Reliability and Validity of PhonePass Spoken English Tests

Seoul, Korea
10 May 2000

Jared Bernstein
jared@ordinate.com

Ordinate Corporation
Menlo Park, California, USA

Outline

- Thesis and definitions
- Language use and score use
- Scoring method
- Evidence for valid score use
- Conclusion

Thesis

PhonePass tests provide a reliable and valid measure of facility in spoken English

Target Use of Scores

- ***Facility in spoken English:*** the ability to understand spoken English and respond intelligibly at a conversational pace on everyday topics

- ***Score use***: for a person or a group of people, determine the level of *facility in spoken English*

Definition: Reliable

- A test is ***reliable*** if the scores are repeatable and consistent

  Reliability is a test-retest correlation

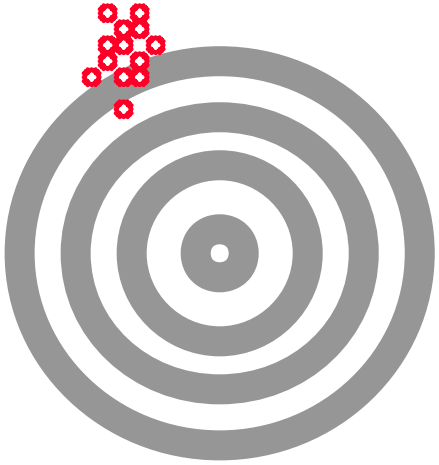
  Reliability ranges from 0.0 to 1.0

Definition: Valid

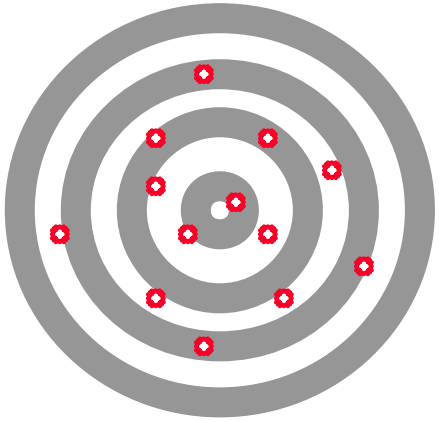
- A test is ***valid*** for a particular use

  If test scores measure a skill set that matches the target requirement or decision, then the scores are valid in that use.

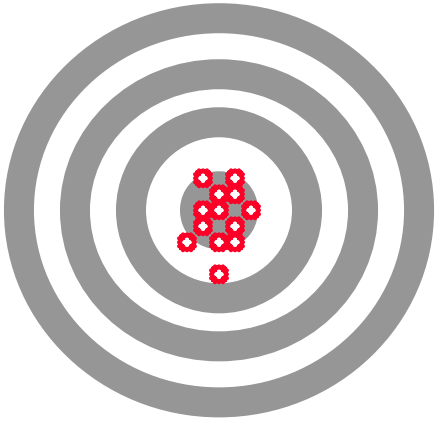
Reliability and Validity Illustrated



Reliable,
but off target



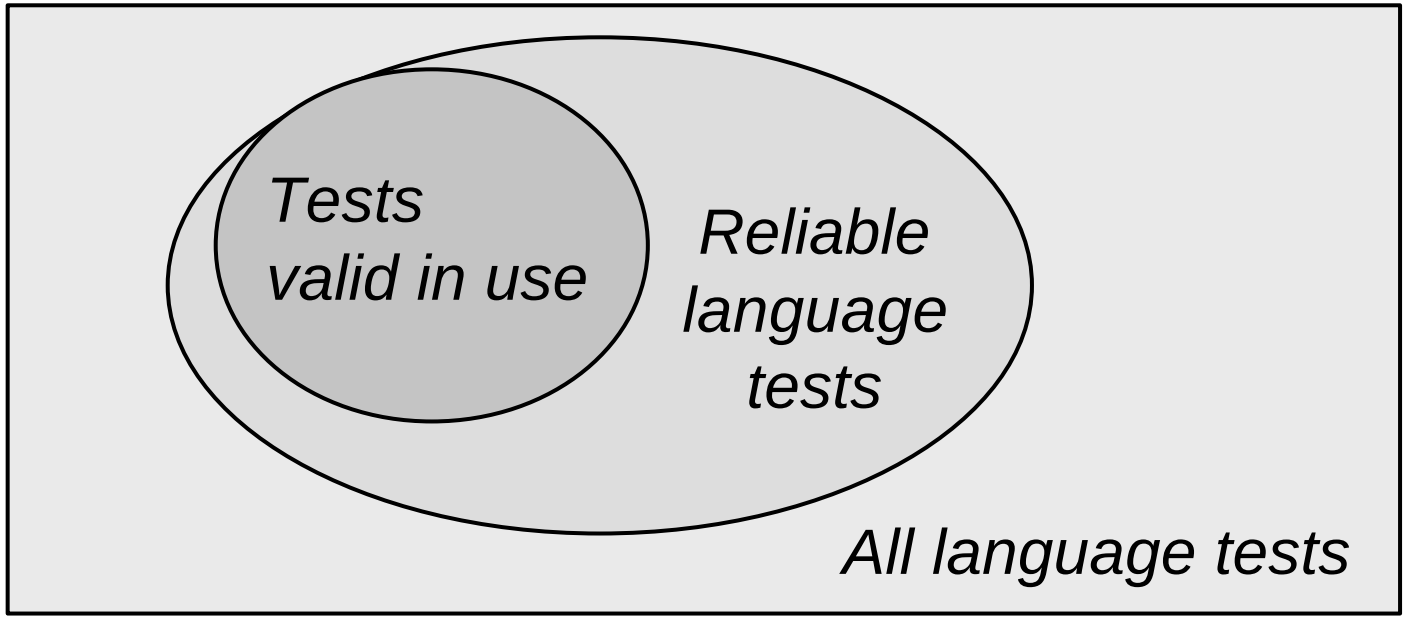
Centered on target,
but not reliable



Reliable,
and on target

Validity implies Reliability

- Tests must be reliable to be valid in use
- Tests can be reliable, but not valid for a particular use



Language Use and Score Use

- Categories of language in use
  - By skill: speaking, listening, reading, writing
  - By function: travel, daily living, work tasks, research, teaching, negotiation, …
- Score uses (decisions)
  - Qualification of individuals for specific tasks
  - Evaluation of people and programs by rank
  - Diagnosis of individual's language skills

Test Selection Procedure

- Skills and measures
  - Skills are relevant to the decision
  - Language test measures those skills
  - Test is reliable and valid
- Test Types
  - Spoken language or written language
  - Language knowledge or language performance
  - Integrated items or discrete point items
  - Objective scoring or subjective scoring

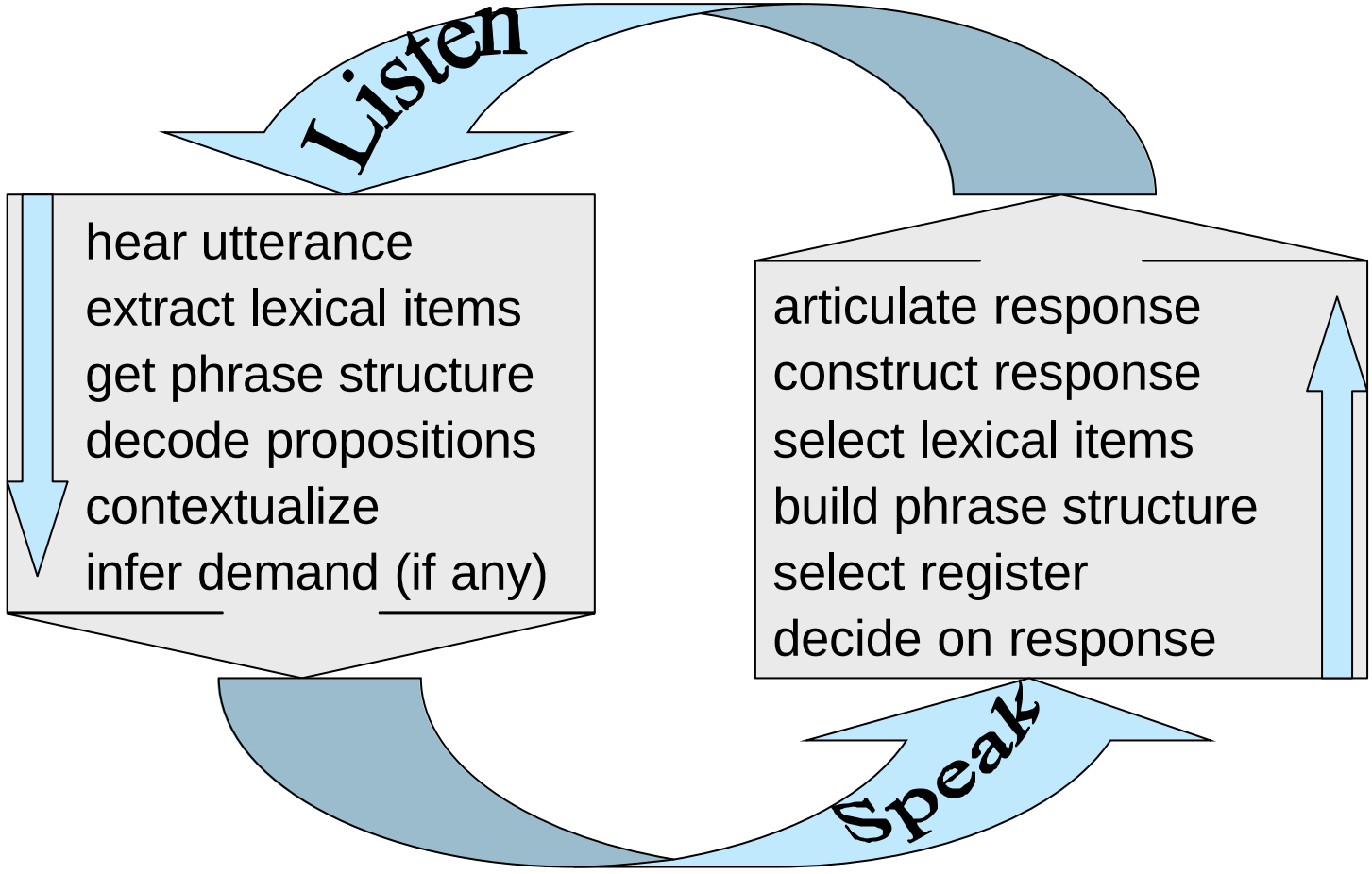
Facility with Spoken Language

The ability to understand spoken language and
respond intelligibly at a conversational pace
on everyday topics

(or)

in a discussion,
the ability to track what's being said,
extract meaning in real time, and
formulate and produce relevant responses,
at a conversational pace.

Conversational Processing Components



Processing Components in Test

Speaking:

a. decide on response	all items
b. select register	— (available for user rating)
c. build phrase structure	long repeats, questions
d. select lexical items	opposites, questions
e. construct response	repeats, opposites, questions
f. articulate response	all items

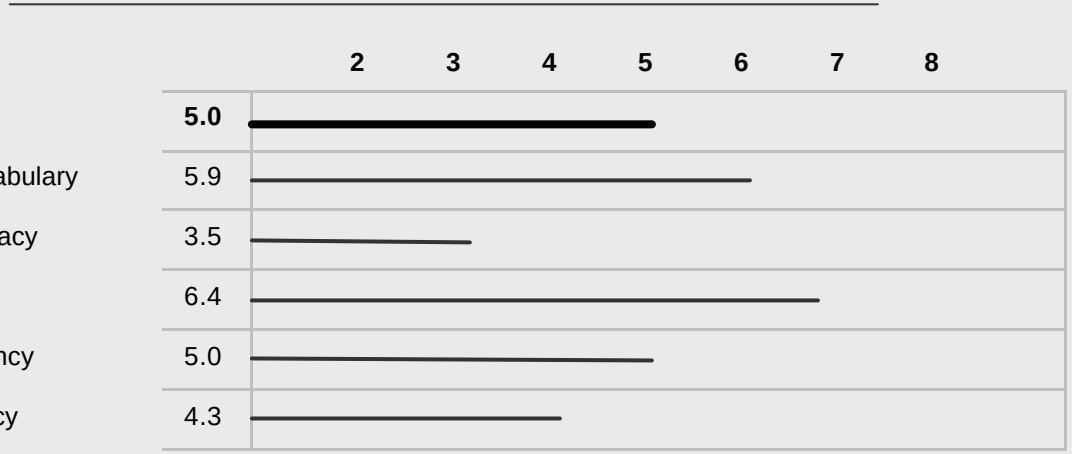
Score Report

PhonePass | **Score Report** | **Spoken English Test**

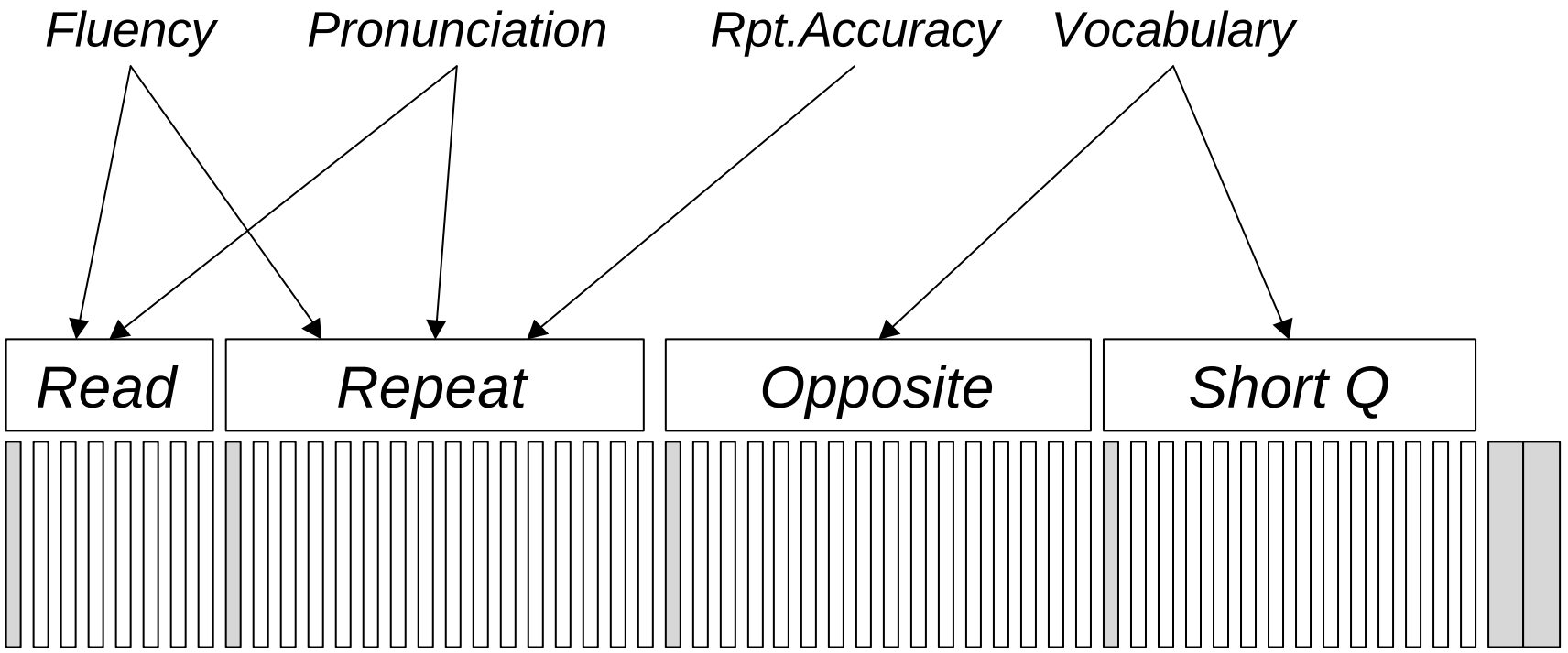
Test: SET-10 (24.33.43)

Test Paper Number: 12345678

Administration Date: 06 JAN 1999, 20:05 - 20:15 (PST)



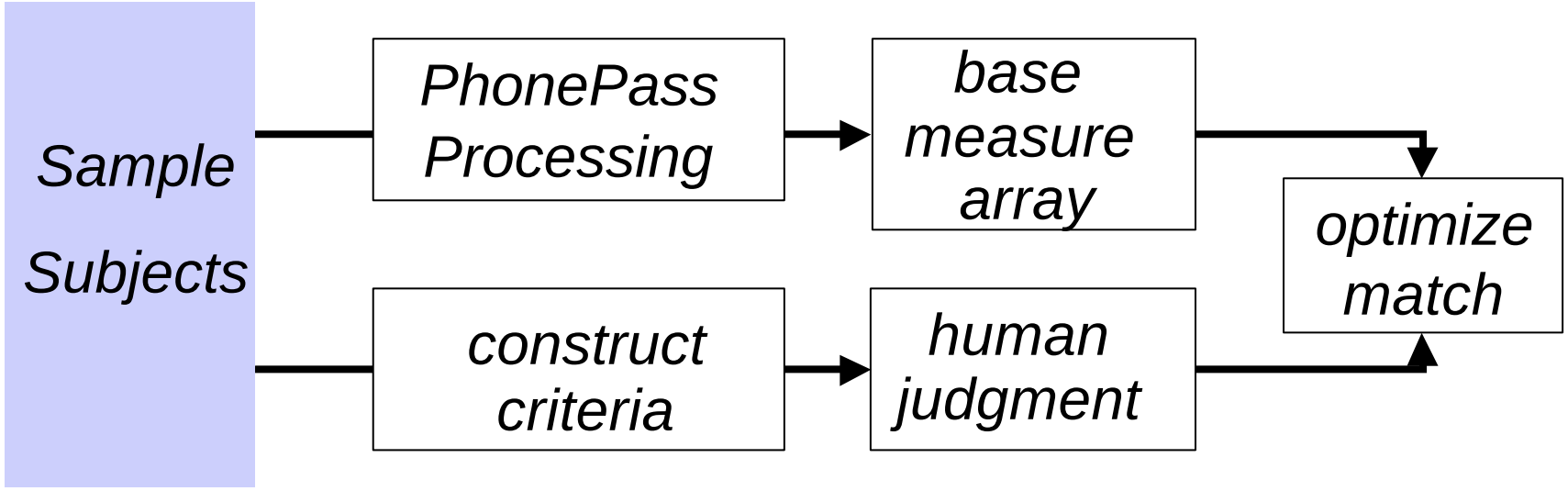
Scoring


Fluency
Pronunciation
Rpt.Accuracy
Vocabulary
Read
Repeat
Opposite
Short Q

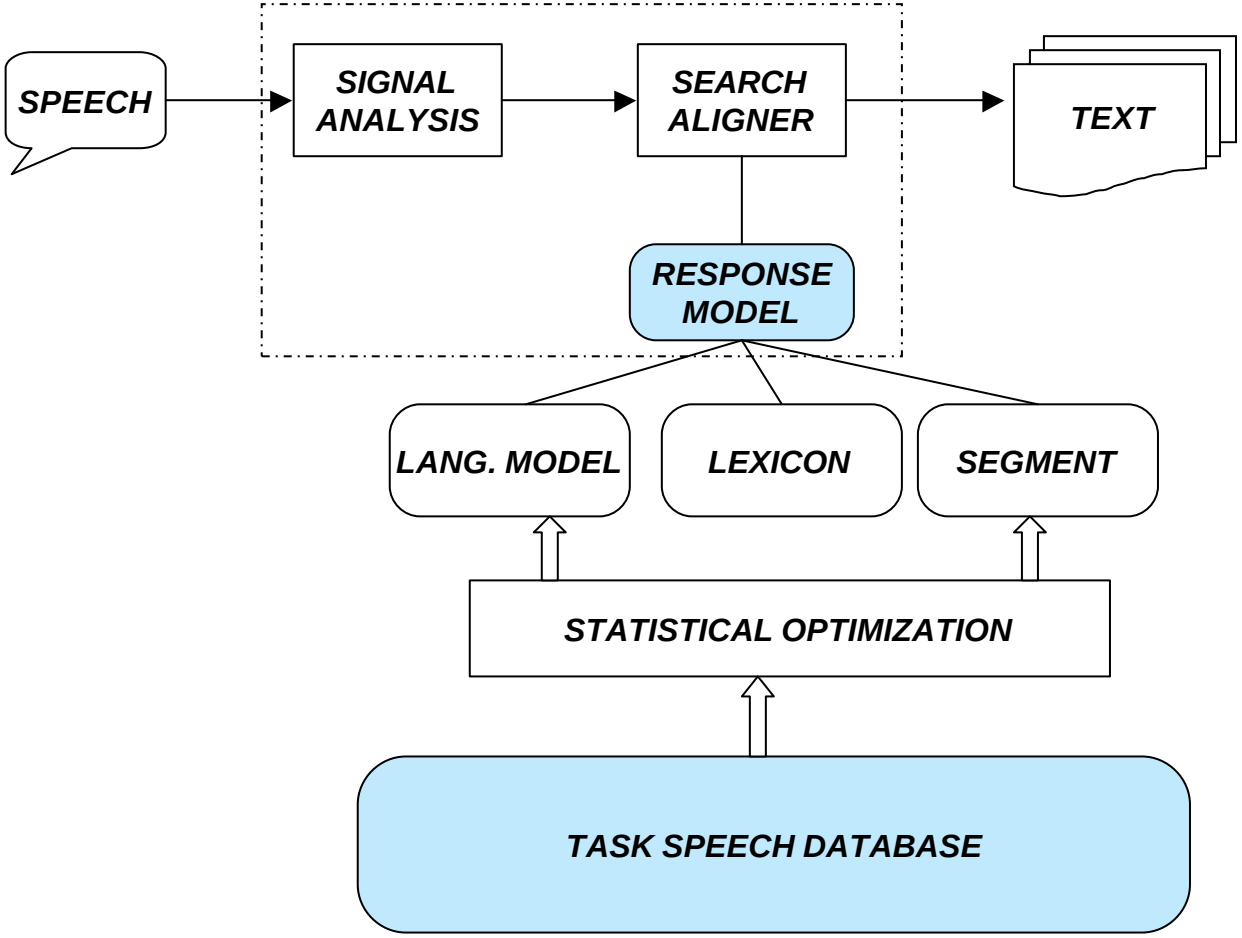
Scoring Method

Subscore	Section	Type	Wgt.
Repeat Accuracy	Repeats	discrete	30%
Listening Vocabulary	Opposites, Short Qs	discrete	30%
Pronunciation	Reading, Repeats	continuous	20%
Reading Fluency	Reading	continuous	10%
Repeat Fluency	Repeats	continuous	10%

Development Process



Speech Recognition Design



Task Speech Data Base

400,000 spoken responses

247,000 hand transcribed responses

123,000 human grades

> 100 different native languages

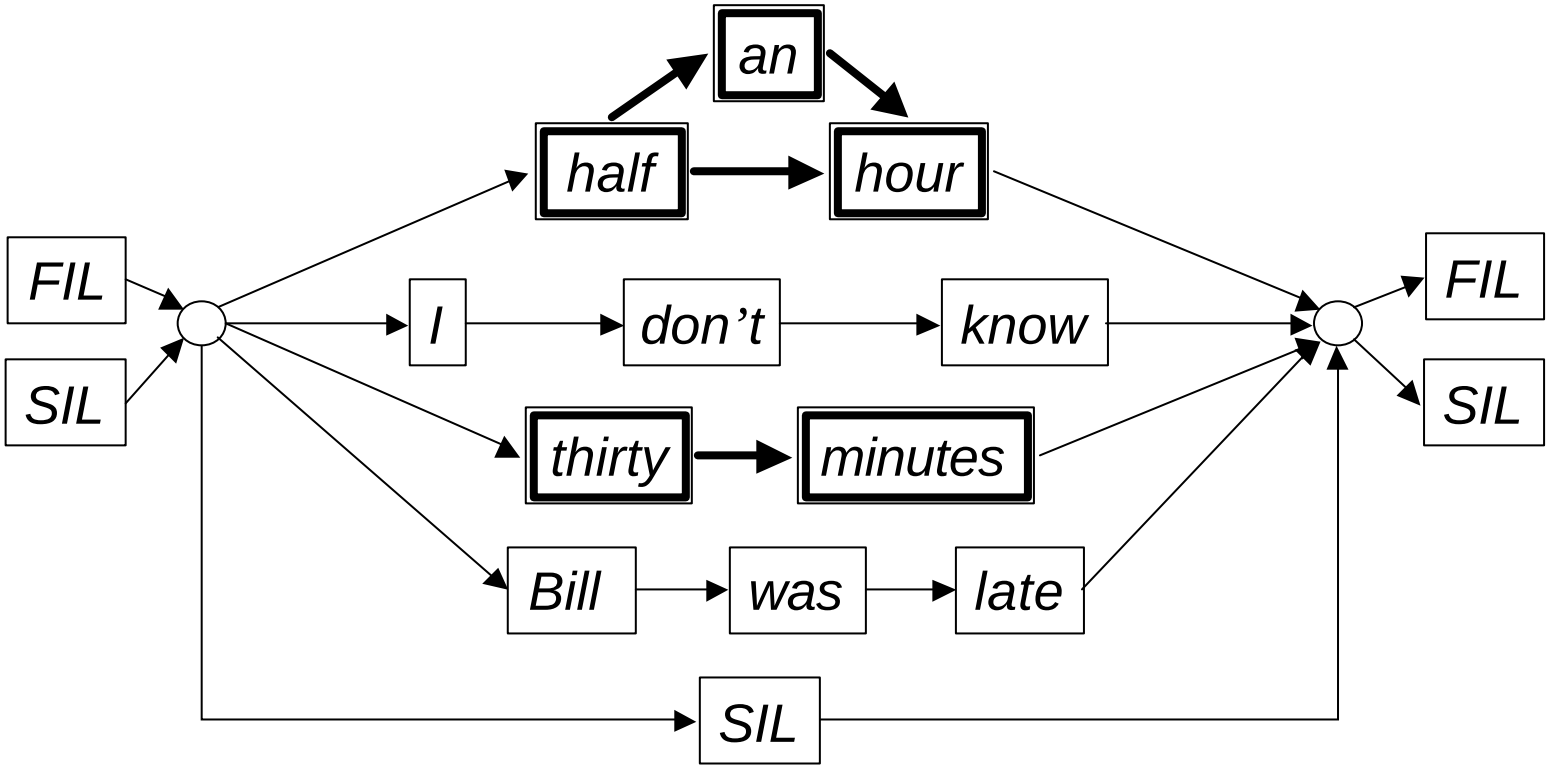
Modeling & Scoring Resources

- Custom acoustic models, response models, and scoring logic were configured from native and non-native databases
- Linear and non-linear processes scale and combine response characteristics in scoring
- Off-the-shelf ASR systems do not accurately recognize non-native speech, will not produce alignments or production scores, and offer limited configuration tools

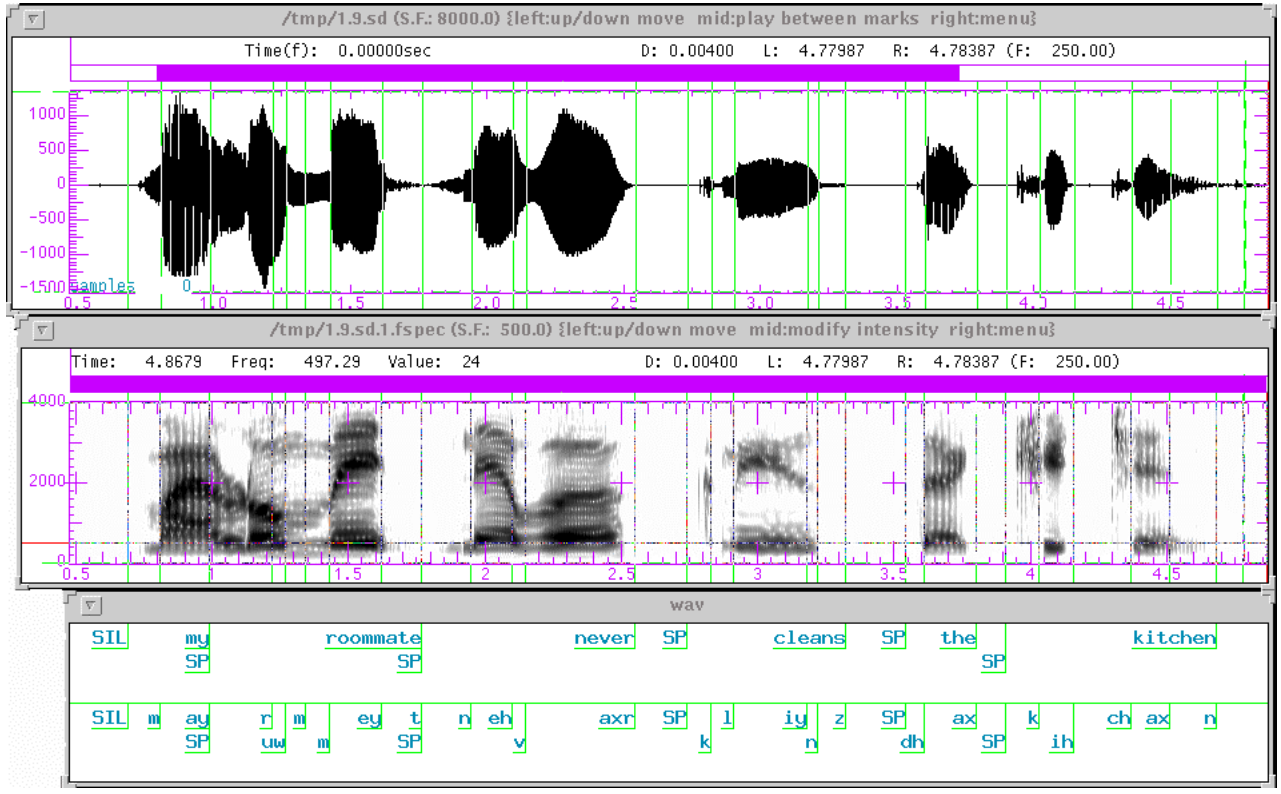
Human Rater Criteria (Example)

5. **ADVANCED Pronunciation**: Vowels and consonants are pronounced clearly and unambiguously. A few minor C, V, or stress distortions do not affect intelligibility. All words are easily understandable; a few consonants or C-sequences may be distorted. Stress is placed correctly in all common words, and sentence level stress is reasonable.
2. **INTRUSIVE Pronunciation**: Many consonants and vowels are mispronounced, resulting in a strong intrusive foreign accent. Listener may have difficulty understanding *a significant portion of the words (>33%).* Secondarily, many consonants may be distorted or omitted; many C-sequences may be simplified. Stress placement is unclear; unstressed vowels may be unreduced or omitted; *a few* syllables may sometimes be added or skipped.

Response Network



Automatic Phone/Word Alignment



waveform

spectrum

words

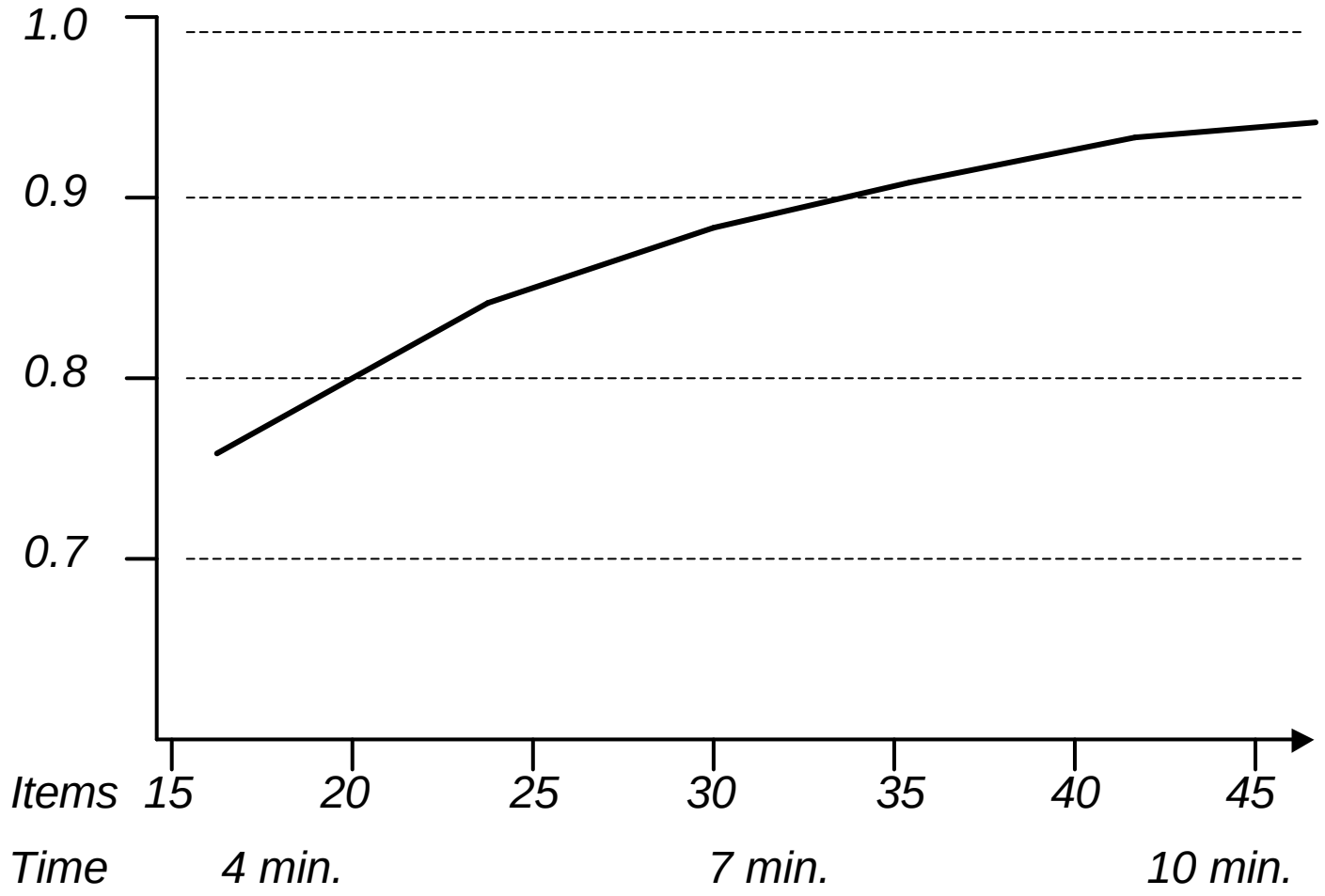
segmentation

call 14126 item 1.9

Validity

- Is the test reliable?
- Does it measure the right thing?
- Does it predict more general performance?
- Are scores in accordance with expectation?

PhonePass Reliability



PhonePass Score Reliabilities

Score	Human	Machine
Overall	**0.94**	**0.94**
Sub-Score		
Listen Vocabulary	0.82	0.75
Repeat Accuracy	0.92	0.85
Pronunciation	0.93	0.94
Read Fluency	0.86	0.92
Repeat Fluency	0.84	0.82

N = 288

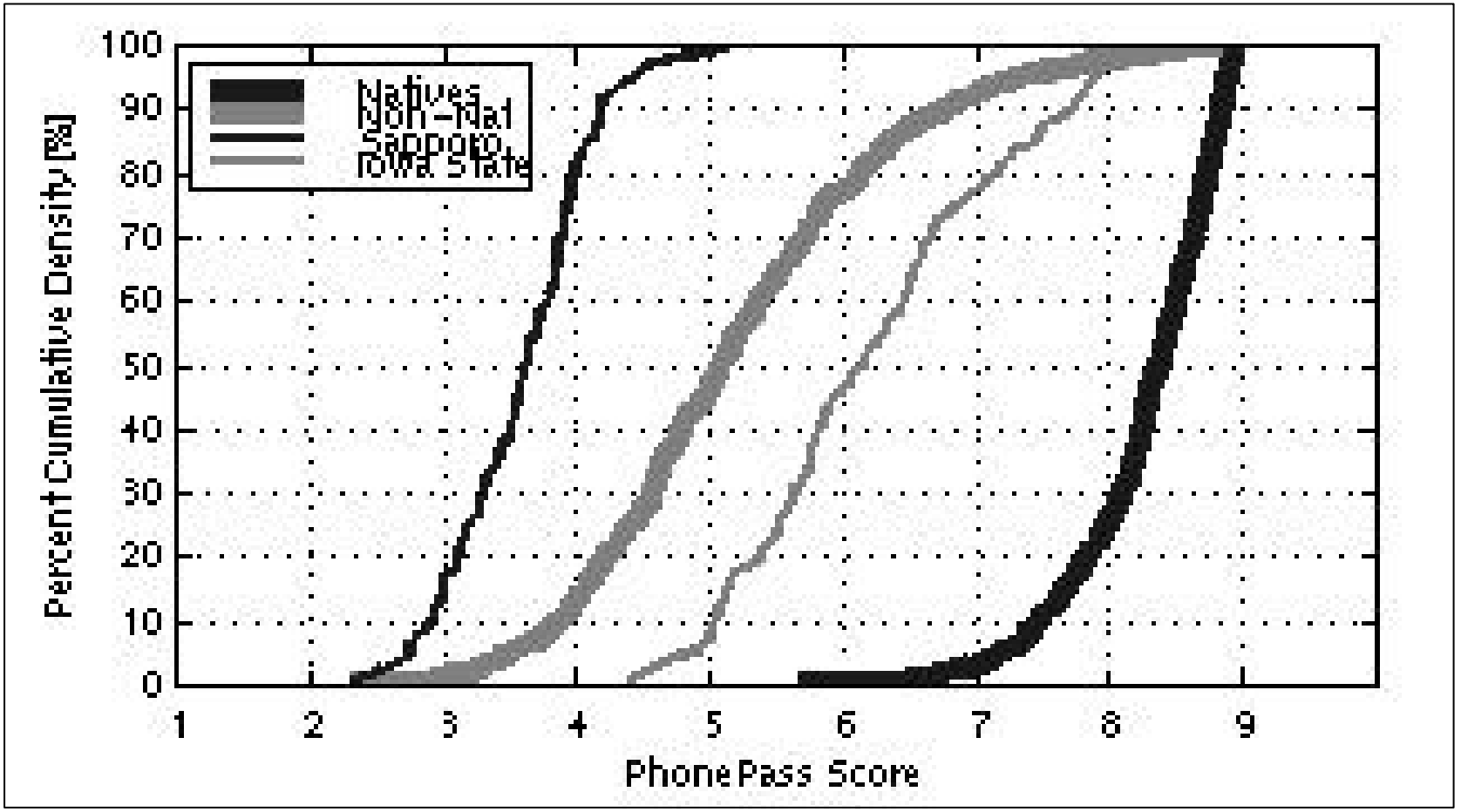
2 hours :: 5 minutes

Correlations: Subscores, Overall

SubScore	2	3	4	5	6
1. Listen Vocabulary	.73	.63	.51	.59	.89
2. Repeat Accuracy	...	.63	.49	.67	.89
3. Pronunciation		…	.73	.80	.85
4. Reading Fluency			...	.62	.72
5. Repeat Fluency				...	.79
6. PhonePass Overall					...

N = 288 non-natives

Native-NonNative Overall CDF

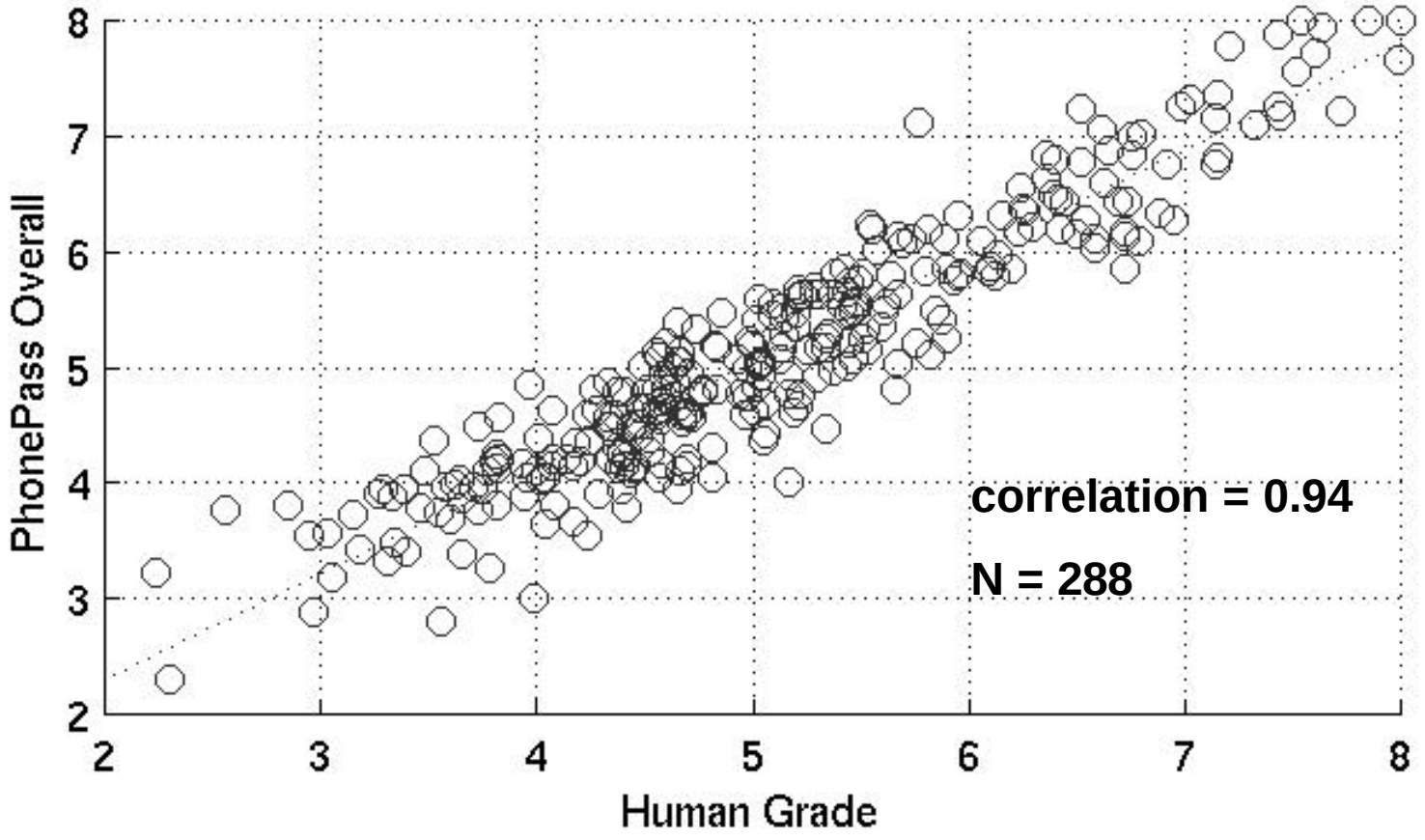


Correlation: Human, SET-10

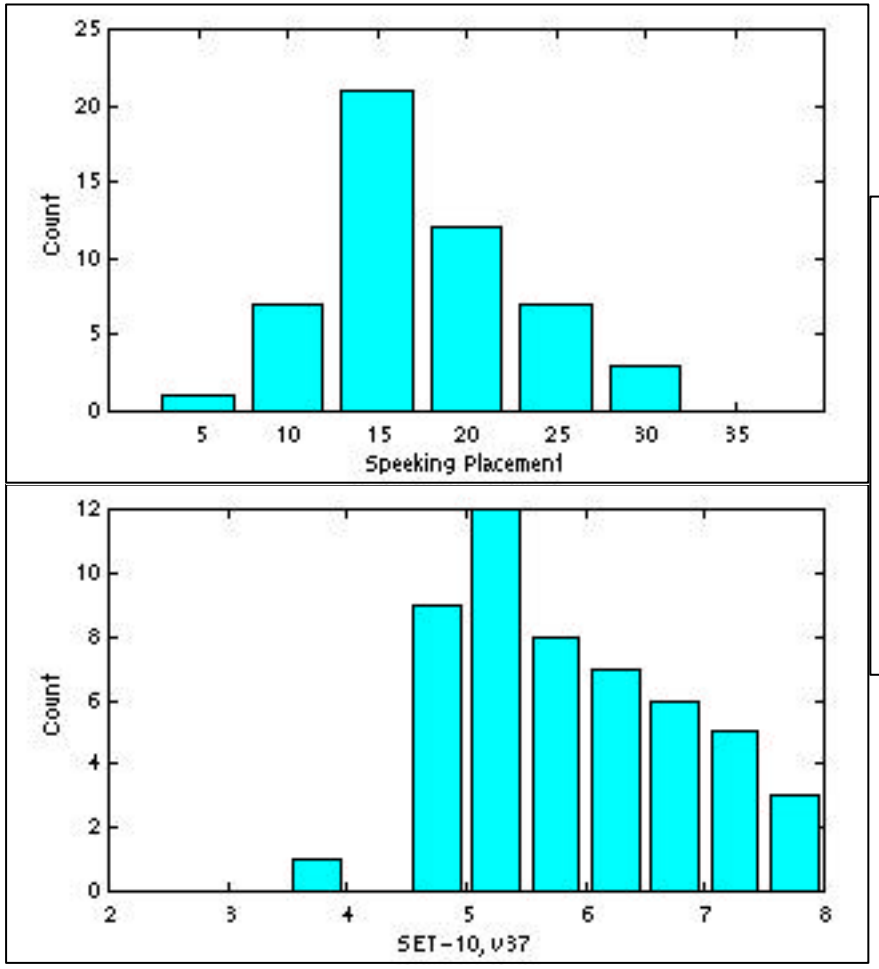
Score	Correlation
Overall	**0.94**
Sub-Score	
Listen Vocabulary	0.89
Repeat Accuracy	0.89
Pronunciation	0.79
Read Fluency	0.86
Repeat Fluency	0.87

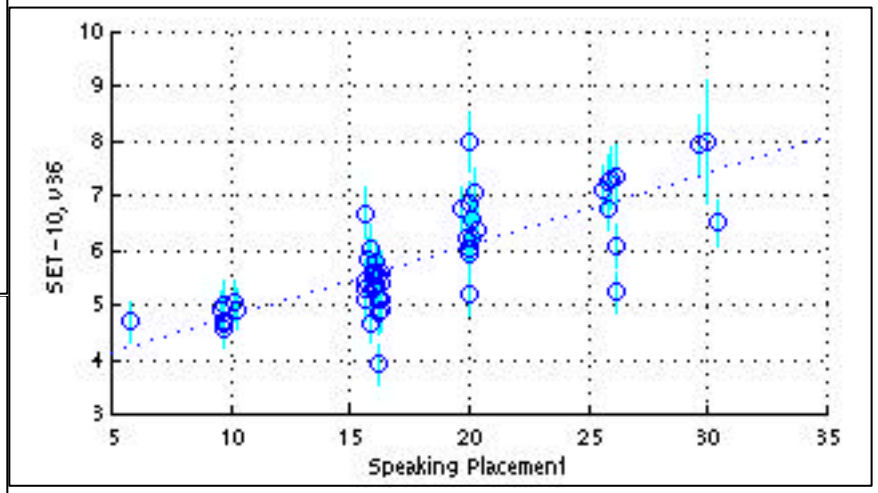
N = 288

Machine-Human Comparison



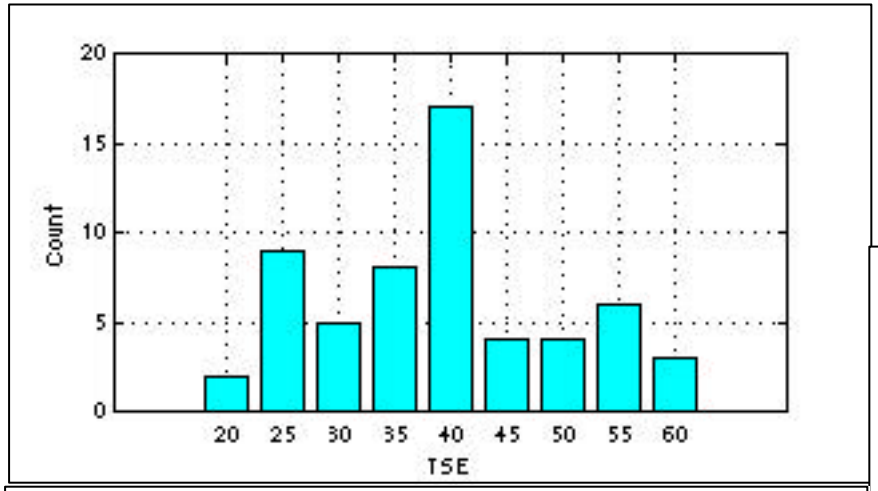
ILR Speaking at DLI ~ PhonePass

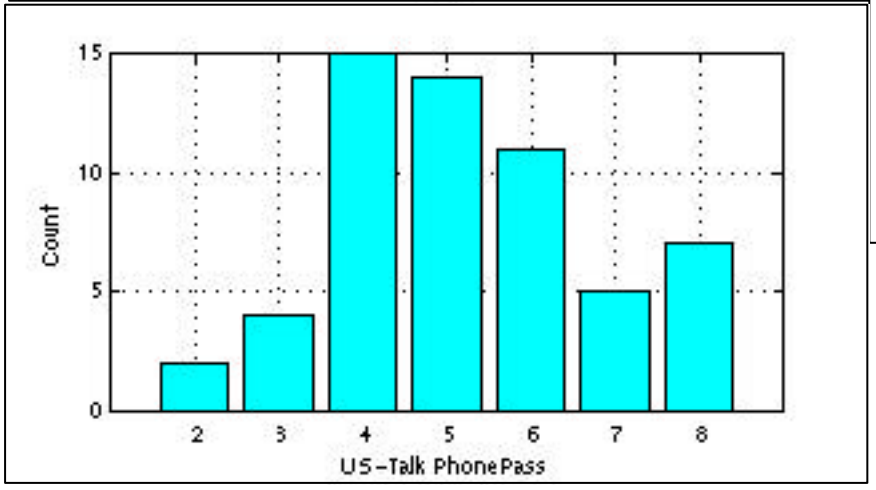


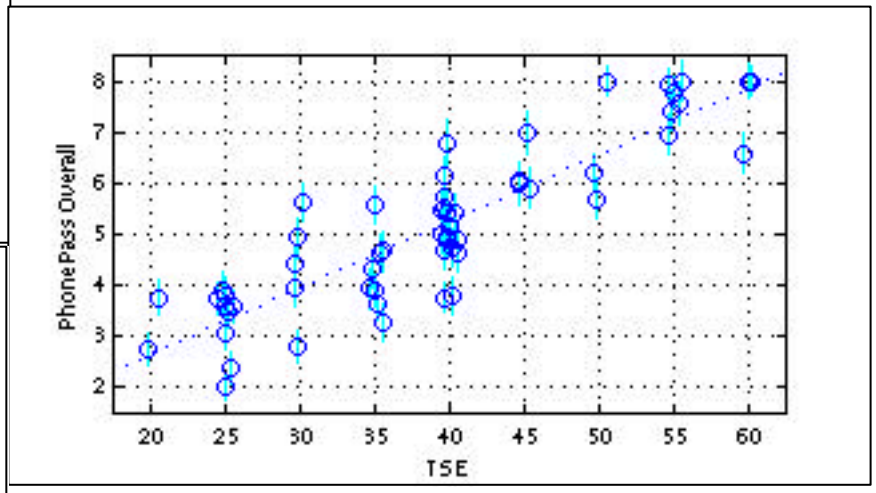


N = 51
r = 0.75

TSE ~ PhonePass Scoring







N = 58
r = 0.88

PhonePass ~ Human Measures

	Rel.	Correlation	N
TSE	0.89	0.88	58
ILR speaking	0.75	0.75	51

Evidence of Validity

- Listeners can estimate examinee skill from responses
- Scores have good precision and reliability
- Scores show adequate native/non-native separation
  Uniform non-native groups get similar scores
- Sub-scores are distinct
- Scores correlate closely with human scores
- Concurrent correlations are reasonable

Reality and Technology

Oral Proficiency Interviews

- OPI tests reliability is very variable
  - U.S. Gov't OPIs reliabilities from 0.35 to 0.75
- OPI calibration shifts across time and location

PhonePass testing matches the best human-rated tests in their best implementations and provides this service consistently

Summary

- PhonePass SET-10 is a reliable test that is valid as a measure of facility in spoken English
- SET-10 scores predict other more elaborate measures at their level of reliability

- SET-10 crystalizes a careful human scoring, and delivers it reliably across time and location