Using an Integrated Approach to Improve the Development and Scoring of Technology-based Speaking Tests

Marlene Johnston (Arizona Department of Education)
Gabriela Finn (Arizona Department of Education)
Jonathan Wolfe (Pearson)
Masa Suzuki (Pearson)
Hiro Fukuhara, Ph.D. (Pearson)

28 June 2017
Introduction
AZELLA

Arizona English Language Learner Assessment

- A standards-based English language proficiency assessment that meets both state and federal requirements
- Placement Test (year-round, July - June, appx. 25K)
  - Administered to students identified as L2 learners on Home Language Survey
- Reassessment Test (annual summative, February-March, appx. 75K)
  - Students who have been placed into an ELL program take the reassessment once a year until they achieve proficiency
  - Reassessment scores used for accountability measures
Arizona English Language Learner Assessment

- Four domains (Listening, Speaking, Reading, Writing)
  - Listening, Reading, Writing: paper-based
  - Speaking: delivered on 1-on-1 by an administrator on a speaker phone and scored automatically by AI scoring

- Five stages
  - Stage I (Kindergarten)
  - Stage II (grade 1-2)
  - Stage III (grade 3-5)
  - Stage IV (grade 6-8)
  - Stage V (grade 9-12)
Project Partnership

- Project Sponsor
- Owner of the Test
- Item development

- Item production
- Booklet design

- Human scoring service
- Automated scoring service
- Psychometric service
- Test Printing
- Test Administration
Motivation for the Project

Before 2012-2013

- Paper/Pencil Test
- Test administered inconsistently
- Hand scored by Test Administrator
- Test scored inconsistently
- Assessment of Speaking proficiency was not a priority

How could we achieve a speaking test that could be administered and scored **consistently** and **reliably** for accountability purposes with about 100K ELLs every year?
Solution

ADE and Pearson agreed in 2011 to initiate the project to develop an **automatically scored speaking test**

- **Development Work: 2011-2012**
  - 1st Field Test (November, 2011)
    - 1st-12th grade, n = 31,685
  - 2nd Field Test (April, 2012):
    - Kindergarten, n = 13,141

- **1st Operational Test: January 2013**
  - K-12, n = 140k
AZELLA Speaking administration

“Benji was a happy dog who liked to play in the garden”

10-12 items per test
12-15 min per test
## Item Types

<table>
<thead>
<tr>
<th>Item Types</th>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give short responses</td>
<td>Speaking</td>
</tr>
<tr>
<td>Questions about an image</td>
<td>Speaking</td>
</tr>
<tr>
<td>Questions about a statement</td>
<td>Speaking</td>
</tr>
<tr>
<td>Give directions from a map</td>
<td>Speaking</td>
</tr>
<tr>
<td>Give instructions to do something</td>
<td>Speaking</td>
</tr>
<tr>
<td>Detailed responses to a topic</td>
<td>Speaking</td>
</tr>
<tr>
<td>Open questions about a topic</td>
<td>Speaking</td>
</tr>
<tr>
<td>Repeat</td>
<td>Speaking</td>
</tr>
<tr>
<td>Similarities &amp; Differences</td>
<td>Speaking</td>
</tr>
<tr>
<td>Naming</td>
<td>Reading/literacy</td>
</tr>
<tr>
<td>Word list</td>
<td>Reading/literacy</td>
</tr>
</tbody>
</table>
Sample Items
Naming (Syllabification)

Look at the picture. This is a strawberry. Say the three syllables of the word strawberry.

Short Response

Look at the picture. The boy is hungry. What does he say to his teacher?

I’m hungry. Can I have a sandwich and a glass of water, please?
Question about Image

Look at the picture. What is happening in the picture? Remember to use a complete sentence.

**Directions**

*Look at the map. Please answer in two or more complete sentences. Use the path marked on the map to tell how to get to from school to the park.*

Chart Description

What can you tell us about the chart?

Follow along as I read the statement. "Walking is the best exercise." Please answer in complete sentences. Do you agree or disagree with the statement?

Explain why you agree or disagree that walking is the best exercise.
### Structural Reliability

<table>
<thead>
<tr>
<th>Stage</th>
<th>Human Cronbach $\alpha$</th>
<th>Machine Cronbach $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>II</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>III</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>IV</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>V</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.96</strong></td>
</tr>
</tbody>
</table>
## Machine scores to Human scores: Test-Level

<table>
<thead>
<tr>
<th>Stage</th>
<th>Human-Human Correlation r</th>
<th>Machine-Human Correlation r</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td>II</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>III</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>IV</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>V</td>
<td>0.98</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.96</strong></td>
<td><strong>0.92</strong></td>
</tr>
</tbody>
</table>
Scatterplot by stage

**Stage II**

**Stage III**

**Stage IV**

**Stage V**
Accomplishments before 2016-2017

- Successfully delivered **2 field tests** and **3 operational tests**:
  - Nov. 2011 - 1<sup>st</sup>-12<sup>th</sup> grade    \( n = 31K \) (Field Test)
  - Apr. 2012 - Kindergarten            \( n = 13K \) (Field Test)
  - Jan. 2013 - K-12                   \( n = 140k \)
  - Feb. 2014 - K-12                   \( n = 90k \)
  - Feb. 2015 - K-12                   \( n = 85k \)

- Trained and optimized automated scoring engines

- Built strong validation evidence for scoring young population

- Received positive feedback from the field
Not all stories are good.....
Lessons Learned from 2011-2015

- Test design and item types
- Development schedule
- Modification of rating rubrics after the system training
  - Lack of understanding on the impact of scoring criteria in the context of automated scoring
- Initial scoring process was atypical and caused difficulty
  - Draft rubrics
  - Lack of familiarity with items
  - No clear idea of how a student might respond
  - Items were difficult to evaluate
Development Project
2016-2017
2016-2017 Project Objectives

Goals:

● Keep the core test structure and items, but partially refresh - current tests in use for 4 years
● Review item statistics from past years and replace mediocre items for better reliability and discrimination
● Include new and more engaging item types
● Enhance the coverage of AZ’s ELP standards
● Prepare for the move to online testing
Development Process

- Review human rating insight
- Review psychometric data
- Insight from machine scoring
- Review AZ ELP coverage
- New item types
- New items for existing item types
- Retry old items for existing item types
- Content Development
- Field Testing
- Transcription
- Human rating
- Rubric dev
- Range-finding
- Automated Scoring Development
- Pyschometric analysis & Validation analysis

ELs Non-ELs
Psychometric Data Review

Quantitative Analysis: Review Psychometric Data

- Classical/IRT statistics
  - P-value
  - Point-biserial correlation
  - Score distribution
  - Item fit
- Percentage of exact and adjacent agreement
  - between human ratings
  - between automated score engine scores and human rating
- Interrater reliability coefficients
Identifying candidates for replacements

Qualitative Analysis: Review Student Performance

- Students don’t know what to say
- Students don’t speak in complete sentences
- Students don’t use academic language
New Item Types
New Item Types

- Further develop item types with demonstrated performance strengths
- Add Listen and Retell - successfully used with other tests and performing well
- Add Oral Reading Fluency - an important reading component that could now be included in a standardized, summative assessment
Listen and Retell

- Retelling is an important academic language function
- Covers an AZ ELP standard
- Provides substantial contextual info for richer language production
- Allow a wider proficiency range
- Present the passage twice; about language production, not listening comprehension
- Decision to score descriptions of the pictures as valid responses
- Allows broader range of topics from social to academic domains
Oral Reading Fluency (ORF)

- ORF is included in the English Language Proficiency Standards but previously not tested
- Arizona EL and ELA practitioners consider ORF an important skill
- Scoring data confirms ORF is a good item type
- ORF rubric development was a collaboration with K-12 Standards and OELAS
- Previous ORF studies were helpful for setting criteria and ranges
Performance Scoring
Performance Scoring
How do we ensure scoring accuracy?

Resources

- **Staffing**
  - 1 Senior Content Specialist
  - 6 Scoring Directors
  - ~200 Expert scorers

- **Items and number of responses**
  - 81 unique items
  - ~95,000 scores to assign

- **Time and Schedule**
  - ~25,000 hours
  - 12 week schedule
Performance Scoring

Processes

- **Item Review**
  - Familiarize yourself with the tasks and/or prompts

- **Rubric Review**
  - Understand standards for scoring and expectations of students

- **Pre-Rangefinding/Rangefinding**
  - Review responses to understand how students respond
  - Meeting(s) to address scoring criteria and how to score responses accurately

- **Training Development**
  - The process by which scorers/raters are introduced to the content and expectations
  - Construct training materials based on approved responses and criteria

- **Training and Scoring**
  - Qualifying scorers/raters to apply standards
  - Applying scores to student work
Performance Scoring

Quality Monitoring

● Reporting
  – IRR/Agreement
  – Validity
  – Frequency Distribution

● Backreading
  – Content support staff reading behind scorers/raters to check quality

● Review
  – A repository for those responses that need clarification to be scored accurately

● Item Evaluations
  – Written feedback about the overall success of a unique item
Some thoughts on future
Some thoughts on future directions

1. Adaptive Testing with Automated Scoring
   – for more efficient and student-friendly test-taking experience

   **Multi-stage adaptive:**
   An initial set of MC items, then present an easy form or a difficult form of MC items (without an immediate feedback by automated scoring, OE items are not a part of adaptation)
   An interactive assessment where each module (testlet) is composed of Listening, Writing, Reading, Speaking (with an immediate feedback by automated scoring)

   **Fully adaptive:** Each response will get scored, and the next item will be determined based on the ability estimate at the point

1. Multi-dimensionality
   – e.g, Listen and Retell contributing both listening and speaking scores
Summary
Considerations in Test Development with technology

- Technology integration is most successful when both human processes and technological capabilities (and limitations) are integrated throughout all stages of development
  - Balancing construct definition, standards coverage and machine’s capability
  - Item development should involve automated scoring experts to increase the likelihood of scoreability
Considerations in Test Development with technology

● Ensure reliable human rating with rigorous training. Quality human rating is a foundation for quality scoring system
  • Formalized process for scoring/rating
  • Lead time for item, prompt, and rubric review

● Ensure field testing yields good data
  • Usability testing / Cog labs
  • Quality hardware
  • Testing rooms / seating arrangement
  • ‘How to take the test” prep material
Stakeholder Priorities
A Balancing Act and an Opportunity

- Technology Experts – Better ways and means to test
- Psychometricians – Optimize item and test statistics
- Content Experts – Adhere to the construct – broad, relevant sampling
- Test Developers – Apply Standards and deliver intended functionality
- Test Administrators – Testing logistics
- Students – Fair opportunity to demonstrate their skills
- Federal and State regulations – Meet testing requirements
- Scorers - Unambiguous items that score reliably
- Users – Useful and meaningful score reporting
- Clients – Managing cost, time, and other resource limitations
The End - Questions?