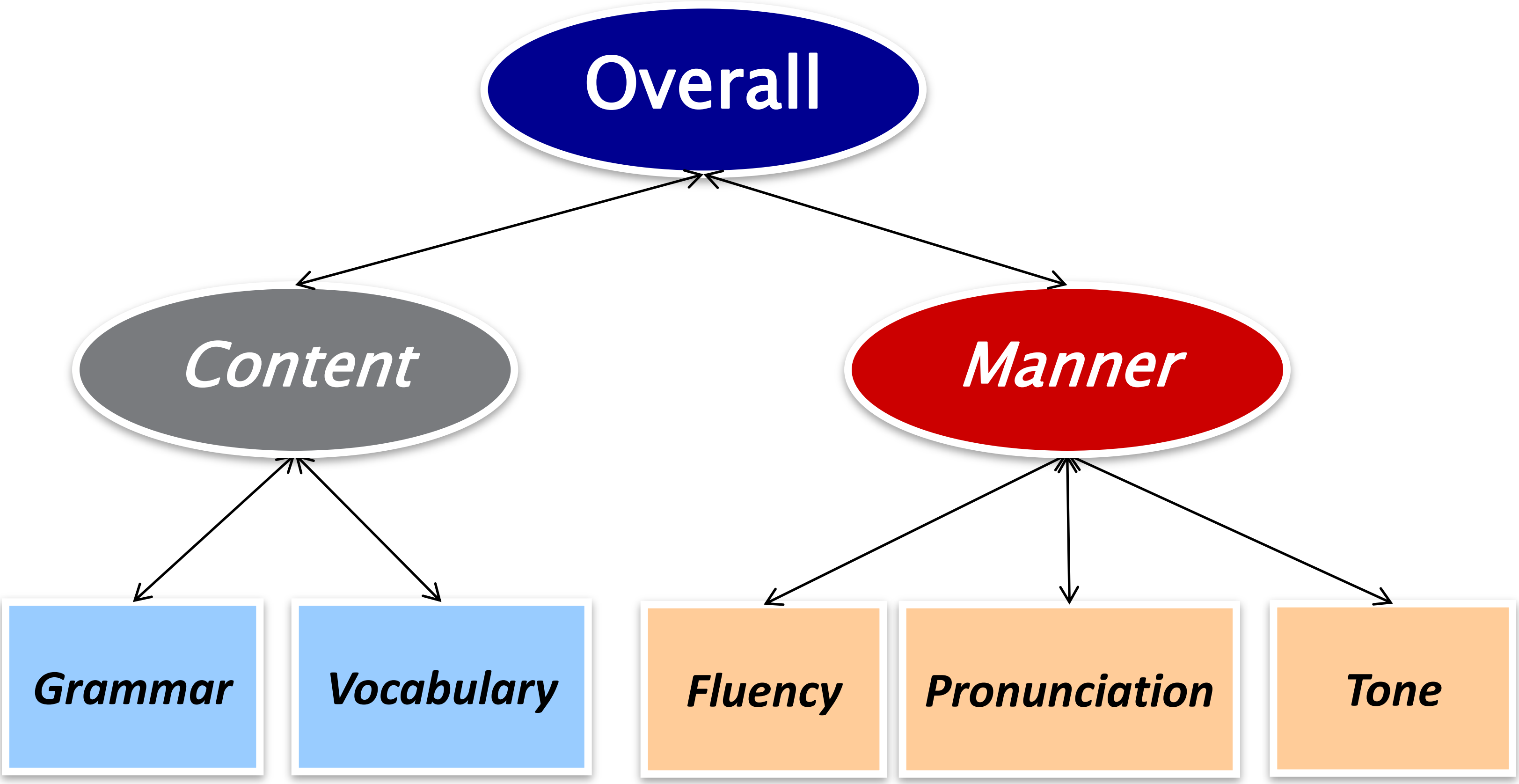
Passage Retelling

手机太好用了。不管你在哪儿，都能随时打电话，发短信。现在的手机功能更多，不但可以听音乐，还能上网呢。

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.

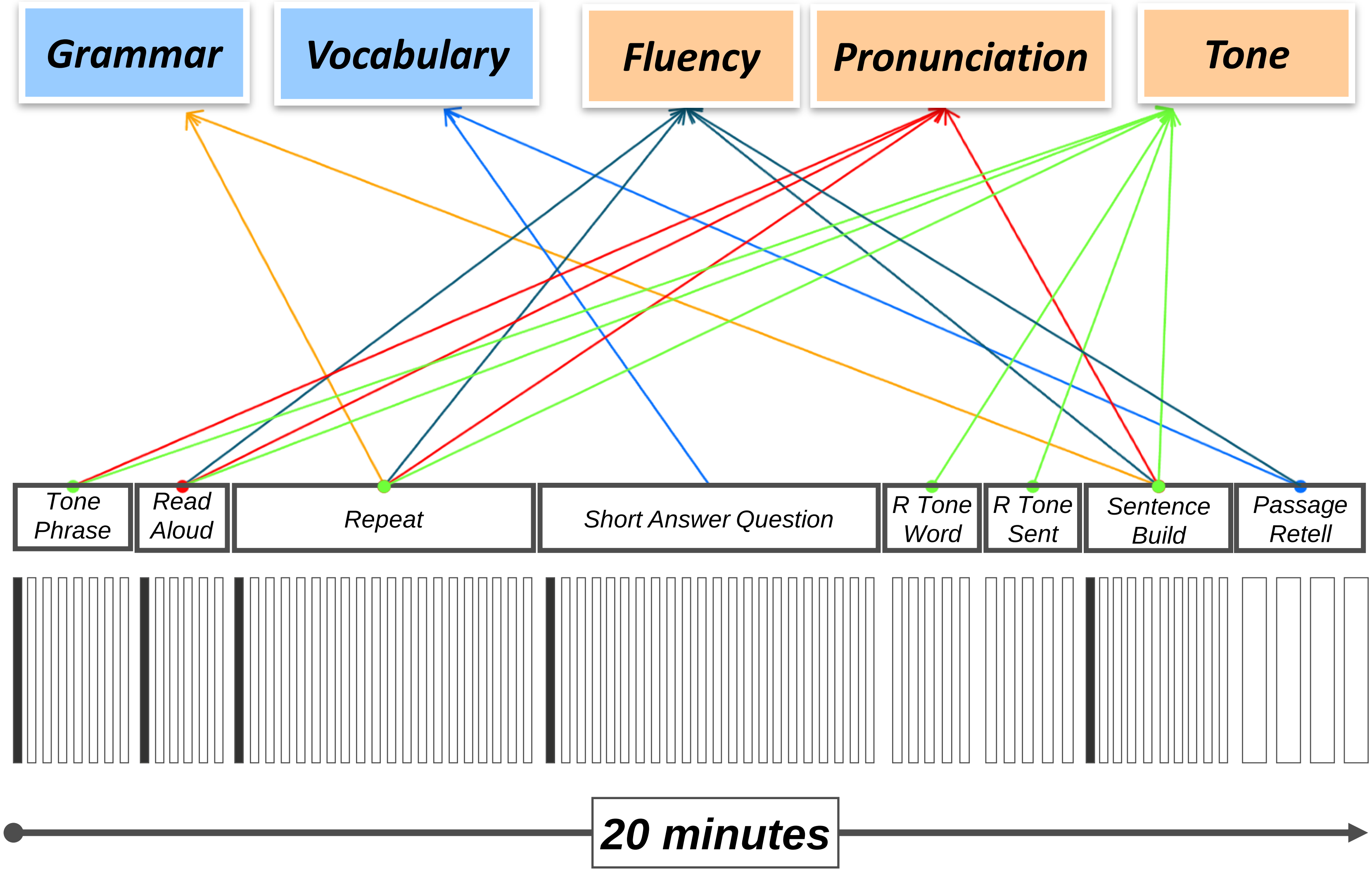
Scored traits: Vocabulary, Fluency

Score reporting design



Five analytic subscores

Multi-Trait, Multi-Method



Five Analytic Subscores

Grammar ***Vocabulary*** ***Fluency*** ***Pronunciation*** ***Tone***

- 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

Dialect	%
Mandarin	57%
Wu	13%
Min	9%
Yue	6%
Xiang	3%
Gan	2%
Hakka	2%
Jing	0.4%

Learner Data Set 1

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

L1	%
English	17%
Japanese	13%
Korean	9%
Russian	7%
Spanish	6%
Arabic	5%
Thai	5%
Bengali	5%

Learner Data Set 2

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

L1	%
English	24%
Korean	13%
Cantonese	11%
Japanese	8%

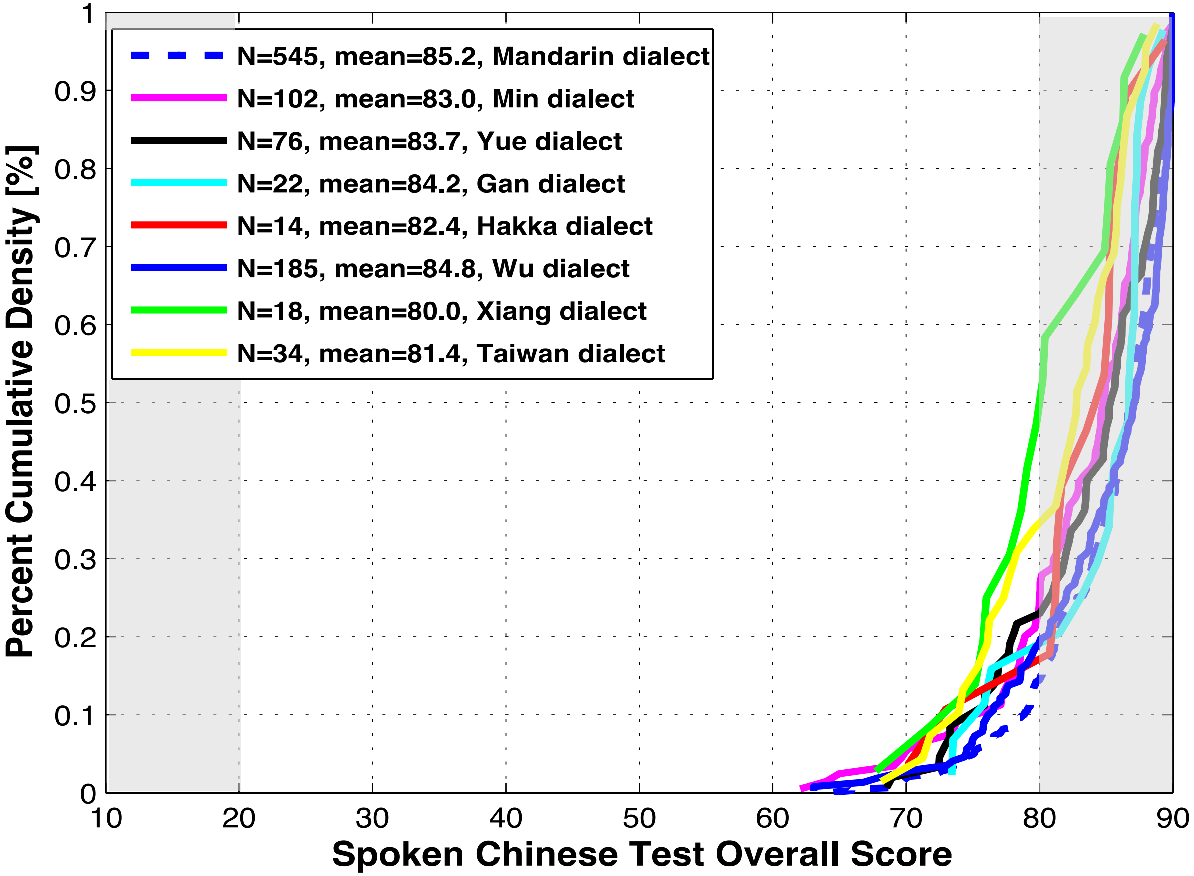
Research Questions

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

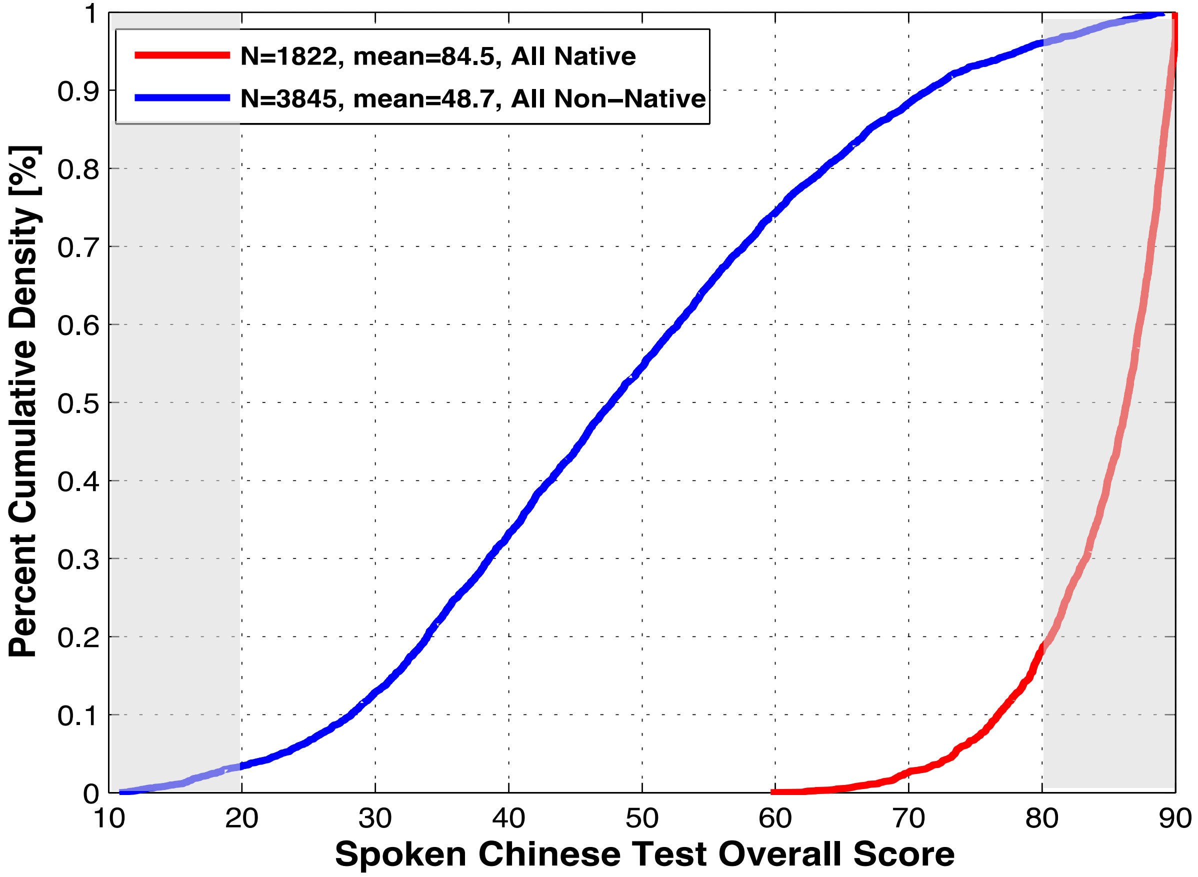
RQ 2: What is the factor structure of the SCT test scores?

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

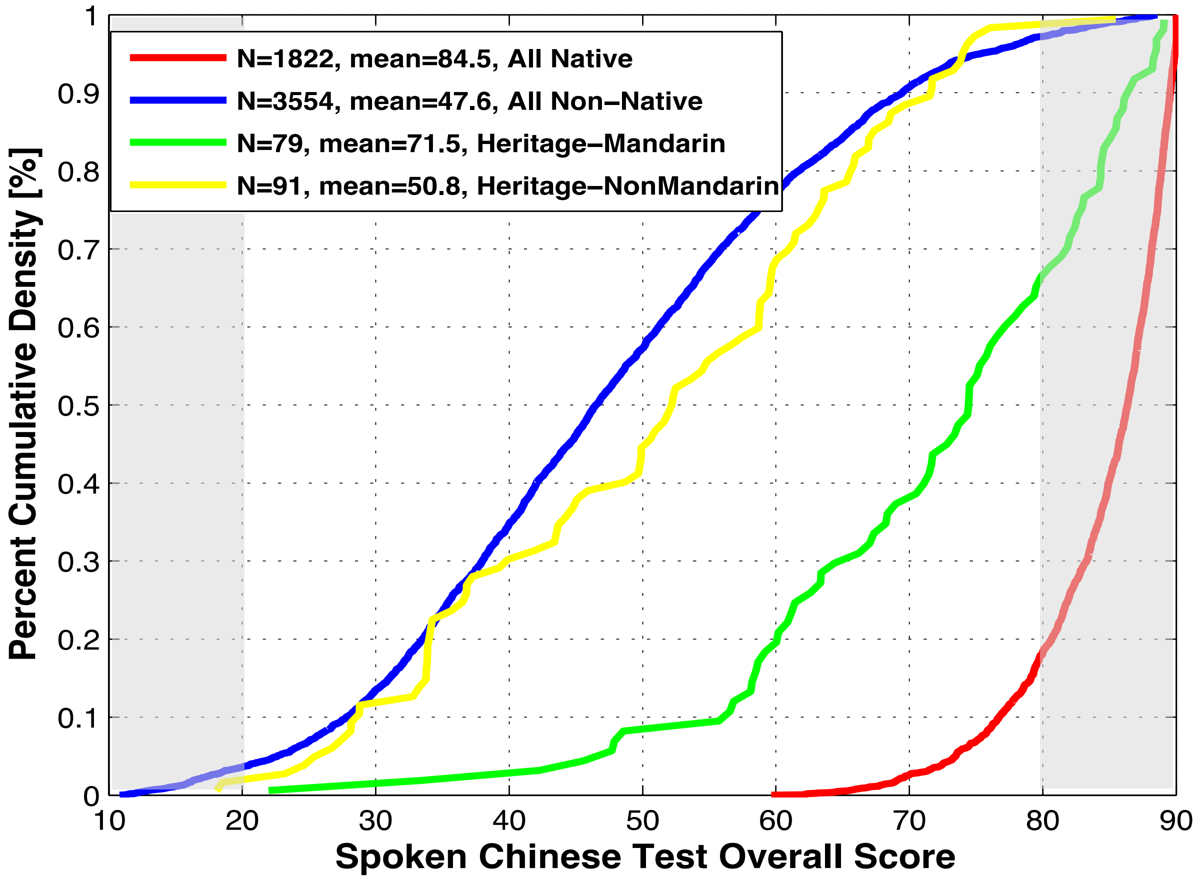
Educated native Chinese speakers of different dialects



Educated native speakers and learners of Chinese



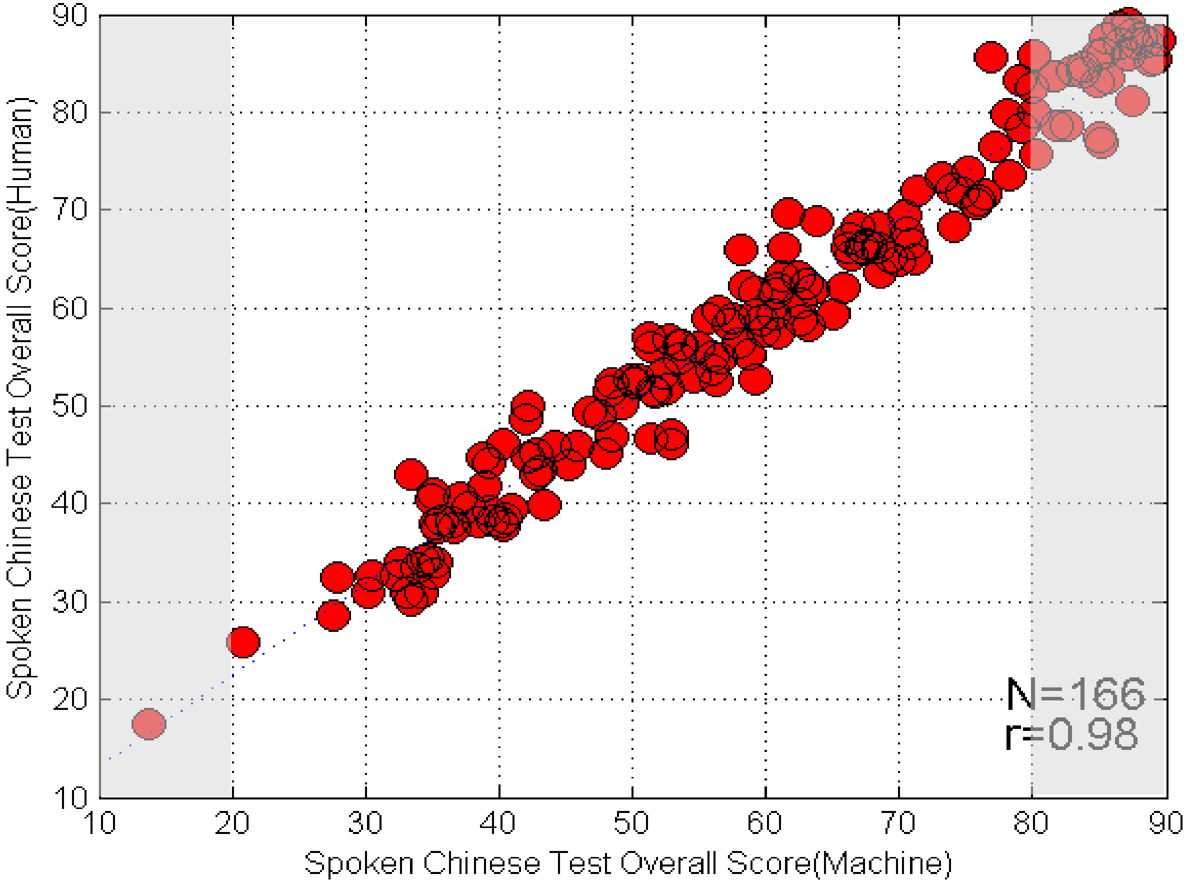
Educated native speakers, learners, heritage-Mandarin speakers, and heritage-NonMandarin speakers



Accuracy of SCT's Automated Scoring

Score	Machine – Human Correlation (n=166)
Overall	**0.98**
Grammar	0.97
Vocabulary	0.97
Fluency	0.93
Pronunciation	0.90
Tone	0.92

Automated Scoring vs. Human Scoring



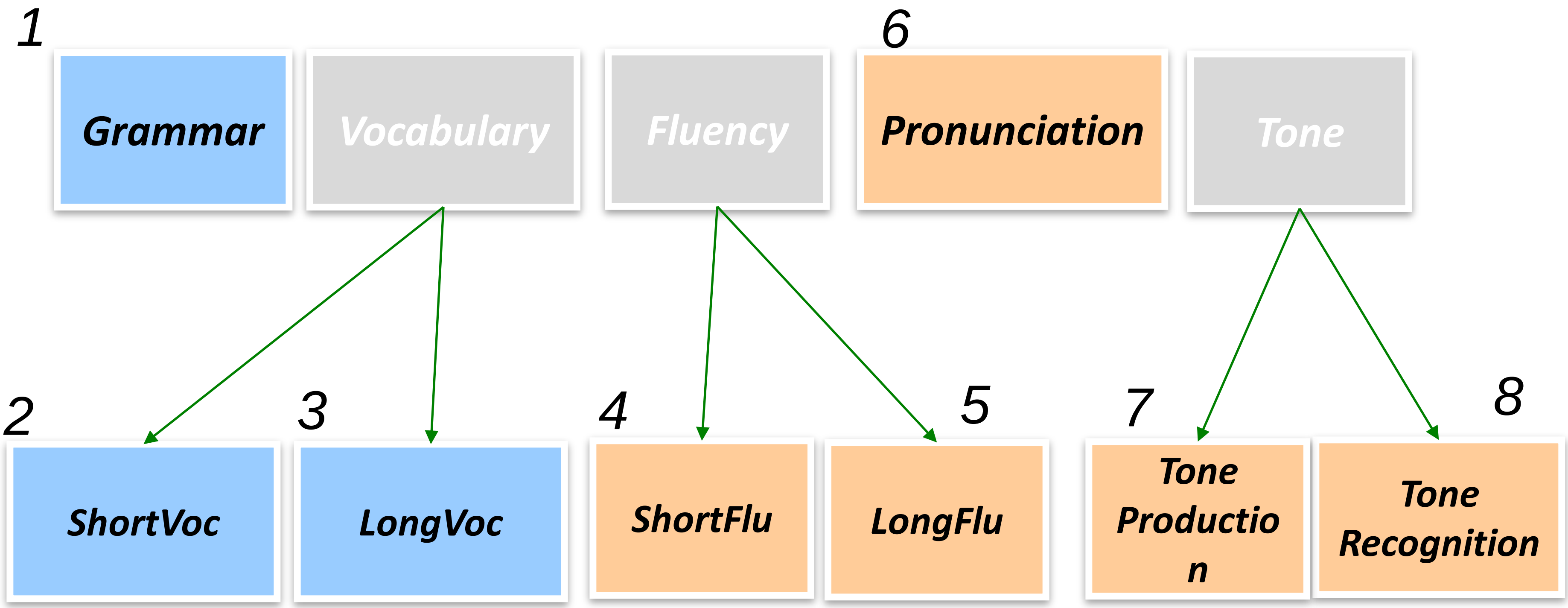
Research Questions

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



Data Analysis

- Confirmatory Factor Analysis

  Four CFA models were hypothesized and tested in this study

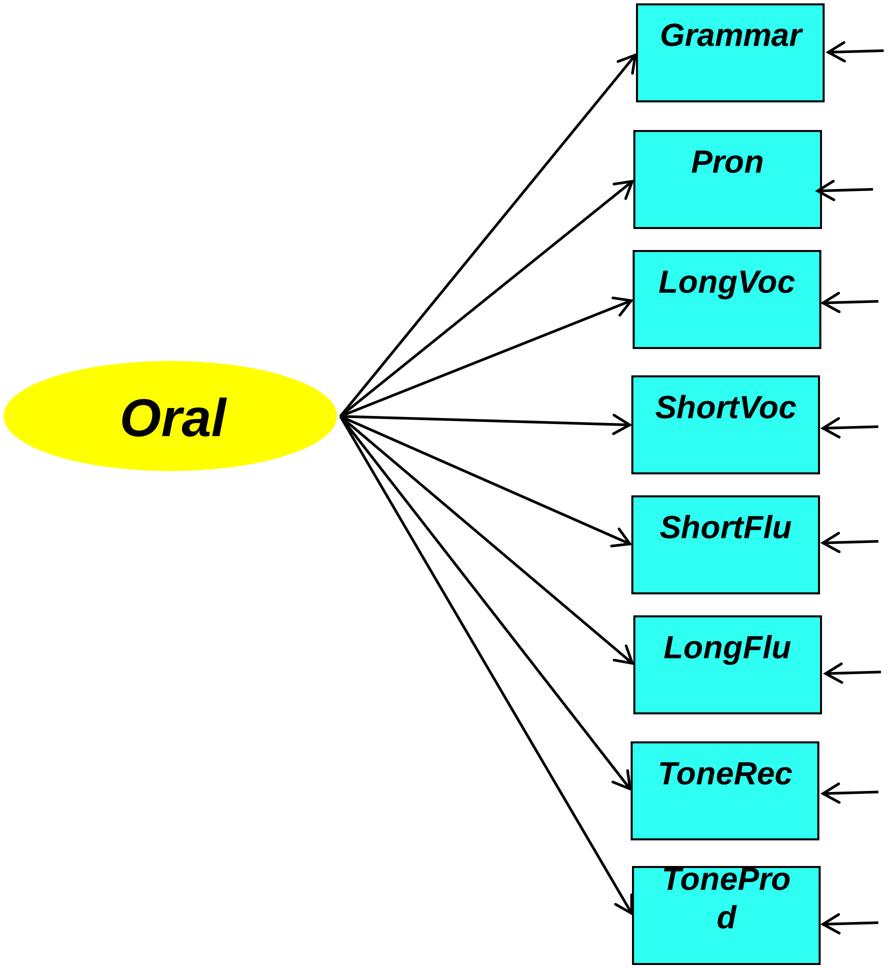
  1) One-factor model (Unidimensional construct)

  2) Correlated two-factor model

  3) Bi-factor model

  4) Correlated trait-uncorrelated method model

One-Factor Model



Oral

Grammar

Pron

LongVoc

ShortVoc

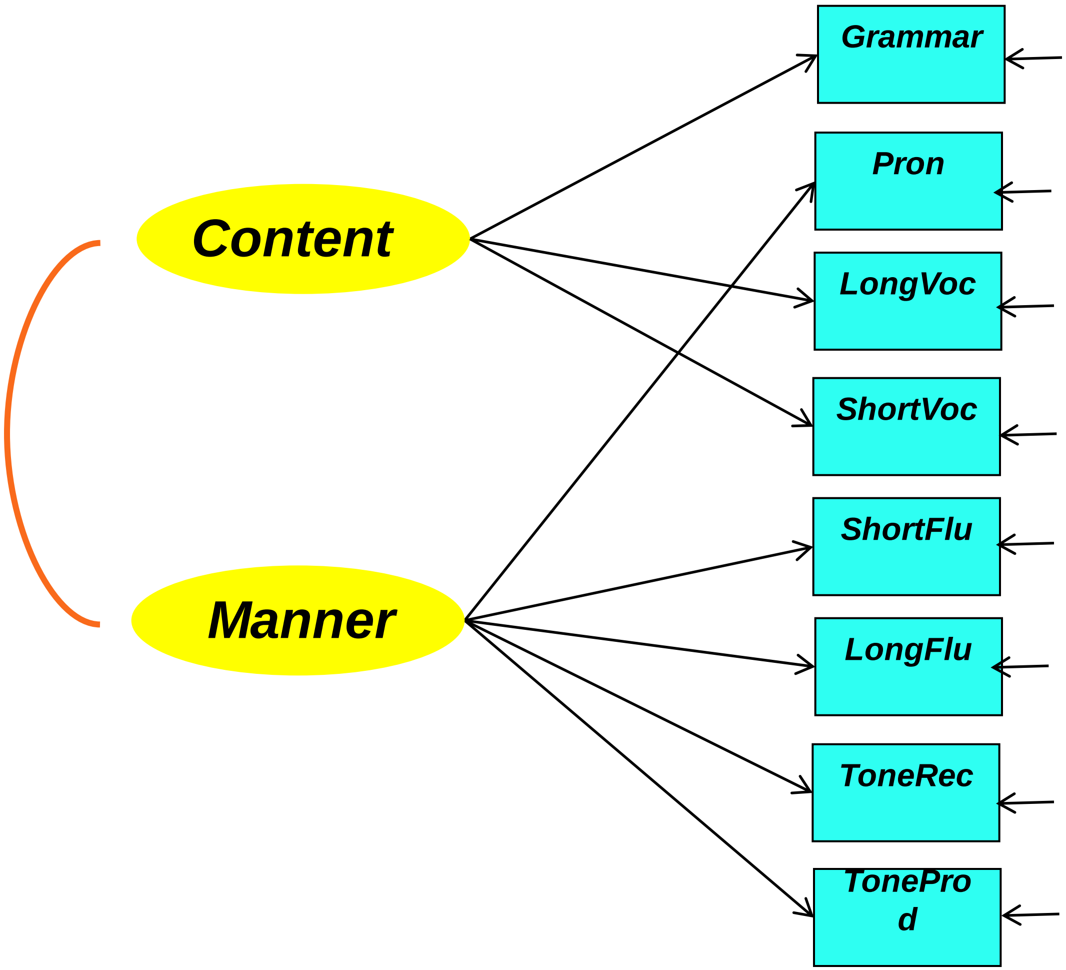
ShortFlu

LongFlu

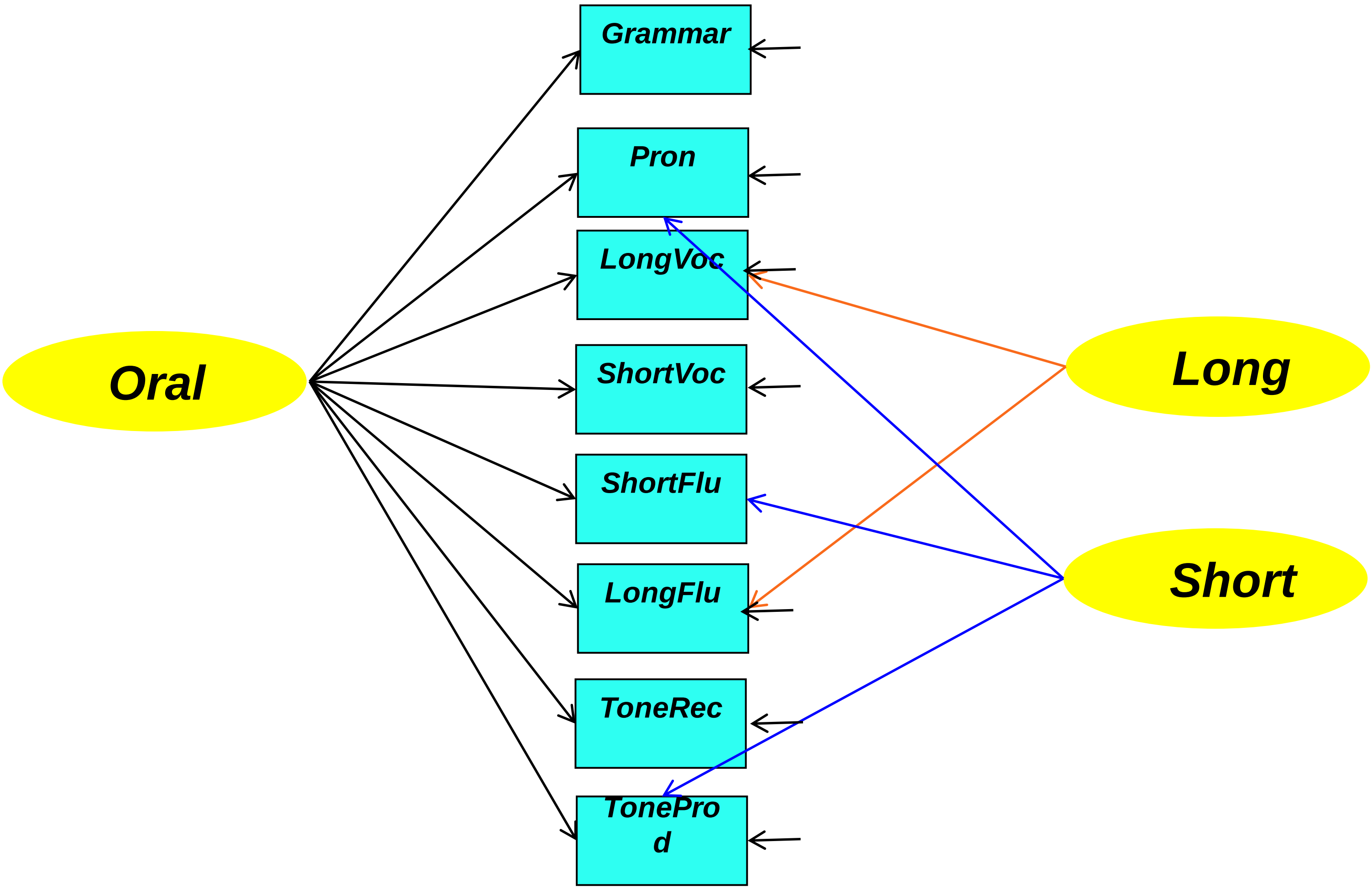
ToneRec

TonePro
d

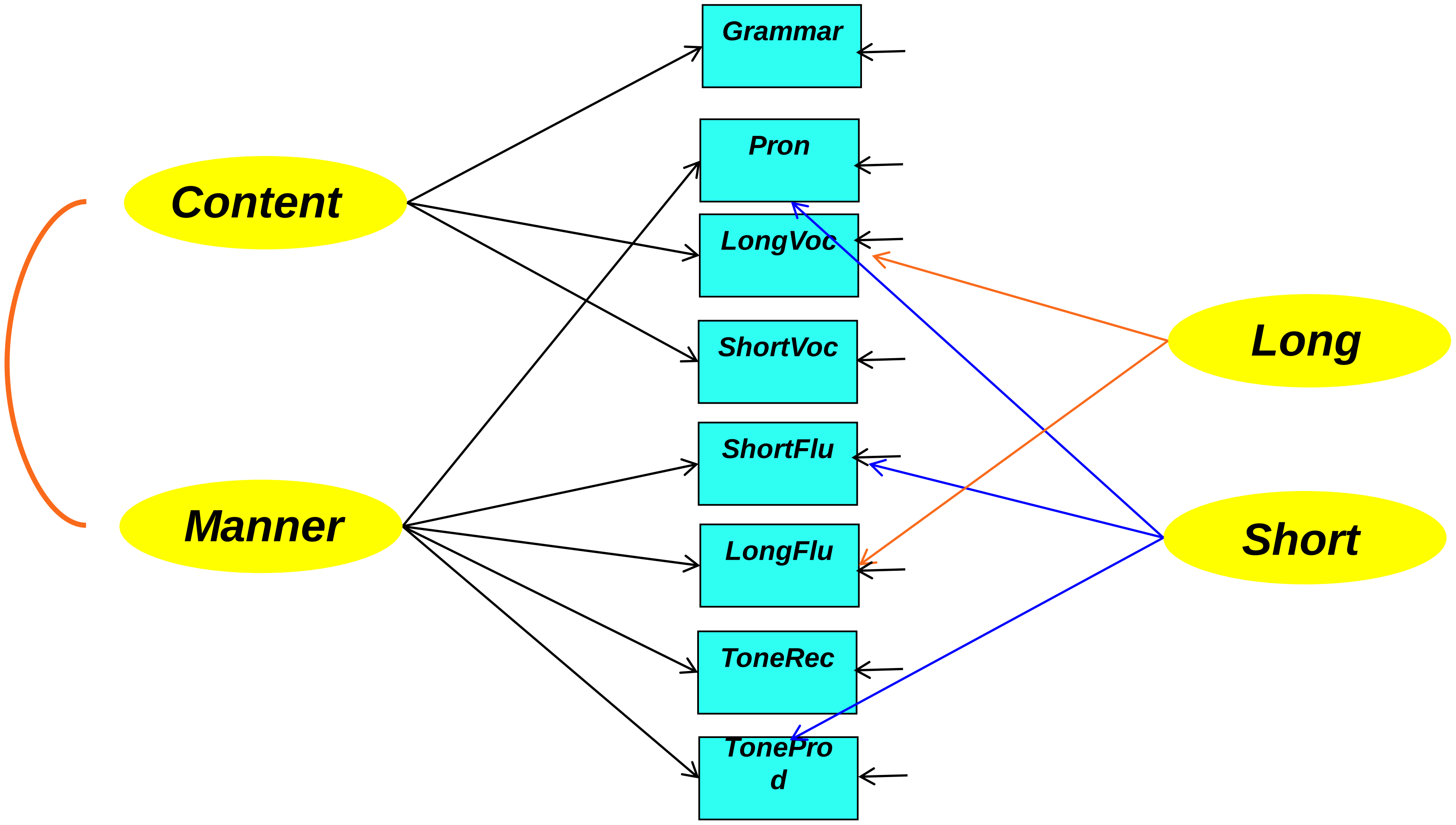
Correlated Two-Factor Model



Bi-Factor Model



Correlated-Traits, Uncorrelated-Method



Results

Model Description	One-factor	Correlated two-factor	Bi-factor	Corrrelated trait-uncorrelated method
df	20	19	16	15
Minimum fit function chi-square	3256.17	2151.25	336.62	194.35
P value	<0.001	<0.001	<0.001	<0.001
RMSEA	0.273	0.227	0.148	0.074
CFI	0.836	0.892	0.984	0.991
NFI	0.835	0.891	0.983	0.990
NNFI	0.770	0.840	0.971	0.983

Model Comparison

Models compared	df difference	Chi-square difference	Significance (p<.05)
One-factor vs. Correlated two-factor	1	1104.92	Significant
Correlated two-factor vs. Bi-factor	3	1814.63	Significant
Bi-factor vs. Correlated trait-uncorrelated method	1	142.27	Significant

Factor Loadings and Correlations

Variables	Content	Manner	Long	Short
Grammar	0.88*			
LongVoc	0.73*		0.57*	
ShortVoc	0.76*			
Pronunciation		0.83*		0.71*
ShortFluency		0.80*		0.18*
LongFluency		0.73*	0.57*	
ToneReception		0.33*		
ToneProduction		0.80*		0.43*

	Manner
Content	0.915*

*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

Score	Split-half Method (n=166)	Test – Retest Method (n=158)	Human Scoring (n=166)
Overall	**0.97**	**0.95**	**0.98**
Grammar	0.92	0.91	0.96
Vocabulary	0.94	0.93	0.96
Fluency	0.97	0.93	0.96
Pronunciation	0.96	0.91	0.95
Tone	0.93	0.87	0.96

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