Estimating Person Characteristics from Voice, Speech, Language and Touch

The 17th Annual Maryland Conference:
APPLICATION OF ARTIFICIAL INTELLIGENCE TO ASSESSMENT

2 November 2017

Sponsors: Maryland State Department of Education & Maryland Assessment Research Center

Jared Bernstein
Stanford University, Analytic Measures Inc.
Jian Cheng
Analytic Measures Inc.

DEFINITIONS (PRO TEM)

Artificial intelligence: the ability of automated systems to perform tasks that recently required human or other biological information processing.

Machine learning: algorithmic process that operates on data sets and then can cluster, classify, recognize, or identify patterns in new data.

… but take a warning:

there has been a huge expansion in the use of the term AI, so many things that used to be called "data science" or "internet of things" are now just called ***AI***.

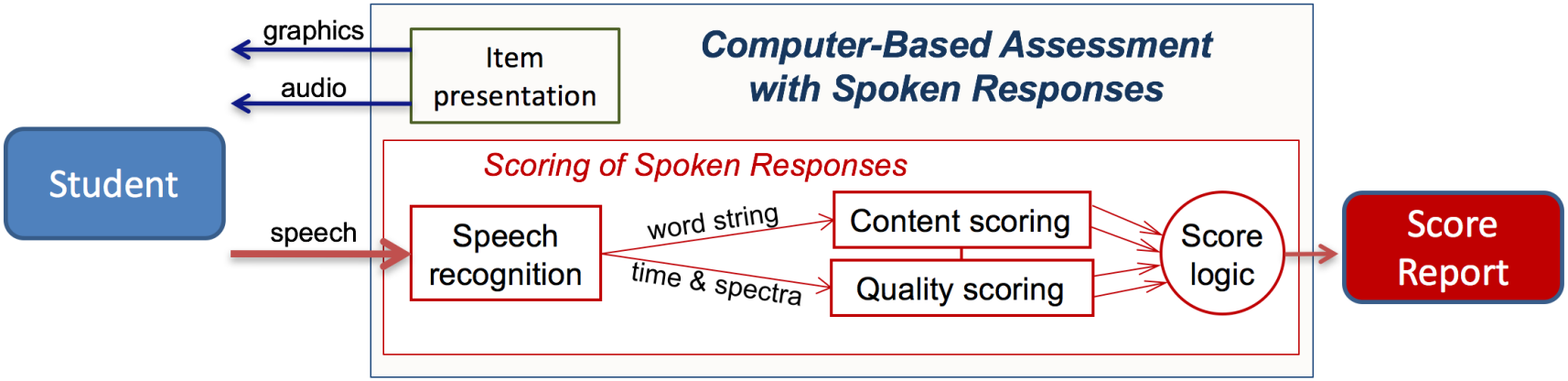
Lately, AI seems to mean "*anything complicated that computers do with external data*".

Underlying Tech: Regression, then … Clustering, HMMs, SVMs, DNNs, …

To build: ASR, NLP, Dialog Systems; face recognition, …

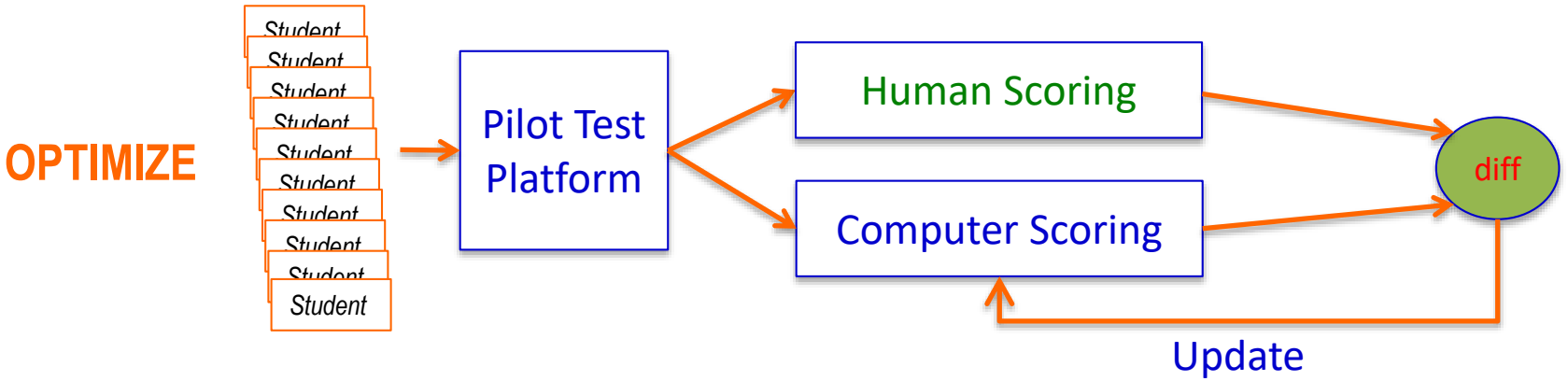
ami

System with Spoken Response Scoring



SCORING OPTIMIZATION

Automatic Scoring Development



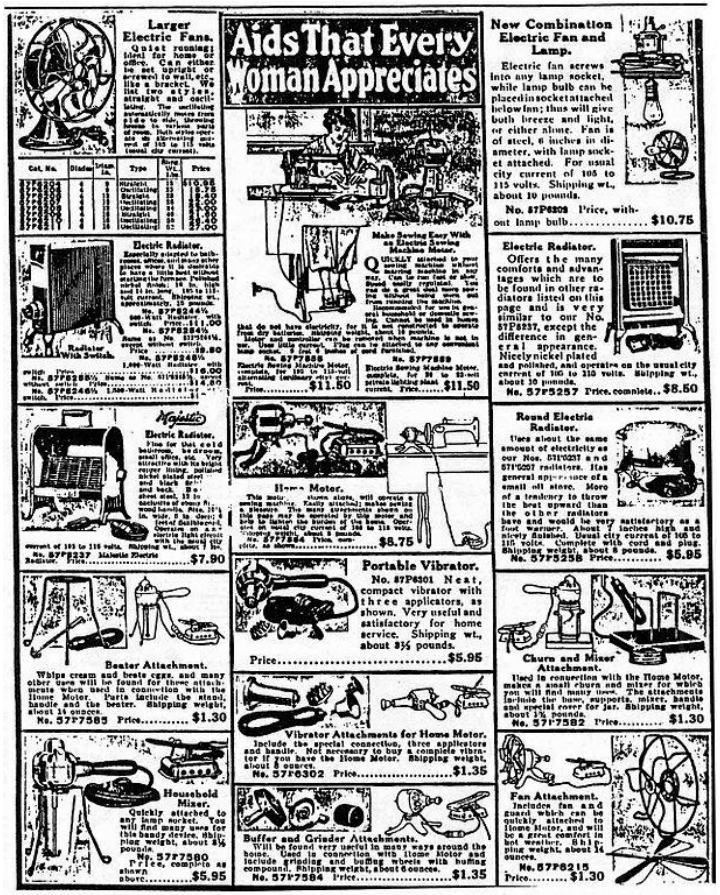
ami

How, When, Where to Use AI?

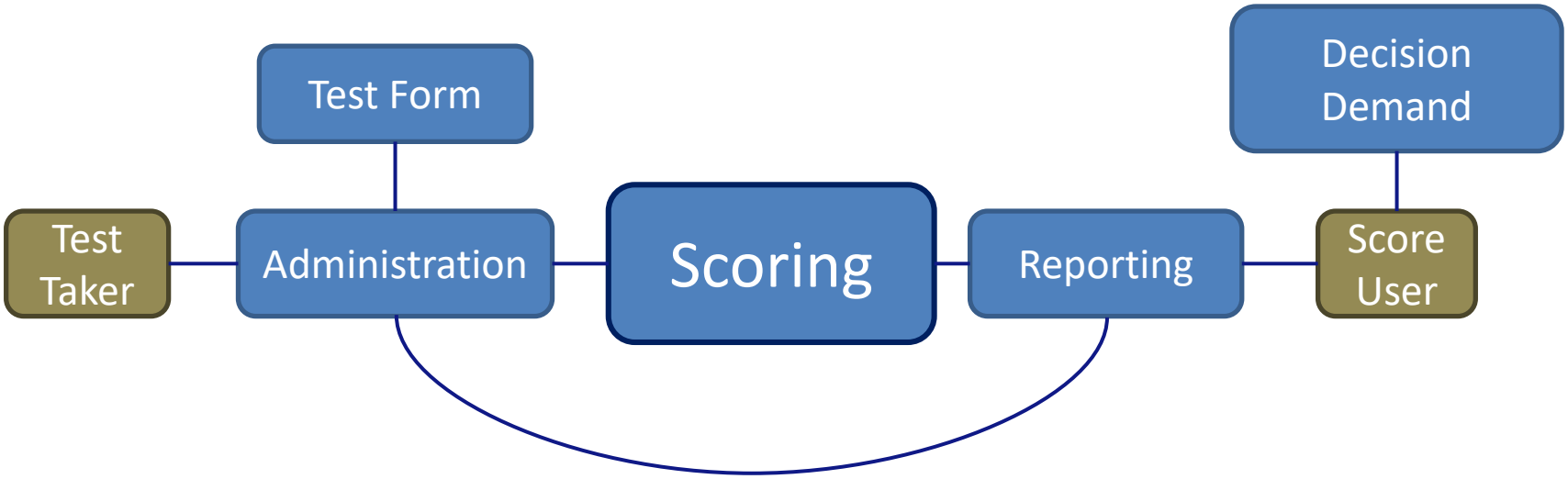
Ubiquitous computing and machine learning have applications that impact every aspect of assessment.

Electric Motors in the kitchen.

(Norman 1998, Sears, 1918)



Testing Elements



Impact on Elements of Testing

Decision Demand: program eval, demographics, selection

AI renders some skills irrelevant or obsolete

Test-Taker: adults, children, special populations

Assumed knowledge and skills are changing

Test Forms: task presentation, response types

New instructions, more task integration, skill isolation

Impact on Elements of Testing (2)

Administration: security, group, self, platforms

Common platforms enable secure self-administration

Scoring: speech, language, voice, touch, video

Automatic scoring of constructed performance

Reporting: states, traits, scores, examples

Scores in time context, with performance samples

ami

Decision Demand

Do our French students reach *ACTFL Superior* in 4 years?

Do our 2018 high school grads read as well as 2010 grads?

What should be the right cut-off score for our CPA exam?

Which response patterns justify sending a worker home?

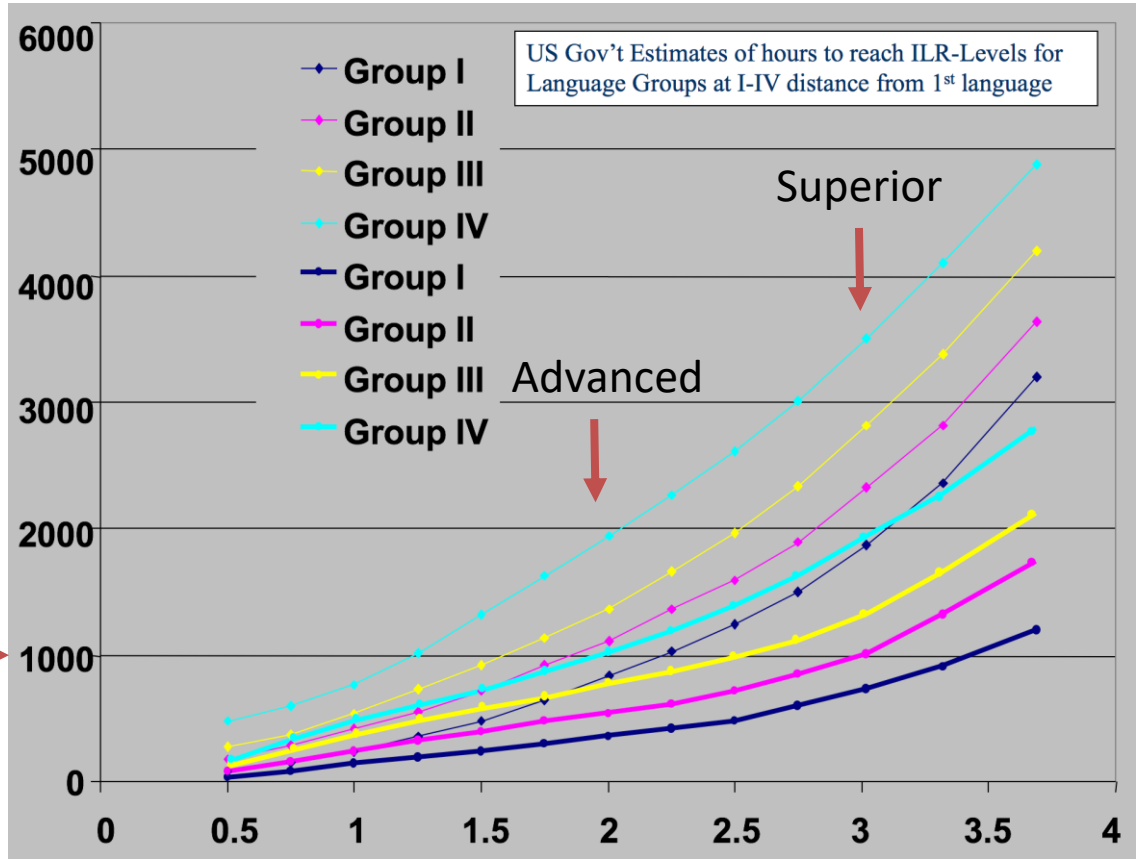
Which outpatient response patterns warrant a home visit?

ami

Cost of Proficiency in French or Chinese

In the long run, automatic spoken language interpretation may obviate L2 testing

6 mo. FSI
9 quarters (Univ.)



ami

AI Scoring Applications

AI scoring enables more efficient, reliable, and authentic assessment.

LANGUAGE

- **Reading:** Students read passages aloud and speech processing tech captures and analyzes speech for words correct per minute, comprehension, and prosody; then finds error patterns.
- **Writing:** Students draft prompt-specific essays or short answer responses and NLP tech yields content scores, feedback on grammar and mechanics, and overall writing scores.
- **L2 Language Proficiency:** English Language Learners (ELLs) provide written and/or spoken responses to short answer tasks; speech and text evaluation technologies return diagnostic and comprehensive measures of language skills (reading, writing, speaking, listening).

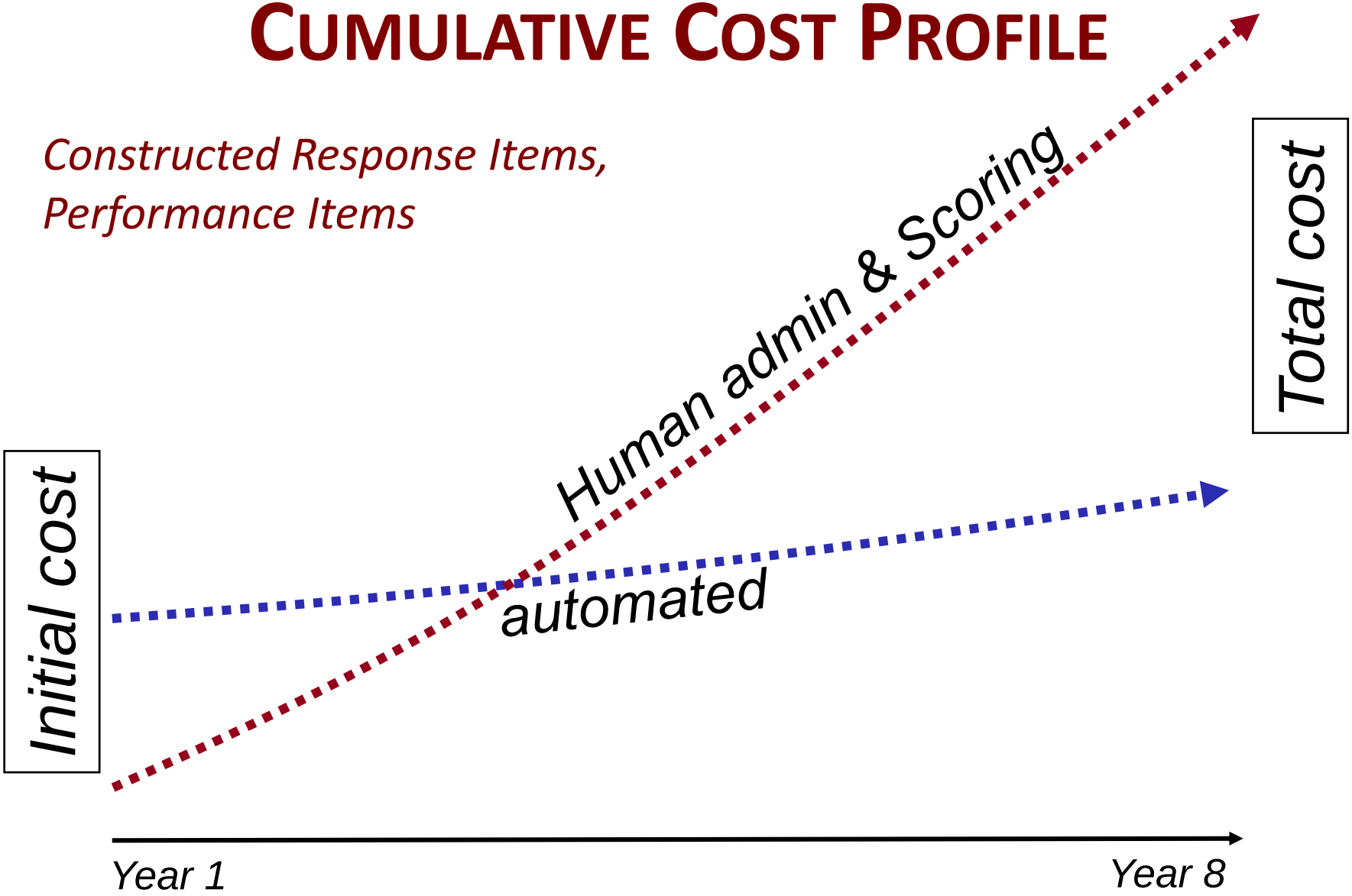
CONTENT KNOWLEDGE

- **Interactive Formative Practice:** Students read and/or watch material on a key STEM topic and then provide written or spoken short answer responses to demonstrate content knowledge. ML technologies can be applied to any content area, including science, social studies and math.

ami

Cumulative Cost Profile

Constructed Response Items,
Performance Items



ami

DEFINE, DEVELOP, SCORE, EVALUATE

Versant - Adult L2 Speaking & Listening

TTELL - K-6 L2 Listen, Speak, Read, Write

AZELLA - K-12 Listen, Speak, Read, Write

dMSE - delta Mental State Estimate (cog. & affect)

PACES - Profile of Attitude, Comm., Energy, Skill

Moby.Read - Self-admin. Oral Reading Fluency

eORF - Special study instrument for 2018 NAEP

ami

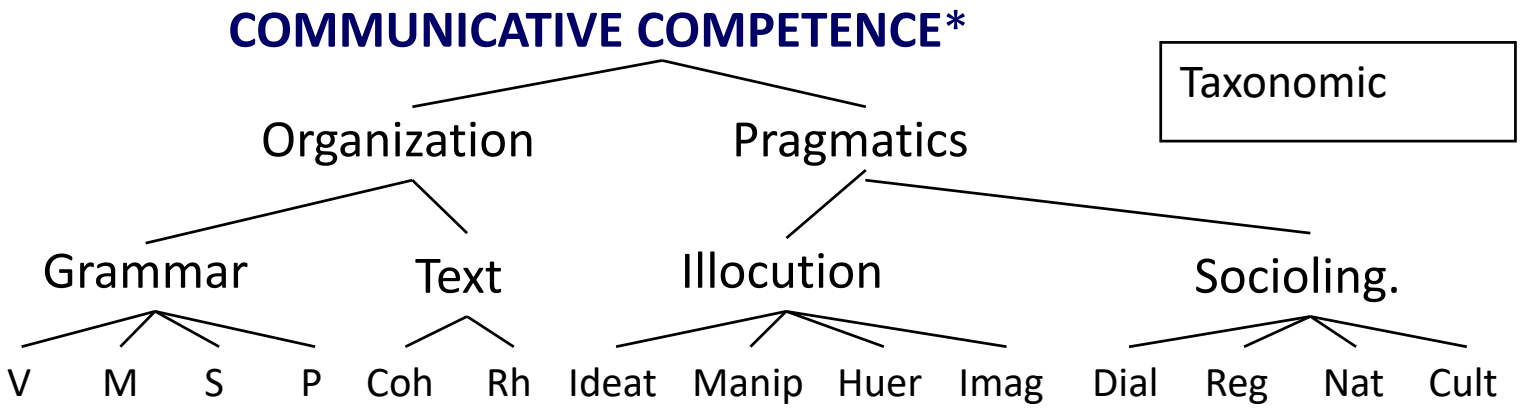
PHONEPASS/VERSANT

Fully automatic spoken language test

- Construct: *facility in spoken English* –
  the ability to understand spoken English and speak appropriately in response at a native-like pace on everyday topics.
- 1998 English
- 2003 Spanish
- 2008 Arabic
- 2012 Chinese

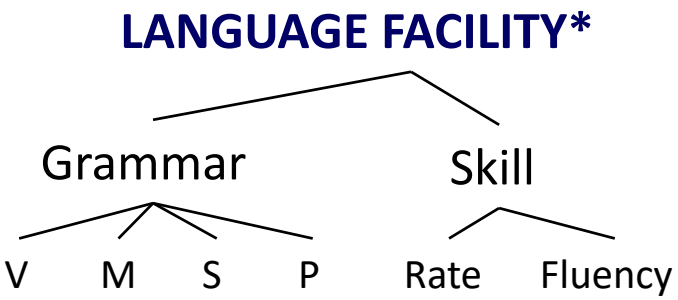
ami

CONSTRUCT COMPARISON



Taxonomic

*Bachman



FSMs, HMMs
Metric in time

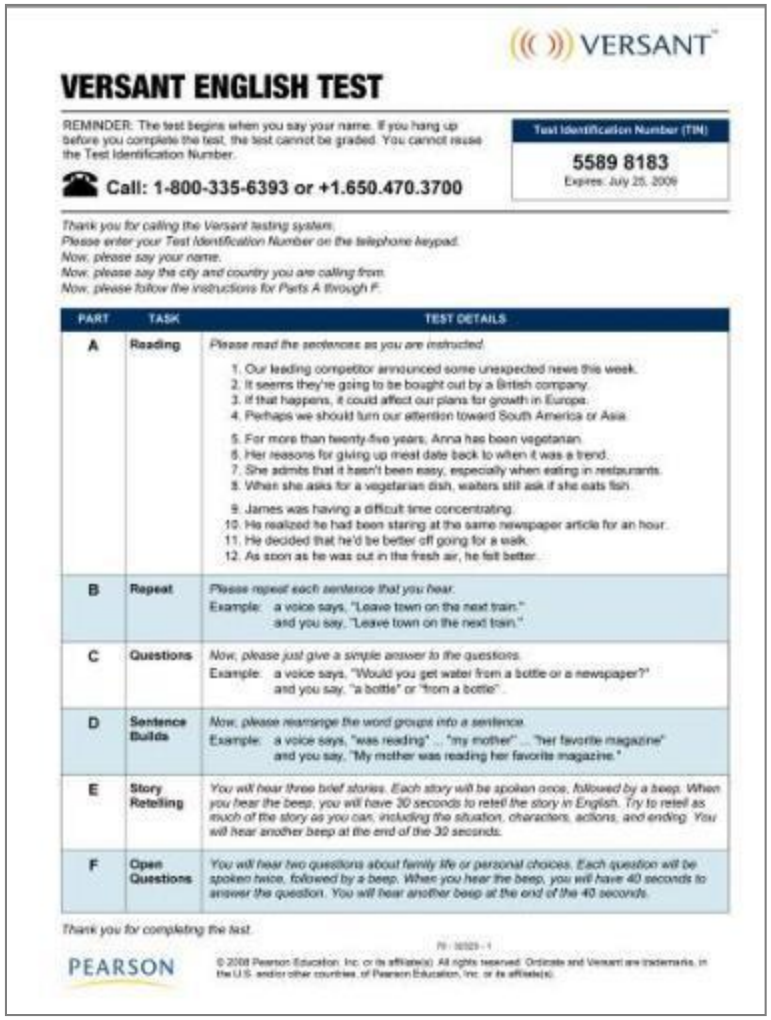
*SET-10

ami

AUTOMATED PROFICIENCY TESTING

- Versant English Test
  - Part A: Reading
  - Part B: Repeat
  - Part C: Short Questions
  - Part D: Sentence Builds
  - Part E: Story Retellings
  - Part F: Open Questions

- Total 63 Questions
- ~14 minutes



VERSANT

VERSANT ENGLISH TEST

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: 1-800-335-6393 or +1.650.470.3700

Test Identification Number (TIN)
5589 8183
Expires: July 25, 2009

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 you are calling from.
Now, please follow the instructions for Parts A through F.

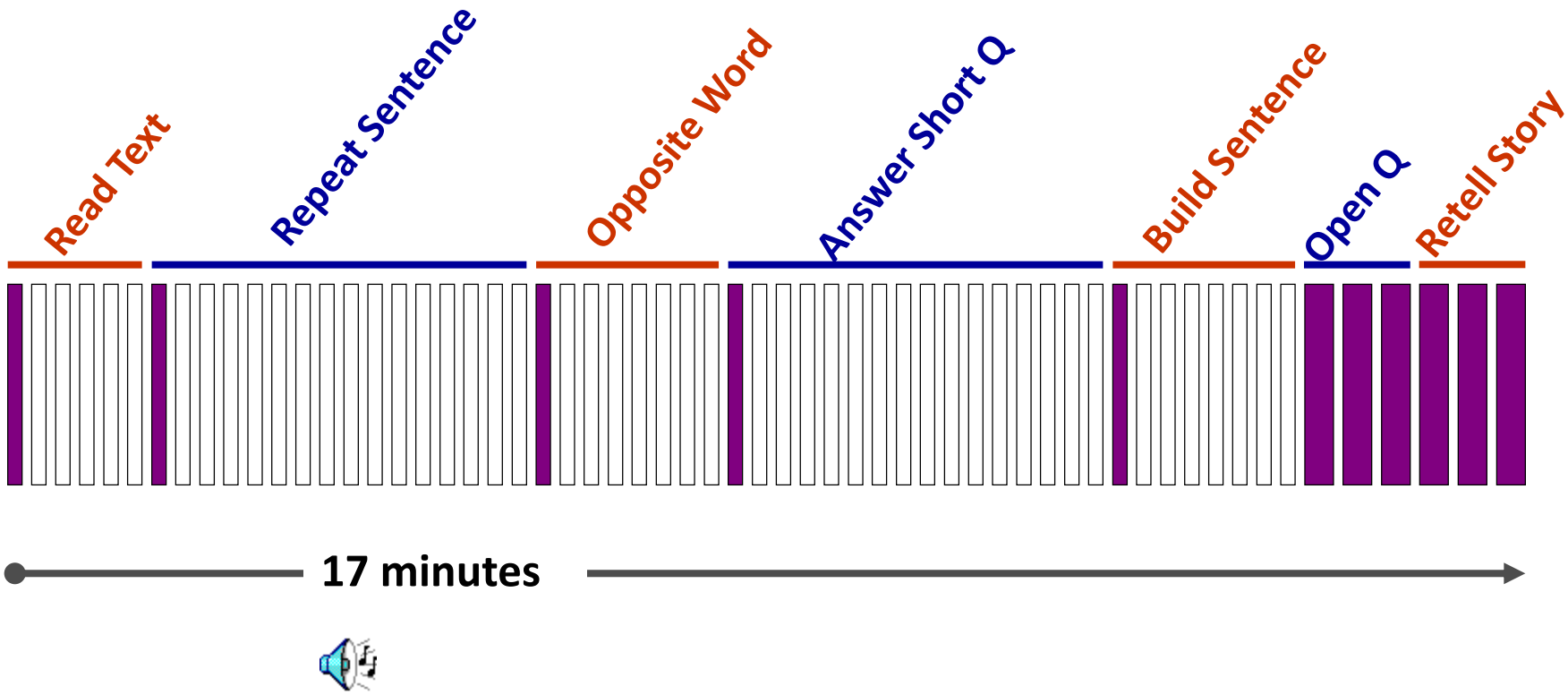
PART	TASK	TEST DETAILS
A	Reading	*Please read the sentences as you are instructed.* 1. Our leading competitor announced some unexpected news this week. 2. It seems they're going to be bought out by a British company. 3. If that happens, it could affect our plans for growth in Europe. 4. Perhaps we should turn our attention toward South America or Asia. 5. For more than twenty-five years, Anna has been vegetarian. 6. Her reasons for giving up meat date back to when it was a trend. 7. She admits that it hasn't been easy, especially when eating in restaurants. 8. When she asks for a vegetarian dish, waiters still ask if she eats fish. 9. James was having a difficult time concentrating. 10. He realized he had been staring at the same newspaper article for an hour. 11. He decided that he'd be better off going for a walk. 12. As soon as he was out in the fresh air, he felt better.
B	Repeat	*Please repeat each sentence that you hear.* Example: a voice says, "Leave town on the next train." and you say, "Leave town on the next train."
C	Questions	*Now, please just give a simple answer to the questions.* Example: a voice says, "Would you get water from a bottle or a newspaper?" and you say, "a bottle" or "from a bottle".
D	Sentence Builds	*Now, please rearrange the word groups into a sentence.* Example: a voice says, "was reading" ... "my mother" ... "her favorite magazine" and you say, "My mother was reading her favorite magazine."
E	Story Retelling	*You will hear three brief stories. Each story will be spoken once, followed by a beep. When you hear the beep, you will have 30 seconds to retell the story in English. Try to retell as much of the story as you can, including the situation, characters, actions, and ending. You will hear another beep at the end of the 30 seconds.*
F	Open Questions	*You will hear two questions about family life or personal choices. Each question will be spoken twice, followed by a beep. When you hear the beep, you will have 40 seconds to answer the question. You will hear another beep at the end of the 40 seconds.*

Thank you for completing the test.

PEARSON

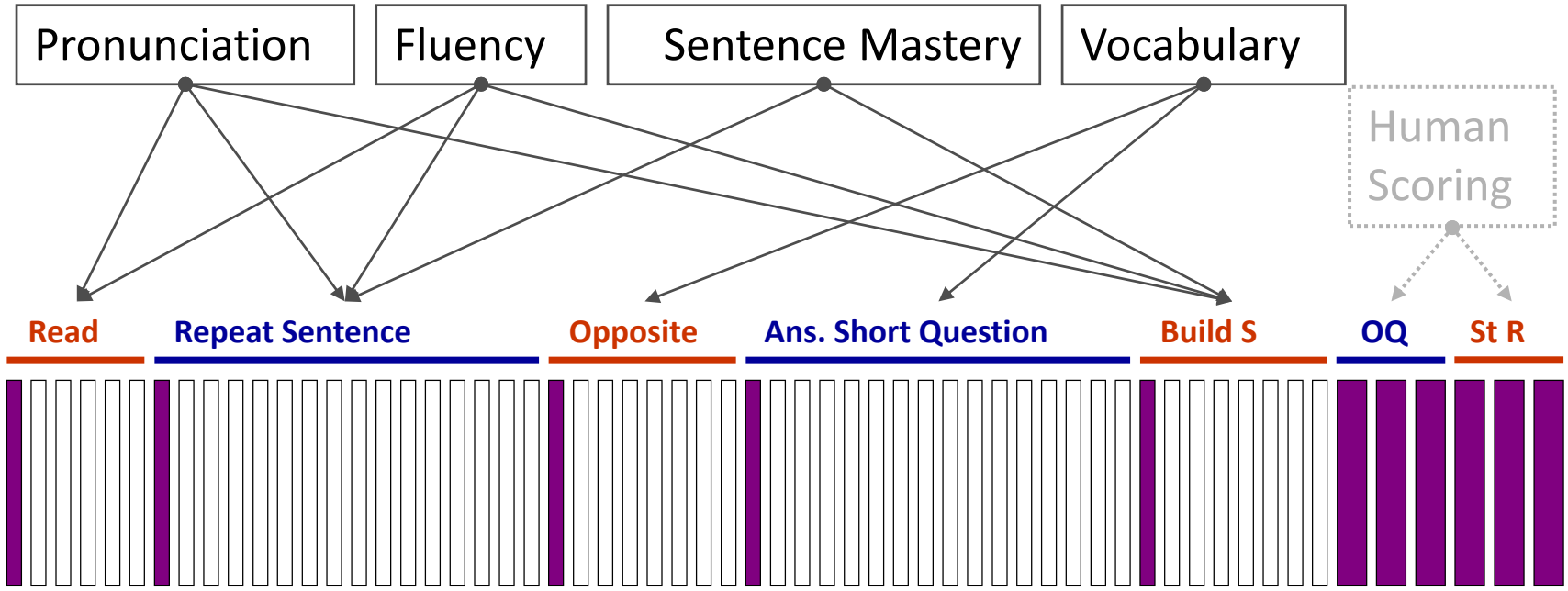
ami

SST: 60-Item Sequence



ami

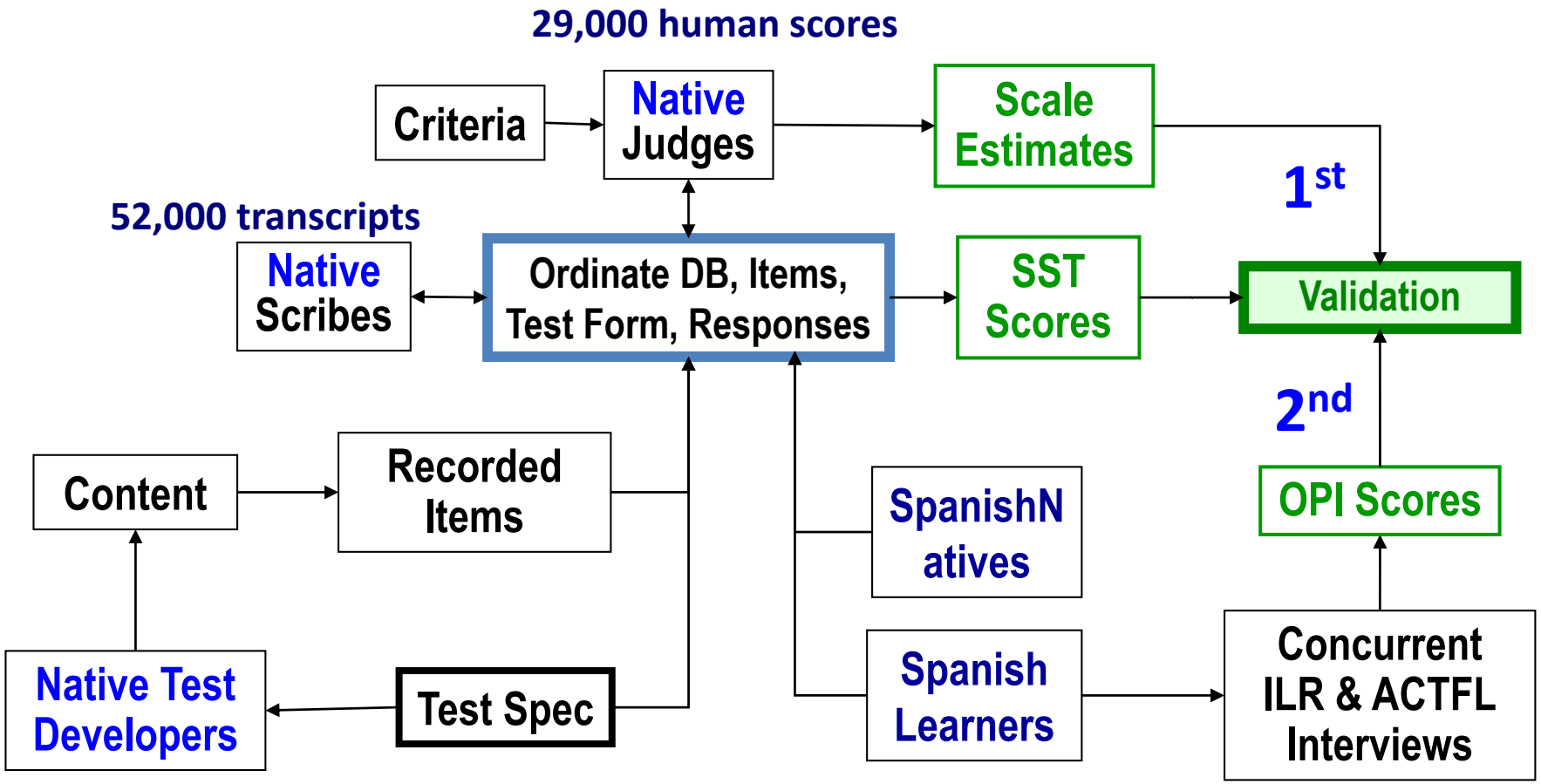
SST Machine Scoring Logic



SST = **(**30% Sent.M, 20% Vocab, 30% Fluency, 20% Pron**)**

ami

VERSANT SPANISH DEVELOPMENT AND VALIDATION



ami

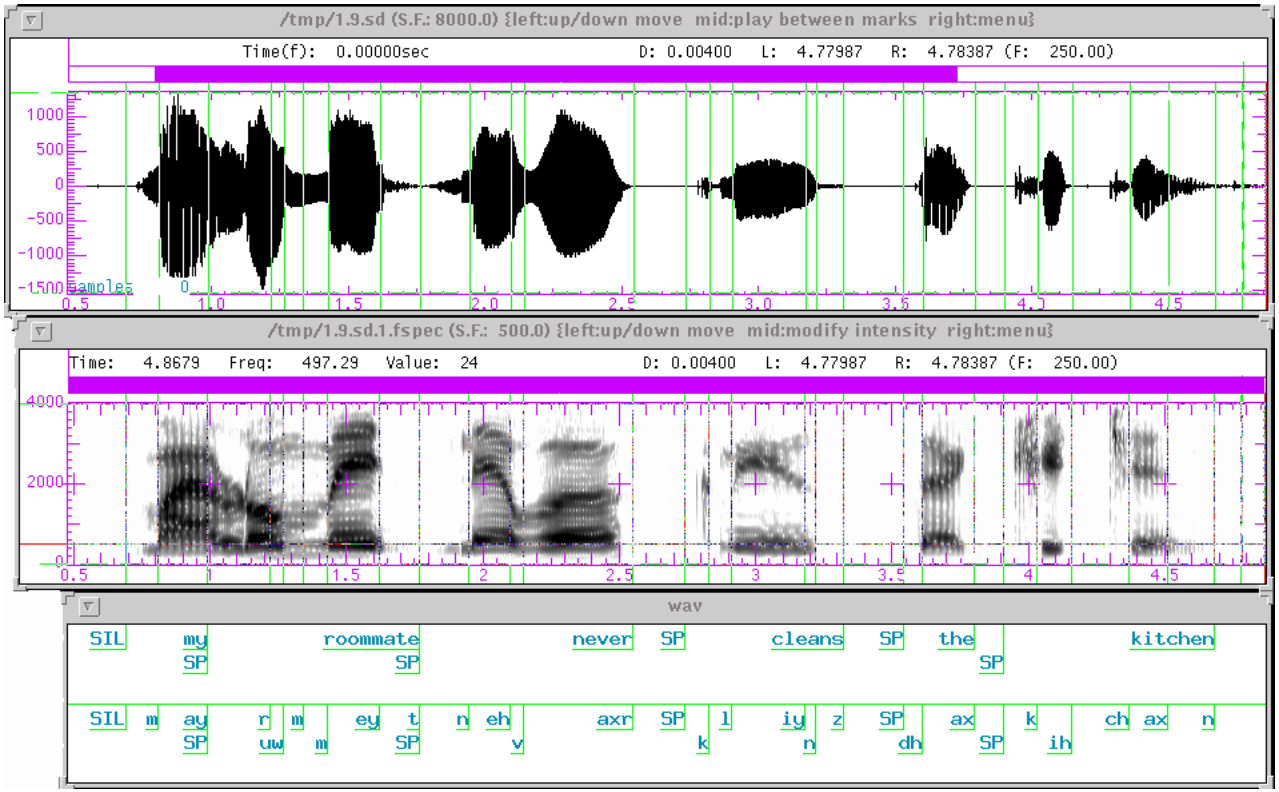
PHONEME & WORD ALIGNMENT

w1 w2 w3 w4 w5 w6

p p pppp p | p p p p | pp ppp | pp | p p p p p

75-90 Words/Min

5.8 Phones/Sec



waveform

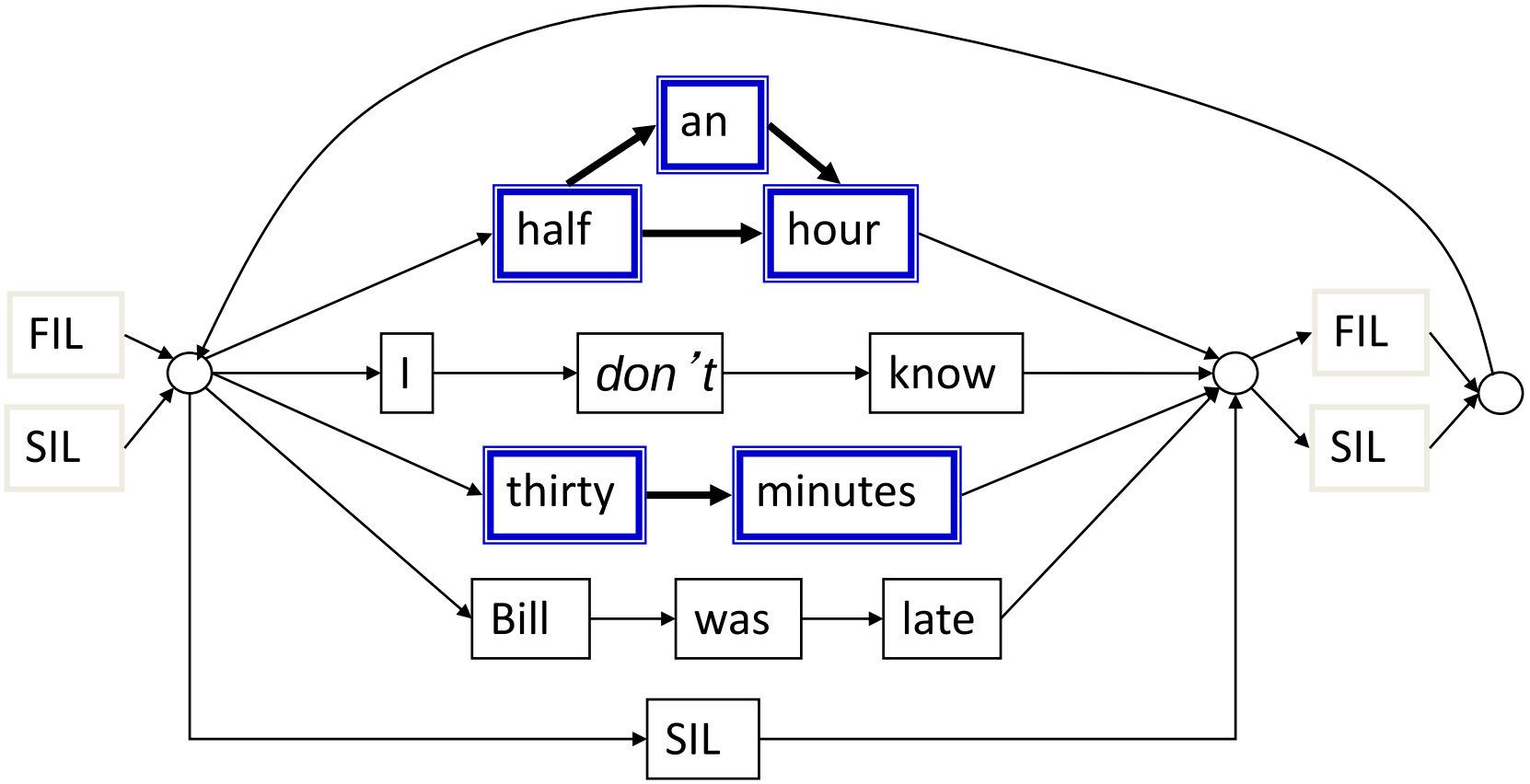
spectrum

words

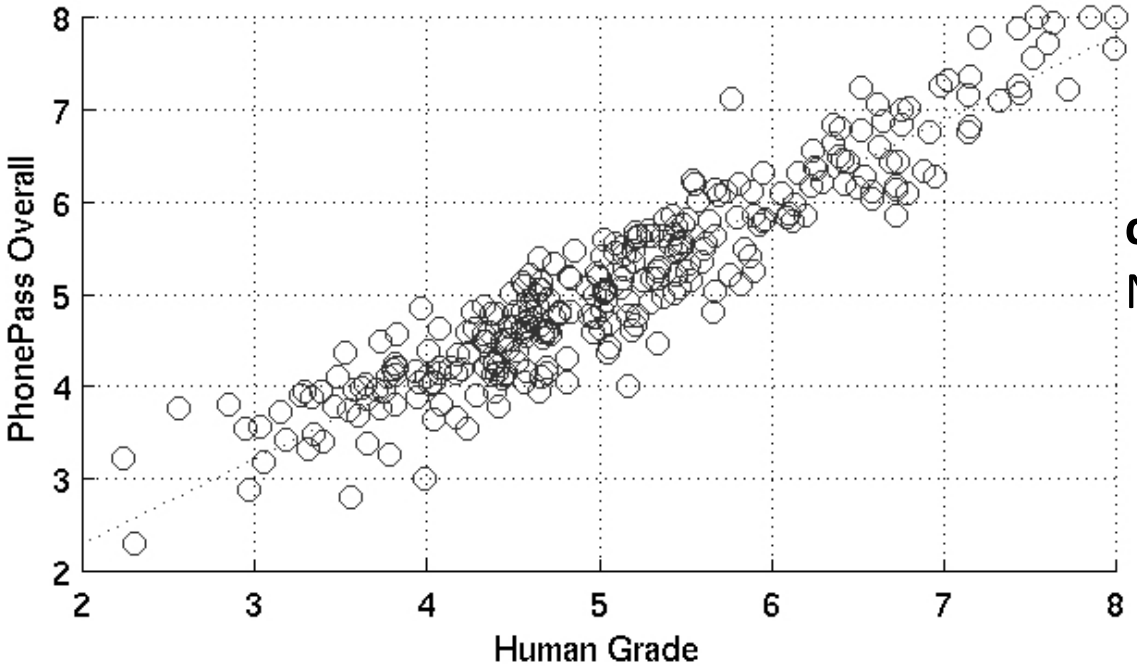
segmentation

ami

Simplified Response Network



1ST MACHINE-HUMAN COMPARISON

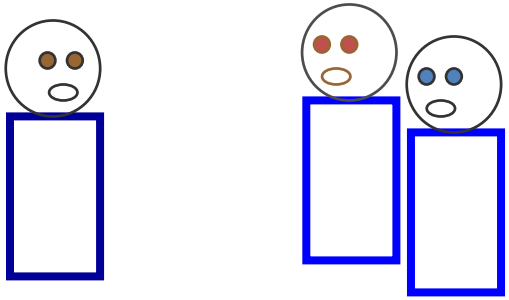


correlation = 0.94
N = 288

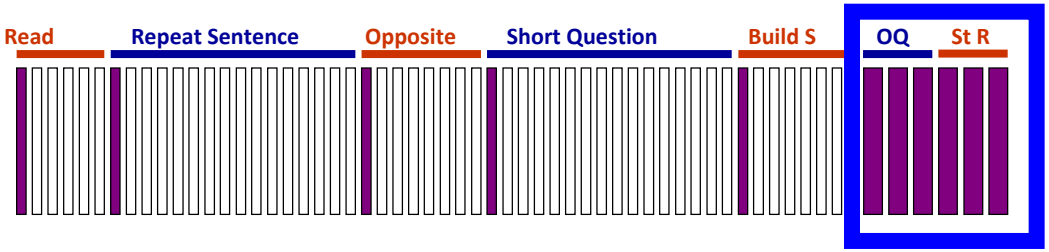
Human scoring compared to machine-scoring (2003)

ami

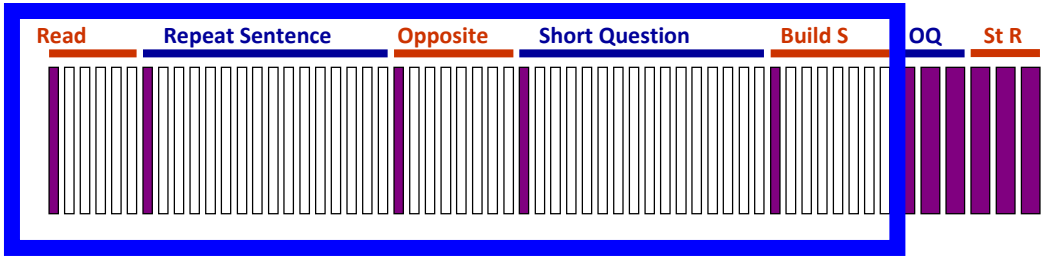
HUMAN AND MACHINE SCORES



ILR-FBI and ACTFL
Human Interview Scores



ILR-FBI, ILR-DLI, CEF
Scale Estimates
(2 human raters per)



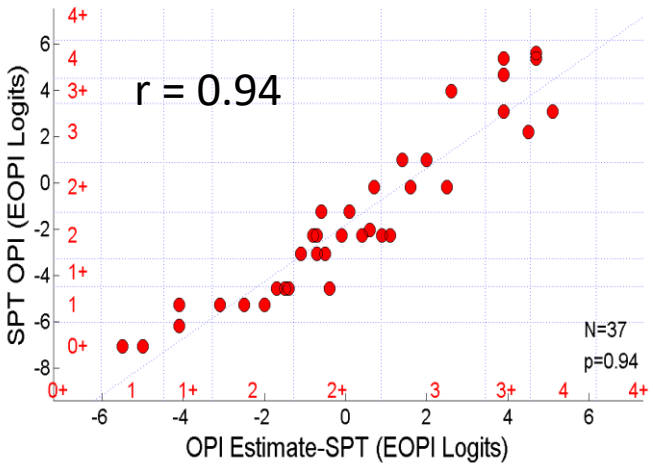
SST
Machine Scores

ami

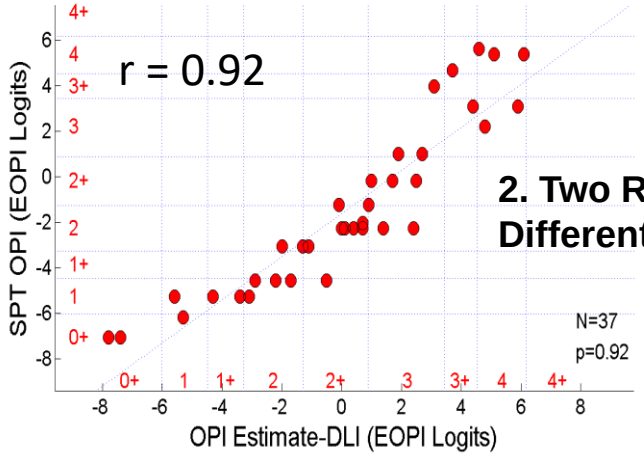
2nd Validation: Spanish Data (SST)

U.S. Government OPI Interviews

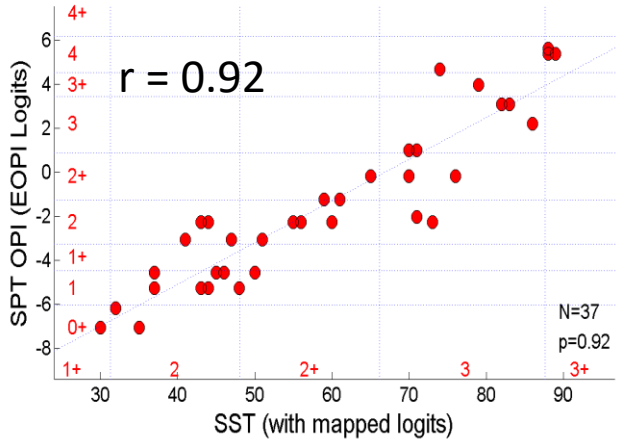
1. OPI A-Raters ~ A-Raters Estimate
2. OPI A-Raters ~ B-Raters Estimate
3. OPI A-Raters ~ Machine score



1. Same Raters
Different Material



2. Two Rater Pairs
Different Material

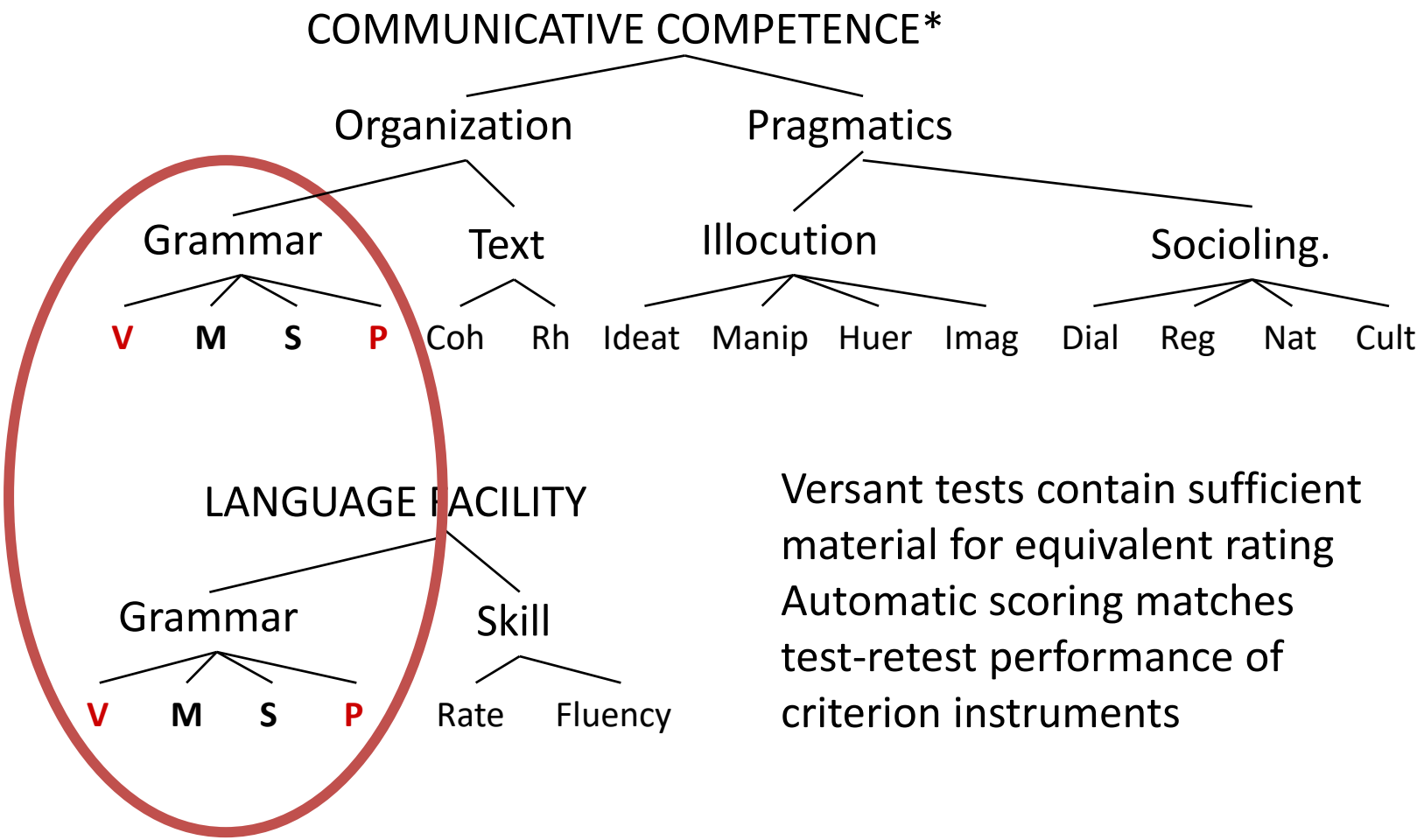


3. Machine ~ Two Raters
Different Material

ami

2ND VALIDATION → PERFORMANCE PUZZLE

~80%
of variance



Versant tests contain sufficient material for equivalent rating
Automatic scoring matches test-retest performance of criterion instruments

ami

SLP Paradigm in Versant Tests

- Integrated model of linguistic performance
  - embedded phoneme, word, and phrase networks
  - quantitative models of criterion judgment and data-driven performance criteria
- Corpus-based content and scoring
  - Content is restricted by corpus occurrence
  - Explicit model of target interlocutor
  - Explicit, metric combination score elements

ami

Assessment Design Space

Scoring Focus	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge	+	+	+	+
Language Skills	+	+	+	+

TTELL

Touch Tablet English Language Learner (2012-13)

Exercise many feasible task formats

K-6 students self-administer ELL assessment on iPad

Four Skills

Automatically Scored

(now ***TELL***™ K-12 product from Pearson)

ami

TOUCH TABLET COMPUTER

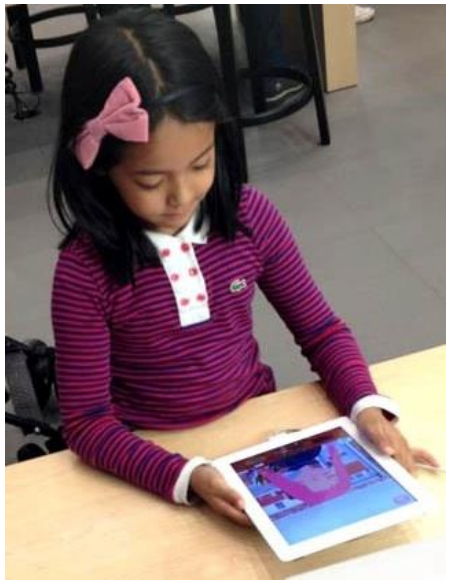
DESIGN, IMPLEMENT, TEST

Touch tablet language tasks that elicit & monitor an ELL's language performance

Keep the best traditions

Aim to improve:

- Engagement
- Independence: self-administered
- Efficiency: more information per time
- Consistency across location
- Fidelity to the new performance standards



ami

FOUR-SKILLS TEST

Listen

– Touch, move, draw path, … as instructed by voice

Speak

– Repeat, describe, retell, read aloud …

Read

– Touch, move, draw path, word recognition, … from text

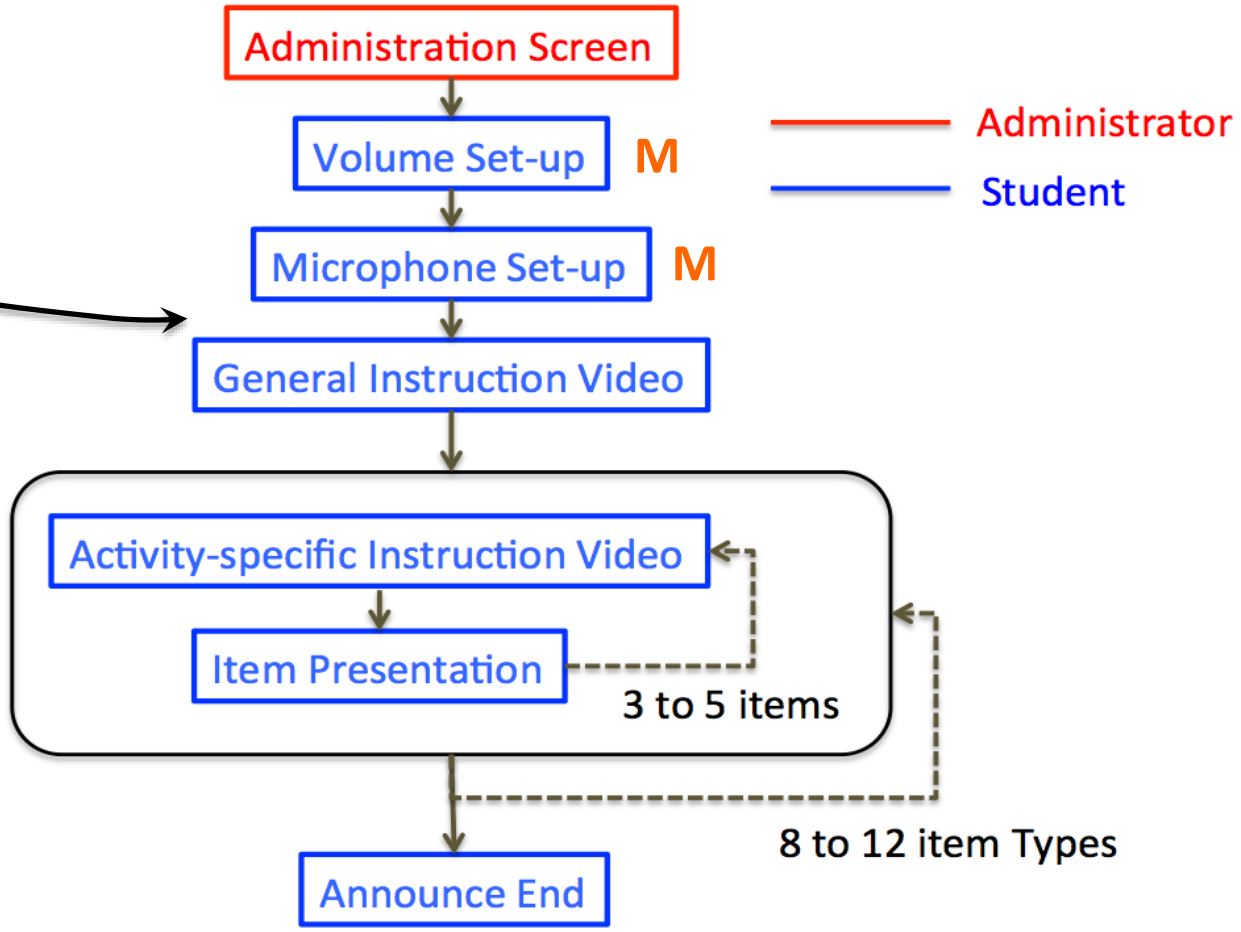
Write

– Describe, relay/summarize, spell, cloze, find error

ami

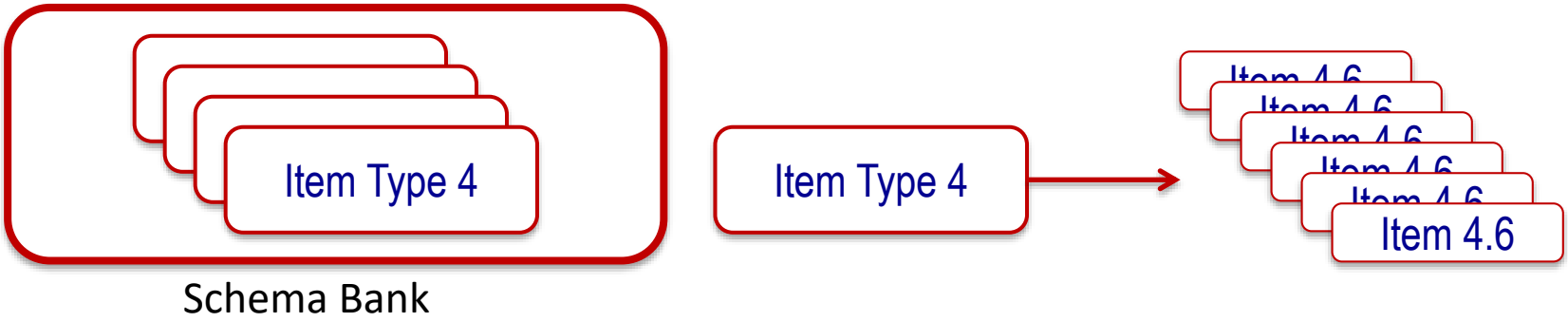
ADMINISTRATION & PRESENTATION

No teacher assistance from here on.



ami

ITEMS < ITEM TYPES <SCHEMAS



Example Schema:

Present: illustration & audio
Capture: touches & gestures

Example Item type

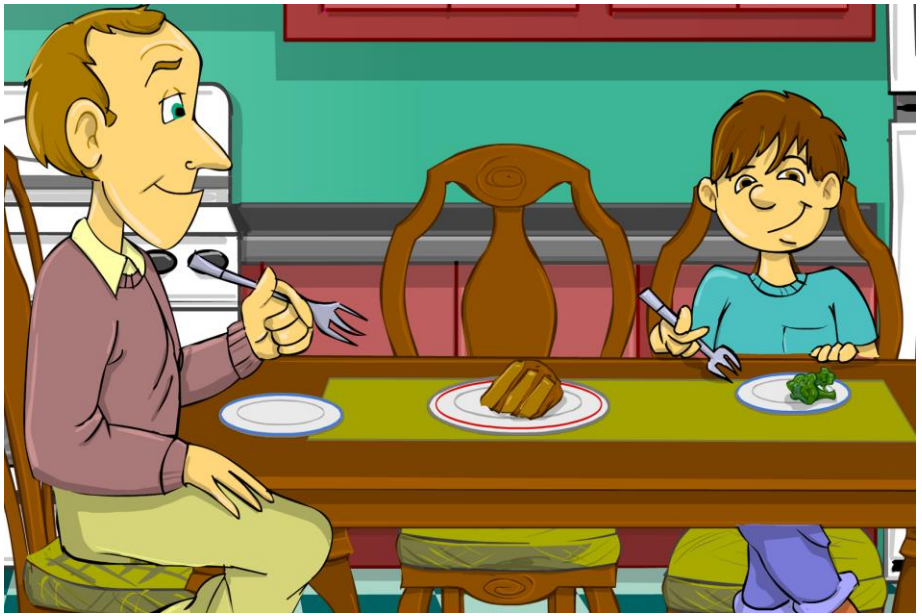
Present: drawing of a scene with familiar objects
recorded dialog mentioning select objects

Capture: touches that highlight objects

ami

Item of the "Affirm Referents" Type

Touch objects as mentioned



Same Schema also works for

"Arrange the assemblies in implicit order of completion"

or

"Write the equation for this line"

or

"Draw 2 circles – one that intersects the given figure at x=4 and one tangent to the figure at y=2"

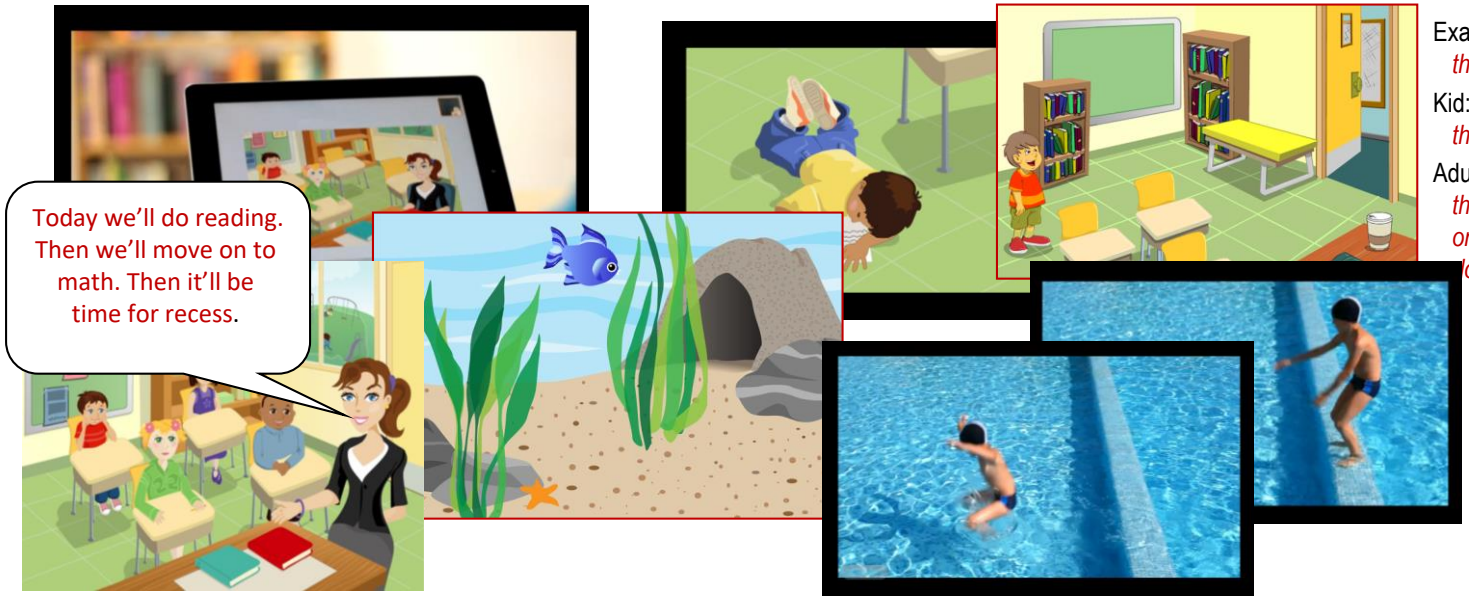
ami

Early Item Types

Narrate action

Follow spoken or written instructions

Re-tell passages



Examiner: *Touch the things they talk about.*

Kid: *You want me to close the door now?*

Adult Female: *Yes, please, then just sit back down on the bench there and don't get up again.*

Presentation and Response Modes

Tablet presents:

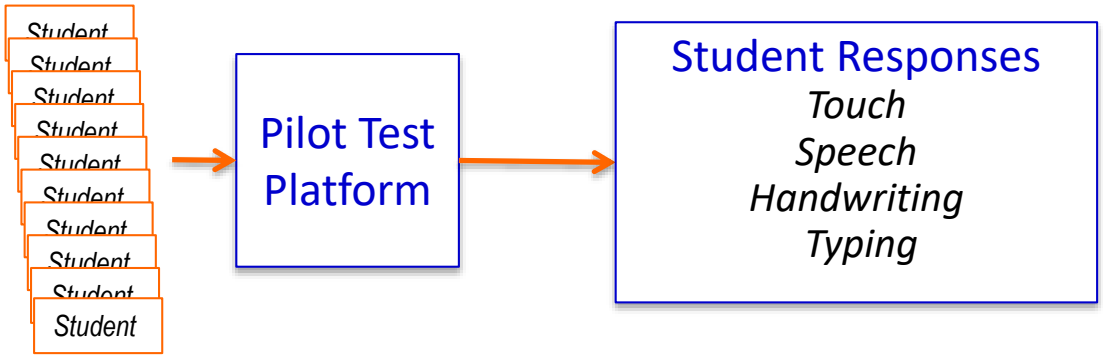
- Speech
- Drawing
- Figure sequence
- Text
- Video
- Animation

Student responds:

- Speech
- Touch
- Drawing
- Typing
- Handwriting
- Gesture

ami

Task Compliance in Pilot Testing



784 students produced 28,000 responses

activities are modeled by a single short video example (8-15 sec.)

In this sample, by age 8 years, children respond meaningfully to almost all these items about **95%** of the time, regardless of first language.

ami

HUMAN SCORING

Human raters score recordings wrt Standards

Machine scoring of content and fluency trained to match

Tools for judging dynamic graph

Delimit what's correct

How to score a fish path or track-t

For speak and listen, we had full c

For move, draw, track-touch, we r

Video example

ami

TTELL SUMMARY

Working prototype of TTELL system (2013)

Pilot Results

- Implements “Next-Generation” activities
- Engages low-SES English learners
- Enables self administration by young students
- Automatic presentation & scoring can yield the data needed for assessment to standards

ami

MOBY.READ

K-5 Early Reading Assessment

- Oral Reading Fluency (WCPM, Expression)
- Reading Level (Comprehension, Accuracy)

Features

- Self-administered
- On-Device Scoring & Reporting
- iOS or Chrome (HTML5)

ami

PREQUEL: 2004 NAAL

Situation

Measure reading fluency of 18,000 adults at home

Requirement

Instrument demonstrably accurate and fair

Method

Compare traditional vs. machine scoring

Results

Both human and machine scores: no detectable bias

Conclusion

Use appropriate validations of machine performance

ami

2004 NAAL PREQUEL

National Assessment of Adult Literacy (NAAL)

- Fluency Addition to NAAL (FAN)

Representative sample of 18,000 U.S. residents

- Test of oral reading fluency administered 1-on-1

Politically sensitive survey of skill distribution

- e.g. headlines: "30% of U.S. adults can't read"

Sample too large for human scoring

Machine scoring must be

- **accurate**
- **free of bias**

ami

Traditional Reading Fluency Method

Mark reading errors

Count the number of words read correctly in one minute (stop watch)

Report WCPM as the parameter of reading fluency

Pedro had just moved from Mexico when
1 2 3 4 5 6 7

he saw an accident. A little boy had
8 9 10 a 11 12 13 14 15

fallen into an open manhole, and now his
16 17 18 19 20 21 22 23

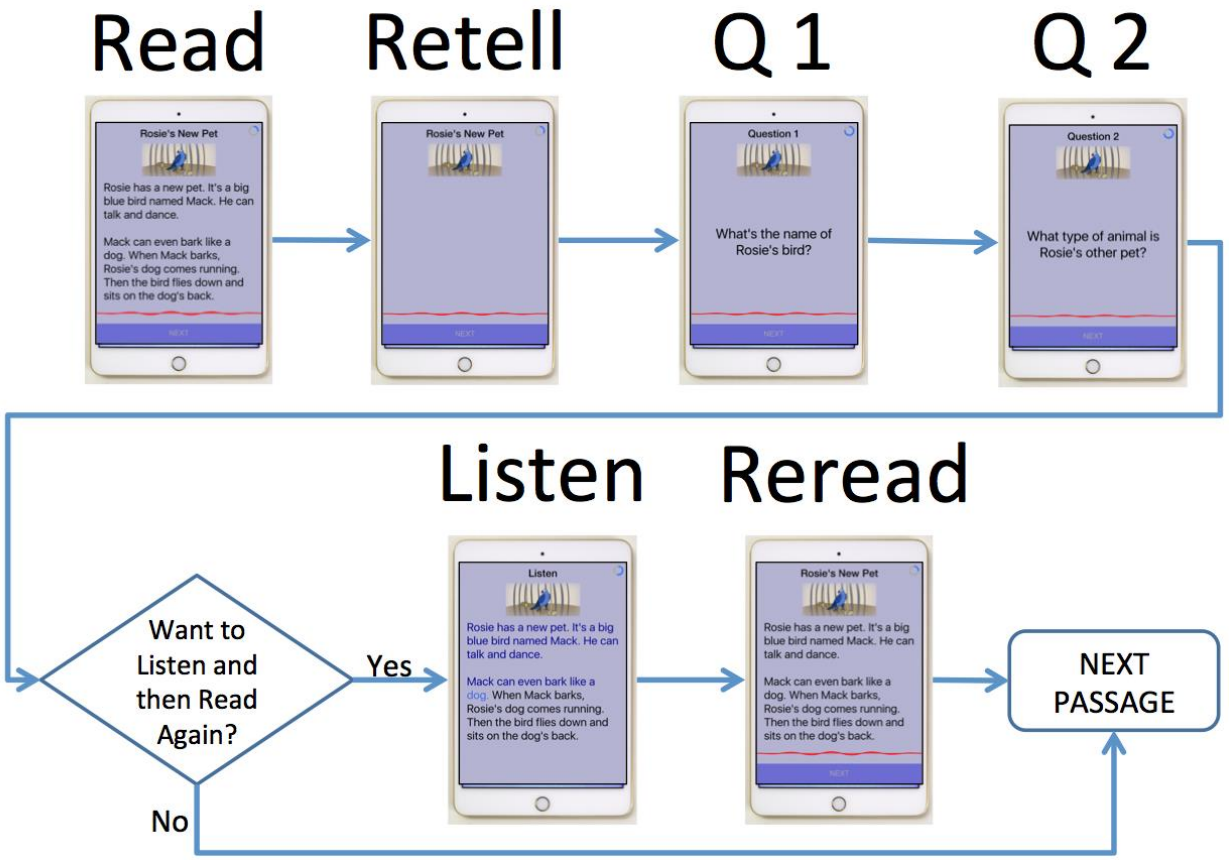
leg was caught between two pipes.
24 25 26 27 28 29 that

Pedro was just 10. He didn't think he
30 31 32 34 35 36 37 38

could rescue the boy alone.
39 40 41 42 43

ami

MOBY.READ PROTOTYPE (STUDENT)

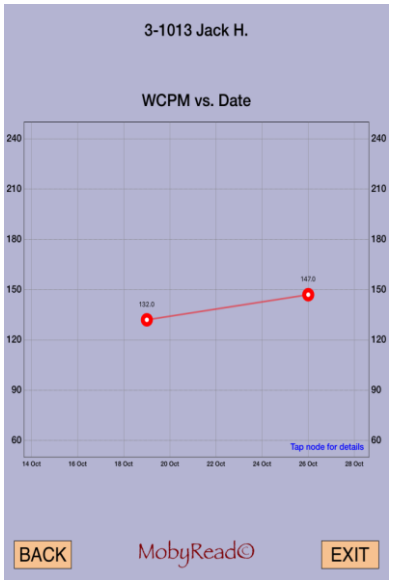


MOBY.READ PROTOTYPE (TEACHER)

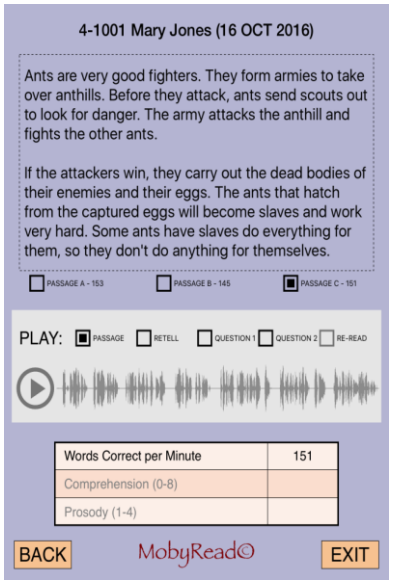
Teacher Interfaces

Class Roster

Progress Graph



Audio & Scores



Reports Words Correct Per Minute (WCPM), reading comprehension, and expression.

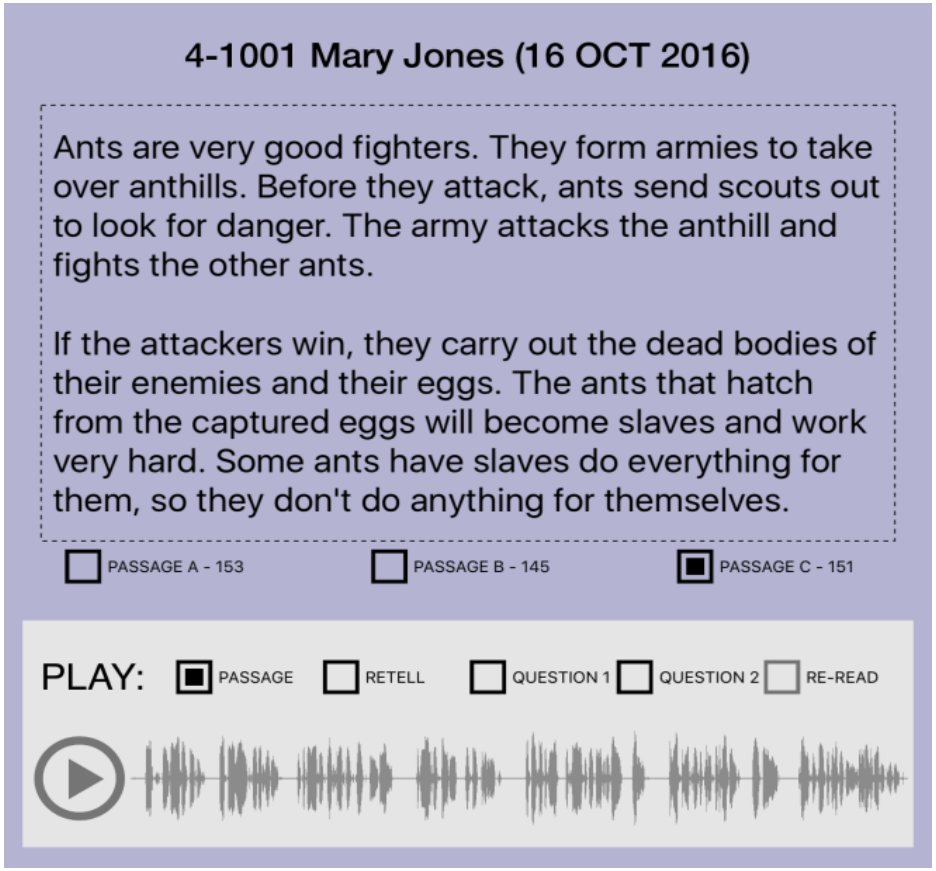
Human-Machine Method

Two human raters tallied reading errors from each recording (n=297).

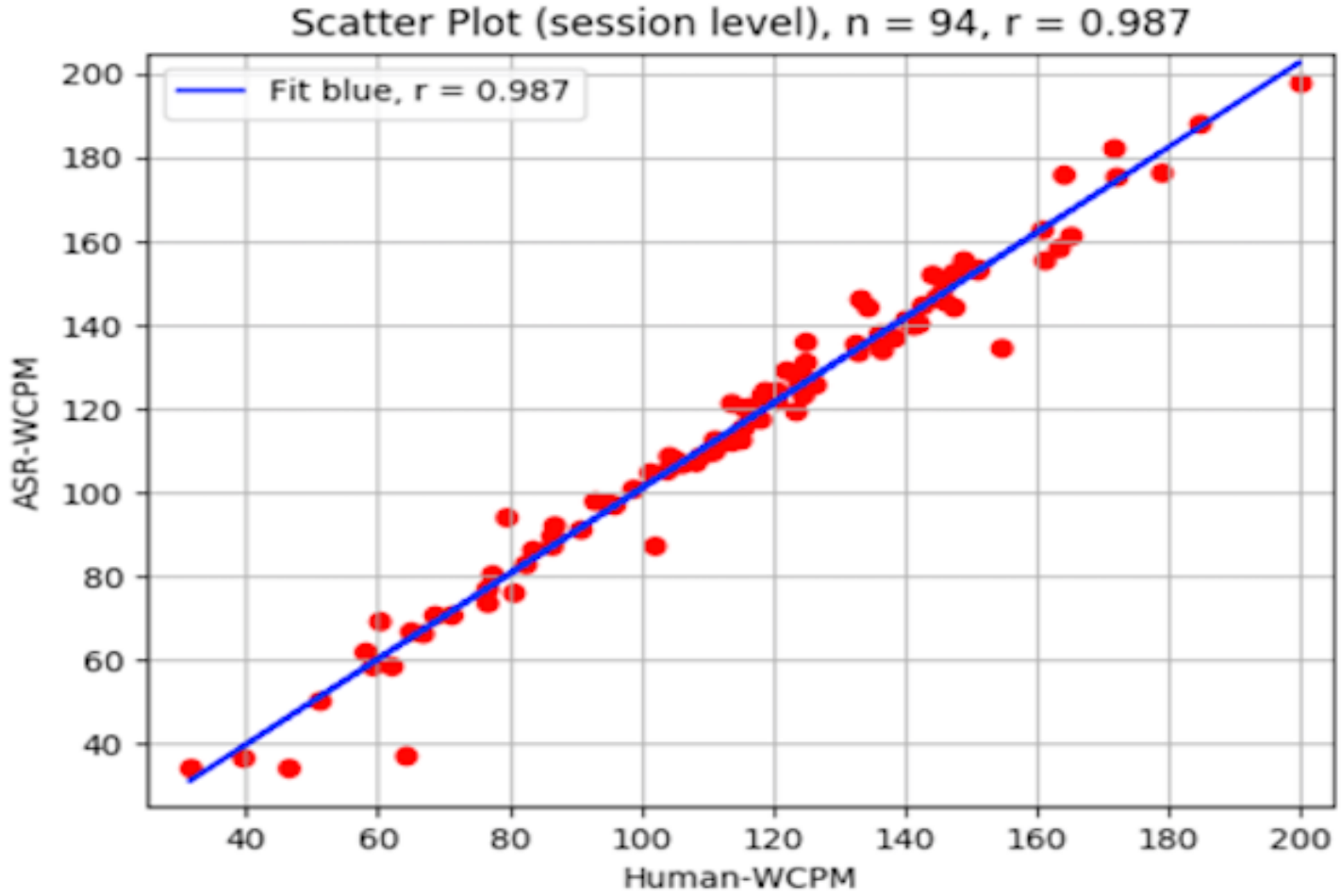
Human raters measured timing.

Output: WCPM

Inter-rater reliability = 0.99



Session-level scatter of median WCPM: server-based scores vs. Human scores



ami

Task Accuracy

Are Moby.Read scores similar to human+paper ORF scores?

(2nd Validation type)

Yes. (r = 0.88) The correlation between Moby.Read scores and DIBELS NEXT scores was 0.88. Published studies of DIBELS report a test-retest reliability of 0.82 and an inter-rater reliability of 0.85.

ami

MOBY.READ OUTCOMES

Alpha & Beta outcomes:

Students prefer Moby.Read self-administration.

95% of students self-administered successfully without any help;

Moby.Read rate & accuracy scores match double human scores;

Moby.Read scores match DIBELS scores at limit of DIBELS reliability.

New Scales and Analytics

Passage-fluid Vertical Scale: students levelled cross-grade.

Difficulty clusters to guide teaching. (Reinvent the running record.)

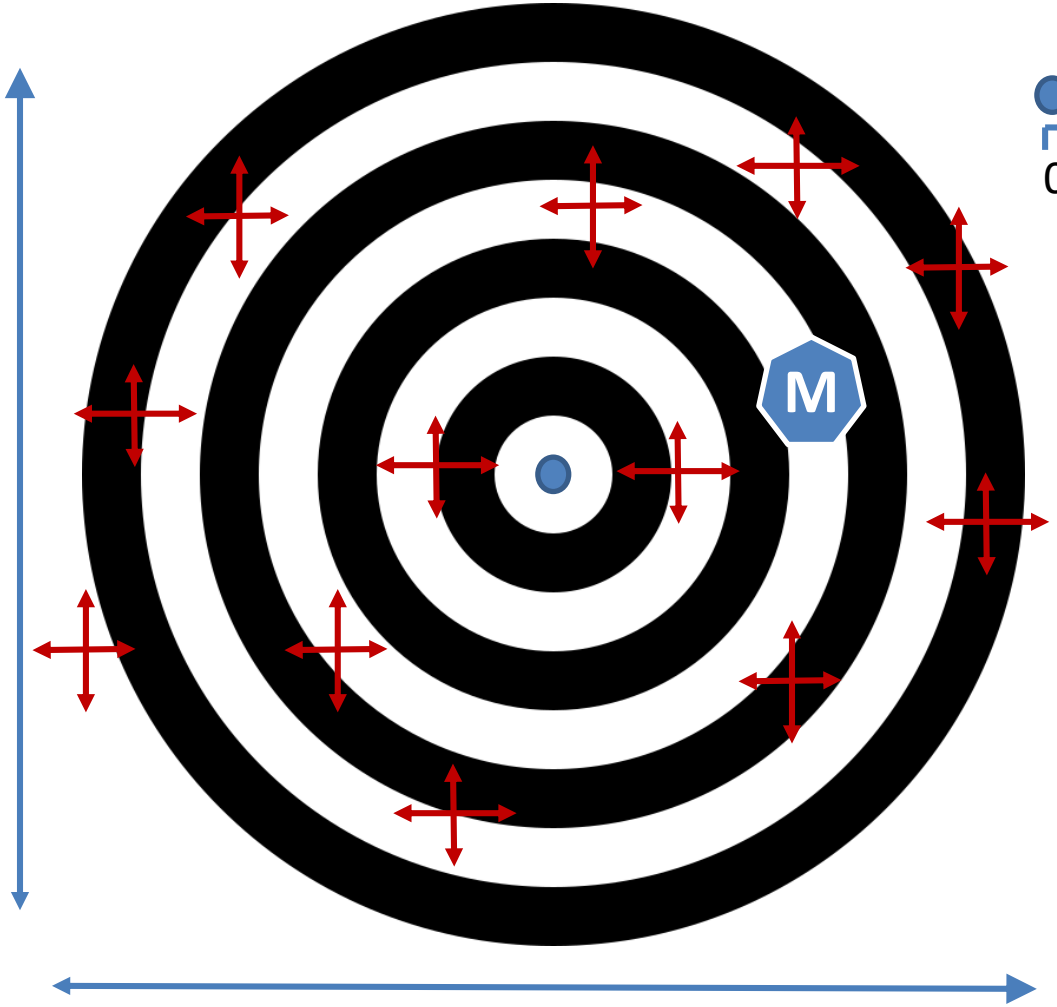
ami

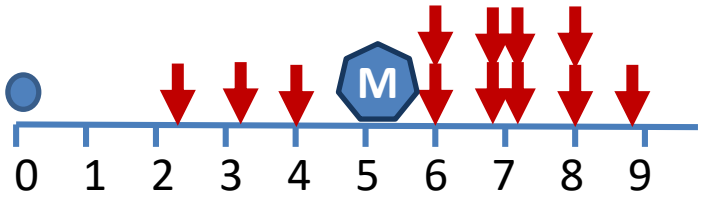
EVALATION PUZZLE: MACHINE > HUMAN

If a machine is trained to match human scores,

can the machine scores be more accurate than human scores?

Machine Compared to Human





ami

OBSERVATIONS

Kids (or adults) handle new items with little instruction.

Authenticity: performance items come under control.

Participant testing times can be reduced.

Many noisy measures combined for high reliability.

Response timing contributes to scores.

New Items Types: Integrate tasks and isolate skills.

Machine scoring can refine construct definitions.

ami

Thanks.

Questions?

Funders/Partners

National Center for Education Statistics (NCES): Oral reading fluency for NAEP
Institute of Education Sciences (IES): Moby.Read app: grant funding from IES
Educational Testing Service (ETS): Diagnostic measures of reading progress
Research Council of Norway: Remote Mental Status Monitor
Pearson: TTELL

ami