CONTRIBUTING TO A VALIDATION ARGUMENT FOR THE ILR SKILL LEVEL DESCRIPTIONS FOR PROFICIENCY

Rachel L. Brooks, FBI
Clayton Leishman, DLIELC
REVISING THE ILR SLDs FOR PROFICIENCY

Process
- Interagency committee met monthly from 2014 – 2020
- Developed a crosswalk matrix and prose forms, with iterative feedback from agencies
- Circulated to USG stakeholders for feedback

Goals
- To clarify and update the SLDs
- To retain the underlying construct of the SLDs without shifting the difficulty of the levels
- To complete the SLDs with consistency across the modalities and levels
- To incorporate current research and updated language testing concepts
- To develop a validation framework for US Government use
SELECTED RESEARCH REFERENCES


VALIDITY AS A UNITARY CONCEPT (MESSICK, 1989)

Traditional validity
- Face validity
- Content validity
- Context validity
- Concurrent validity

Scoring validity
- Rater reliability

Test fairness and validity

Rater responses are central to test validity (Weir, 2005)
ARGUMENT-BASED VALIDITY (KANE, 2006)

Scoring validity
• Rater reliability

Scoring inference
• Appropriate administration conditions
• Effective rater training
• Sound rating criteria
### Inferences and Their Associated Claims Expressing Their Meanings (Knoch & Chapelle, 2018, p.35)

<table>
<thead>
<tr>
<th>Inference</th>
<th>Claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
<td>Observations are evaluated using procedures that provide observed scores with intended characteristics.</td>
</tr>
<tr>
<td>Generalization</td>
<td>Observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters.</td>
</tr>
<tr>
<td>Explanation</td>
<td>Expected scores are attributed to the defined construct.</td>
</tr>
<tr>
<td>Extrapolation</td>
<td>The construct of the assessment sufficiently accounts for the quality of linguistic performances in the target language use (TLU) domain.</td>
</tr>
<tr>
<td>Decision</td>
<td>Decisions made based on the estimates of the quality of the performance are appropriate and well communicated.</td>
</tr>
<tr>
<td>Consequence</td>
<td>Test consequences are beneficial to users.</td>
</tr>
</tbody>
</table>
BUILDING A VALIDITY ARGUMENT FOR THE ILR SLDS

• To be determined by the USG organization

• To be determined by the USG organization

• To be determined by the USG organization

• Support given from the ILR SLD Revisions Committee on theory and processes of revisions

• Evidence gathered through the ILR SLD Pilot Study

• Evidence gathered through the ILR SLD Pilot Study
The scale properties are as intended by the developers.

Assumption:
- Scale criteria (in case of an analytic scale) can be shown to be assessing separate abilities as hypothesized.
- Scale steps are adequate to distinguish among the levels that appear in the scale.
- The scale is able to spread test takers into different levels as needed for the test purpose.

Claim:
- The ILR SLD Speaking abilities and sub-abilities assess separate features of speaking.
- The main levels of the ILR SLDs (0, 1, 2, 3, 4, 5) are distinct from each other.
- USG speaking tests cover the spread of proficiency abilities of USG personnel and speakers exist at all levels.
EVALUATION INFERENCE

Assumption

- Raters are able to identify differences in performances across score levels.
- Raters can consistently apply the scale to test tasks.
- Raters are comfortable when applying descriptors and confident in their decisions.
- Detectable rater characteristics do not introduce systematic construct-irrelevant variance into task ratings above acceptable levels set by the test designer.

Claim

- USG speaking raters can reliably identify examinees' levels across the entire ILR SLDs.
- USG speaking raters rate reliably using the ILR SLDs for Speaking.
- USG speaking raters are comfortable and confident in giving ratings using the ILR SLDs for Speaking.
- USG speaking raters' characteristics do not introduce construct-irrelevant variance to ILR ratings.

Warrant

- Raters rate reliably at task level.
GENERALIZATION INFERENCE

**Warrant**
Different raters assign the same ratings to responses.

**Assumption**
- Raters rate consistently at the whole test level.
- The number of raters is sufficient to arrive at a reliable score.

**Claim**
- USG speaking raters are reliable at assigning final ratings.
- Using 2 individual ratings is sufficient for ILR Speaking ratings.
ILR SPEAKING PILOT STUDY
ILR SPEAKING PILOT STUDY

Organizers
- ODNI FLEXCOM TAEG ISS participants in 2018

Purpose
- To help build a validation argument for the ILR SLD speaking revisions.
- To examine whether there is a score shift resulting from the updates of the ILR SLD speaking revisions.

Research Questions
1. Can USG speaking raters reliably identify examinees' levels across the entire ILR SLDs?
2. Are USG speaking raters more confident in giving ratings using the revised ILR SLDs for Speaking?
3. Do the ILR SLD Speaking abilities and sub-abilities, as outlined in the ILR Crosswalk Matrix, assess separate features of speaking?
4. Is there any patterned shift of scores on tests rated on the current scale vs the revised scale?
PROCEDURE

Speaking testers from USG agencies will rate speaking tests from their own agency including a wide range of ILR levels.

Prior to rating, they will undergo interagency training to familiarize them with the new ILR SLDs, including:

- Six hours of an online, live workshop with testers from other organizations
- One practice test in their test language with the USG organization trainer.

After rating each test, they will complete the Individual Rater Report, which includes:

- Holistic ratings
- Ability ratings
- Sub-ability ratings
- Justifications
- Rating confidence levels
- Comparison to current SLD ratings

After all ratings are complete, final questionnaires and interviews are conducted.
<table>
<thead>
<tr>
<th>Language</th>
<th>ILI</th>
<th>DLIELC</th>
<th>DLIFLC</th>
<th>FBI</th>
<th>FSI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese – Mandarin</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Spanish</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>English</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>10</td>
<td>6</td>
<td>32</td>
</tr>
</tbody>
</table>

N = 32 experienced speaking test raters
N = 30 for Mandarin Chinese and Spanish raters, n = 40 for English raters

<table>
<thead>
<tr>
<th>Mandarin Chinese</th>
<th>DLI</th>
<th>FLC</th>
<th>FBI</th>
<th>FSI</th>
<th>ILI</th>
<th>Total Tests</th>
<th>Total Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Level</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>60</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>(0+-2+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper-Level</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>60</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>(3-5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>120</td>
<td>360</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English</th>
<th>DLI</th>
<th>ELC</th>
<th>FBI</th>
<th>Total Tests</th>
<th>Total Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower-Level</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>(0+-2+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper-Level</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>(3-5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>40</td>
<td>80</td>
<td>320</td>
<td></td>
</tr>
</tbody>
</table>
INDIVIDUAL RATER REPORT

ILR SPEAKING PILOT STUDY

Rater Information
- Rater ID
- Agency
- Rating Date
- Test ID
- Language

Preliminary Score
Complete the bracketing exercise using the prose form of the SLDs and then enter a holistic ILR Speaking rating.

How confident are you of the rating?
- Not confident
- Somewhat confident
- Confident
- Very confident

Score Justification
Set the SLDs aside and in your own words, describe aspects of the examinee’s performance that justify the rating.

Provide a statement from the next higher SLD that illustrates how the sample does not meet or fully meet that next higher SLD.
**Ability Review**

Without referring to the SLDs, **in your own words**, describe the examinee’s performance in the following ability and sub-abilities.

<table>
<thead>
<tr>
<th>Functional Ability</th>
<th>Discourse Management</th>
<th>lexical Control</th>
<th>Structural Control</th>
<th>Phonetic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision of Forms and Meanings</strong></td>
<td>Range</td>
<td>Relevance</td>
<td>Substantive Coverage</td>
<td></td>
</tr>
<tr>
<td><strong>Content Meaningfulness</strong></td>
<td>Cultural Appropriateness</td>
<td>Social Appropriateness</td>
<td>Interactional Appropriateness</td>
<td></td>
</tr>
</tbody>
</table>
**INDIVIDUAL RATER REPORT**

Consult the ILR Speaking Matrix and assign a rating for each ability and sub-ability.

<table>
<thead>
<tr>
<th>Functional Ability</th>
<th>Precision of Forms and Meanings</th>
<th>Content Meaningfulness</th>
<th>Contextual Appropriateness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse Management</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonetic Features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substantive Coverage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultural Appropriateness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Appropriateness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriateness</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Final Score**

Complete the bracketing exercise using the prose form of the SLDs and then enter a holistic ILR Speaking rating.

- **Final Holistic ILR Rating**
- **How confident are you in the rating?**
  - Not confident
  - Somewhat confident
  - Confident
  - Very confident

Would your rating be different if you used the current ILR SLDs – Speaking?

- **Yes**
- **No**

If yes, what would your rating be?

**Any comments?**
QUESTIONNAIRES

Rater:

Perceptions:
1. Which version is clearer?
2. Which version is more complete?
3. Which version is easier to use?

Background Information:
Organization, gender, languages, education, experience in testing and teaching

Trainer Interview:
1. What was your experience in using the revised 2020 ILR SLDs for rating?
2. Are the categories and subcategories of the Matrix clear?
3. Did the training sufficiently prepare you to use the revised 2020 ILR SLDs?
4. On a scale of 1 to 5, with 1 being not well at all and 5 being very well, how well do the revised 2020 ILR SLDs differentiate between ILR base levels?
5. Level by level, compare the current 1985 version of the ILR SLDs with the revised 2020 version.
LINKING THE DATA TO THE CLAIM

The ILR SLD Speaking abilities and sub-abilities assess separate features of speaking.

1. IRR: Ability and sub-ability ratings; factor analysis
2. Which version is more complete? Are the categories and subcategories of the Matrix clear?

The main levels of the ILR SLDs (0, 1, 2, 3, 4, 5) are distinct from each other.

1. IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; chi square
2. How well do the revised 2020 ILR SLDs differentiate between ILR base levels? Plus levels?

USG speaking tests cover the spread of proficiency abilities of USG personnel and speakers exist at all levels.

IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; factor analysis
LINKING THE DATA TO THE CLAIM

USG speaking raters can reliably identify examinees' levels across the entire ILR SLDs.

1. IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; G-theory, factor analysis
2. IRR: Rating confidence levels

USG speaking raters rate reliably using the ILR SLDs for Speaking.

1. IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; rater reliability
2. IRR: Rating justifications

USG speaking raters are comfortable and confident in giving ratings using the ILR SLDs for Speaking.

1. IRR: Rating confidence levels
2. Overall, how would you describe your level of confidence when assigning a score using the new ILR SLDs by the end of the study?
<table>
<thead>
<tr>
<th>Claim</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>USG speaking raters' characteristics do not introduce construct-irrelevant variance to ILR ratings.</td>
<td>1. Rater characteristics and IRR: final ILR ratings;</td>
</tr>
<tr>
<td>USG speaking raters' characteristics do not introduce bias to ILR ratings.</td>
<td>2. IRR: Rating justifications</td>
</tr>
<tr>
<td>USG speaking raters are reliable at ILR base levels.</td>
<td>IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; rater reliability statistics</td>
</tr>
<tr>
<td>IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5</td>
<td>IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; G-theory</td>
</tr>
</tbody>
</table>
# ILR Speaking Validation Study Timeline

<table>
<thead>
<tr>
<th>Dates</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 2020</td>
<td>Raters attend the two-day, six-hour training and rate one practice test..</td>
</tr>
<tr>
<td>November 2020 — January 2021</td>
<td>Raters rate all tests individually and submit reports. Then, questionnaires are completed.</td>
</tr>
<tr>
<td>February — March 2021</td>
<td>Data is compiled and analyzed.</td>
</tr>
<tr>
<td>April 2021</td>
<td>Results and conclusions completed.</td>
</tr>
</tbody>
</table>
QUESTIONS?

Thank you!