

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

Xiaoqiu Xu, Masa Suzuki

Pearson Knowledge Technologies

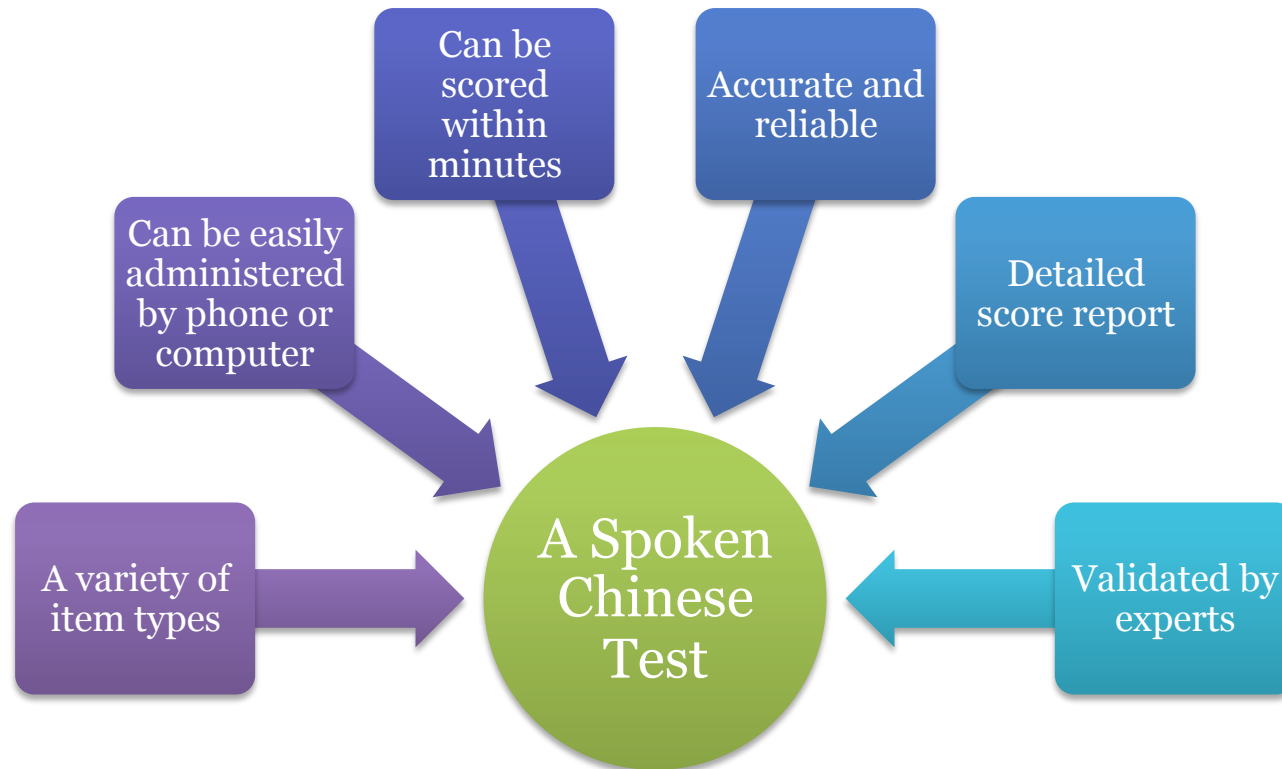
& Xiaoqi Li

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...



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)

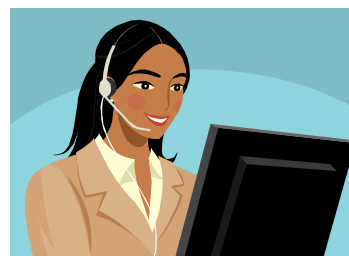
- Face-to-face interview
- Paired or group testing



Admin: Human
Score: Human

- Semi-direct testing ('80s)

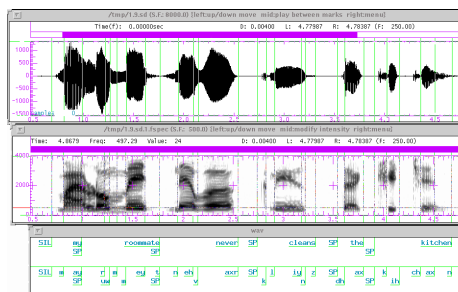
- Computer-mediated
- Tape-mediated interview



Admin: Technology
Score: Human





- Automated testing ('90s)

- Computer-mediated
- Speech processing

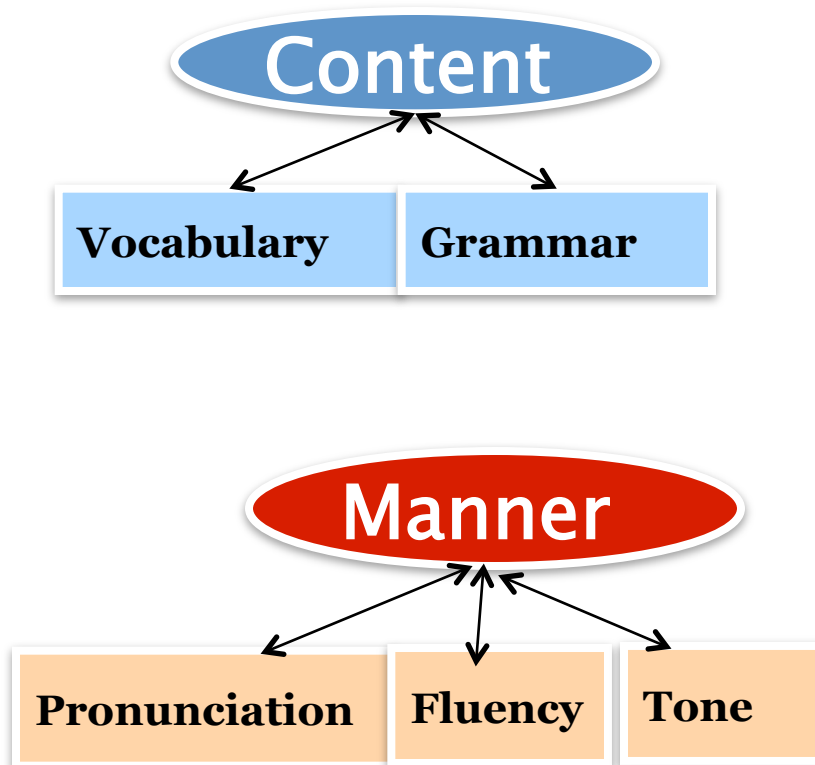


Admin: Technology
Score: Technology

SCT: Test Structure


Part (部分)	Item Type (试题)
A	Tone Phrases (声调词语)
B	Read Aloud (朗读) 
C	Repeats (重复) 
D	Short Answer Questions (问答) 
E	Recognize Tones – Word (声调识别 – 词)
F	Recognize Tone – Sentence (声调识别 – 句子)
G	Sentence Builds (组句) 
H	Passage Retellings (短文重述)

SCT: Score Report



成绩报告单 SCORE REPORT

汉语口语考试 Spoken Chinese Test


北京大学
 PEKING UNIVERSITY

考试号码 (Test Identification Number): 1234 5678

考试结束日期 (Test Completion Date): 2019-11-01

考试结束时间 (Test Completion Time): 09:35 上午 (UTC)

总分 (OVERALL SCORE)
58

技能领域 (SKILL AREA)	分数 (SCORE)	20	30	40	50	60	70	80	
总分 (Overall)	58	<div style="width: 58%;"></div>							
语法 (Grammar)	65	<div style="width: 65%;"></div>							
词汇 (Vocabulary)	55	<div style="width: 55%;"></div>							
流利度 (Fluency)	63	<div style="width: 63%;"></div>							
发音 (Pronunciation)	73	<div style="width: 73%;"></div>							
声调 (Tone)	68	<div style="width: 68%;"></div>							

总分	定义 (DEFINITION)
Overall Score	<p>本测试的总分代表考生理解汉语口语的能力。该分数由考生理解汉语口语的能力（即听懂）和发音（即发音）两部分组成。总分由两部分得分相加得出，总分在20到80分之间。</p> <p>The Overall Score of the test represents the ability to understand spoken Chinese (Jiutonghua) as it is spoken internationally and speak it intelligibly at a native-like conversational pace on everyday topics. Scores are a weighted combination of five diagnostic subskill scores. Scores are reported in a range from 20 to 80.</p>


说明: 该测试由北京大学和培生教育联合开发。

Note: The Spoken Chinese Test was developed jointly by Peking University and Pearson Education.

© 2019 Pearson Education, Inc. or its affiliate(s). All rights reserved.
 Ordina and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.

PEARSON

For more information, visit us online at www.speakingtest.com



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

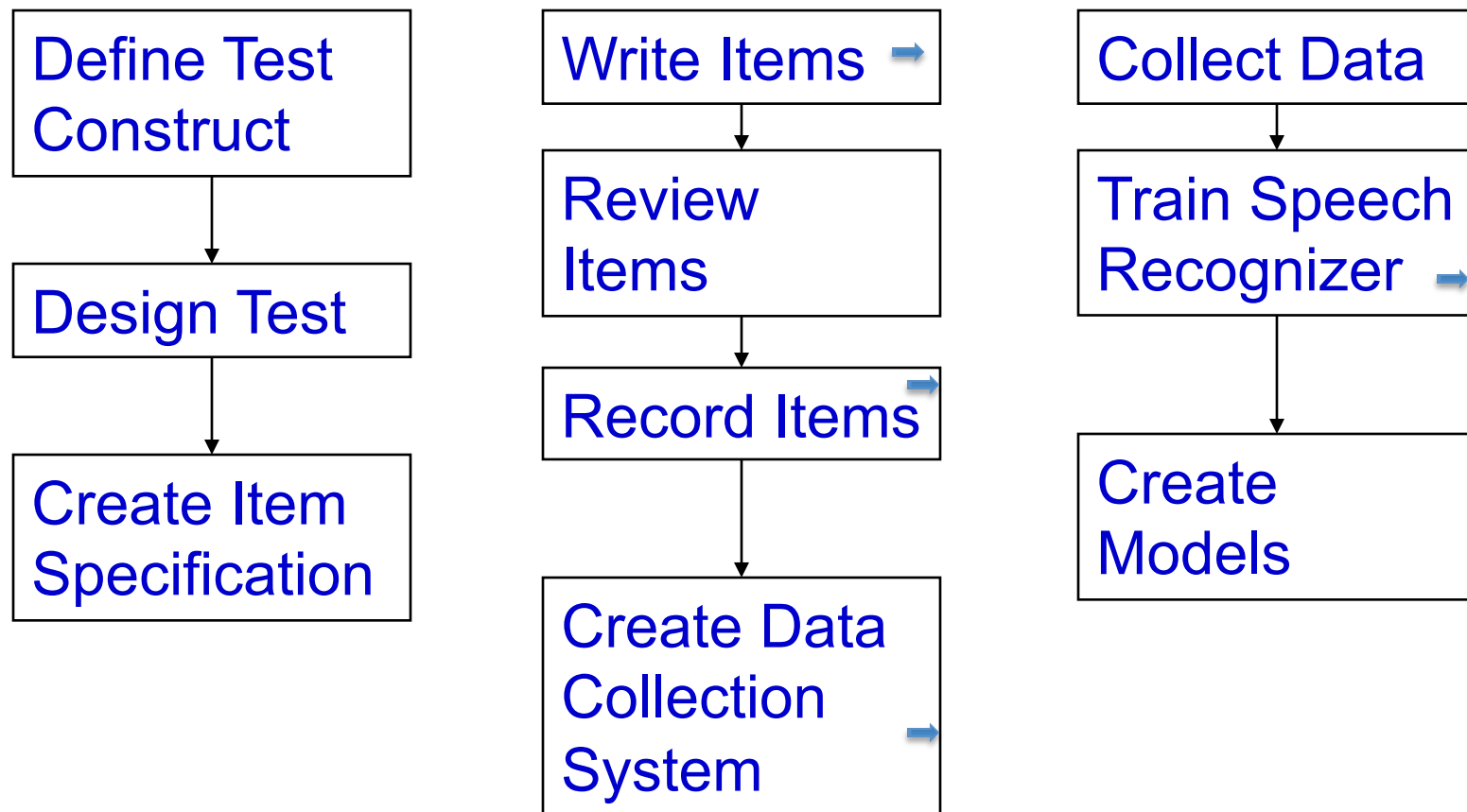
Testing
System
scores the
test and
posts scores
to the web
reporting
tool



SCT takes 20 minutes; results available within 10 minutes

SCT: Test Development

PLAN → DEVELOP → SCORE



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

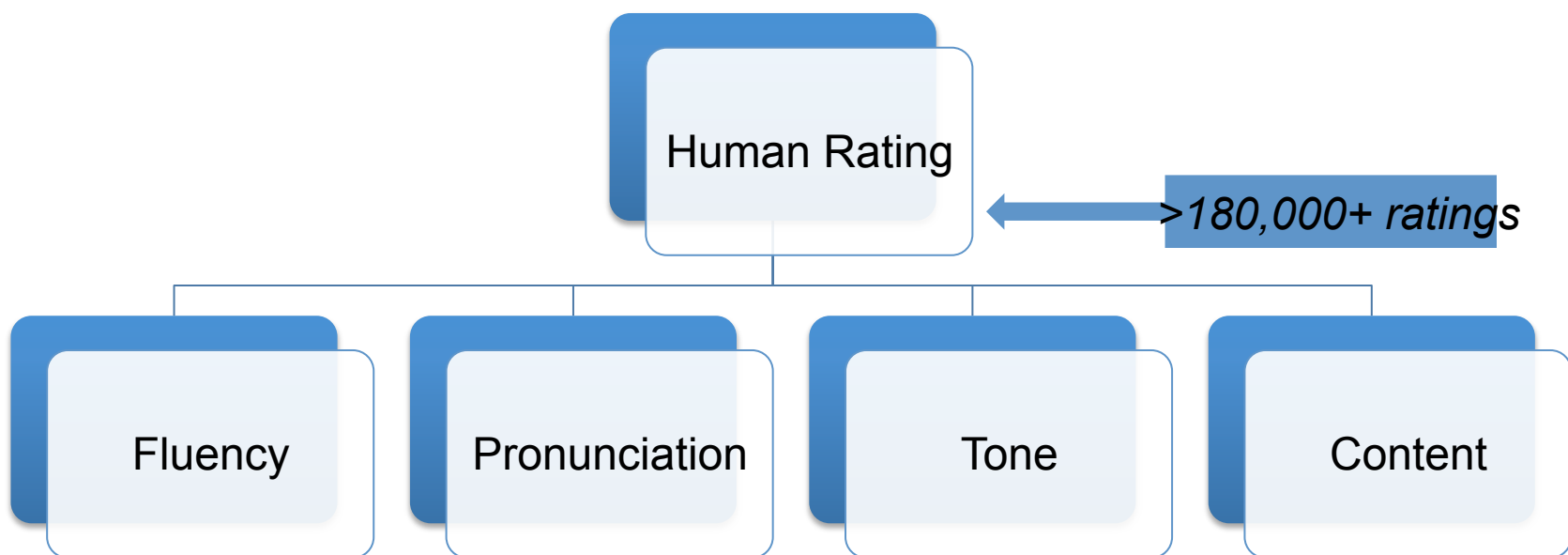
Mandarin Variety	Percentage
Mandarin (官话)	57.2
Wu (吴)	13.0
Min (闽)*	9.3
Yue (粤)	5.5
Xiang (湘)	3.1
Gan (赣)	1.9
Hakka (客家)	1.8
Jin (晋)	0.4
Taiwan Mandarin (台湾国语)	5.6
Singapore Mandarin (新加坡华语)	1.8
Unidentified Mandarin	0.5

2,459 non-native speakers -> 4,142 tests

Country	%
English	15%
Japanese	13%
Korean	7%
Russian	6%
Spanish	4%
Arabic	4%
Thai	4%
Other L1	47%



SCT: Human Rating



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

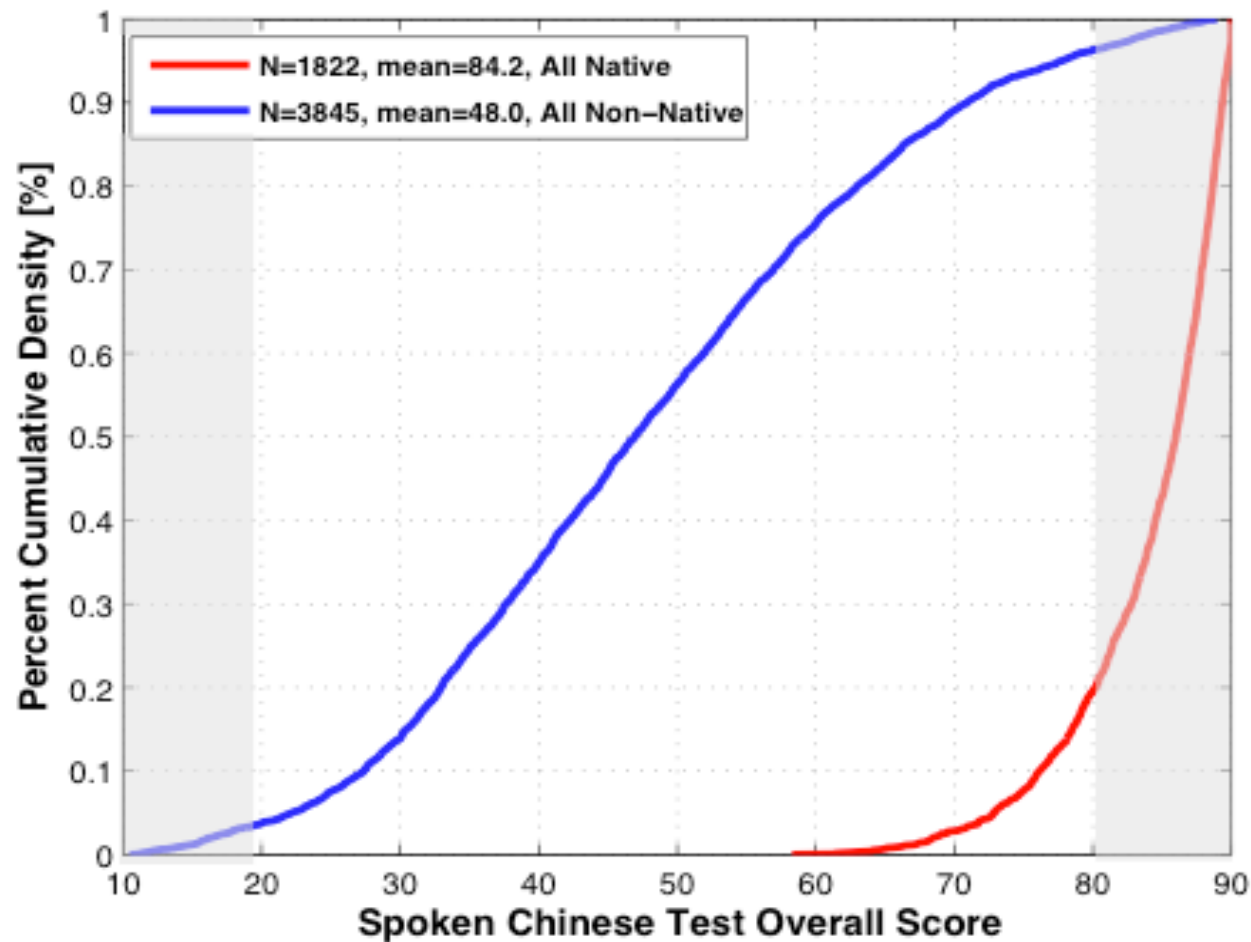
Score	Split-half Reliability (N=166)	Test – Retest Reliability (N=158)
Overall	0.97	0.95
Grammar	0.92	0.91
Vocabulary	0.94	0.93
Fluency	0.97	0.94
Pronunciation	0.96	0.90
Tone	0.92	0.87

SCT Machine – Human Correlation

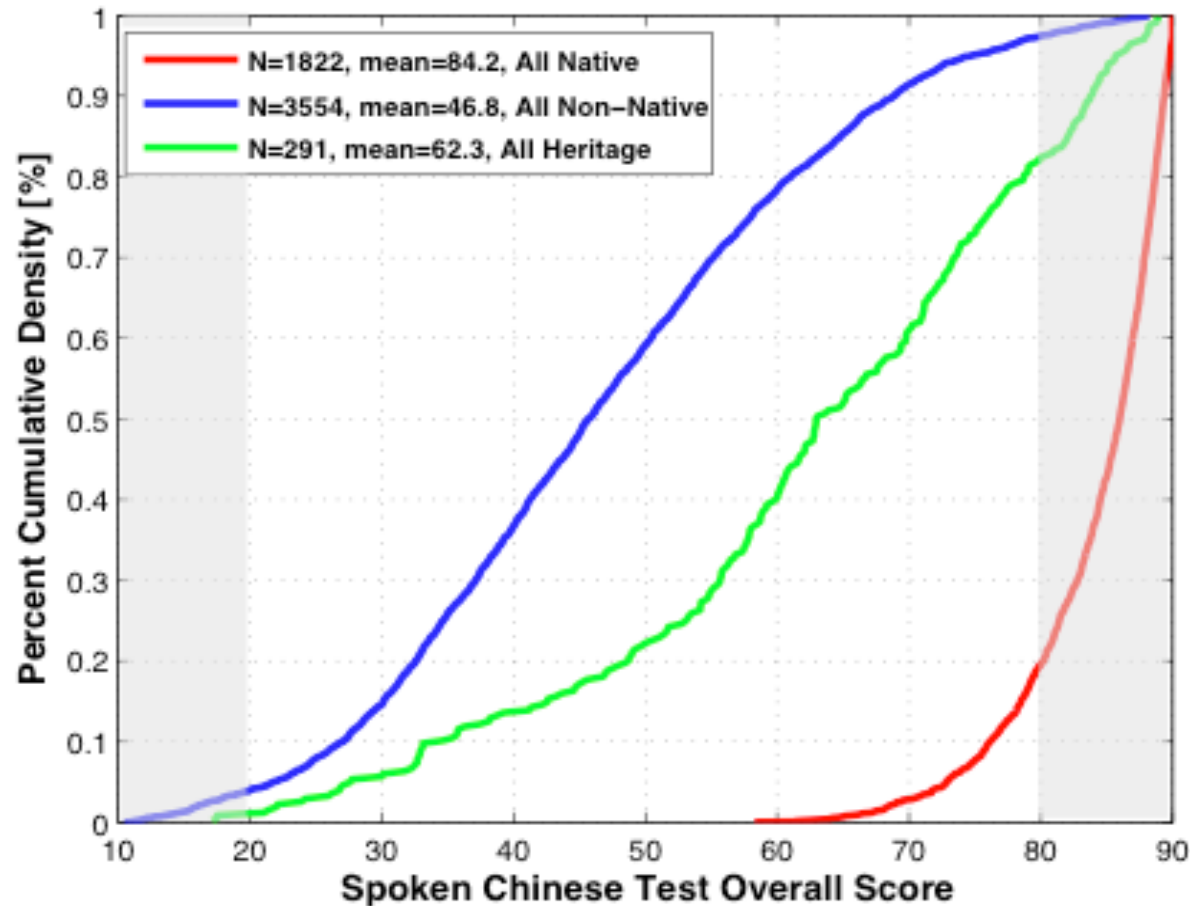
How well do machine scores line up with human scores?

Score	Correlation
Overall	0.98
Grammar	0.96
Vocabulary	0.96
Fluency	0.96
Pronunciation	0.95
Tone	0.96

Native vs. Non-Native



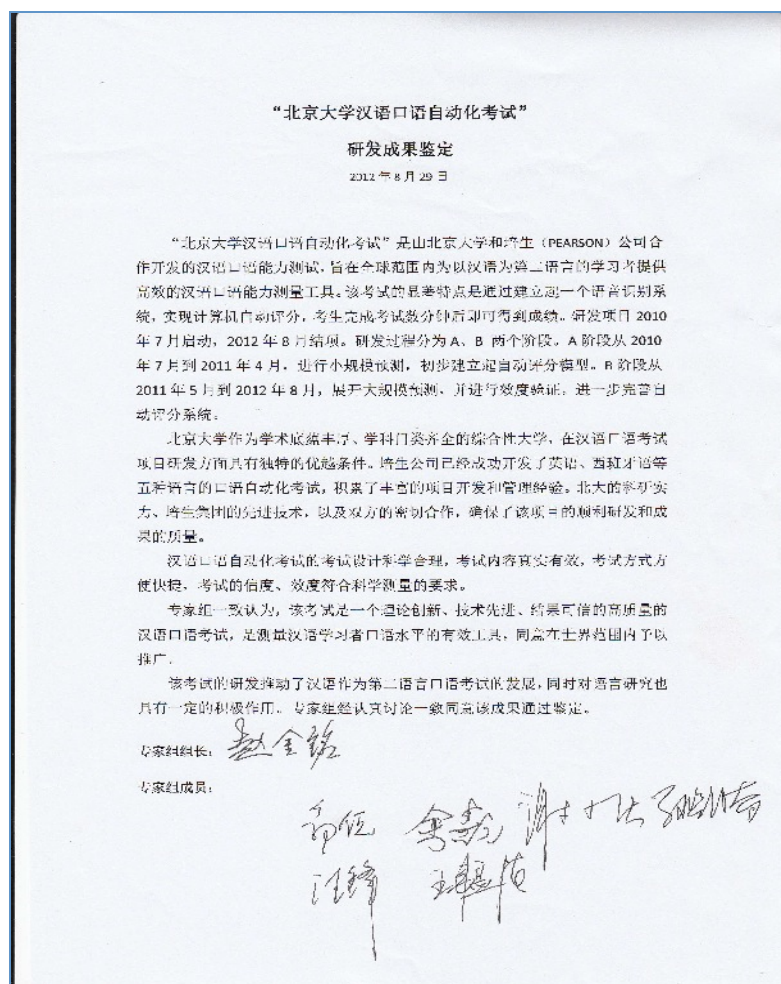
Native, Non-Native, & Heritage



SCT vs. Other Speaking Tests

Test	Correlation	Number of test-takers
HSK Oral Test – Intermediate Level vs. SCT	0.86	148
ILR-OPI	0.87	169

SCT Expert Validation



How is the SCT Used?

Proficiency Test	<ul style="list-style-type: none">• Determine test takers' proficiency levels in a standardized manner
Placement Test	<ul style="list-style-type: none">• Accurate placement for targeted levels• Verifying Chinese for study abroad programs• Graduate and move up one level in program
Monitor Progress	<ul style="list-style-type: none">• Monitor student language progress over a certain period of time• Ability to share results and utilize results as learning tool

Questions?