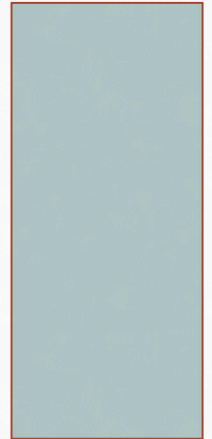


AZELLA DEVELOPMENT

THE INSIDE STORY

OELAS CONFERENCE
DECEMBER 12, 2012



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 Learners 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

School	ELL	Non-ELL	Total
10220101	34	25	59
30201112	12	9	21
3201114	21	6	27
30201118	8	10	18
30201121	6	13	19
30201122	2	10	12
70402101	0	37	37
70403122	31	49	80
70479101	10	10	20
70479104	29	10	39
70479109	24	10	34
....			

EVALUATING AN ITEM

What other data is used to evaluate the items on a test?

Descriptive Statistics for Total and Subtest Scaled Scores by ELL Category, and by Performance Level based on Total for ELL Students

	ELL Students								Non-ELL Students		
	Total (N=1040)			PL = 1 (N=207)	PL = 2 (N=160)	PL = 3 (N=601)	PL = 4 (N=63)	PL = 5 (N=9)	Total (N = 481)		
Subtest	Mean	SD	Alpha	Mean	Mean	Mean	Mean	Mean	Mean	SD	Alpha
01.Listening	488.26	54.49	0.71	424.31	473.71	506.11	552.03	579.44	511.17	43.48	0.52
03.Reading	475.66	53.78	0.73	417.66	467.41	487.76	553.35	604.78	501.00	43.29	0.66
04.Prewriting	429.71	73.02	0.90	361.86	394.86	447.36	545.63	619.78	461.08	72.49	0.90
05.Speaking	499.53	61.85	0.92	414.99	469.97	528.07	564.00	612.89	554.10	49.62	0.81
10.Comprehension	480.58	47.41	0.81	421.54	471.59	494.58	547.37	596.11	503.78	37.30	0.73
11.Oral	495.47	51.72	0.91	420.78	471.78	519.38	557.22	605.22	537.48	37.86	0.79
13.Total	476.78	43.95	0.93	411.99	457.34	494.98	546.57	608.89	508.34	32.08	0.90

EVALUATING AN ITEM

(continued)

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

PL Total	ELL Students					Non-ELL Students				
	PreEmergent	Emergent	Basic	Intermediate	Proficient	PreEmergent	Emergent	Basic	Intermediate	Proficient
1.Listening	5.48	6.92	39.13	37.12	11.35	0.83	3.33	27.03	50.10	18.71
3.Reading	11.06	22.60	52.50	9.52	4.33	1.25	17.05	55.30	14.97	11.43
4.Prewriting	52.21	15.77	28.75	2.88	0.38	35.97	13.93	41.37	8.11	0.62
5.Speaking	18.27	10.87	42.40	16.44	12.02	0.83	2.29	35.14	26.61	35.14
10.Comprehension	7.98	12.40	62.79	13.75	3.08	1.04	4.99	62.99	22.45	8.52
11.Oral	10.67	14.52	45.58	23.65	5.58	0.42	1.87	37.21	38.67	21.83
13.Total	19.90	15.38	57.79	6.06	0.87	1.25	6.86	70.69	19.54	1.66

EVALUATING INDIVIDUAL ITEMS

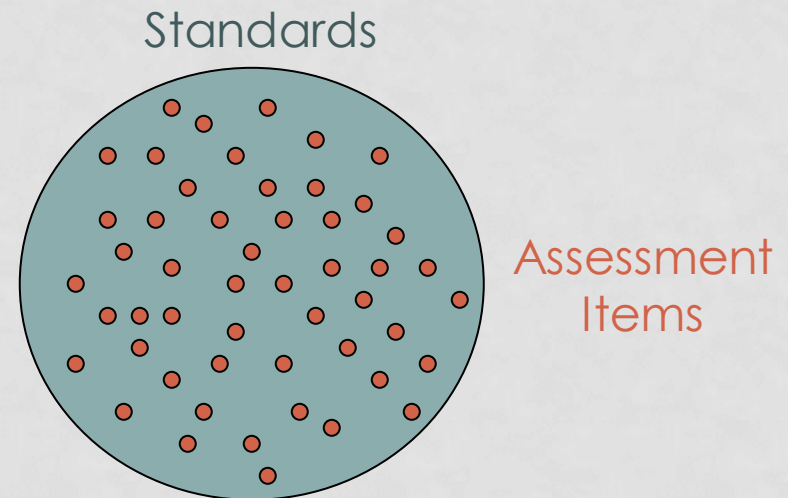
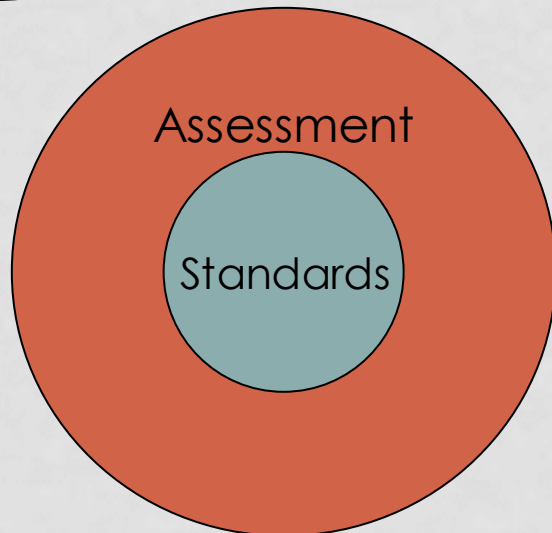
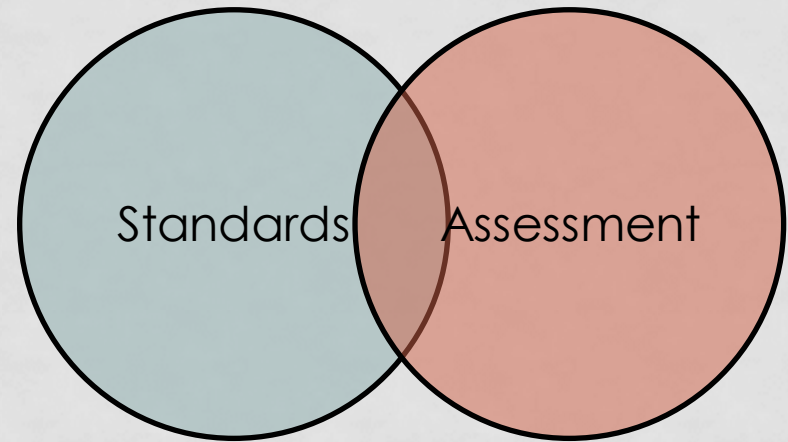
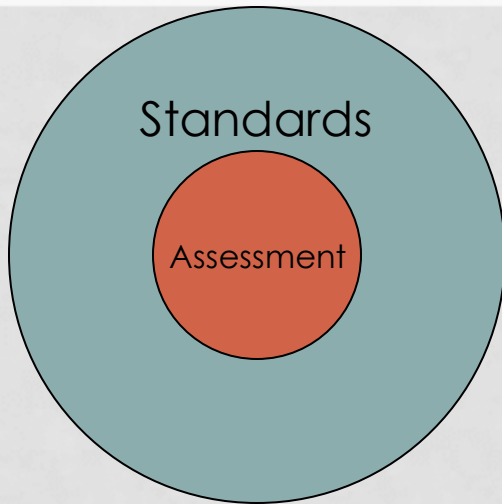
P-value for ELL Students, ELL Students by Performance Level by Total and Non-ELL Students

Item	Max Possible	ELL Students						SE	Non-ELL Students				
		Total ELL (N=1040)	PL = 1 (N=207)	PL = 2 (N=160)	PL = 3 (N=601)	PL = 4 (N=63)	PL = 5 (N=9)		Total Non-ELL (N=481)	NonELL - ELL	Non-PL5	Non-PL4	Non-PL3
i10	1	0.71	0.35	0.65	0.82	0.92	1.00	0.00	0.81	0.10	-0.19	-0.11	-0.01
i11	1	0.46	0.21	0.39	0.54	0.67	0.89	0.10	0.48	0.02	-0.40	-0.18	-0.05
i12	1	0.44	0.19	0.40	0.50	0.75	0.78	0.14	0.58	0.14	-0.20	-0.17	0.07
i13	1	0.58	0.29	0.53	0.66	0.86	1.00	0.00	0.67	0.09	-0.33	-0.19	0.01
i14	1	0.58	0.33	0.59	0.63	0.87	0.89	0.10	0.72	0.15	-0.17	-0.15	0.10
i15	1	0.56	0.34	0.56	0.61	0.76	0.89	0.10	0.59	0.03	-0.30	-0.17	-0.02
i16	1	0.33	0.15	0.29	0.37	0.56	0.89	0.10	0.41	0.08	-0.48	-0.15	0.04
i17	1	0.51	0.21	0.46	0.60	0.81	0.89	0.10	0.69	0.18	-0.20	-0.12	0.10
i18	1	0.23	0.08	0.16	0.23	0.73	1.00	0.00	0.31	0.08	-0.69	-0.42	0.08
i19	1	0.67	0.38	0.60	0.75	0.92	1.00	0.00	0.78	0.11	-0.22	-0.14	0.03
i20	1	0.68	0.27	0.64	0.80	0.84	1.00	0.00	0.82	0.15	-0.18	-0.02	0.02
i21	1	0.39	0.16	0.29	0.43	0.89	1.00	0.00	0.56	0.17	-0.44	-0.33	0.13
i22	1	0.28	0.06	0.19	0.31	0.79	1.00	0.00	0.40	0.13	-0.60	-0.39	0.10
i23	1	0.33	0.10	0.23	0.37	0.87	1.00	0.00	0.41	0.08	-0.59	-0.46	0.04
i24	1	0.33	0.10	0.25	0.36	0.84	1.00	0.00	0.43	0.11	-0.57	-0.41	0.07
i25	2	0.91	0.76	0.88	0.97	0.99	1.00	0.00	0.97	0.05	-0.03	-0.03	0.00
i26	2	0.83	0.61	0.77	0.91	0.97	1.00	0.00	0.91	0.08	-0.09	-0.06	0.00

EVALUATING AN ITEM

Arizona Fall 2011		Grade: KPT						
AZID: 7812P108		Domain: Listening			N-Count: 358			
Item Type: OR1		Form: R			Item #: 6			Key:
Standard: 1		Prof Level: B			Indicator: 6			
Passage:								
The student will demonstrate understanding of oral communications by: responding to comments and questions in social conversations by s haring one's experiences and expressing one's thoughts.								
Item Statistics								
P-Value: 0.80		Item Mean: 0.80			PtBis : 0.63			
Non-ELL P-Value: 0.97		Non-ELL Item Mean: 0.97			Non-ELL PtBis: 0.42			
Rasch: -0.02		Infit: 0.80			Outfit: 0.59*			
	N Count	%0	%1	%2	%3	%4	%5	%Omit
Total ELL	358	19.90	80.10	0.00	0.00	0.00	0.00	0.00
Non-ELL	144	2.80	96.50	0.00	0.00	0.00	0.00	0
High	102	0.00	100.00	0.00	0.00	0.00	0.00	0.00
Medium	170	12.90	87.10	0.00	0.00	0.00	0.00	0.00
Low	84	58.30	41.70	0.00	0.00	0.00	0.00	0.00
Bias: MHzB- SMD-Delta (Gender, Race, Ethnicity, SES)								
	M-F	NH-H	NAI-AI	N-FRL	NEL-ELL	FEP-ELL		
MHzB	0.001	0.152			2.013			
Bias Catg Flag	A	A			A			
SMD	0.005	-0.048			-0.254			
Delta	0.027	-0.373			-1.789			
Bias: MHzB- SMD-Delta (based on Language)								
	NSpE-SpE	NS-S	E-S	E-NS	E-AI			
MHzB								
Bias Catg Flag								
SMD								
Delta								
Population (GRES)		Count			Population (Language)		Count	
Male		160			Non-Special Ed			
Female		194			Special Ed			
Hispanic		296			Non-Spanish			
Native American					Spanish			
Free/Reduced Lunch					English		144	
English Lang Learner		358			Native American			

ALIGNMENT TO STANDARDS

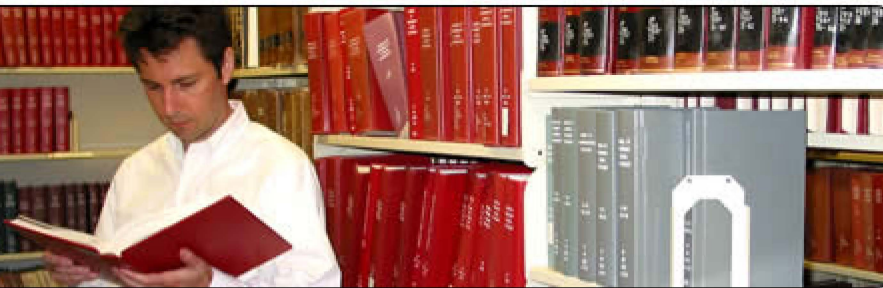


WHAT ALIGNMENT IS GOOD ENOUGH?

Alignment Level	Categorical Concurrence	Depth of Knowledge	Range of Knowledge	Balance of Representation
<i>Acceptable</i>	6 items per standard	50%	50%	70%
<i>Weak</i>	---	40%-49%	40%-49%	60%-69%
<i>Unacceptable</i>	Less than 6 items per standard	Less than 40%	Less than 40%	Less than 60%

ALIGNMENT

- ELP Standards (Highly Instructionally Oriented)
 - V.S.
- Assessable Standards (Major Skills to be Proficient – Able to access content in a Non-ELL Classroom).



Disclaimer:
This material is based upon work supported by the National Science Foundation under contract number EHR-0233445 awarded to the University of Wisconsin-Madison and the Wisconsin Center for Education Research. Any opinions, findings, or conclusions are those of the creator(s) and developers and do not necessarily reflect the views of the supporting agencies.

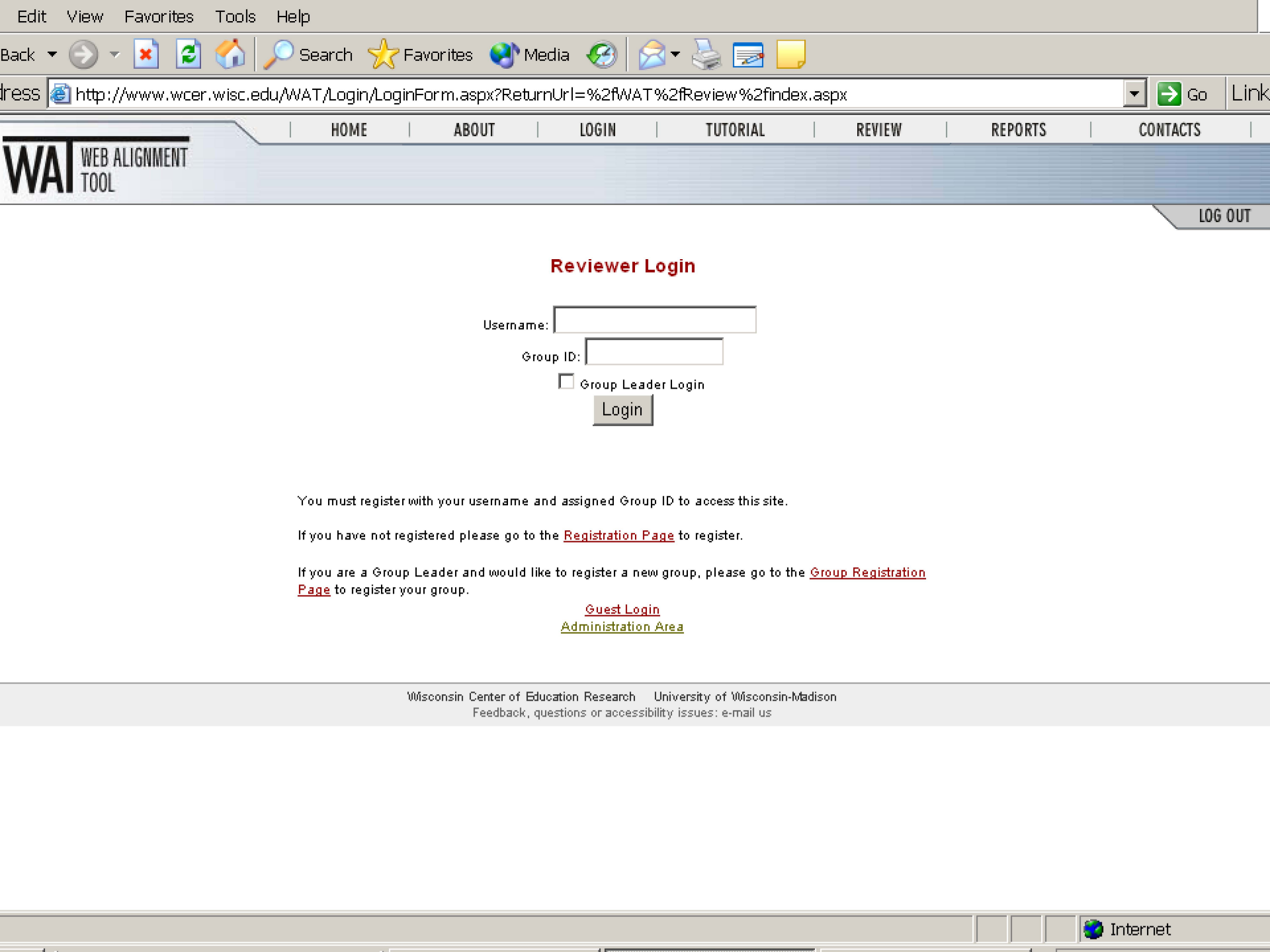
LOG OUT

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:

Group ID:

☐ Group Leader Login

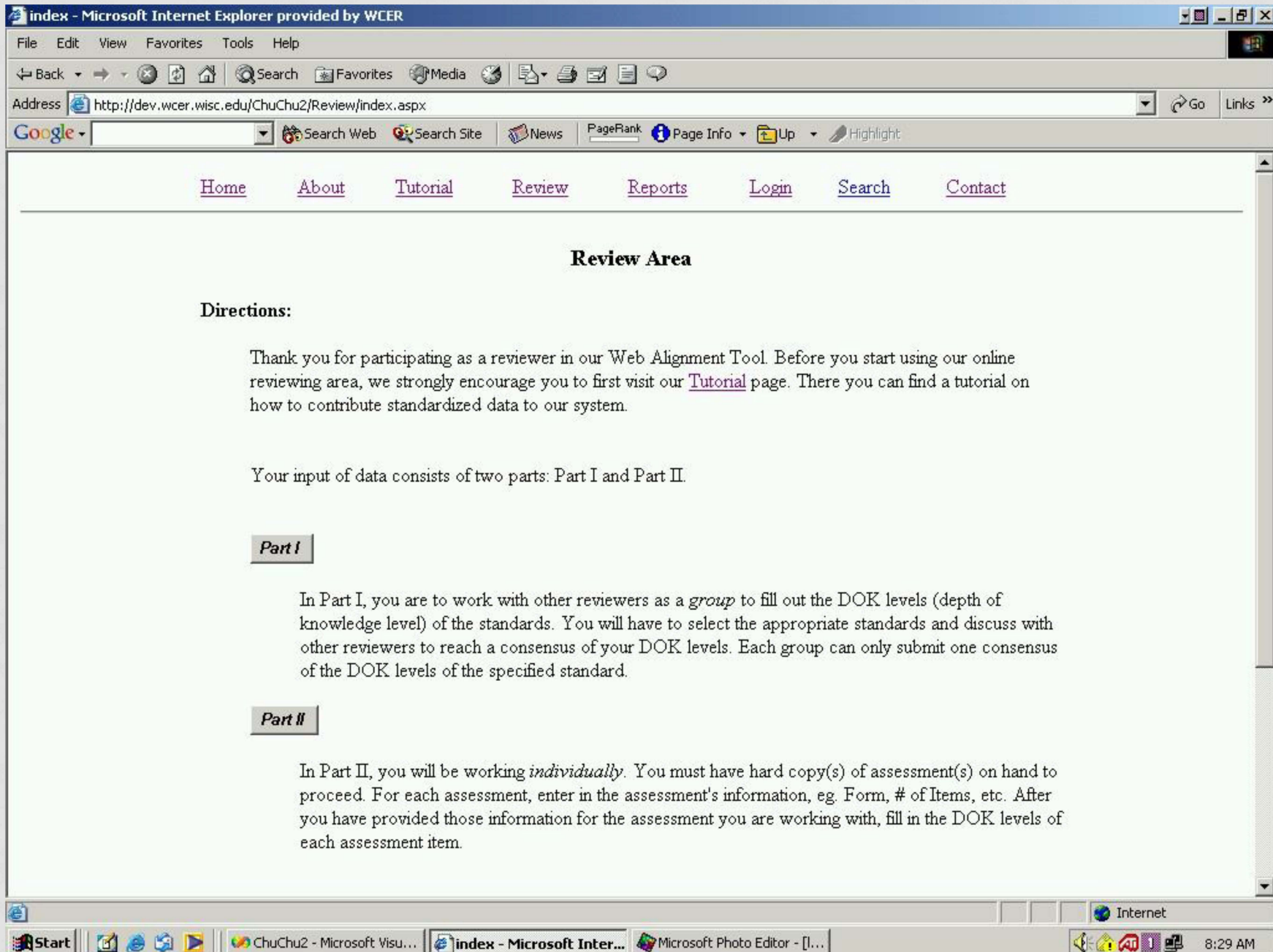
Login

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](#)



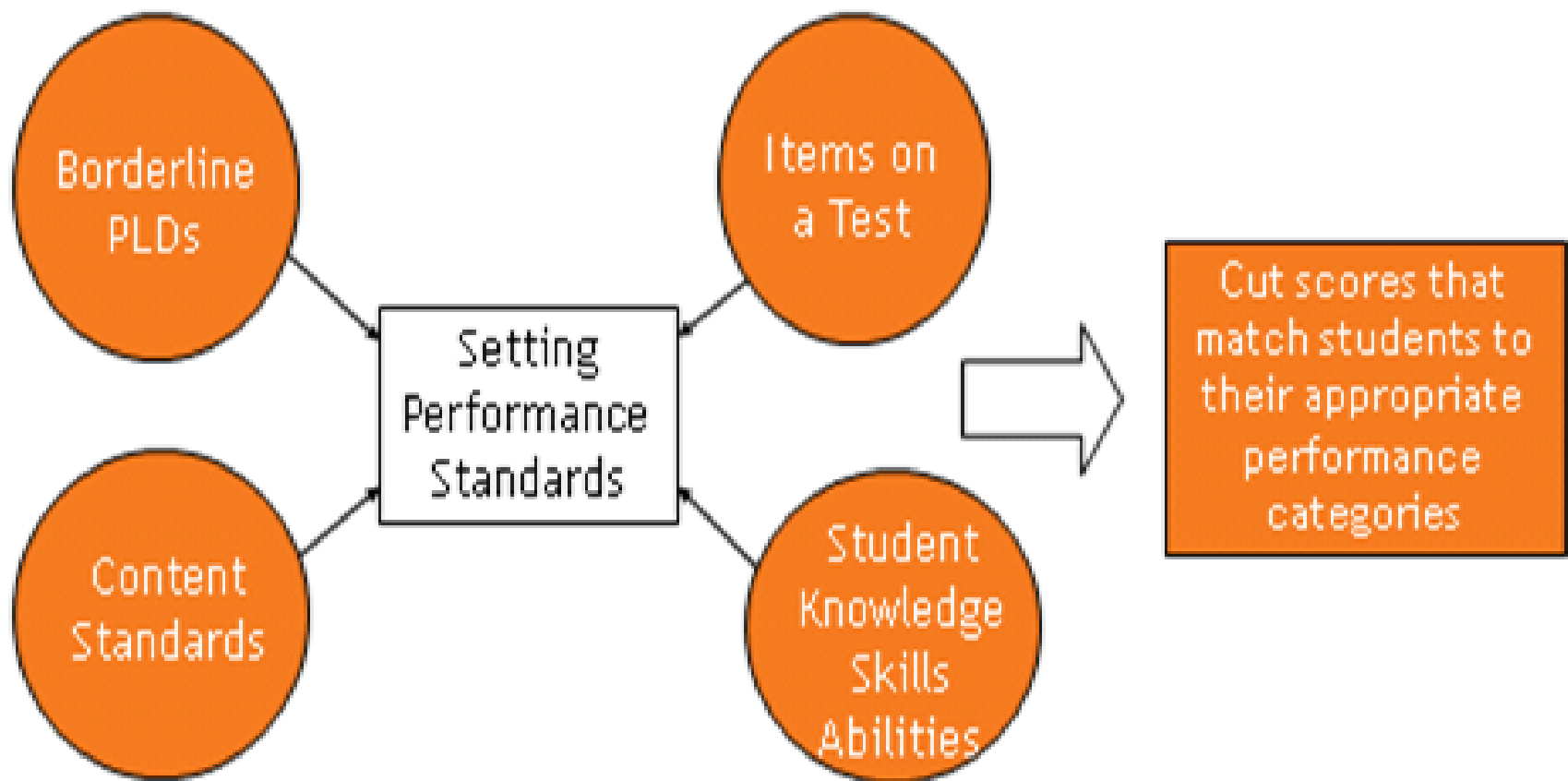
STRUCTURE OF THE AUTOMATED ALIGNMENT PROCESS

- 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



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 (screeners) 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:
 - Keep the item as is
 - Keep the item with some minor changes
 - 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 16th- 20th
- 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

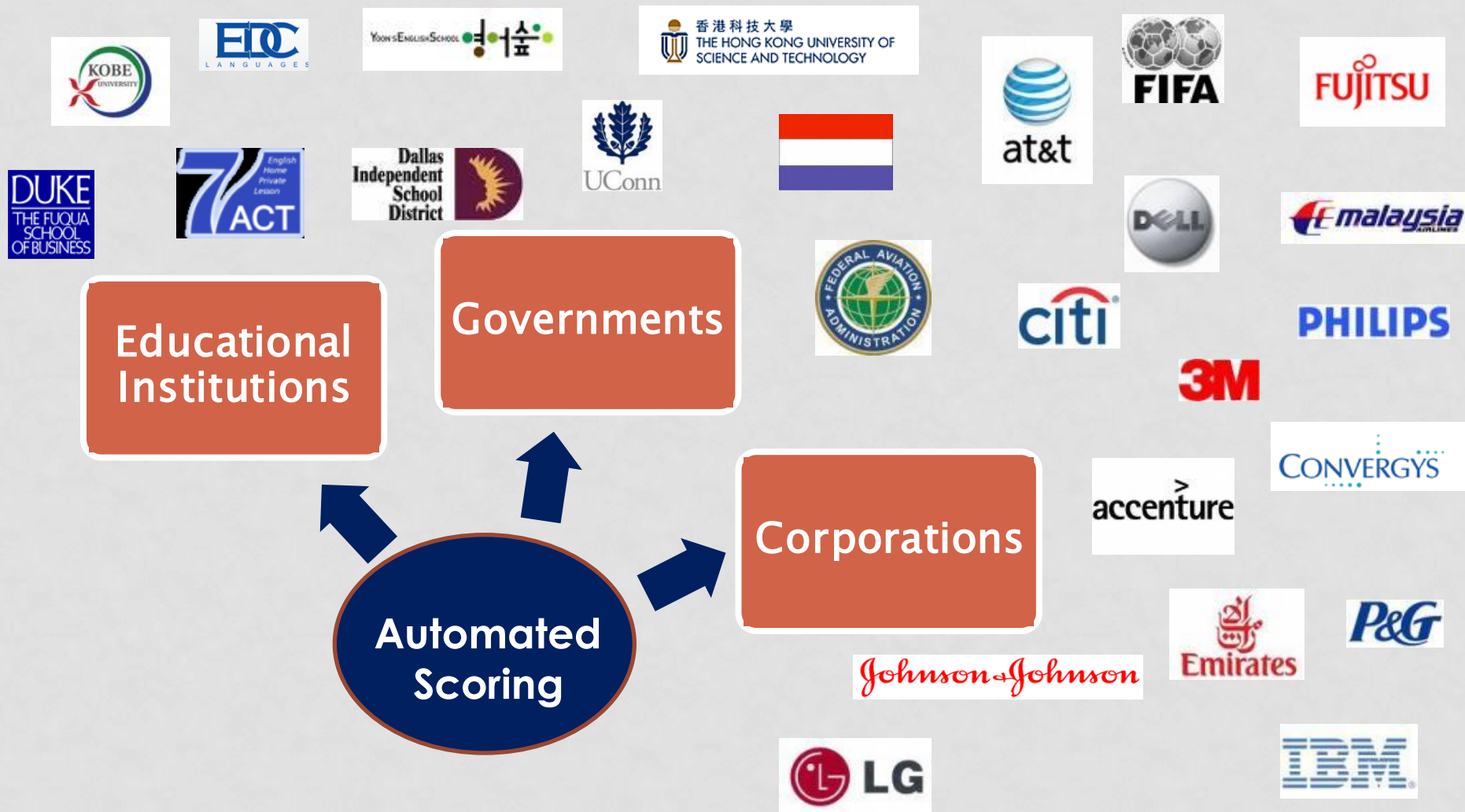
Alistair Van Moere, PhD is the Vice President of Product and Test Development at Knowledge Technologies, Pearson. He has an MA in English Language Teaching from Warwick University and a PhD in Applied Linguistics from Lancaster University. Additionally, he is studying for his Executive MBA.

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.

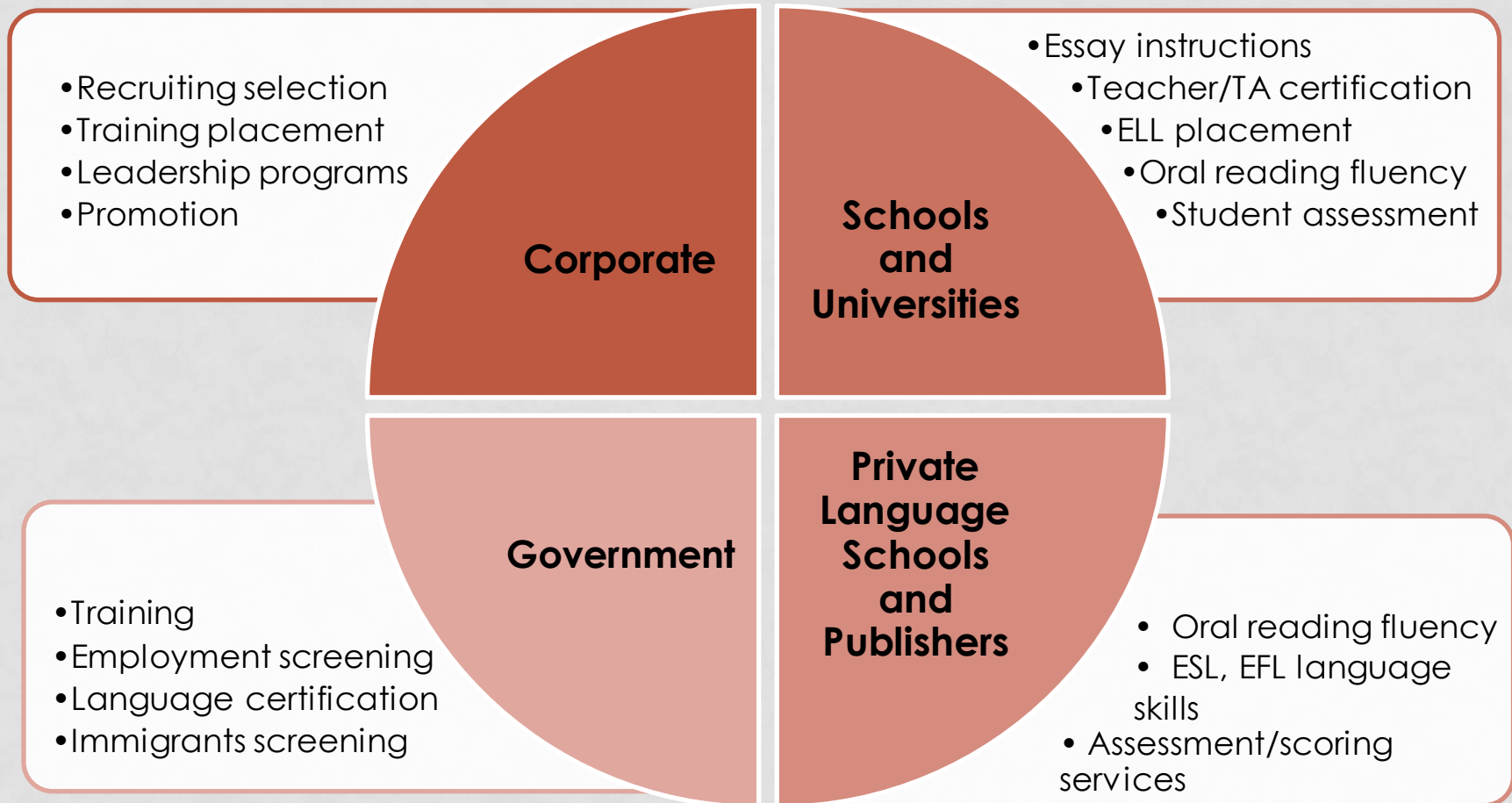
AZELLA SPEAKING TEST

- Who uses automated scoring?
- Why does it work (when Siri doesn't)?
- AZELLA development process
- Students' performances
- Validation evidence

WHO USES AUTOMATED SPOKEN SCORING?



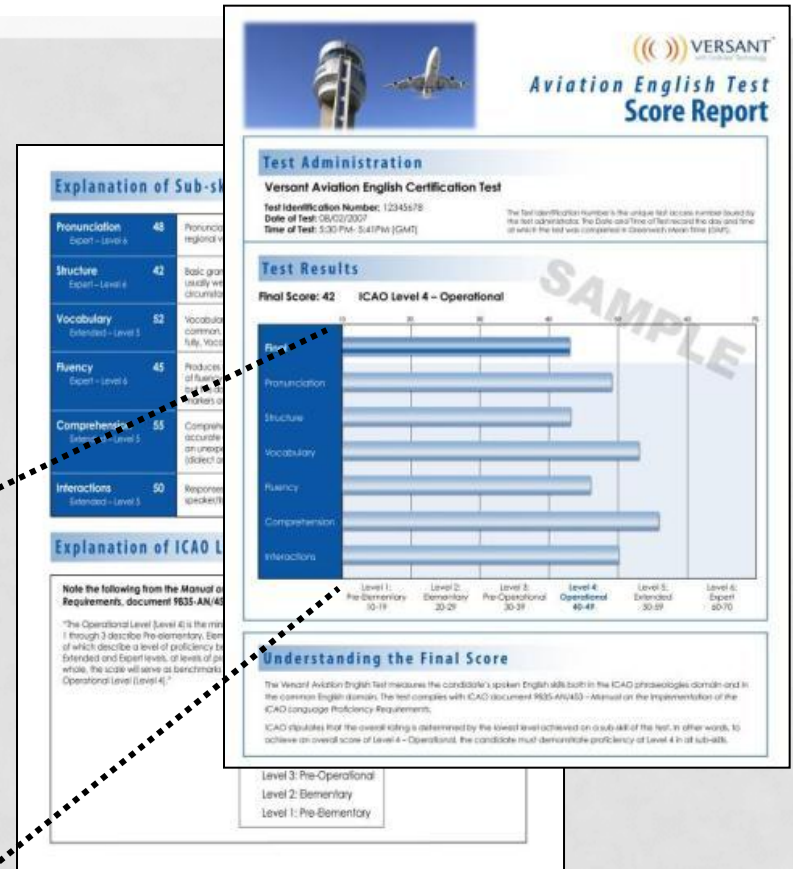
BROADLY APPLICABLE SOLUTIONS



HIGH-STAKES ASSESSMENTS



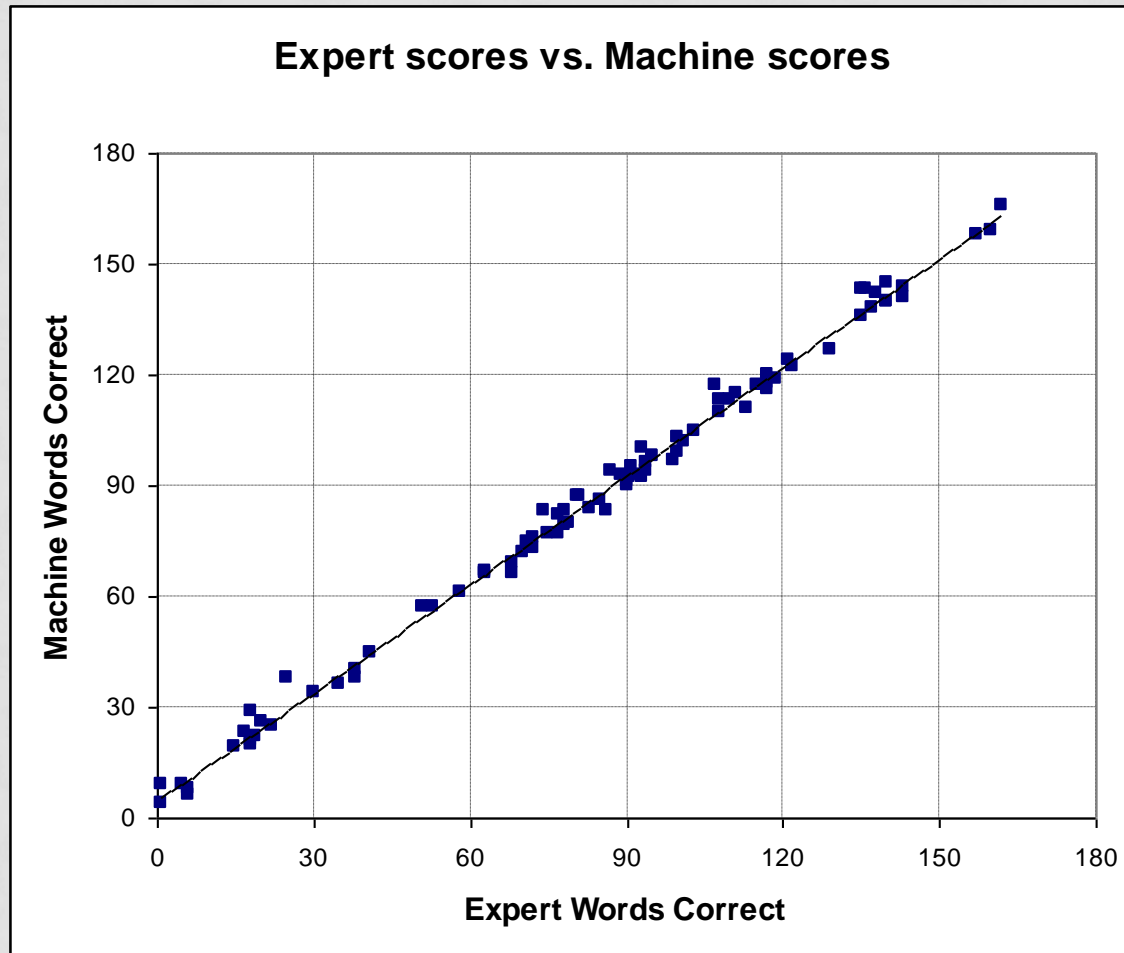
Pronunciation
Structure
Vocabulary
Fluency
Comprehension
Interactions



AUTOMATED TESTS IN USE

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

ACCURACY (ORAL READING)

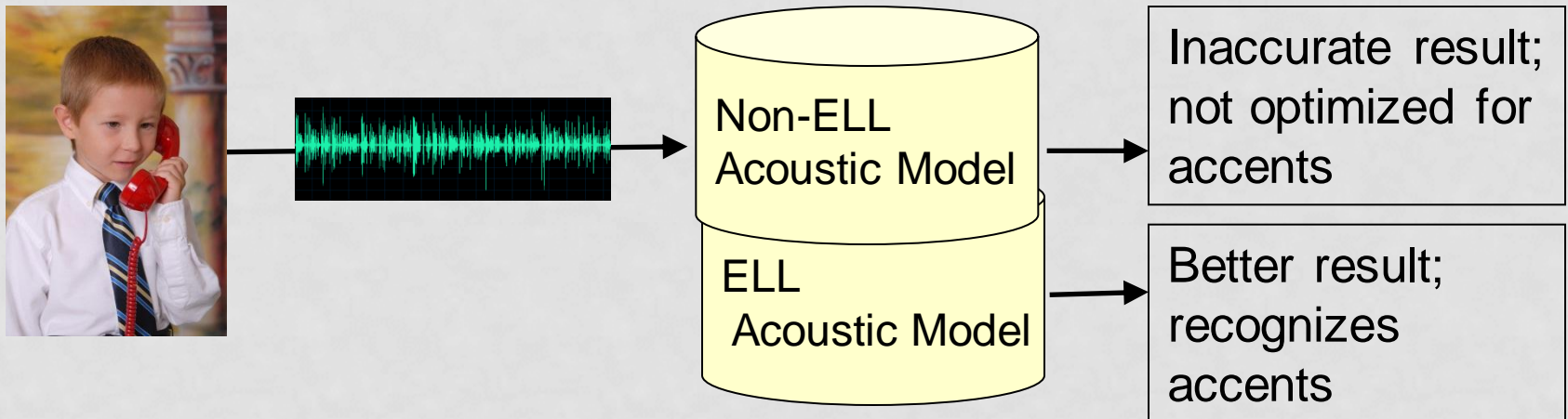


WHY DOES IT WORK?

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

ACOUSTIC MODELS

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

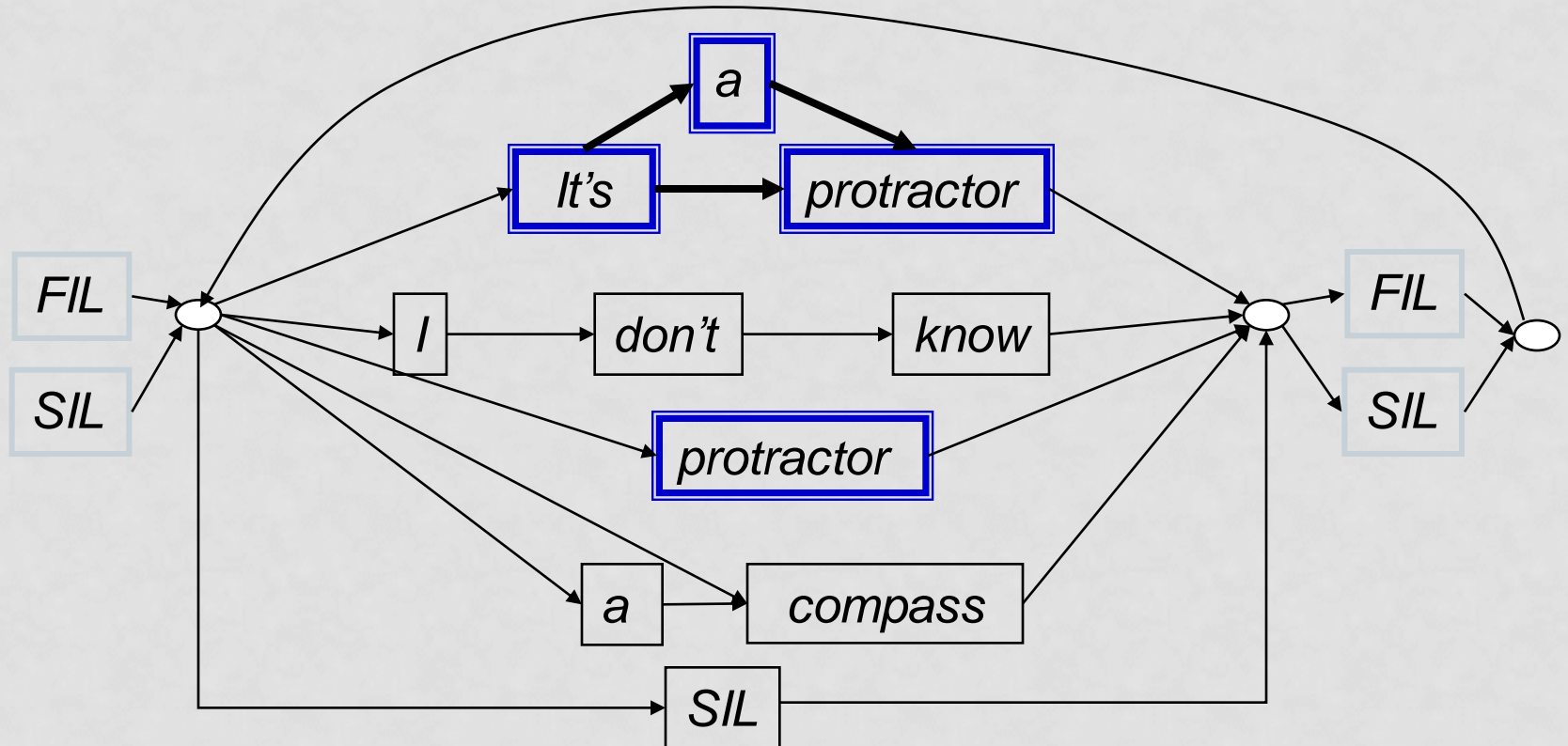
2. The test questions have been modeled from field test data – the system anticipates the various ways that students respond

e.g. *“What is it?”*

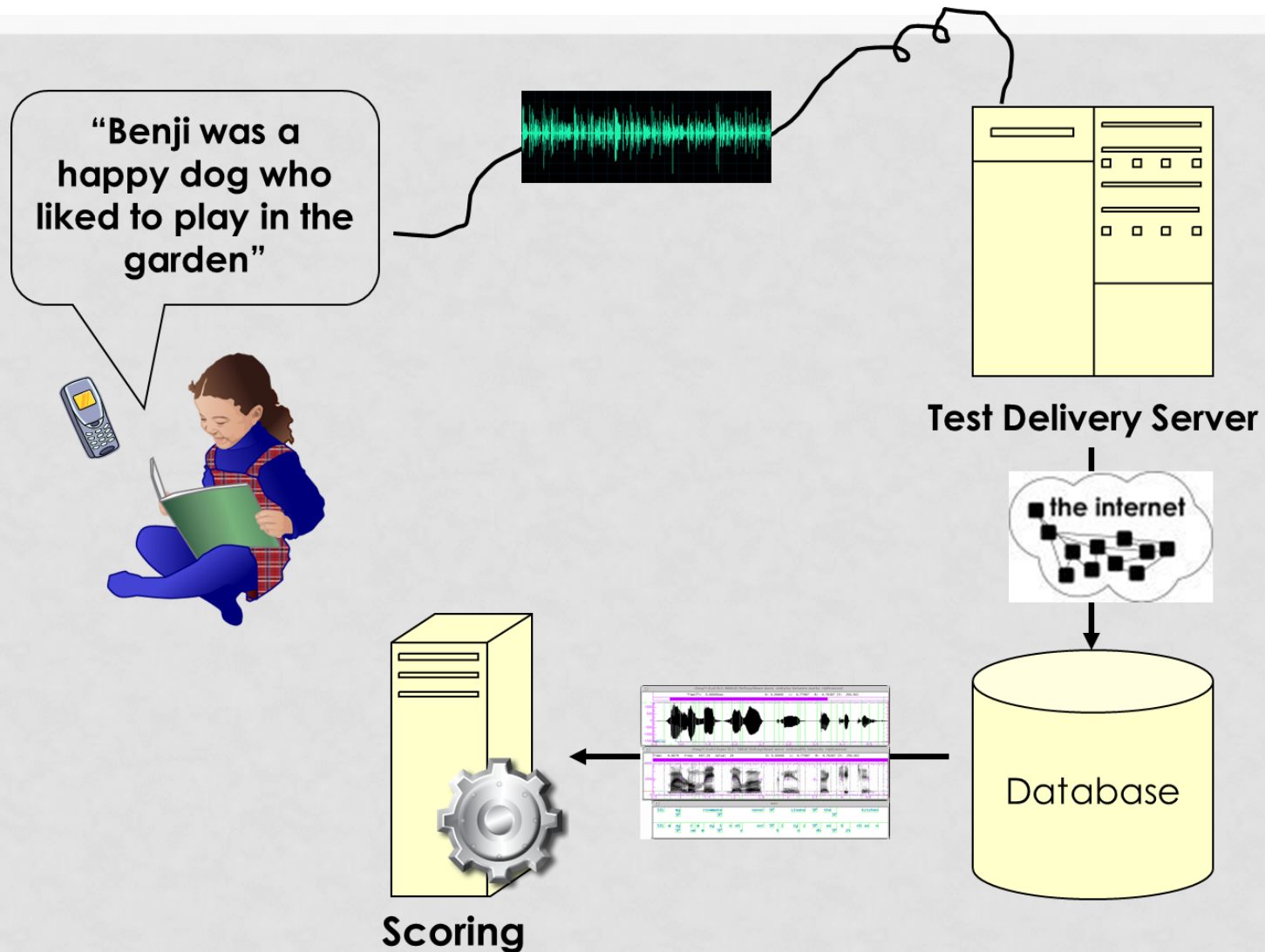


LANGUAGE MODELS

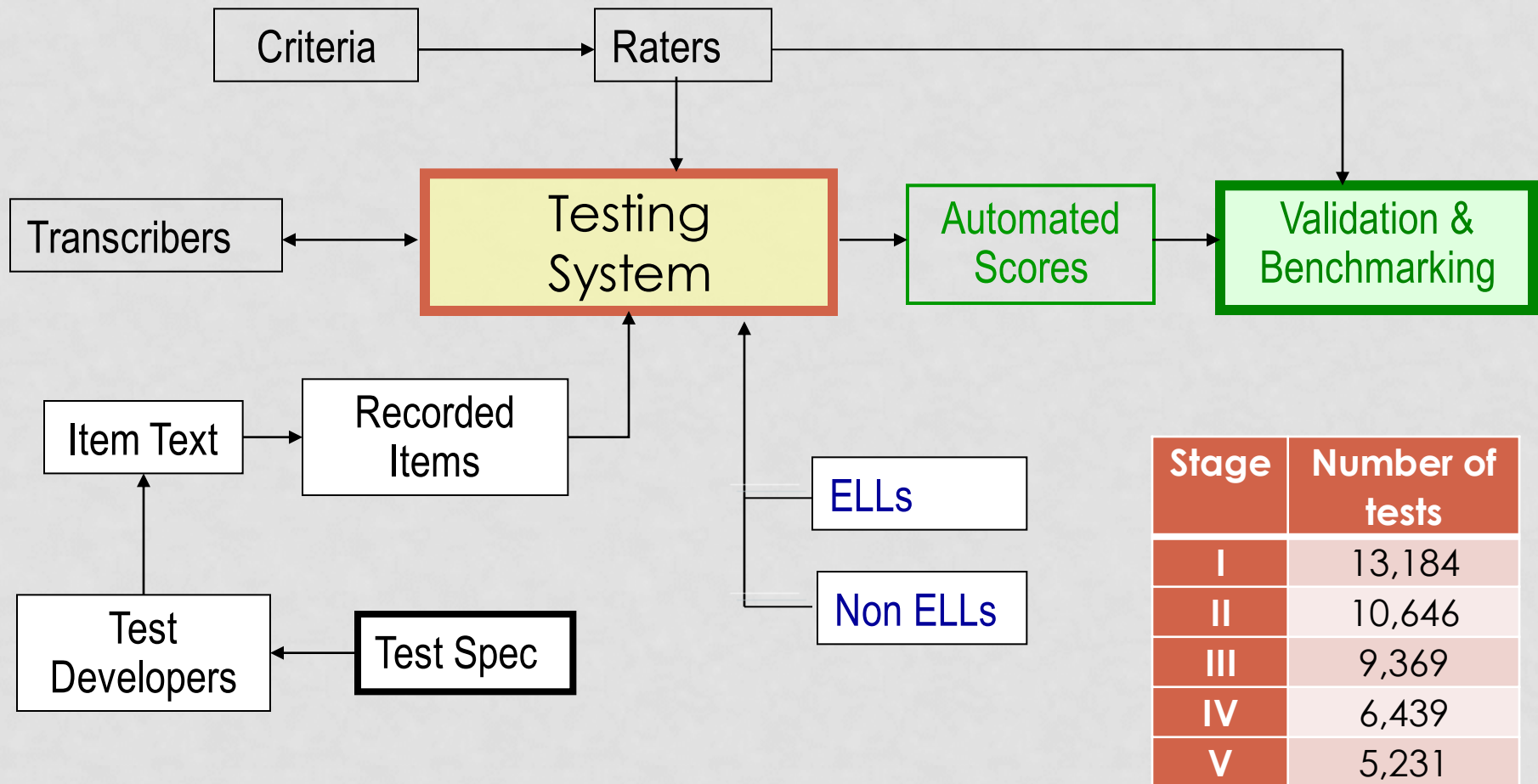
Simplified response model



AZELLA TEST DELIVERY

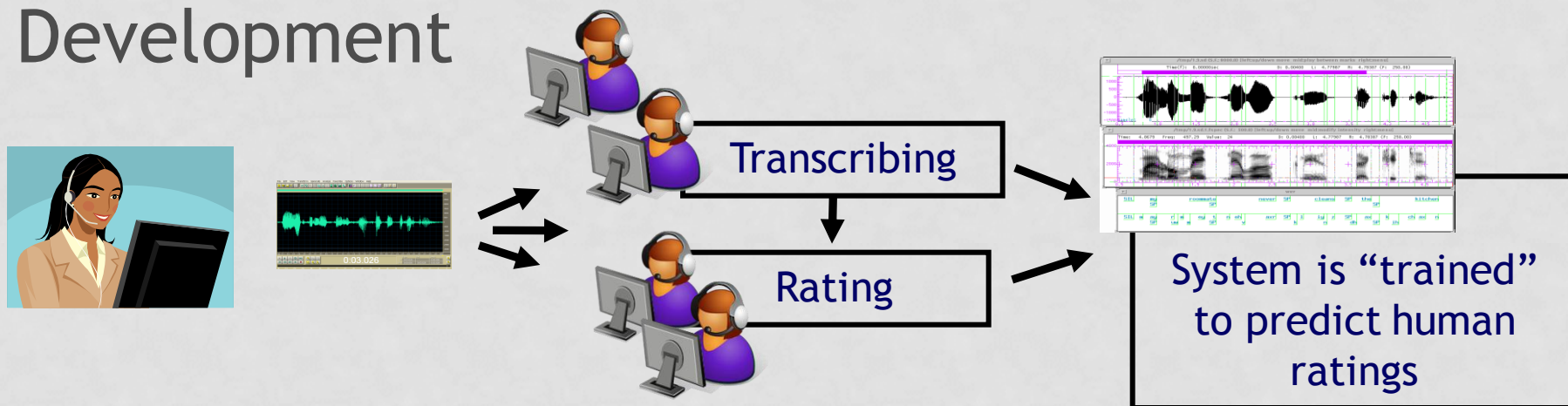


AZELLA TEST DEVELOPMENT

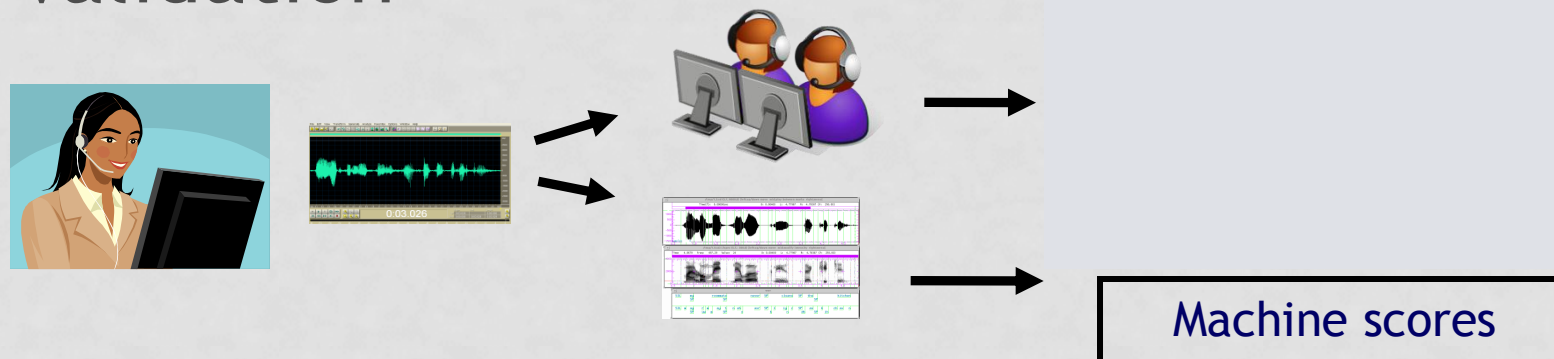


DEVELOPMENT & VALIDATION

Development



Validation

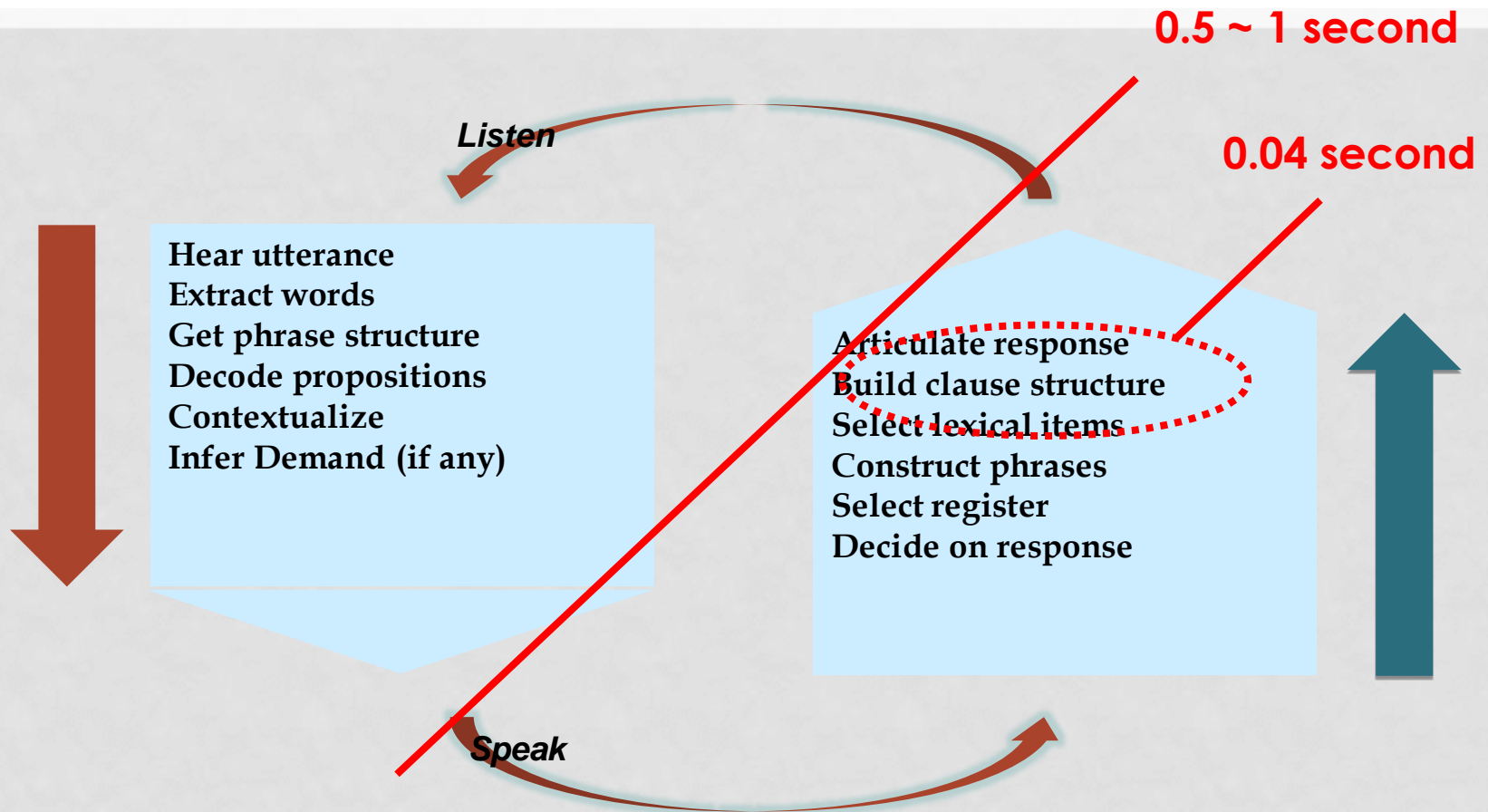




SENTENCE REPEATS

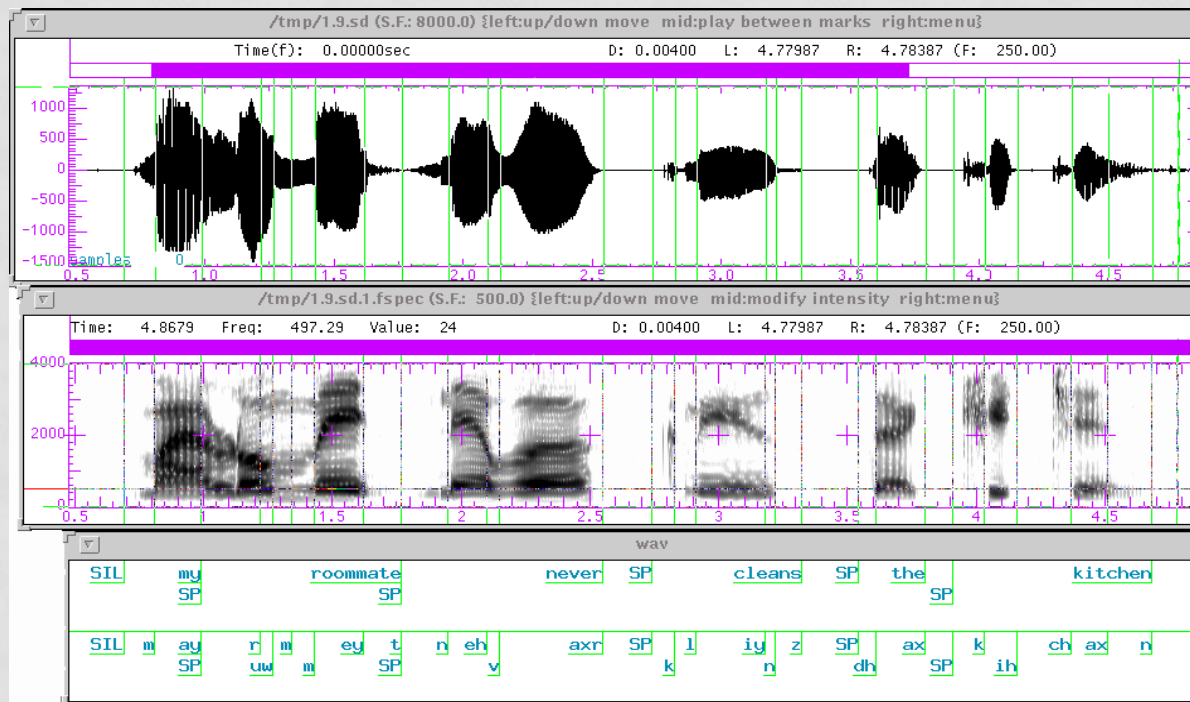
(Audience participation)

SENTENCE REPEATS



Adapted from Levelt, 1989

PHONEME & WORD ALIGNMENT



waveform

spectrum

words

segmentation

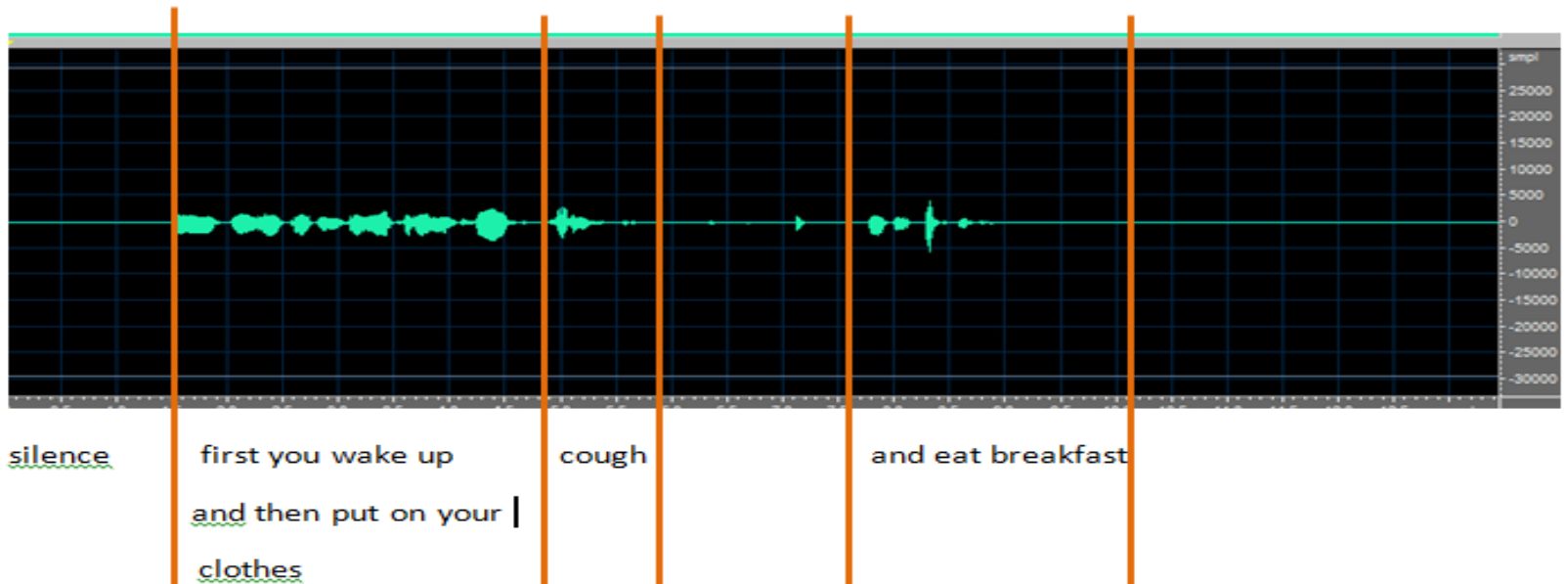
ABBREVIATED DESCRIPTORS



	Descriptors
4	Student formulates a response in correct, understandable English using two or more sentences based on given stimuli.
3	Student formulates a response in understandable English using two or more sentences based on a given stimuli.
2	Student formulates an intelligible English response based on given stimuli.
1	Student formulates erroneous responses based on given stimuli.
0	<ul style="list-style-type: none">• Student formulates responses in non-English.• Student does not respond.

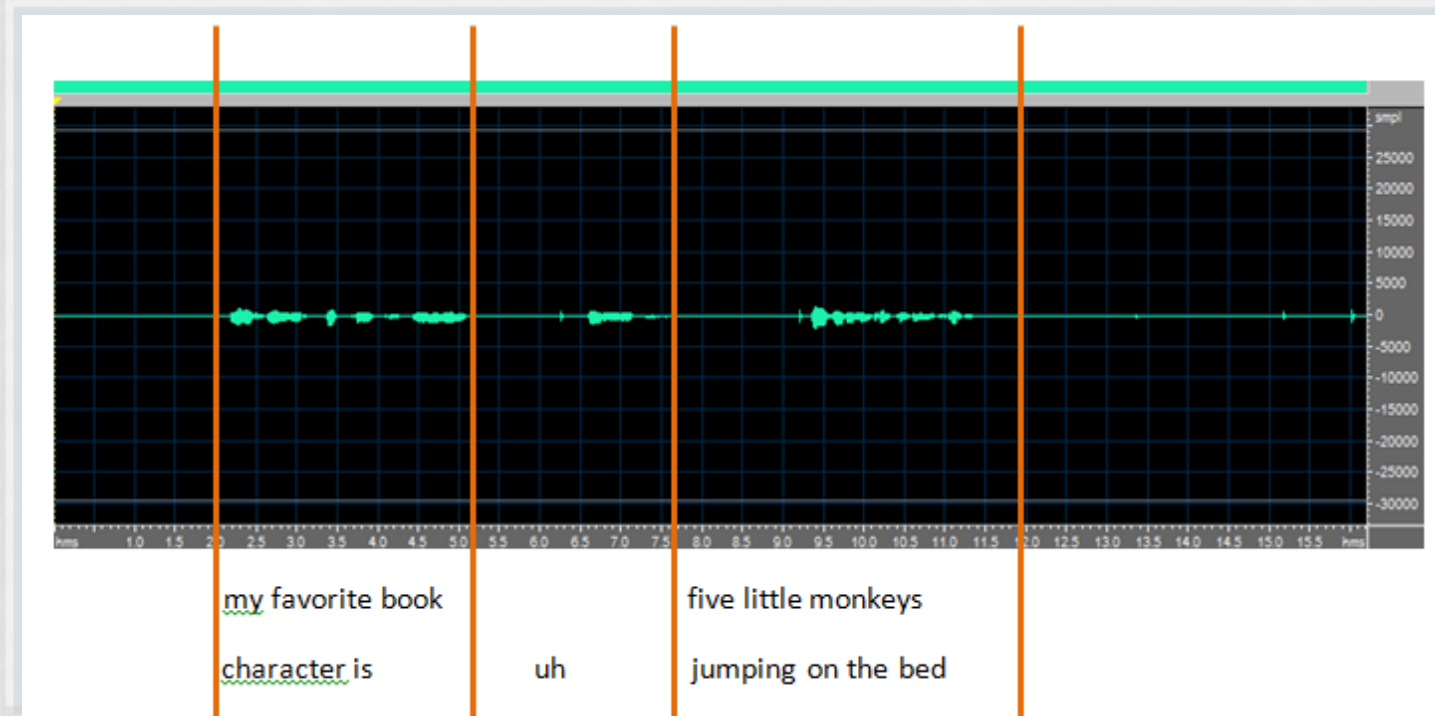


STUDENT RESPONSES

Item	Response Transcript	Human Score	Machine Score
Next, please answer in complete sentences. Tell how to get ready for school in the morning. Include at least two steps.	first you wake up and then you put on your clothes # and eat breakfast	3	3.35



Item	Response Transcript  	Human Score	Machine Score
Next, please answer in complete sentences. Who is your favorite cartoon or book character?	my favorite book character is: [N] uh five [N] little monkeys jumping on the bed # [N] [N]	3	2.85
Why is that character your favorite?	becau:se they always make me laugh # when i read it # [S]		



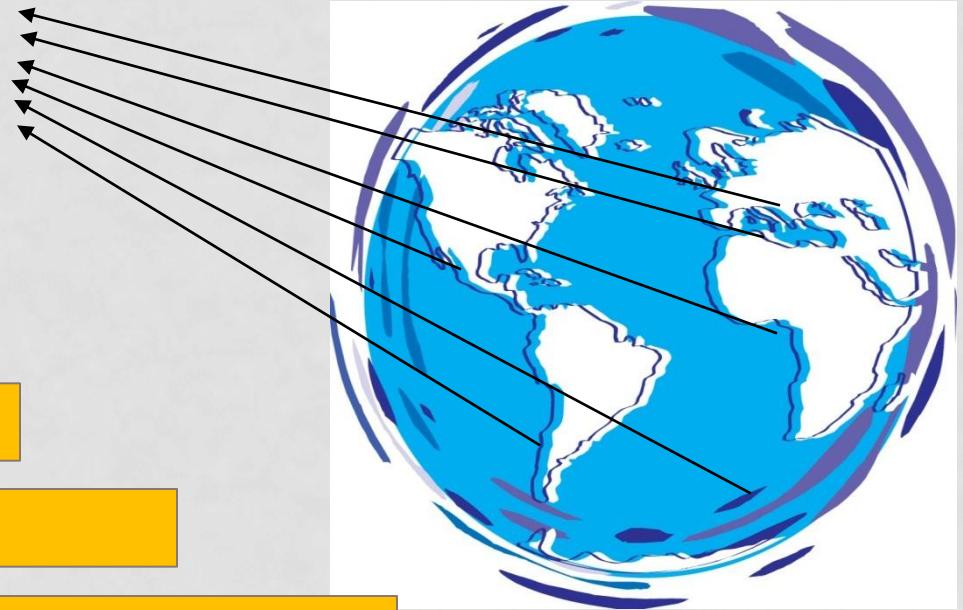
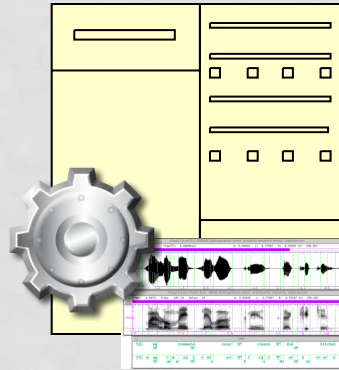
RELIABILITY: TEST LEVEL

Stage	Human-Human Correlation r	Machine-Human Correlation r
I	0.91	0.88
II	0.96	0.90
III	0.97	0.94
IV	0.98	0.95
V	0.98	0.93
Average	0.97	0.93

RELIABILITY: ITEM-TYPE

Item Type	Human-human correlation	Machine-human correlation
Questions about an image	0.87	0.77
Similarities and differences	0.75	0.75
Give directions from a map	0.74	0.85
Questions about a statement	0.79	0.82
Give instructions to do something	0.77	0.81
Open questions about a topic	0.85	0.85
Detailed responses to a topic	0.81	0.80
Repeat	0.97	0.88

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?