INSIDER PANEL

Beverly Nedrow, Senior ELA/ELL Content Specialist, WestEd

Charles Bruen, PhD, Director of Data Analysis, Budget, and Technology, Assessment Section, Arizona Department of Education

Jane Scott, ELD Coordinator, Madison School District

Alistair Van Moere, PhD, VP Product & Test Development, Knowledge Technologies, Pearson

Moderator
Marlene Johnston, Director of English Language Learner Assessments, Arizona Department of Education,
GOALS OF THE REVISION

• Multiple versions of AZELLA.
• Correctly identify ELL students.
• Ensure validity evidence for test development.
• Valid and useful subtest information.
• Useful reporting of the language strand.
• Place Kindergarten students appropriately, quickly, and with sufficient information to guide instruction.
• FEP students are efficiently and effectively assessed
• AZELLA data useful for evaluating program effectiveness.
BALANCING MULTIPLE PRIORITIES

• Federal Laws
• State Laws
• Test Development Standards
• Contract Requirements
• Project Management
• Timelines
• Staffing Needs
• Quality Assurance
• Item Development Needs
• Psychometric Needs

• Scoring Requirements
• IT needs
• Technology Needs
• LEA needs
• Educator Needs
• Student Needs
• Reporting Requirements
• Fiscal Pressures
• Test Administration Changes
• OCR

AZELLA
Beverly Nedrow
Beverly Nedrow is a Senior ELA/ELL Content Specialist with WestEd. She received her M.S. in Curriculum and Instruction from Texas A&M at Corpus Christi, with specializations in English as a Second Language and Reading.

Beverly is the WestEd Content Lead on the AZELLA program. She has taught English Language Learns from the elementary through college levels, and has nearly 25 years experience in developing English Language Arts and English Language Learner assessments.
BLUEPRINT

- Map of the assessment to determine:
  - The set of assessable standards/indicators appropriate for a statewide assessment
  - Degree of emphasis for each domain
  - Item types to be used
ITEM BANK

- Developed sufficient items to produce two operational forms
- Item overage banked for potential use as:
  - Release items
  - Items for refreshing the test
  - Future field or operational test items
PASSAGE DEVELOPMENT

- Genres
  - From the ELP standards
  - Aligns to the Common Core Standards
- Lexiles
  - Range based on Common Core Standards
  - Range covers grades and proficiency levels of each stage
- Reviewed by Arizona teachers and revised based on their feedback
ITEM DEVELOPMENT

- Item Specifications
  - Developed by WestEd experts
  - Reviewed by Advisory Committee
  - Revised based on committee feedback
- Item Writing Workshop
  - Facilitated by WestEd content staff
  - Items written by Arizona teachers
- Item Review and Revision by WestEd content experts
- Content and Bias Review by Arizona teacher committees
CONTENT & BIAS COMMITTEE

- Arizona teachers and Educators
- Bias and Sensitivity Training
  - Provided to item writing committee
  - Provided to teacher review committee
- Reviewed passages, items, and graphics
Charles Bruen, PhD
Charles Bruen, PhD is the Director of Data Analysis, Budget, and Technology in the Assessment Section of the Arizona Department of Education. He received his doctorate in Mathematics from Columbia University.

Charles has 34 years of teaching experience and has been supporting assessments for 12 years at the Arizona Department of Education.
PSYCHOMETRICS

The focus should be:

1. Choosing items for the test
2. AZELLA alignment to the English Language Proficiency Standards
3. The Standard Setting
CREATING A NEW ASSESSMENT

There are several steps necessary to bring life to a new assessment:

• Start with an idea of what to test and create a blueprint. The blueprint identifies the most important content from the standards.
• Tests are composed of items; next, create item specifications.
• Write the items to match the blueprint and item specs.
• Edit the items and pass them through a content, bias, and sensitivity committee.
• Field test the items, either stand-alone or embedded.
• Gather the data on each item and evaluate the effectiveness.
WHERE DID THE DATA COME FROM?
FREQUENCY OF SCHOOLS IN THE STUDY

<table>
<thead>
<tr>
<th>School</th>
<th>ELL</th>
<th>Non-ELL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10220101</td>
<td>34</td>
<td>25</td>
<td>59</td>
</tr>
<tr>
<td>30201112</td>
<td>12</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>3201114</td>
<td>21</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>30201118</td>
<td>8</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>30201121</td>
<td>6</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>30201122</td>
<td>2</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>70402101</td>
<td>0</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>70403122</td>
<td>31</td>
<td>49</td>
<td>80</td>
</tr>
<tr>
<td>70479101</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>70479104</td>
<td>29</td>
<td>10</td>
<td>39</td>
</tr>
<tr>
<td>70479109</td>
<td>24</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What other data is used to evaluate the items on a test?

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Total (N=1040)</th>
<th>ELL Students</th>
<th>Non-ELL Students (N = 481)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>01.Listening</td>
<td>488.25</td>
<td>54.49</td>
<td>0.71</td>
</tr>
<tr>
<td>03.Reading</td>
<td>475.66</td>
<td>53.78</td>
<td>0.73</td>
</tr>
<tr>
<td>04.Prewriting</td>
<td>429.71</td>
<td>73.02</td>
<td>0.90</td>
</tr>
<tr>
<td>05.Speaking</td>
<td>499.53</td>
<td>61.85</td>
<td>0.92</td>
</tr>
<tr>
<td>10.Comprehension</td>
<td>480.58</td>
<td>47.41</td>
<td>0.81</td>
</tr>
<tr>
<td>11.Oral</td>
<td>495.47</td>
<td>51.72</td>
<td>0.91</td>
</tr>
<tr>
<td>13.Total</td>
<td>476.78</td>
<td>43.95</td>
<td>0.93</td>
</tr>
</tbody>
</table>
EVALUATING AN ITEM

The Percent of Students at Performance Levels based on Total by ELL Category

<table>
<thead>
<tr>
<th>PL Total</th>
<th>ELL Students</th>
<th>Non-ELL Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtest</td>
<td>PreEmergent</td>
<td>Emergent</td>
</tr>
<tr>
<td>List</td>
<td>5.48</td>
<td>6.92</td>
</tr>
<tr>
<td>Read</td>
<td>11.06</td>
<td>22.60</td>
</tr>
<tr>
<td>Write</td>
<td>52.21</td>
<td>15.77</td>
</tr>
<tr>
<td>Speak</td>
<td>18.27</td>
<td>10.87</td>
</tr>
<tr>
<td>Compreh</td>
<td>7.98</td>
<td>12.40</td>
</tr>
<tr>
<td>Oral</td>
<td>10.67</td>
<td>14.52</td>
</tr>
<tr>
<td>Total</td>
<td>19.90</td>
<td>15.38</td>
</tr>
</tbody>
</table>
# Evaluating Individual Items

## P-value for ELL Students, ELL Students by Performance Level by Total and Non-ELL Students

<table>
<thead>
<tr>
<th>Item</th>
<th>Max Possible</th>
<th>Total ELL (N=1040)</th>
<th>PL = 1 (N=207)</th>
<th>PL = 2 (N=160)</th>
<th>PL = 3 (N=601)</th>
<th>PL = 4 (N=61)</th>
<th>PL = 5 (N=9)</th>
<th>SE</th>
<th>Total Non-ELL (N=481)</th>
<th>NonELL-ELL</th>
<th>Non-PL5</th>
<th>Non-PL4</th>
<th>Non-PL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>i10</td>
<td>1</td>
<td>0.71</td>
<td>0.35</td>
<td>0.65</td>
<td>0.82</td>
<td>0.92</td>
<td>1.00</td>
<td>0.00</td>
<td>0.81</td>
<td>0.10</td>
<td>-0.19</td>
<td>-0.11</td>
<td>-0.01</td>
</tr>
<tr>
<td>i11</td>
<td>1</td>
<td>0.46</td>
<td>0.21</td>
<td>0.39</td>
<td>0.54</td>
<td>0.67</td>
<td>0.89</td>
<td>0.10</td>
<td>0.48</td>
<td>0.02</td>
<td>-0.40</td>
<td>-0.18</td>
<td>-0.05</td>
</tr>
<tr>
<td>i12</td>
<td>1</td>
<td>0.44</td>
<td>0.19</td>
<td>0.40</td>
<td>0.50</td>
<td>0.75</td>
<td>0.78</td>
<td>0.14</td>
<td>0.58</td>
<td>0.14</td>
<td>-0.20</td>
<td>-0.17</td>
<td>0.07</td>
</tr>
<tr>
<td>i13</td>
<td>1</td>
<td>0.58</td>
<td>0.29</td>
<td>0.53</td>
<td>0.66</td>
<td>0.86</td>
<td>1.00</td>
<td>0.00</td>
<td>0.67</td>
<td>0.09</td>
<td>-0.33</td>
<td>-0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>i14</td>
<td>1</td>
<td>0.58</td>
<td>0.33</td>
<td>0.59</td>
<td>0.63</td>
<td>0.87</td>
<td>0.89</td>
<td>0.10</td>
<td>0.72</td>
<td>0.15</td>
<td>-0.17</td>
<td>-0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>i15</td>
<td>1</td>
<td>0.56</td>
<td>0.34</td>
<td>0.56</td>
<td>0.61</td>
<td>0.76</td>
<td>0.89</td>
<td>0.10</td>
<td>0.59</td>
<td>0.03</td>
<td>-0.30</td>
<td>-0.17</td>
<td>-0.02</td>
</tr>
<tr>
<td>i16</td>
<td>1</td>
<td>0.33</td>
<td>0.15</td>
<td>0.29</td>
<td>0.37</td>
<td>0.56</td>
<td>0.89</td>
<td>0.10</td>
<td>0.41</td>
<td>0.08</td>
<td>-0.48</td>
<td>-0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>i17</td>
<td>1</td>
<td>0.51</td>
<td>0.21</td>
<td>0.46</td>
<td>0.60</td>
<td>0.81</td>
<td>0.89</td>
<td>0.10</td>
<td>0.69</td>
<td>0.18</td>
<td>-0.20</td>
<td>-0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>i18</td>
<td>1</td>
<td>0.23</td>
<td>0.08</td>
<td>0.16</td>
<td>0.23</td>
<td>0.73</td>
<td>1.00</td>
<td>0.00</td>
<td>0.31</td>
<td>0.08</td>
<td>-0.69</td>
<td>-0.42</td>
<td>0.08</td>
</tr>
<tr>
<td>i19</td>
<td>1</td>
<td>0.67</td>
<td>0.38</td>
<td>0.60</td>
<td>0.75</td>
<td>0.92</td>
<td>1.00</td>
<td>0.00</td>
<td>0.76</td>
<td>0.11</td>
<td>-0.22</td>
<td>-0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>i20</td>
<td>1</td>
<td>0.68</td>
<td>0.27</td>
<td>0.64</td>
<td>0.80</td>
<td>0.84</td>
<td>1.00</td>
<td>0.00</td>
<td>0.82</td>
<td>0.15</td>
<td>-0.18</td>
<td>-0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>i21</td>
<td>1</td>
<td>0.39</td>
<td>0.16</td>
<td>0.29</td>
<td>0.43</td>
<td>0.89</td>
<td>1.00</td>
<td>0.00</td>
<td>0.56</td>
<td>0.17</td>
<td>-0.44</td>
<td>-0.33</td>
<td>0.13</td>
</tr>
<tr>
<td>i22</td>
<td>1</td>
<td>0.28</td>
<td>0.06</td>
<td>0.19</td>
<td>0.31</td>
<td>0.79</td>
<td>1.00</td>
<td>0.00</td>
<td>0.40</td>
<td>0.13</td>
<td>-0.60</td>
<td>-0.39</td>
<td>0.10</td>
</tr>
<tr>
<td>i23</td>
<td>1</td>
<td>0.33</td>
<td>0.10</td>
<td>0.23</td>
<td>0.37</td>
<td>0.87</td>
<td>1.00</td>
<td>0.00</td>
<td>0.41</td>
<td>0.08</td>
<td>-0.59</td>
<td>-0.46</td>
<td>0.04</td>
</tr>
<tr>
<td>i24</td>
<td>1</td>
<td>0.33</td>
<td>0.10</td>
<td>0.25</td>
<td>0.36</td>
<td>0.84</td>
<td>1.00</td>
<td>0.00</td>
<td>0.43</td>
<td>0.11</td>
<td>-0.57</td>
<td>-0.41</td>
<td>0.07</td>
</tr>
<tr>
<td>i25</td>
<td>2</td>
<td>0.91</td>
<td>0.76</td>
<td>0.88</td>
<td>0.97</td>
<td>0.99</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.05</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>i26</td>
<td>2</td>
<td>0.83</td>
<td>0.61</td>
<td>0.77</td>
<td>0.91</td>
<td>0.97</td>
<td>1.00</td>
<td>0.00</td>
<td>0.91</td>
<td>0.08</td>
<td>-0.09</td>
<td>-0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>
EVALUATING AN ITEM

**Arizona Fall 2011**  
Grade: KPT

<table>
<thead>
<tr>
<th>AZID: 7612P106</th>
<th>Domain: Listening</th>
<th>N-Count: 350</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Type: OR1</td>
<td>Form: R</td>
<td>item #: 6</td>
</tr>
<tr>
<td>Std: 1</td>
<td>Prof Level: B</td>
<td>Indicator: 6</td>
</tr>
</tbody>
</table>

**Passage:**
The student will demonstrate understanding of oral communications by: responding to comments and questions in social conversations by sharing one’s experiences and expressing one’s thoughts.

**Item Statistics**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Value</td>
<td>0.80</td>
</tr>
<tr>
<td>Item Mean</td>
<td>0.80</td>
</tr>
<tr>
<td>PtBis</td>
<td>0.63</td>
</tr>
<tr>
<td>Non-ELL P-Value</td>
<td>0.97</td>
</tr>
<tr>
<td>Non-ELL Item Mean</td>
<td>0.97</td>
</tr>
<tr>
<td>Non-ELL PtBis</td>
<td>0.42</td>
</tr>
</tbody>
</table>

**Rasch**

<table>
<thead>
<tr>
<th>Level</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch</td>
<td>-0.02</td>
</tr>
<tr>
<td>Infit</td>
<td>0.80</td>
</tr>
<tr>
<td>Outfit</td>
<td>0.58*</td>
</tr>
</tbody>
</table>

**Bias: MHzb-SMD-Delta (Gender, Race, Ethnicity, SES)**

<table>
<thead>
<tr>
<th>Bias</th>
<th>M-F</th>
<th>NH-H</th>
<th>NAI-AI</th>
<th>N-FRL</th>
<th>NEL-ELL</th>
<th>FEP-ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHzb</td>
<td>0.001</td>
<td>0.152</td>
<td></td>
<td></td>
<td>2.013</td>
<td></td>
</tr>
<tr>
<td>Bias Catg Flag</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMD</td>
<td>0.005</td>
<td>-0.048</td>
<td></td>
<td></td>
<td>-0.254</td>
<td></td>
</tr>
<tr>
<td>Delta</td>
<td>0.027</td>
<td>-0.373</td>
<td></td>
<td></td>
<td>-1.789</td>
<td></td>
</tr>
</tbody>
</table>

**Bias: MHzb-SMD-Delta (based on Language)**

<table>
<thead>
<tr>
<th>Bias</th>
<th>NSpE-SpE</th>
<th>NS-S</th>
<th>E-S</th>
<th>E-NS</th>
<th>E-AI</th>
</tr>
</thead>
</table>

**Population (GRES)**

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>100</td>
</tr>
<tr>
<td>Female</td>
<td>194</td>
</tr>
<tr>
<td>Hispanic</td>
<td>296</td>
</tr>
<tr>
<td>Native American</td>
<td></td>
</tr>
<tr>
<td>Free/Reduced Lunch</td>
<td></td>
</tr>
<tr>
<td>English Lang Learner</td>
<td>350</td>
</tr>
</tbody>
</table>

**Population (Language)**

<table>
<thead>
<tr>
<th>Language</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Special Ed</td>
<td></td>
</tr>
<tr>
<td>Special Ed</td>
<td></td>
</tr>
<tr>
<td>Non-Spanish</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>144</td>
</tr>
<tr>
<td>Native American</td>
<td></td>
</tr>
</tbody>
</table>
ALIGNMENT TO STANDARDS
## WHAT ALIGNMENT IS GOOD ENOUGH?

<table>
<thead>
<tr>
<th>Alignment Level</th>
<th>Categorical Concurrence</th>
<th>Depth of Knowledge</th>
<th>Range of Knowledge</th>
<th>Balance of Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable</td>
<td>6 items per standard</td>
<td>50%</td>
<td>50%</td>
<td>70%</td>
</tr>
<tr>
<td>Weak</td>
<td>---</td>
<td>40%-49%</td>
<td>40%-49%</td>
<td>60%-69%</td>
</tr>
<tr>
<td>Unacceptable</td>
<td>Less than 6 items per standard</td>
<td>Less than 40%</td>
<td>Less than 40%</td>
<td>Less than 60%</td>
</tr>
</tbody>
</table>
ALIGNMENT

- ELP Standards (Highly Instructionally Oriented)
  - V.S.

- Assessable Standards (Major Skills to be Proficient – Able to access content in a Non-ELL Classroom.)
Welcome to Web Alignment Tool

This tool is designed to produce reports on the alignment of curriculum standards and student assessments.

The process requires a group of reviewers first to assign depth-of-knowledge (DOK) levels to standards/objectives (Part I). Then reviewers are to code assessment items by identifying the depth-of-knowledge for each item and the corresponding standard/objective (Part II).

1. The steps in using this tool and the process include
2. Training on DOK levels for content area
3. Logging on
4. Selecting a state, content area, and grade
5. Individually coding DOK for each objective
6. Group reaching consensus on the DOK for each objective
7. Coding independently the DOK for each assessment item and corresponding objective(s)
8. Recording Source of Challenge and Notes
Reviewer Login

Username: [input field]
Group ID: [input field]

Group Leader Login

Login

You must register with your username and assigned Group ID to access this site.

If you have not registered please go to the Registration Page to register.

If you are a Group Leader and would like to register a new group, please go to the Group Registration Page to register your group.

Guest Login
Administration Area
Review Area

Directions:

Thank you for participating as a reviewer in our Web Alignment Tool. Before you start using our online reviewing area, we strongly encourage you to first visit our Tutorial page. There you can find a tutorial on how to contribute standardized data to our system.

Your input of data consists of two parts: Part I and Part II.

**Part I**

In Part I, you are to work with other reviewers as a *group* to fill out the DOK levels (depth of knowledge level) of the standards. You will have to select the appropriate standards and discuss with other reviewers to reach a consensus of your DOK levels. Each group can only submit one consensus of the DOK levels of the specified standard.

**Part II**

In Part II, you will be working *individually*. You must have hard copy(s) of assessment(s) on hand to proceed. For each assessment, enter in the assessment's information, e.g. Form, # of Items, etc. After you have provided those information for the assessment you are working with, fill in the DOK levels of each assessment item.
STRUCTURE OF THE AUTOMATED ALIGNMENT PROCESS

• Registration
  • Group Leader
  • Reviewers
• Domain/Standards/Performance Indicators
• Entry Process
• Training on Linguistic Difficulty Levels
• Phase I Consensus Process on Assigning LDL
• Phase II Coding of Assessment Tasks
• Phase III Analysis of Coding
• Phase IV Reporting
WHAT IS A STANDARD SETTING?

- An advisory process where judges make recommendations for performance level cuts
- A process of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons will be made (Cizek, 2006)
- Test scores can be used to group students into meaningful performance levels
- Standard setting is the process whereby we “draw the lines” that separate the test scores into various performance levels
SETTING THE STANDARD

- Borderline PLDs
- Items on a Test
- Content Standards
- Student Knowledge Skills Abilities

Cut scores that match students to their appropriate performance categories
STANDARD SETTING

Bookmark Standard Setting:

• Ordered Item Booklet
  • Overall
  • By Domain
• Place Bookmark at the Item Indicating Proficiency
• Several Rounds to Identify the KSAs Needed
• Set Additional Bookmarks
• Translated to Scale Score Cut Scores
• Remain in place until New Standards or Test
THE SCORES

- After the Cut Scores are determined, student scores can be calculated.
- Reports can be generated.
- Students can be placed in classes for the following year.
THINGS TO REMEMBER

• Realize that this has been a whirlwind view of a quite lengthy process.

• Some of you may be asked to be a part of the process as a member of one of the numerous committees mentioned earlier.

• Please consider being a part of this process since it can only benefit your students and all ELL students in the State of Arizona.

• You are the best judges of the achievements of your students.
Jane Scott
Jane Scott is the English Language Development Coordinator and Instructional Coach for Madison School District. Jane received her BA in Early Childhood Education from Northern Arizona University. She holds a teaching Certificate in Elementary Education.

Jane has been an educator at the Madison Elementary School District for 26 years and has served on many AZELLA development committees.
PILOT TESTING

Stage I-IV AZELLA - October 2011

• 80% ELL 20% Non-ELL
• Listening, Reading, and Writing administered in small groups
• Speaking administered individually on phone

Stage I Speaking Assessment - November 2011

• Done on the Phone
• 10 Kindergarten students
  (Pre. Emergent-Intermediate AZELLA levels and Non-ELL)
• Dialogue after each student

KG Placement Test (screener) - April and May 2012

• Pre-school students and non-pre-school students
• 80% ELL 20% Non-ELL
• Administered individually
• Approximately 20 minutes each
PLACEMENT TEST & STAGE I FOCUS GROUPS

- May 2012
- Educators who administered the same test were grouped together
- Went through every item and discussed
DATA ANALYSIS

• June 28th, 2012

• Groups of educators for each stage

• Educators looked at all of the data collected for their stage to determine the validity of the item, then decided whether to:
  o Keep the item as is
  o Keep the item with some minor changes
  o Take the item out all together
KINDERGARTEN PLACEMENT TEST
STANDARDS SETTING

• July 11 and 12
• Kindergarten and ELD teachers/coordinators
• Looked at how many students answered each question successfully
• Decided what the cut score should be for each of the AZELLA levels
RANGE FINDING FOR WRITING

- July 16\textsuperscript{th}- 20\textsuperscript{th}
- 2 groups of 5 or 6 educators per Stage
- Look at what the directions required, what a student would need in order to get 3, 2,1, or, 0 points
- Read papers and scored them. Went back and picked papers that would be high, average and low of each given point value.
- Met with like Stage groups to compare papers
ALIGNMENT STANDARDS/
PERFORMANCE LEVEL DESCRIPTORS

- October 2012
- 3 educators per Stage
- Reviewed standards for that Stage
- Read each question and found the standard that matched
- Input information in a data analysis system
Alistair
Van Moere, PhD
Alistair Van Moere, PhD is the Vice President of Product and Test Development at Knowledge Technologies, Pearson. He has an MA in English Language Teaching from Warwick University and a PhD in Applied Linguistics from Lancaster University. Additionally, he is studying for his Executive MBA.

Alistair has worked in language training and assessment for over 20 years, and has published 20 research articles in peer-reviewed journals on the subjects of oral language assessment and automated scoring.
AZELLA SPEAKING TEST

- Who uses automated scoring?
- Why does it work (when Siri doesn’t)?
- AZELLA development process
- Students’ performances
- Validation evidence
WHO USES AUTOMATED SPOKEN SCORING?

Educational Institutions

Governments

Corporations

Automated Scoring

- Duke
- Kobe Languages
- ACT
- Envision English School
- Yonsei English School
- The Hong Kong University of Science and Technology
- FIFA
- FUJITSU
- AT&T
- UConn
- Netherlands
- Dell
- Malaysia
- Philips
- 3M
- Convergys
- Accenture
- Emirates
- P&G
- LG
- Johnson & Johnson
BROADLY APPLICABLE SOLUTIONS

Corporate
- Recruiting selection
- Training placement
- Leadership programs
- Promotion

Schools and Universities
- Essay instructions
- Teacher/TA certification
- ELL placement
- Oral reading fluency
- Student assessment

Government
- Training
- Employment screening
- Language certification
- Immigrants screening

Private Language Schools and Publishers
- Oral reading fluency
- ESL, EFL language skills
- Assessment/scoring services
HIGH-STAKES ASSESSMENTS

- Pronunciation
- Structure
- Vocabulary
- Fluency
- Comprehension
- Interactions

Aviation English Test Score Report

- Test Administration
- Test Results

Understanding the Final Score

Level 1: Pre-Operational
Level 2: Elementary
Level 3: Operational
Level 4: Pre-Operational
## AUTOMATED TESTS IN USE

<table>
<thead>
<tr>
<th>Automated Test</th>
<th>Correlation to human raters</th>
<th>Primary Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>.97 (n=100)</td>
<td>US Government including Department of Homeland Security and US Dept of Defense</td>
</tr>
<tr>
<td>Dutch</td>
<td>.93 (n=139)</td>
<td>Dutch Government as part of immigration and naturalization procedure</td>
</tr>
<tr>
<td>Arabic</td>
<td>.98 (n=134)</td>
<td>US Defense Language Institute in the Arabic training program</td>
</tr>
<tr>
<td>English</td>
<td>.97 (n=150)</td>
<td>AT&amp;T, Dell, IBM, Samsung, P&amp;G, Accenture, Network Rail, CitiBank, LG, Convergys</td>
</tr>
<tr>
<td>Aviation English</td>
<td>.94 (n=140)</td>
<td>Boeing, Emirates Airlines, Belgian Government, Indian Government, Air Asia</td>
</tr>
<tr>
<td>PTE Academic</td>
<td>.97 (n=158)</td>
<td>Students for university entrance; recognized by ~2,000 institutions</td>
</tr>
</tbody>
</table>
ACCURACY (ORAL READING)

Expert scores vs. Machine scores

---

**Graph Description:**
- **X-axis:** Expert Words Correct
- **Y-axis:** Machine Words Correct
- The graph shows a strong positive correlation between expert and machine scores, with most data points falling close to the line of perfect agreement (y = x).
- The linear trend indicates that the machine is accurately predicting the expert's performance.
WHY DOES IT WORK?

1. The acoustic models are optimized for various accents

2. The test questions have been modeled from field test data – the system anticipates the various ways that students respond
1. The acoustic models are optimized for various accents

The system is forgiving of speech-sound errors and recognizes mis-pronounced words.
FIELD-TESTED ITEMS

2. The test questions have been modeled from field test data – the system anticipates the various ways that students respond

e.g. “What is it?”
LANGUAGE MODELS

Simplified response model
“Benji was a happy dog who liked to play in the garden”
AZELLA TEST DEVELOPMENT

Criteria -> Raters

Transcribers

Criteria

Test Developers

Item Text

Test Spec

Testing System

Recorded Items

Automated Scores

Validation & Benchmarking

ELLs

Non ELLs

<table>
<thead>
<tr>
<th>Stage</th>
<th>Number of tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>13,184</td>
</tr>
<tr>
<td>II</td>
<td>10,646</td>
</tr>
<tr>
<td>III</td>
<td>9,369</td>
</tr>
<tr>
<td>IV</td>
<td>6,439</td>
</tr>
<tr>
<td>V</td>
<td>5,231</td>
</tr>
</tbody>
</table>
DEVELOPMENT & VALIDATION

Development

System is “trained” to predict human ratings

Validation

Machine scores
SENTENCE REPEATS

(Audience participation)
SENTENCE REPEATS

Adapted from Levelt, 1989

Listen

Hear utterance
Extract words
Get phrase structure
Decode propositions
Contextualize
Infer Demand (if any)

Speak

Articulate response
Build clause structure
Select lexical items
Construct phrases
Select register
Decide on response

0.04 second

0.5 ~ 1 second
PHONEME & WORD ALIGNMENT

waveform
spectrum
words
segmentation
<table>
<thead>
<tr>
<th>Score</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Student formulates a response in correct, understandable English using two or more sentences based on given stimuli.</td>
</tr>
<tr>
<td>3</td>
<td>Student formulates a response in understandable English using two or more sentences based on a given stimuli.</td>
</tr>
<tr>
<td>2</td>
<td>Student formulates an intelligible English response based on given stimuli.</td>
</tr>
<tr>
<td>1</td>
<td>Student formulates erroneous responses based on given stimuli.</td>
</tr>
<tr>
<td>0</td>
<td>- Student formulates responses in non-English.</td>
</tr>
<tr>
<td></td>
<td>- Student does not respond.</td>
</tr>
</tbody>
</table>
Next, please answer in complete sentences. Tell how to get ready for school in the morning. Include at least two steps.

<table>
<thead>
<tr>
<th>Item</th>
<th>Response Transcript</th>
<th>Human Score</th>
<th>Machine Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>first you wake up and then you put on your clothes and eat breakfast</td>
<td>3</td>
<td>3.35</td>
</tr>
<tr>
<td>Item</td>
<td>Response Transcript</td>
<td>Human Score</td>
<td>Machine Score</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Next, please answer in complete sentences. Who is your favorite cartoon or book character?</td>
<td>my favorite book character is: [N] uh five [N] little monkeys jumping on the bed # [N] [N]</td>
<td>3</td>
<td>2.85</td>
</tr>
<tr>
<td>Why is that character your favorite?</td>
<td>becau:se they always make me laugh # when i read it # [S]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## RELIABILITY: 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>Average</td>
<td>0.97</td>
<td>0.93</td>
</tr>
</tbody>
</table>
## RELIABILITY: ITEM-TYPE

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Human-human correlation</th>
<th>Machine-human correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions about an image</td>
<td>0.87</td>
<td>0.77</td>
</tr>
<tr>
<td>Similarities and differences</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Give directions from a map</td>
<td>0.74</td>
<td>0.85</td>
</tr>
<tr>
<td>Questions about a statement</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Give instructions to do something</td>
<td>0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>Open questions about a topic</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Detailed responses to a topic</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>Repeat</td>
<td>0.97</td>
<td>0.88</td>
</tr>
</tbody>
</table>
ADVANTAGE OF AUTOMATED SCORING

- Standardized administration
- Objective, bias-free scoring
- Data-driven models from 1000s speakers
- Accumulation of measures from multiple expert raters
Sincere gratitude to our Arizona ELL community for their partnership and dedication during this accelerated development process.
“The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route to broader and more equitable access to education and employment.

The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions.”

**STANDARDS for educational and psychological testing**

American Educational Research Association
American Psychological Association
National Council on Measurement in Education
QUESTIONS?