A Guide to Measuring Listening According to the ILR Skill Level Descriptions

Findings and Recommendations of the DLIFLC Listening Summits

Gerd Brendel – DLIFLC
Beth Mackey – Department of Defense
Elvira Swender - ACTFL

ILR Plenary Presentation
National Foreign Language Center
December 13, 2013

Agenda

- Goals and context
- Summit summary
- Sample activity
- Current state of the Listening SLDs
- Preliminary findings
- Future areas for research
- Research since the summits
- Recommendations and next steps

Summit Goals

- Raise Awareness of ILR Listening Descriptions
- Examine the adequacy of these descriptions
- Build a Community of Interest in furthering our Understanding of Listening
- Develop a common understanding of the ILR construct of Listening Comprehension

Contexts

- ILR is indispensable as common metric in measuring foreign language proficiencies of US Government personnel
- ILR provides common metric for test and curriculum development
- ILR serves as basis for testing with authentic texts and real world tasks

Summary of the Summits

- There were three (2009, 2011, 2011)
- Diverse group of attendees representing USG agencies and contractors
- Each Summit had a different focus
  - A critical look at the current ILR SLDs – Listening
  - Identifying factors that impact listening comprehension
  - Produce by consensus a written document

Sample Listening Activity

ILR Summits on Listening
Sample Activity

ACTFL

Directions: Listen to the passage and complete both charts.

Speaker's Purpose	Content/context	Salient Listening Characteristics that affect this text

Level	What will a listener at each level understand about this text?
L1	
L2	
L3	
L4	
L5	

Current State of the Listening SLDs

Usefulness of Listening SLDs

- Provide a common metric for test development and curriculum
- Make the distinction between participatory, non-participatory and overheard listening
- Set expectations for what a listener can and cannot do

Difficulties in Using the Listening SLD's

- They are too connected to speaking yet too derivate of reading
- They lack examples
- Understanding of key terms needs to be tightened in order to more clearly differential levels
- They lack empiricism
- Relationship between authentic texts and noise
- Can-do versus cannot
- Interaction of text length and level

Missing from the Listening SLDs

- Defining different listeners
  - Native, WENL, heritage, etc.
- Accounting for listening-specific factors
  - cognitive load and memory
  - Technology
- Language, dialect, regionalism
- Conditions that impact listening
  - Noise, loudspeakers, mumbling speaker
  - Non-native accent

Factors That Appear at Some Levels but Should Appear at Other Levels

- Noise
- Participative vs. Non-Participative
- Overheard
- Adverse Listening Conditions
- Effect of length of spoken language
- Rate of Speech
- Utterance Length
- Interference from Native Language
- Mastery
- Emotional Overtones
- Dialect
- Tension and pressure
- Socio-Linguistic/Cultural References
- Repetition
- Context of Communication
- Purpose of Text

Preliminary Findings

Preliminary Findings

- The existing descriptions can continue to be used in their current state to describe listening comprehension according to the ILR provided that there is a consensus regarding the definition of terms, qualifiers, and context/content domains
- Findings to serve as a basis for decision making if and when the SLDs-L undergo revision
- The findings can also be interpreted in terms of their relevance for test design specifications.

Salient Features of Each Level

- Salient features characteristic of each level and that distinguished a particular level from other levels
- Organized by task, accuracy, context and content area

Salient features for Listening at ILR Level 1

Tasks	Accuracy	Context	Content Areas
Understand short and simple utterances (e.g., statements, questions) that contain high frequency vocabulary and structural patterns Understand redundant and repetitive language of literal and factual information that is purpose specific Understand a standard dialect and accent	Most accurate with high frequency, predictable speech presented in simple sentences with topical cohesion Accuracy improves when meaning is supported by context Ability to understand may be inconsistent depending on content/context factors	In participative context: requires a sympathetic interlocutor for routine exchanges and simple transactions In non-participative context: announcements, radio or TV broadcast, eavesdropping	Self and immediate world Basic needs, travel, and courtesy requirements Survival related areas: food, shelter, health, travel. security, weather Basic high frequency words, base nouns, action verbs, basic personal information, kinship terms

Level-specific Factors

- Linguistic and extra linguistic factors that may influence a listener's ability to understand at that level

Factors that Influence Listening at L1

- Automaticity
- Clarity of speech
- Cohesion
- Cultural load
- Directness of text
- Familiarity of topic; previous knowledge of topic
- Frequency of vocabulary
- Hesitation and pause
- Information density
- Length of passage and listener fatigue
- Number of speakers
- Purpose of listening text
- Rate of speech
- Redundancy
- Sentence structure
- Syntactic complexity
- Unfavorable listening conditions
- Visual support

Challenges in Using ILR SLDs for Test Development

- The various listening modes are not addressed specifically and across levels
- Differences between proficiency, performance, achievement are not always clear
- The definitions of terms are not unilaterally accepted
- No consistent mention of the impact of factors such as memory load, cultural load, tension and pressure, density of information, length of passage, etc.
- Do not address the types of redundancy that are expected at different levels
- Appropriateness of multiple choice options at the higher levels
- Is there a hierarchy of tasks for Listening?

Future Areas for Research

- Language-specific research
- Listening for different populations and different types of learners
- Impact of length, density, and organization
- Effect of unfavorable or adverse listening conditions
- Relationship between proficiency level and memory load; proficiency level and cultural load
- Is there a hierarchy of listening tasks?
- Best practices for designing test specifications
- Testing culture as a component of listening comprehension

Research Since the Summits

Major Steps in Creating C-R Listening Proficiency Tests - BYU

1. Defining the construct to be tested.
2. Using the proficiency scales to describe Task, Conditions, and Accuracy (TCA) criteria for each level to be tested.
3. Training the items writers to develop C-R, TCA-aligned items that target a specific level.
4. Testing whether the C-R, TCA-aligned item sets cluster in a difficulty hierarchy.
5. Assembling tests by items

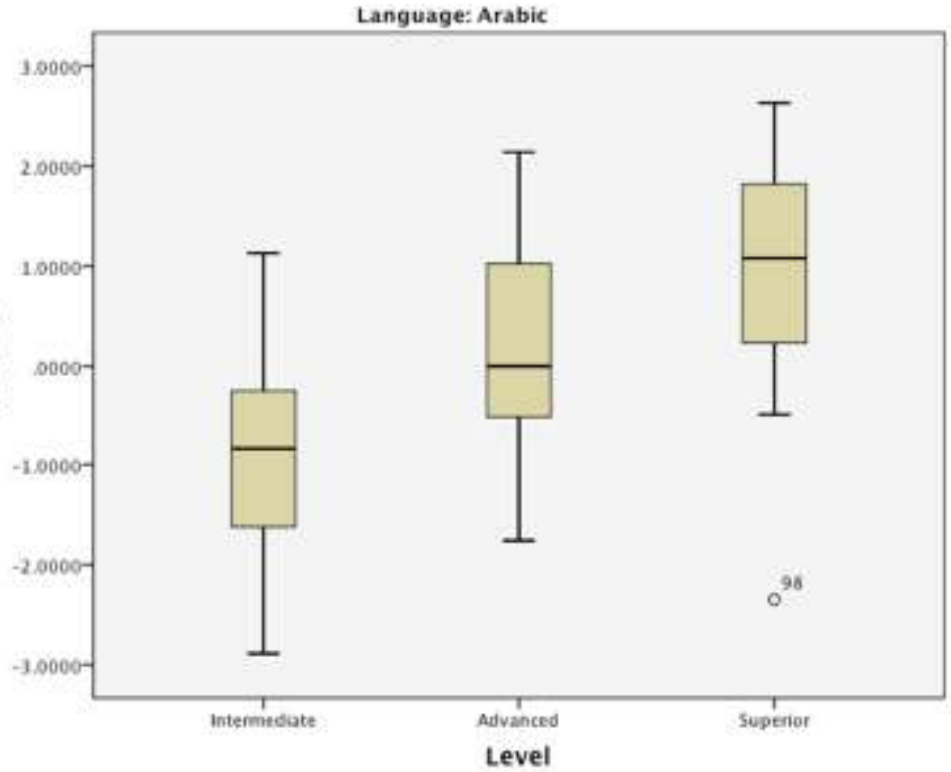
Empirical Evidence

- This item process was used in…
  - Arabic
  - Chinese
  - English
  - Russian
  - Spanish
  - Turkish

The following slides will report the PILOT results of these tests.

Distribution of Arabic Item Difficulty by Intended Proficiency Level

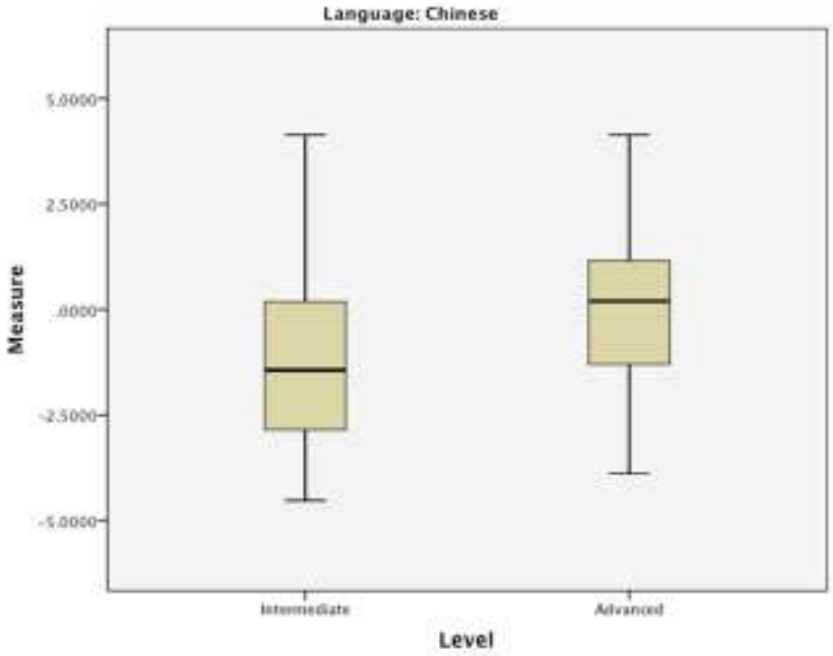
	# of Items in Test Bank	Mean Logit	SD
Intermediate	36	-.92	.94
Advanced	36	.16	.95
Superior	28	1.12	1.12
Total Items	**100**		



Comparisons using Bonferroni's contrasts found statistical differences between the Intermediate and Advanced Items (mean difference =-1.08 logits, a 95% CI between -1.65 and -.51, and $p < .001$); and between the Advanced and Superior items (mean difference = -.83 logits, a 95% CI between -1.44 and -.21, and $p < .01$).

Distribution of Chinese Item Difficulty by Intended Proficiency Level

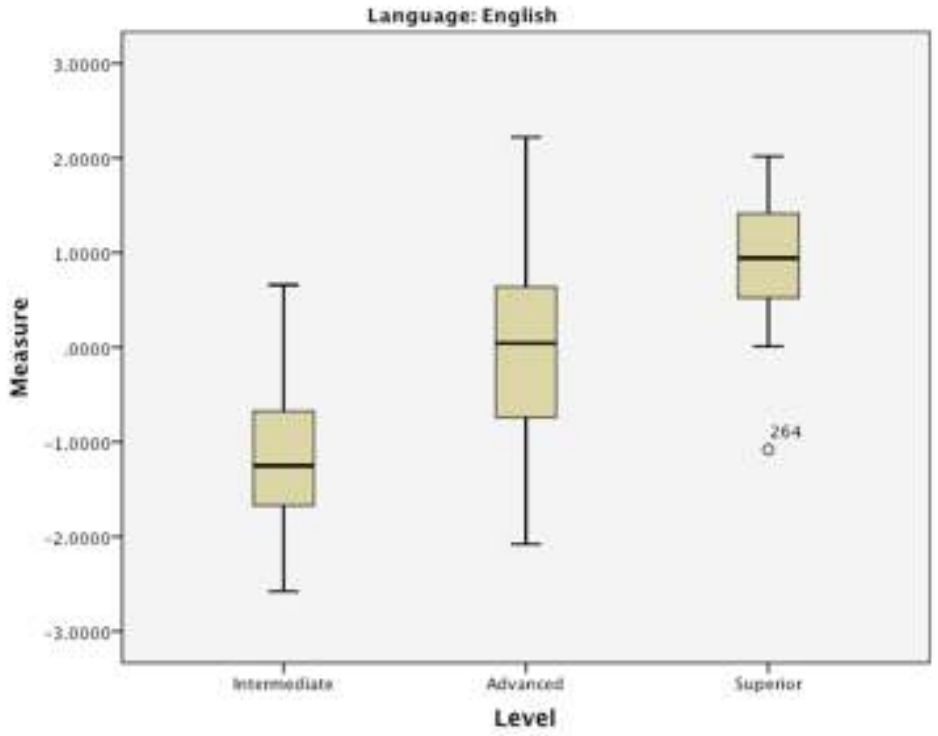
	# of Items	Mean Logit	SD
Intermediate	35	-1.12	2.37
Advanced	47	.07	1.76
Total Items	**82**		



An independent samples T-test between the Intermediate and Advanced Level items was conducted. The 95% CI for the difference in the means was between -3.73 and -2.38 ($t = -9.08$, $p < .001$, $df = 60$).

Distribution of English Item Difficulty by Intended Proficiency Level

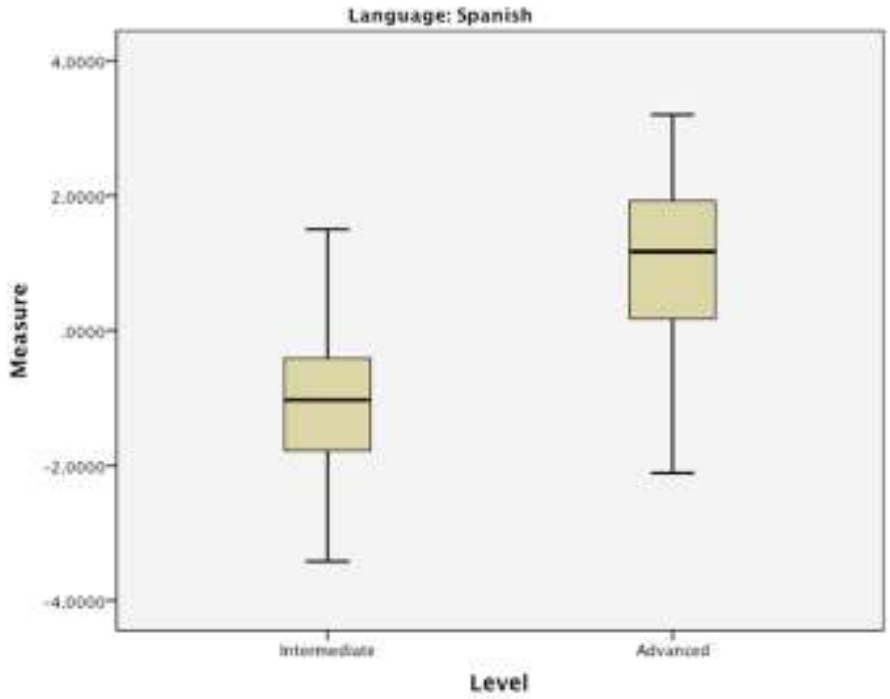
	# of Items in Test Bank	Mean Logit	SD
Intermediate	20	-1.13	.83
Advanced	38	.02	.94
Superior	24	.90	.72
Total Items	**82**		



Comparisons using Bonferroni's contrasts found statistical differences between the Intermediate and Advanced Items (mean difference =-1.15 logits, a 95% CI between -1.73 and -.57, and $p < .001$); and between the Advanced and Superior items (mean difference = -.88 logits, a 95% CI between -1.42 and -.33, and $p < .001$).

Distribution of Spanish Item Difficulty by Intended Proficiency Level

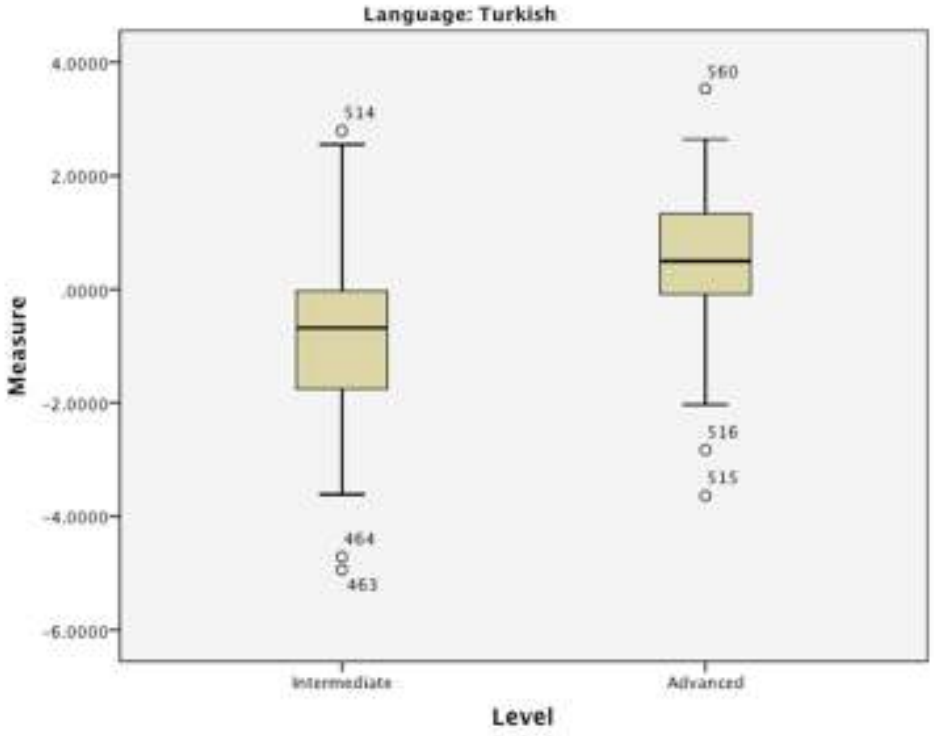
	# of Items	Mean Logit	SD
Intermediate	35	-1.03	1.12
Advanced	39	.93	1.41
Total Items	**74**		



An independent samples T-test between the Intermediate and Advanced Level items was conducted. The 95% CI for the difference in the means was between -3.54 and -2.63 ($t = -13.60$, $p < .001$, $df = 53$).

Distribution of Turkish Item Difficulty by Intended Proficiency Level

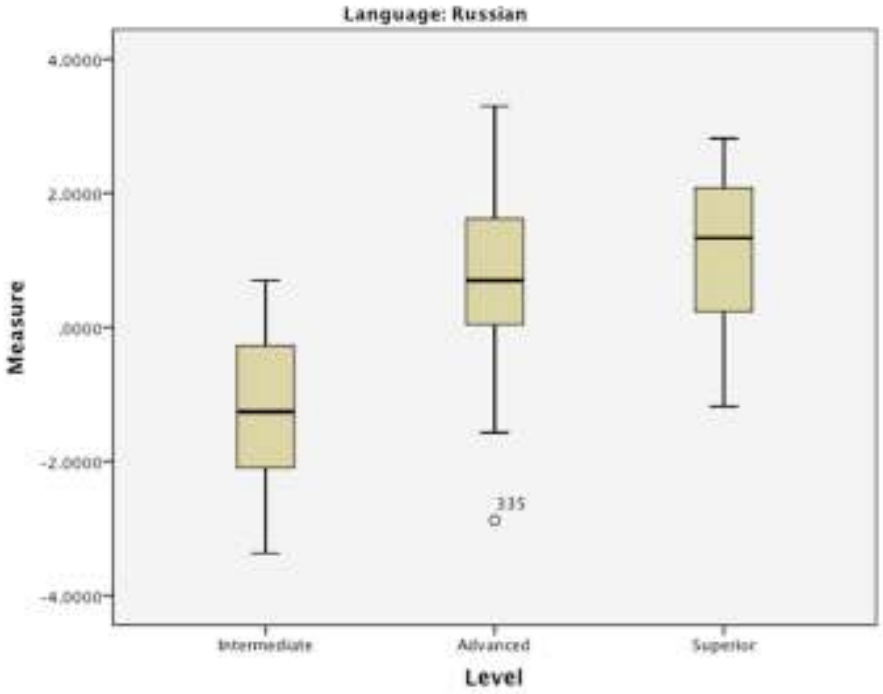
	# of Items in Test Bank	Mean Logit	SD
Intermediate	52	-.99	1.58
Advanced	46	.49	1.30
Total Items	**98**		



An independent samples T-test between the Intermediate and Advanced Level items was conducted. The 95% CI for the difference in the means was between -3.12 and -2.07 ($t = -9.94$, $p < .001$, $df = 71$).

Distribution of Russian Item Difficulty by Intended Proficiency Level

	# of Items	Mean Logit	SD
Intermediate	51	-1.26	1.17
Advanced	49	.79	1.26
Superior	24	1.05	1.10
Total Items	**124**		



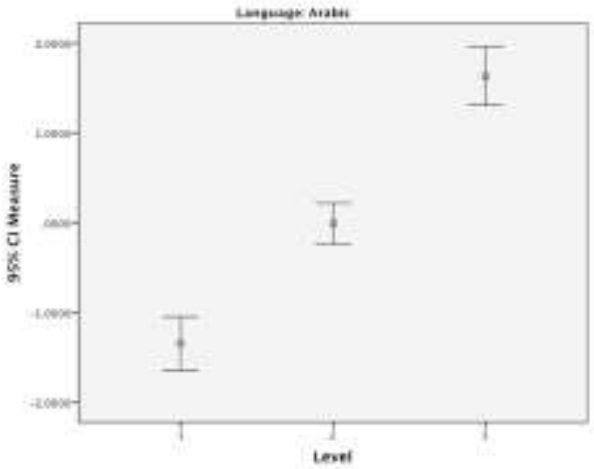
Comparisons using Bonferroni's contrasts found statistical differences between the Intermediate and Advanced Items (mean difference =-2.05 logits, a 95% CI between -2.63 and -1.47, and p < .001). However there was not a difference between the Advanced and Superior items (mean difference = -.25 logits, a 95% CI between -.98 and -.46).

Test Assembly

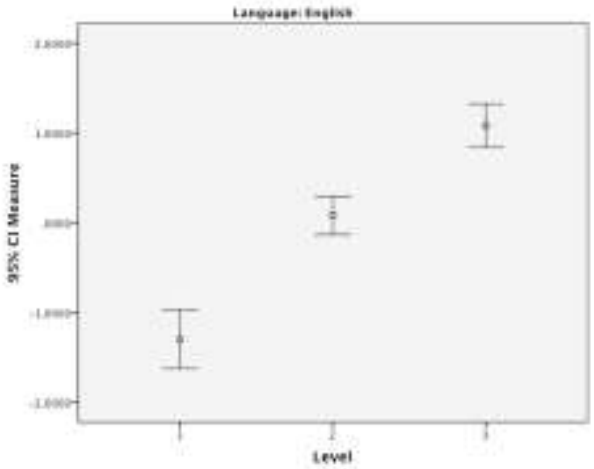
- Send 25% of the poorest functioning items back to item writers
- Re-analyze data and look at means with 95% Confidence Interval Measures

95% CI Mean Logit Item Difficulty by Level in Listening Proficiency Tests

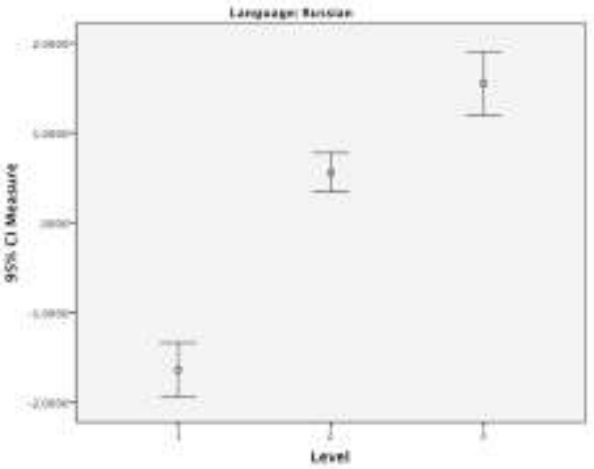
Arabic



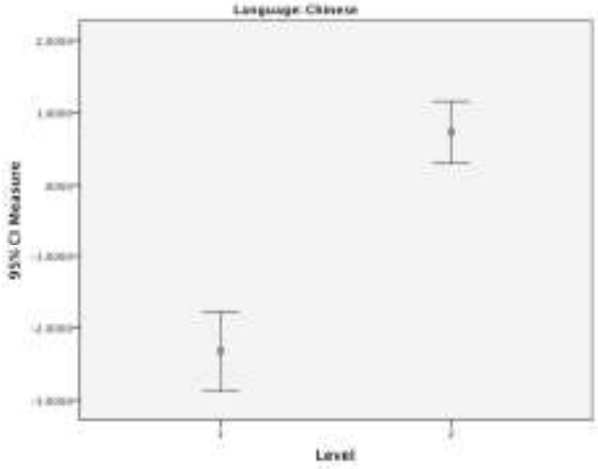
English



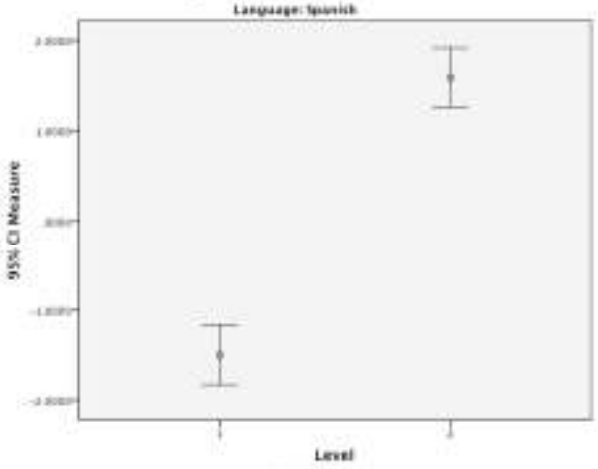
Russian



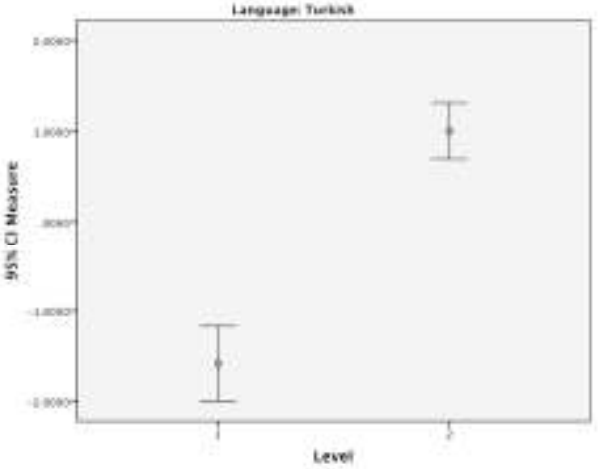
Chinese



Spanish



Turkish



Conclusion

When proficiency test items are constructed so that both the passage selected and the question asked are based on and aligned with the ACTFL Listening Proficiency Guidelines, those test items do cluster in an ascending hierarchy of difficulty levels.

CASL's Second Language Listening Research

- **Native Arabic and Spanish speakers learning English**
- **Targeted level = ILR 2/2+**
- **Variables:**
  - **Versant™ score (English listening proficiency)**
  - **Length in syllables**
  - **Type/Token Ratio**
  - **Working Memory capacity**
  - **Accent similarity to L1 (w.r.t. stress patterns)**
  - **Accent familiarity to the listener**
  - **Ability to take notes while listening**

Recommendations and Next Steps

Recommendations

- Plan for revision of the Listening SLDs
- Prioritize and co-select of all of the information generated from the Summits with the goal of determining what information is essential and should be included in a revision
- Reference the salient features that differentiate one level from another (i.e., the tasks, accuracy expectations, contexts and content areas that are unique to the level) as well as the conditions and factors that influence listening at each level are a critical starting point

What's Next for the Listening SLDs?

- Revise ILR Listening Skill Level Descriptions
  - Small steering committee to do initial drafting
  - Reviewed by ILR Testing Committee
  - Use ILR community for review and feedback

Questions?