Developing and Validating an Automated Test of Spoken MSA
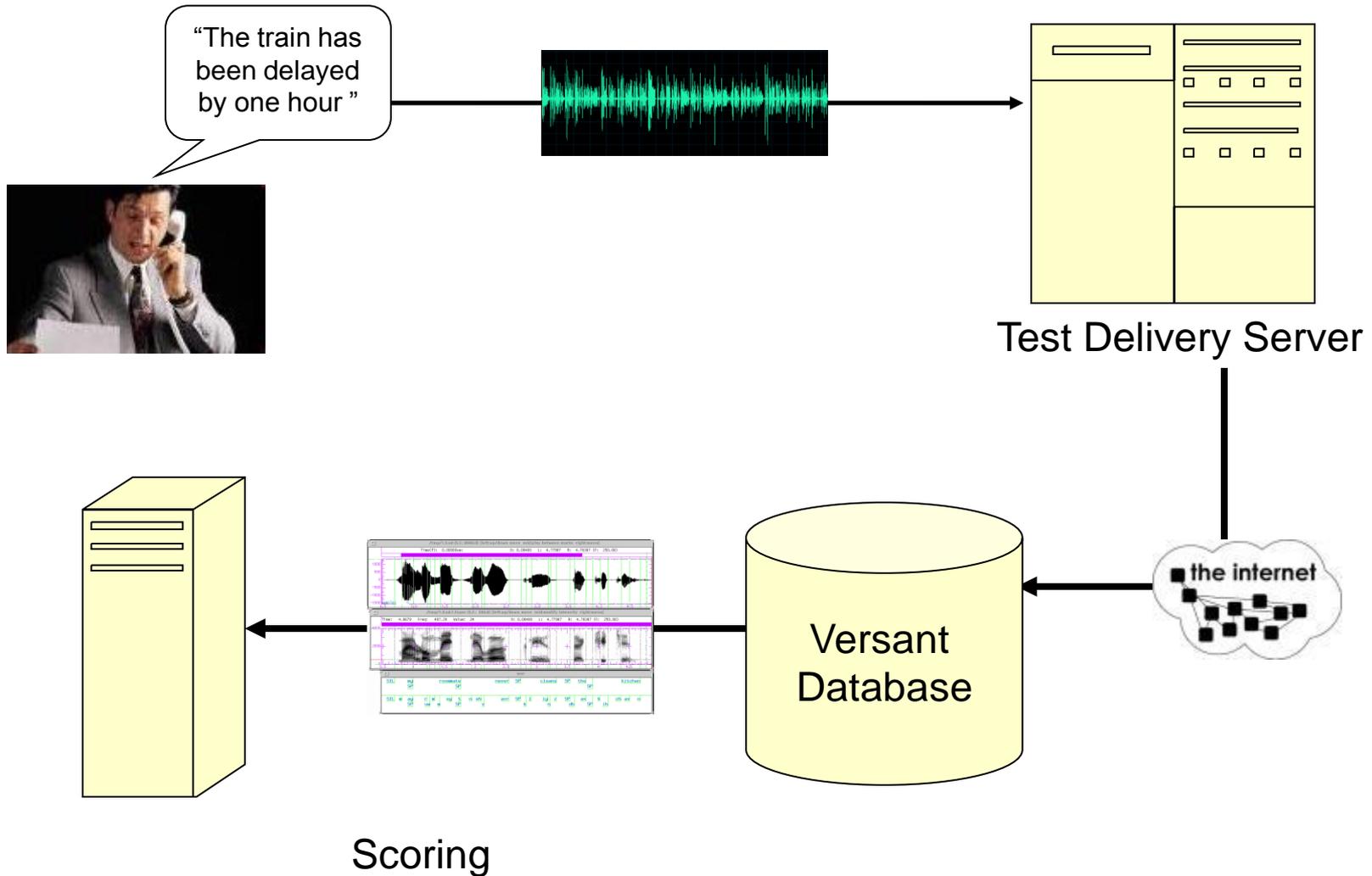
Jared Bernstein

Jian Cheng

Ulrike Pado

Masa Suzuki

Mohamed Al-Saffar

Alistair van Moere

Pearson Knowledge Technologies

Palo Alto, California

Outline

1. How Versant tests work

2. Versant Arabic Test (development)

3. Validation evidence

4. Predictive accuracy

How Versant tests work

Versant Arabic Test

- DLI purpose

  ~1000 students at DLI need predictive speaking tests

- Requirements

  Accurate test of Arabic listening & speaking

  Convenient to use at DLI and worldwide

  Suitable for repeated formative testing

  High peak capacity for mass screening

Versant Arabic Test

Test Structure

Part A: Reading

Part B: Repeat -1

Part C: Short Answers

Part D: Sentence Builds

Part E: Repeat -2

Part F: Passage Retelling

Schematic test structure

Read aloud · Repeat Sentence 1 · Short Question · Sentence Build · Repeat Sentence 2 · Passage

17 minutes
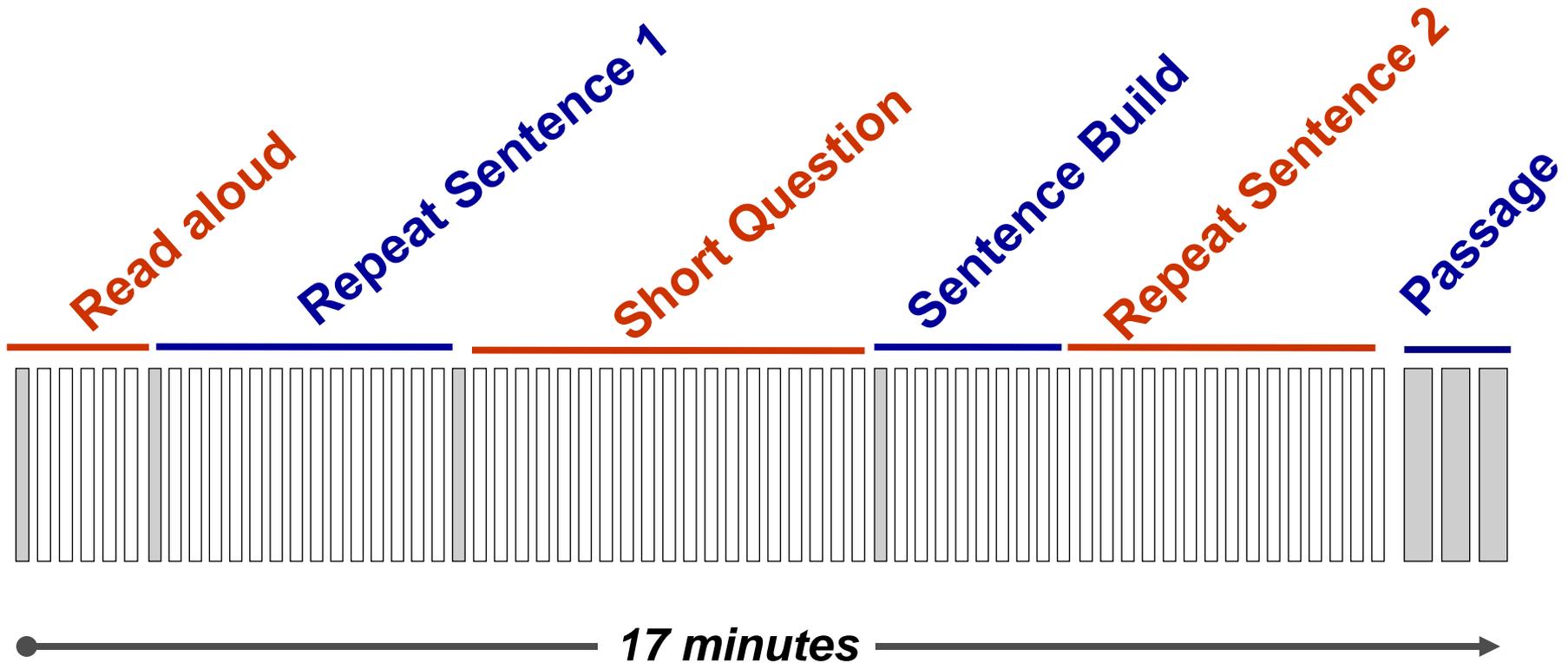
Prompt recordings

Country	Egypt	Iraq	Jordan	Morocco	Lebanon	Palestine	Syria
Voices	F, M	F, M	M	F	M	F, M	F, M

PEARSON ((()))VERSANT™

How Versants are developed
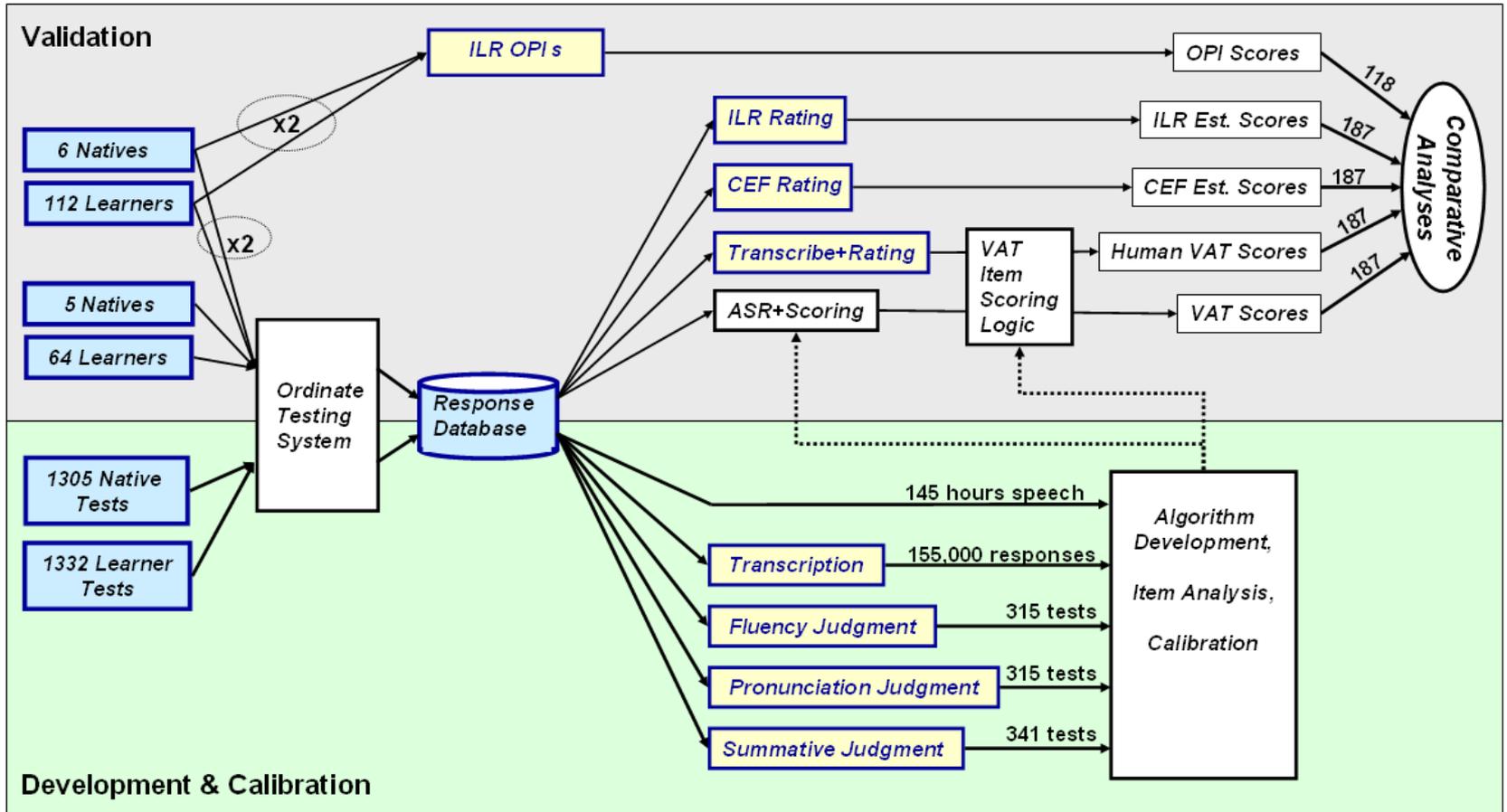
Pilot data sources

Native Data					Total
Egypt	Syria	Iraq	Palestine	Other	
484	281	179	187	517	1648

Learner Data		Total
DLI	Non-DLI	
1120	552	1672

Reliability

Score	Split-Half Reliability (N = 134)	Test – Retest Reliability (N = 100)
Overall	0.98	0.97
Sentence Mastery	0.97	0.96
Vocabulary	0.89	0.82
Fluency	0.97	0.96
Pronunciation	0.96	0.94

Machine – Human Comparison

Score	Correlation (N = 134)
Overall	0.97
Sentence Mastery	0.97
Vocabulary	0.96
Fluency	0.84
Pronunciation	0.83

Native ~ Non-Native Scores

Arabic Overall

N=1309, mean=85.0, Native
N=1337, mean=44.2, Non-Native
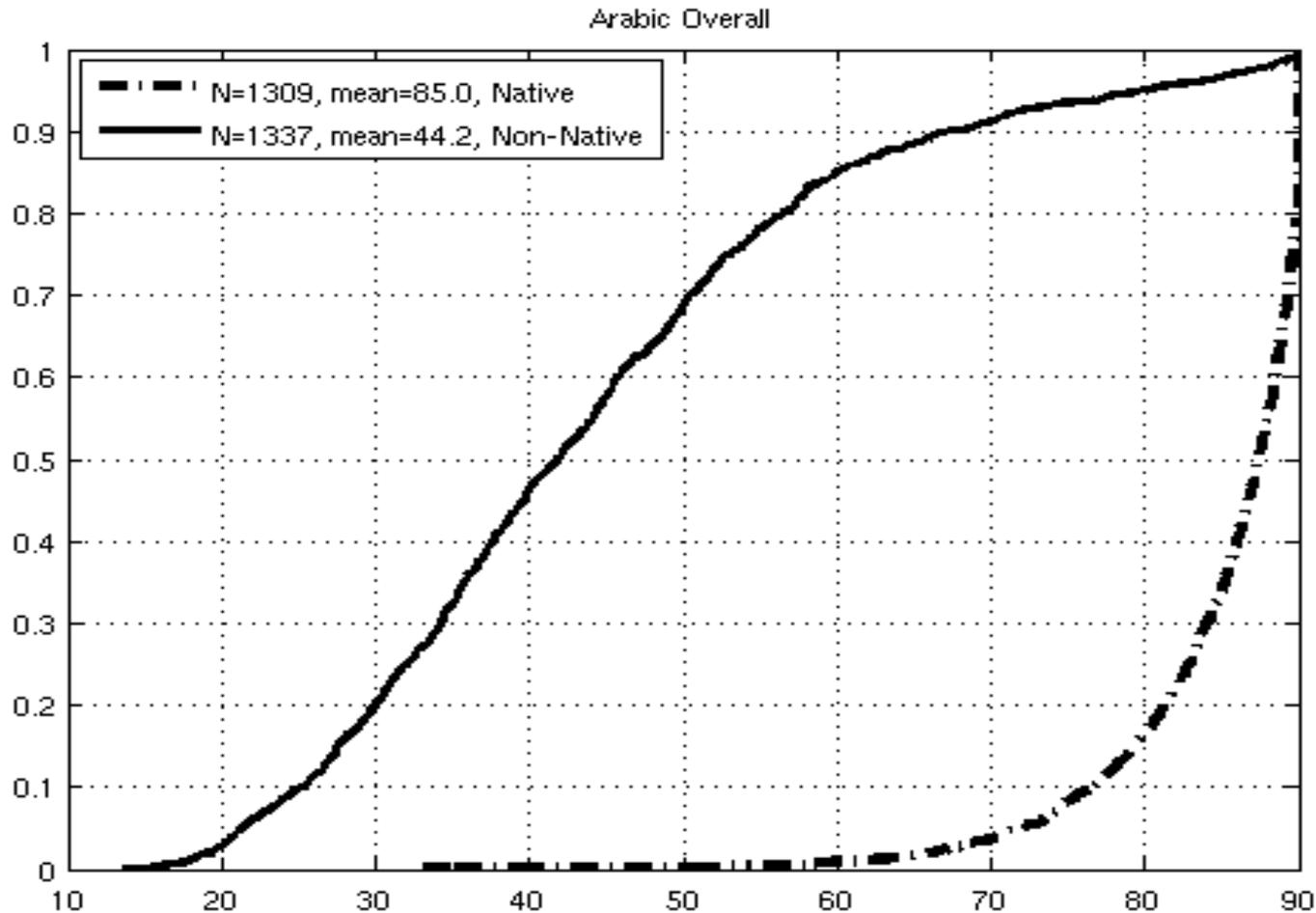
Natives by Countries
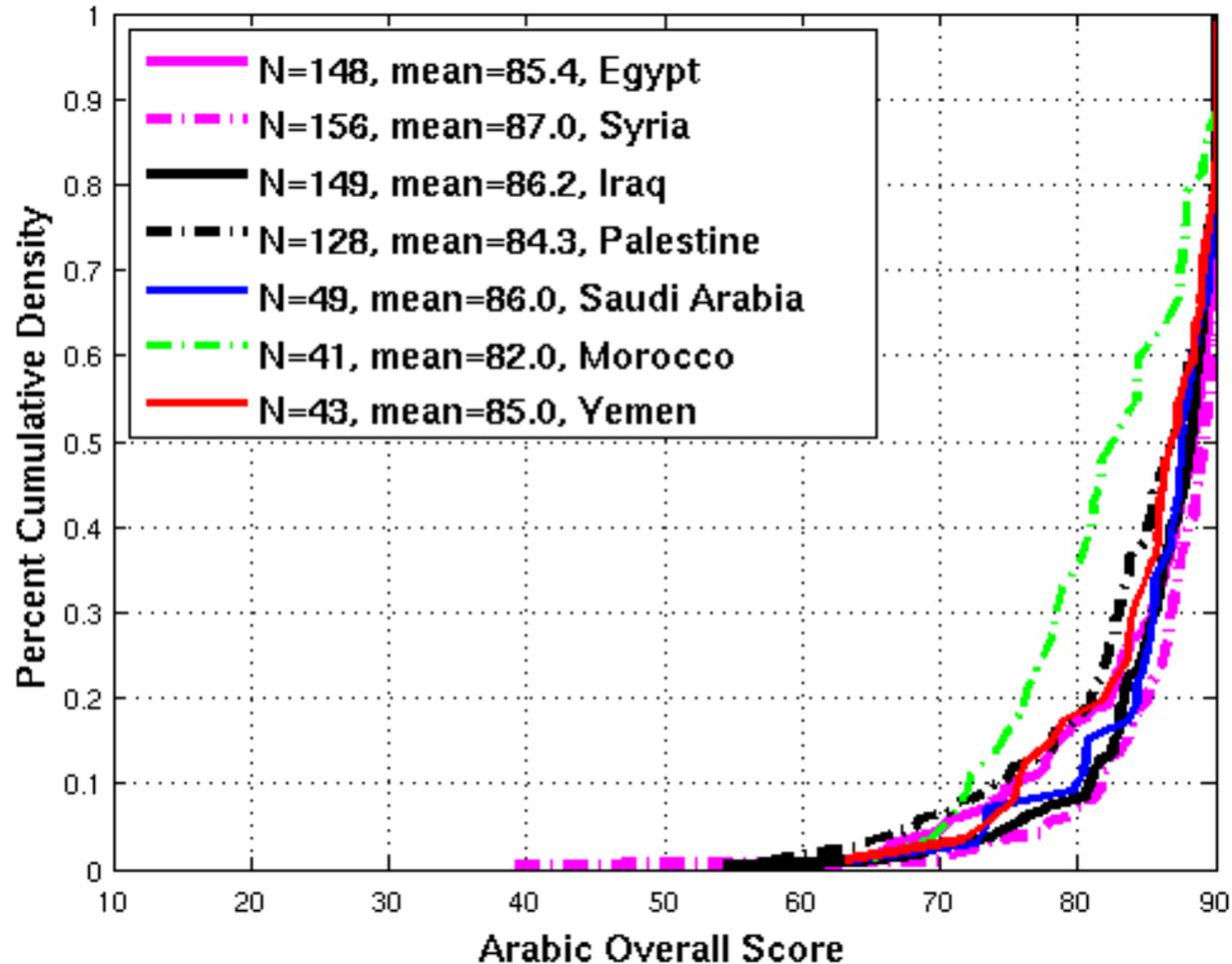
Concurrent Validation

1. Versant
   Machine Scores

Read	Repeat Sentence	Opposite	Short Question	Build S	OQ	St R
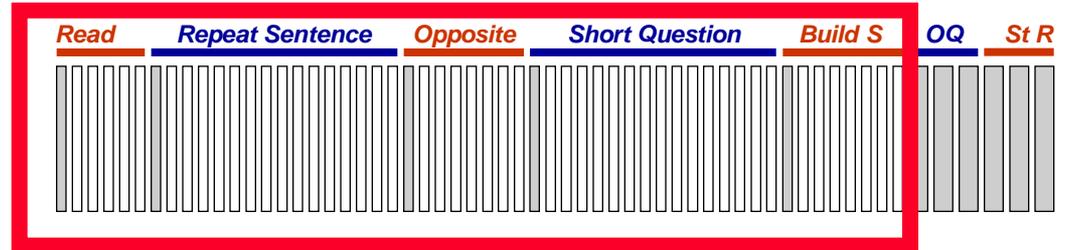
2. ILR-OPI
   Human Interview Scores
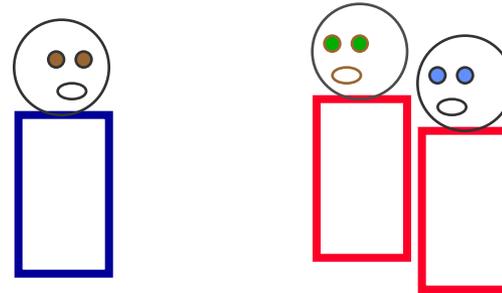
Rater-rater & rater-OPI correlations

Interview/Rater	N	Rater-rater Correlation	Rater-OPI Correlation
A	85	0.88	0.89
B	84	0.86	0.86
C	92	0.88	0.90
D	76	0.89	0.91
E	49	0.92	0.93
F	53	0.90	0.92
G	33	0.84	0.86
Average	472	0.88	0.90

In all cases, rating different OPI performances
Rater-based prediction range: 0.84-0.93

How Versants Compare to OPIs (3)

N = 118
r = 0.87

ILR OPI Score (logits) (y-axis)

Versant Arabic Overall Score (x-axis)
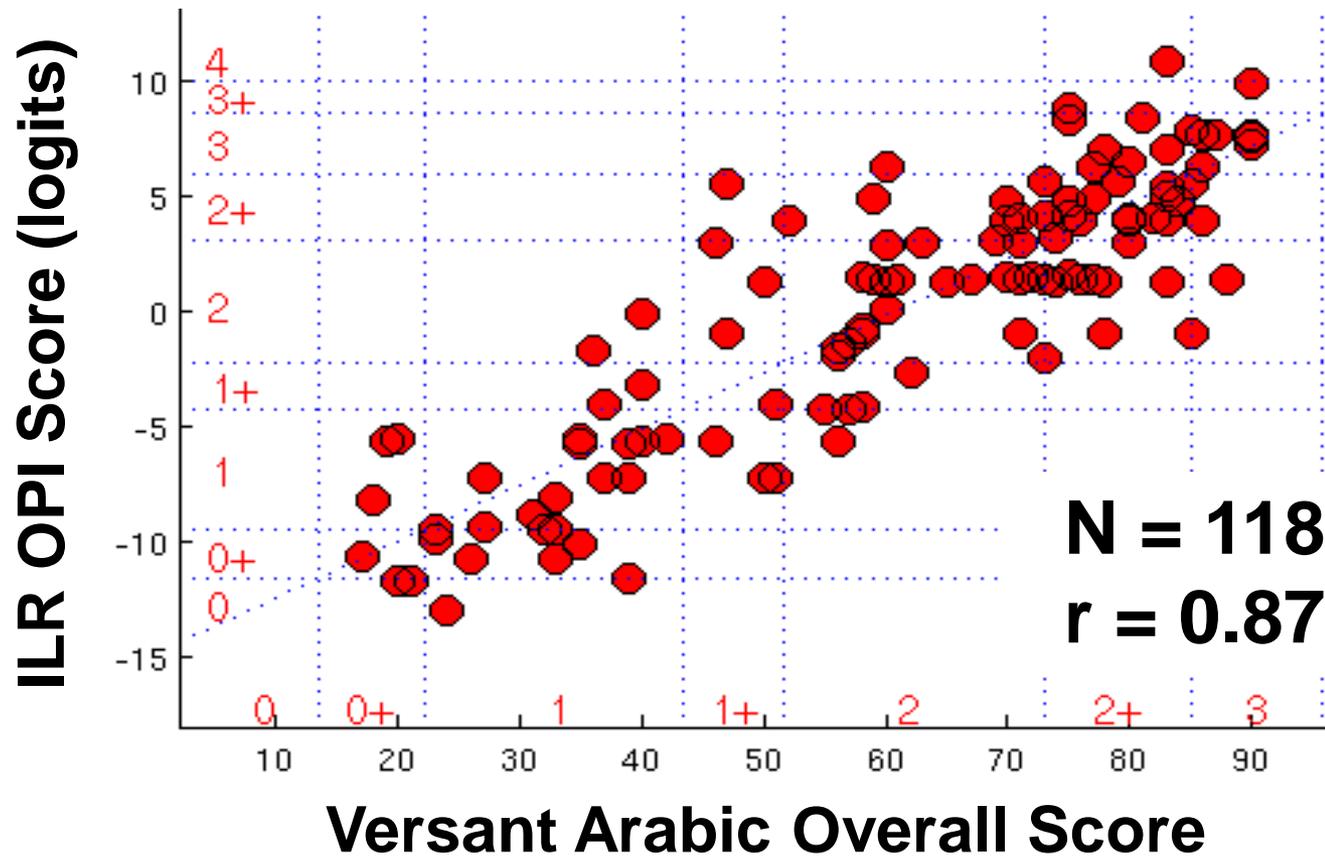
Spanish Data: Versant ~ OPI

N = 37
r = 0.92
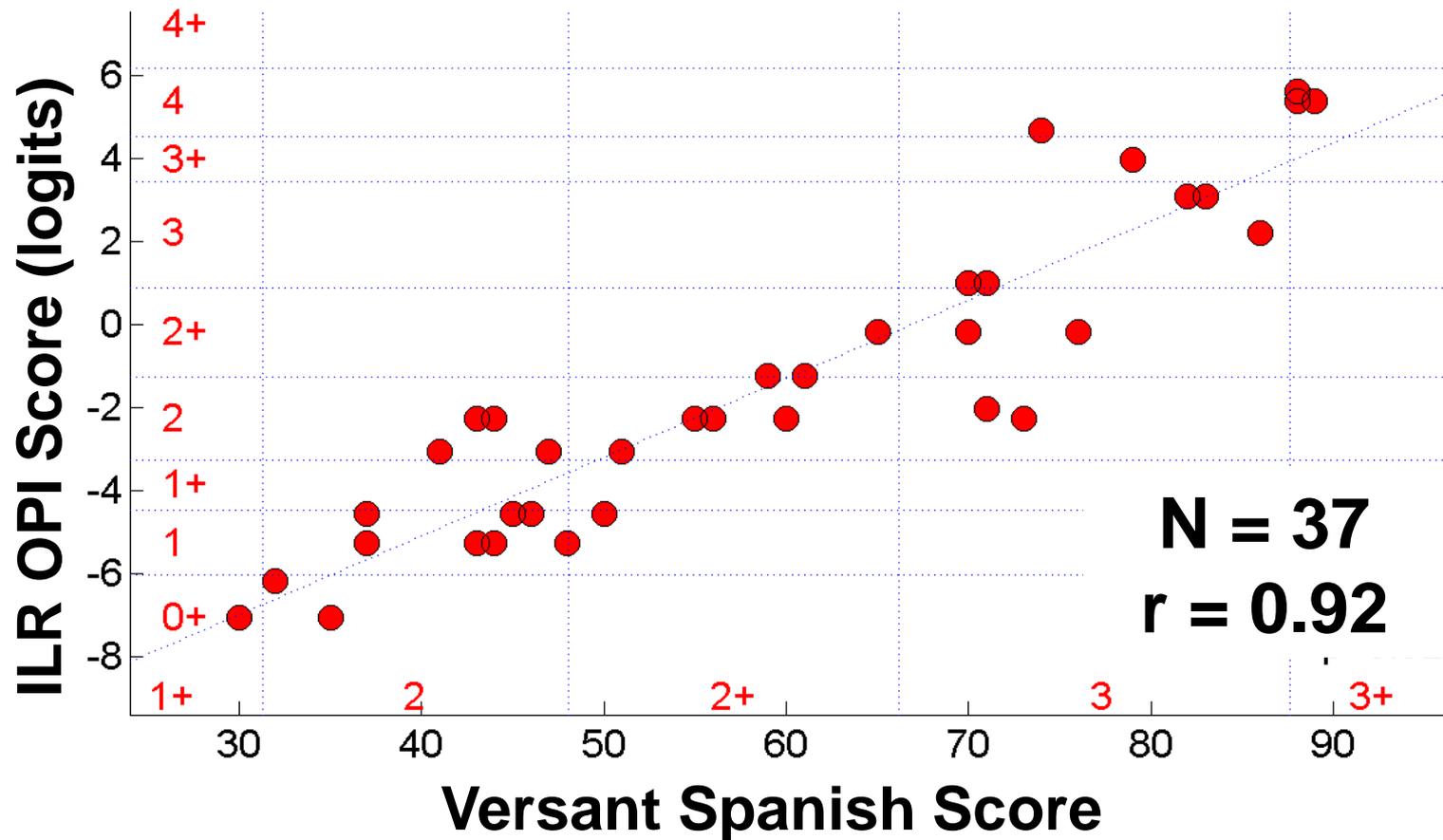
Concordance Table

SST Overall Score	Best Estimate of ILR Score	≥ ILR Score with 80% Confidence
20	0	0
21- 35	0+	At least 0+
36 - 43	1	At least 0+
44 - 49	1+	At least 1
50 - 55	2	At least 1+
56 - 60	2	At least 2
61 - 66	2+	At least 2
67 - 71	2+	At least 2+
72 - 77	3	At least 2+
78 - 80	3	At least 3

ILR triage steps

- Versant scores provide ILR level floors (twice) with 92% accuracy

Versant Arabic Score

Versant Spanish Score
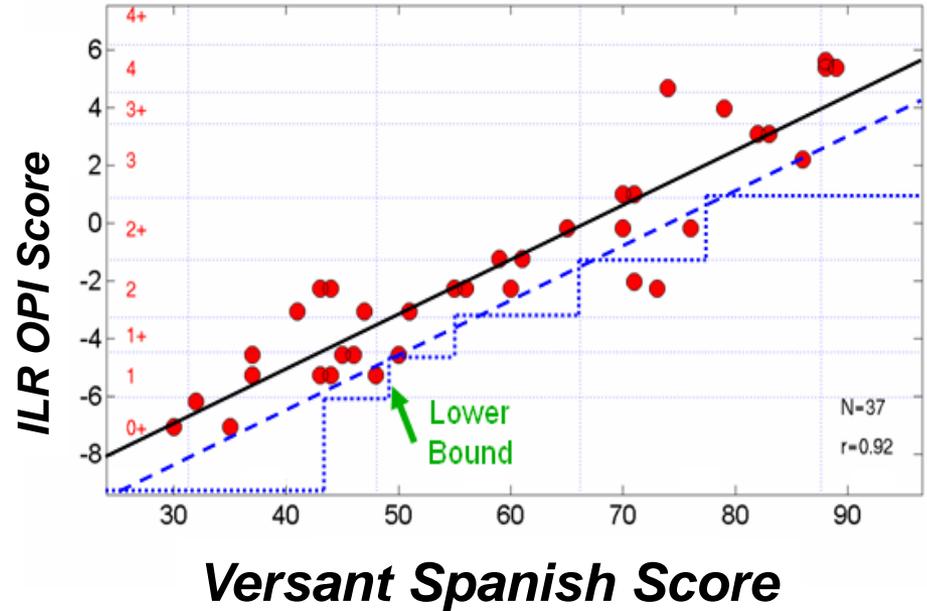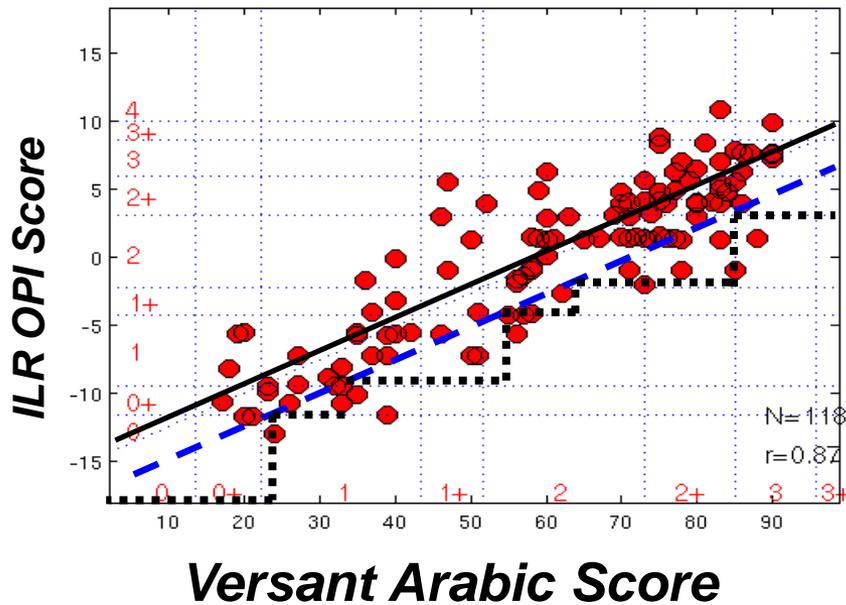
LowerBound(ILR) = ILR - (t-score)(standard error of the estimate)

For 80% confidence, 36 df: t = 0.85 (one tailed)

Construct Comparison

COMMUNICATIVE COMPETENCE*

Organization   Pragmatics

Grammar   Text   Illocution   Socioling.

V   M   S   P   Coh   Rh   Ideat   Manip   Huer   Imag   Dial   Reg   Nat   Cult

Taxonomic

***Bachman**

LANGUAGE FACILITY*

Grammar   Skill

V   M   S   P   Rate   Fluency

Metric in time

***Versant**
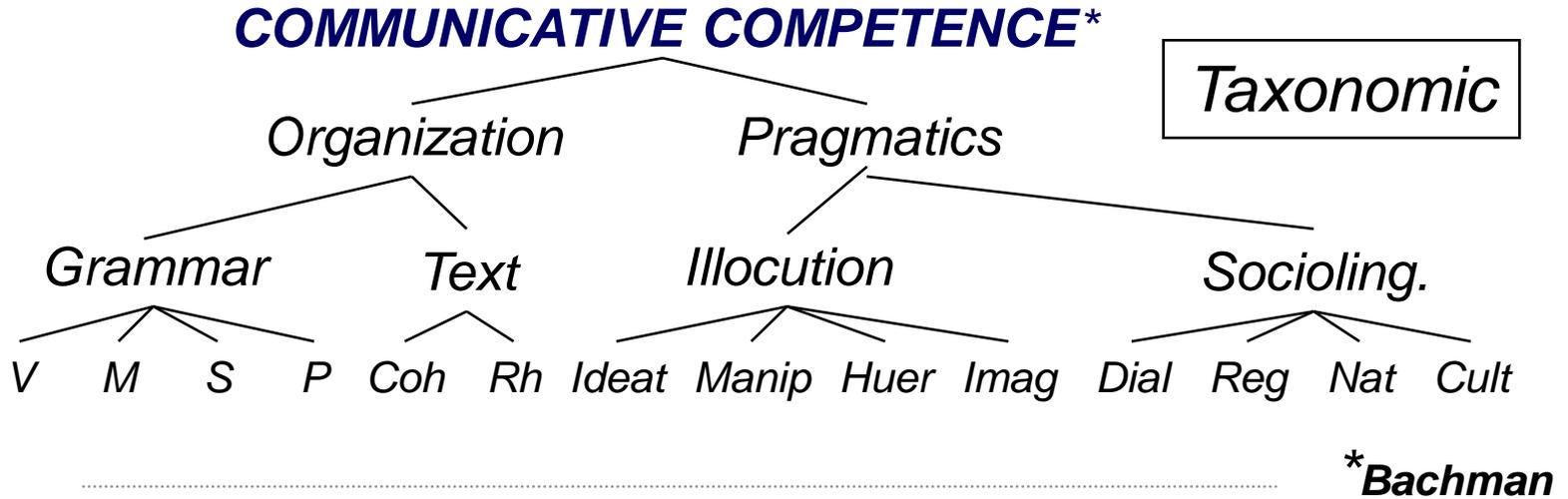
Construct Comparison

OPI Construct: Oral Proficiency as manifest in an Oral Proficiency Interview, is compatible with *communicative competence* as reflected in the functional level and/or complexity of content accurately produced.
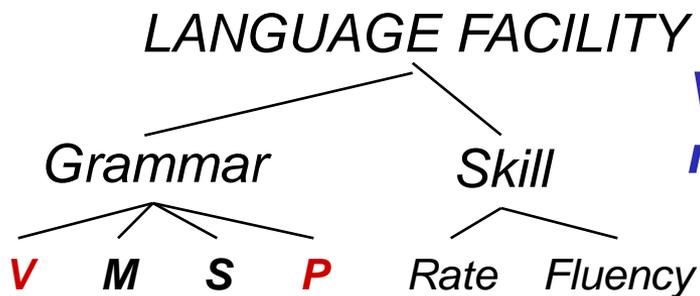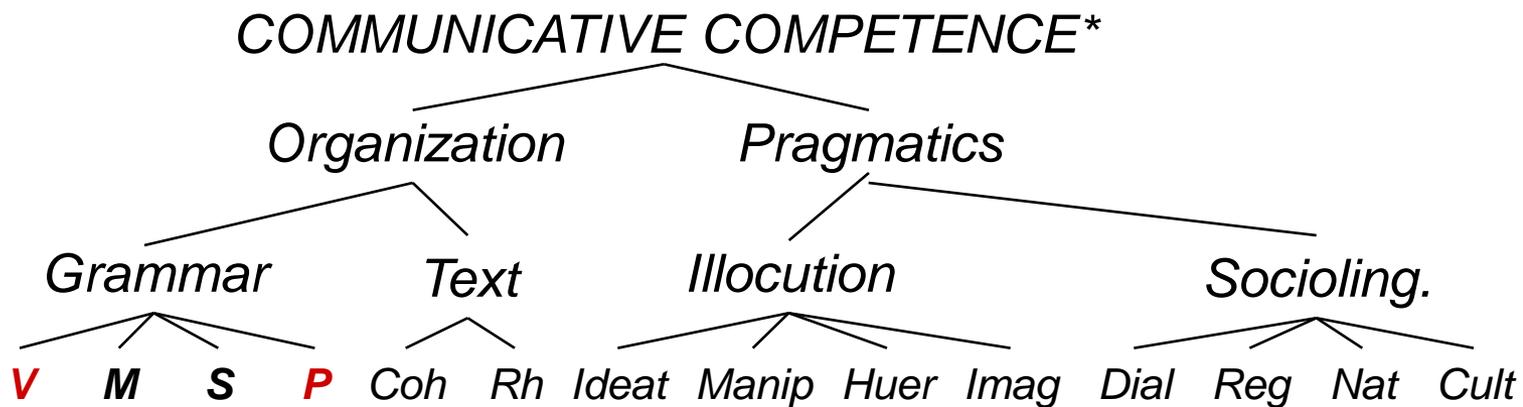
Versant Construct: *facility in spoken language* – the ability to understand spoken language and speak appropriately in response at a conversational pace on everyday topics.

Shared variance

0.87
prediction

COMMUNICATIVE COMPETENCE*

Organization   Pragmatics

Grammar   Text   Illocution   Socioling.

V **M** **S** **P** Coh Rh Ideat Manip Huer Imag Dial Reg Nat Cult

LANGUAGE FACILITY

Grammar   Skill

Versant tests contain sufficient material for predictive scoring

V **M** **S** **P** Rate Fluency

Summary

- Versant Arabic Test (VAT) is in operation
- Based on a large and wide body of transcribed spoken material
- VAT is available on demand
- VAT can triage or screen for OPI tests

- Useful score correspondence may be provided even when constructs differ

النهاية