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, ...
SYSTEM WITH SPOKEN RESPONSE SCORING

Computer-Based Assessment with Spoken Responses

Scoring of Spoken Responses

Speech recognition

Speech recognition

Content scoring

Quality scoring

Score logic

Score Report

Student

graphics

audio

speech

word string
time & spectra
SCORING OPTIMIZATION

Automatic Scoring Development

OPTIMIZE

CHECK NEW SAMPLE

Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student

Pilot Test Platform

Human Scoring

Computer Scoring

diff

Update

Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student

r

Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student
Student

Pilot Test Platform

Human Scoring

Computer Scoring
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

- Test Form
- Administration
- Scoring
- Reporting
- Decision Demand
- Score User
- Test Taker
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
**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?
In the long run, automatic spoken language interpretation may obviate L2 testing.

6 mo. FSI
9 quarters (Univ.)
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.
CUMULATIVE COST PROFILE

Constructed Response Items, Performance Items

Initial cost

Year 1

Year 8

Human admin & Scoring

automated

Total cost

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
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
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
SST: 60-Item Sequence

Read Text
Repeat Sentence
Opposite Word
Answer Short Q
Build Sentence
Open Q
Retell Story

17 minutes
SST Machine Scoring Logic

Pronunciation  Fluency  Sentence Mastery  Vocabulary

Read  Repeat Sentence  Opposite  Ans. Short Question  Build S  OQ  St R

SST = (30% Sent.M, 20% Vocab, 30% Fluency, 20% Pron)
VERSANT SPANISH DEVELOPMENT AND VALIDATION

29,000 human scores

Criteria → Native Judges

Native Scribes → Ordinate DB, Items, Test Form, Responses

Native Judges → Scale Estimates

Ordinate DB, Items, Test Form, Responses → SST Scores

Validation

1st

2nd

52,000 transcripts

Content

Recorded Items

Native Test Developers

Test Spec

52,000 transcripts

Spanish Natives

Spanish Learners

OPI Scores

Concurrent ILR & ACTFL Interviews

52,000 transcripts

Native Test Developers

Test Spec

Criteria
PHONEME & WORD ALIGNMENT

75-90 Words/Min
5.8 Phones/Sec

waveform
spectrum
words
segmentation
I don’t know thirty minutes late was Bill

Simplified Response Network

an
half
hour

thirty
minutes

SIL

FIL

ami
1st MACHINE-HUMAN COMPARISON

Human scoring compared to machine-scoring (2003)

correlation = 0.94
N = 288
**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
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

r = 0.94

r = 0.92

r = 0.92

2. Two Rater Pairs
Different Material

3. Machine ~ Two Raters
Different Material
2\textsuperscript{ND} VALIDATION $\rightarrow$ PERFORMANCE PUZZLE

COMMUNICATIVE COMPETENCE$^*$

- Organization
- Pragmatics

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

LANGUAGE FACILITY

- Grammar
- Text
- Illocution
- Socioling.

~80% of variance

VERSANT tests contain sufficient material for equivalent rating
Automatic scoring matches test-retest performance of criterion instruments
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
## Assessment Design Space

<table>
<thead>
<tr>
<th>Scoring Focus</th>
<th>Presentation</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spoken</td>
<td>Written</td>
</tr>
<tr>
<td>Declarative Knowledge</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Language Skills</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
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)
Draw a path for the cat to go up the tree.
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
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
No teacher assistance from here on.
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
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”
EARLY ITEM TYPES

Narrate action
Follow spoken or written instructions
Re-tell passages

Today we’ll do reading. Then we’ll move on to math. Then it’ll be time for recess.

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
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.
**HUMAN SCORING**

Human raters score recordings wrt Standards

Machine scoring of content and fluency trained to match

Tools for judging dynamic graphics

Delimit what’s correct

How to score a fish path or track-touch?

For speak and listen, we had full development tools

For move, draw, track-touch, we needed to build tools

Video example

---

**ELL Path rating**

Goal: Judge if the path drawn is correct according to the given instruction.

- If the path is correct, give score 2.
- If the path is wrong or there is no video, give score 0.
- If the path is not clearly right or wrong, give score 1.

Candidate Registration ID: 70132114 Call: 4688860 Response: 154851360 Group.Ansitem: 2059

Prompt:

Show how the child walks from the playground to the bench by the street.

Video: If you see no path, hit skip

Grade:

☐ 0  ☐ 1  ☐ 2

Skip
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
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)
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
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
### 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.

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pedro had just moved from Mexico when he saw an accident. A little boy had fallen into an open manhole, and now his leg was caught between two pipes. Pedro was just 10. He didn’t think he could rescue the boy alone.**
MOBY.READ PROTOTYPE (STUDENT)

Read

Retell

Q 1

Q 2

Listen

Reread

Want to Listen and then Read Again?

Yes

Yes

No

NEXT PASSAGE
**Moby.Read Prototype (Teacher)**

Teacher Interfaces

- **Class Roster**
- **Progress Graph**
- **Audio & Scores**

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

Scatter Plot (session level), n = 94, r = 0.987

Fit blue, r = 0.987
Are Moby.Read scores similar to human+paper ORF scores? (2\textsuperscript{nd} 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.
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.)
If a machine is trained to match human scores, can the machine scores be more accurate than human scores?
MACHINE COMPARED TO HUMAN
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.
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