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

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational measurement: issues and practice*, *23*(4), 31-31.

Grabowski, K. C. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking* (Doctoral dissertation, Teachers College, Columbia University).

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of educational Research*, *64*(3), 425-461.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477-499.

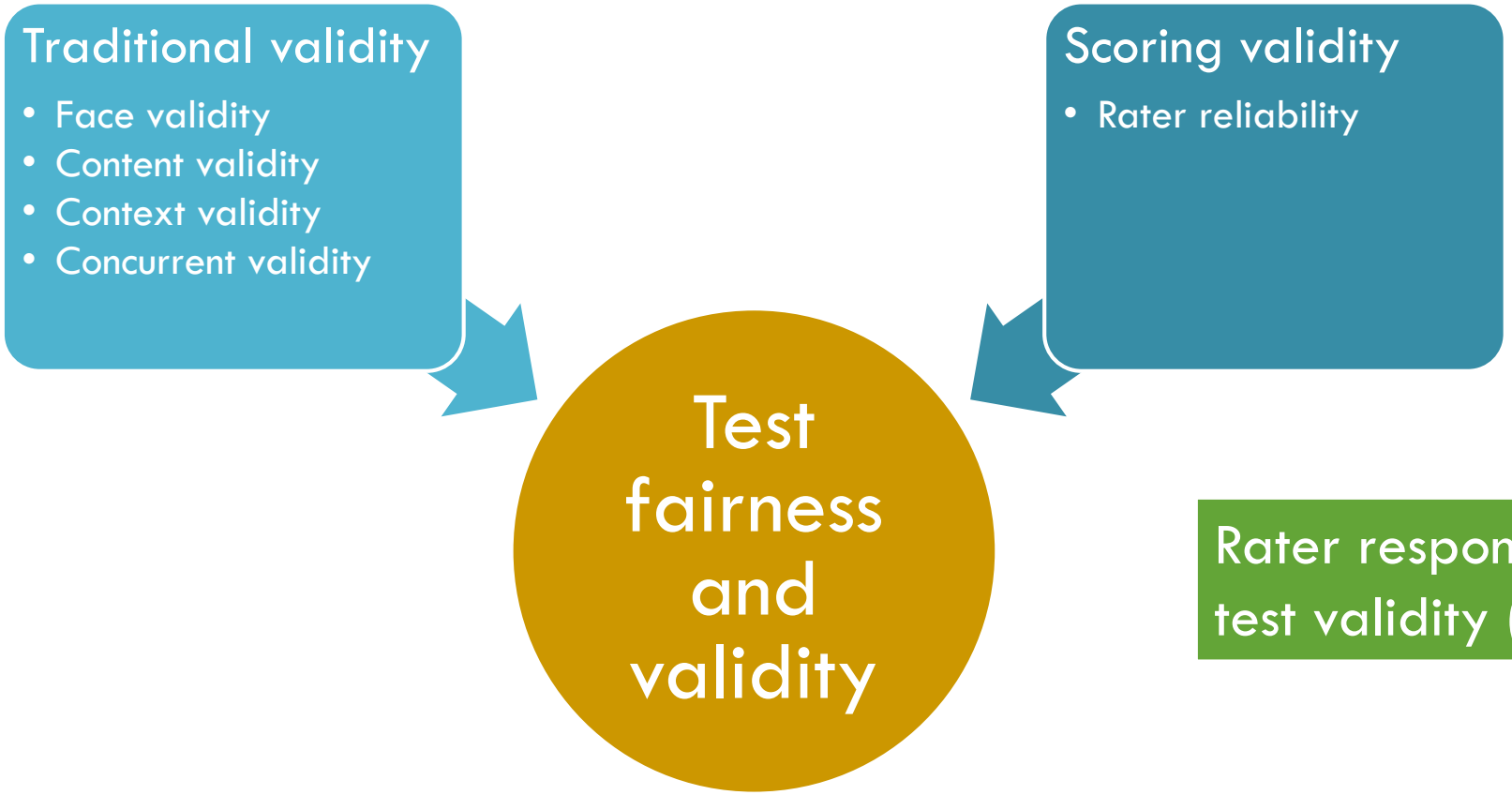
Mislevy, R. J. (2012). Modeling language for assessment. *The encyclopedia of applied linguistics*.

Mislevy, R. J., & Yin, C. (2009). If language is a complex adaptive system, what is language assessment?. *Language Learning*, *59*, 249-267.

Plake, B. S. (2008). Standard setters: Stand up and take a stand!. *Educational Measurement: Issues and Practice*, *27*(1), 3-9.

Purpura, J. E. (2016). Assessing meaning. Shohamy et al. (eds.), Language Testing and Assessment, Encyclopedia of Language and Education, 1-26.

VALIDITY AS A UNITARY CONCEPT (MESSICK, 1989)



Rater responses are central to test validity (Weir, 2005)

ARGUMENT-BASED VALIDITY (KANE, 2006)

Scoring validity

- Rater reliability

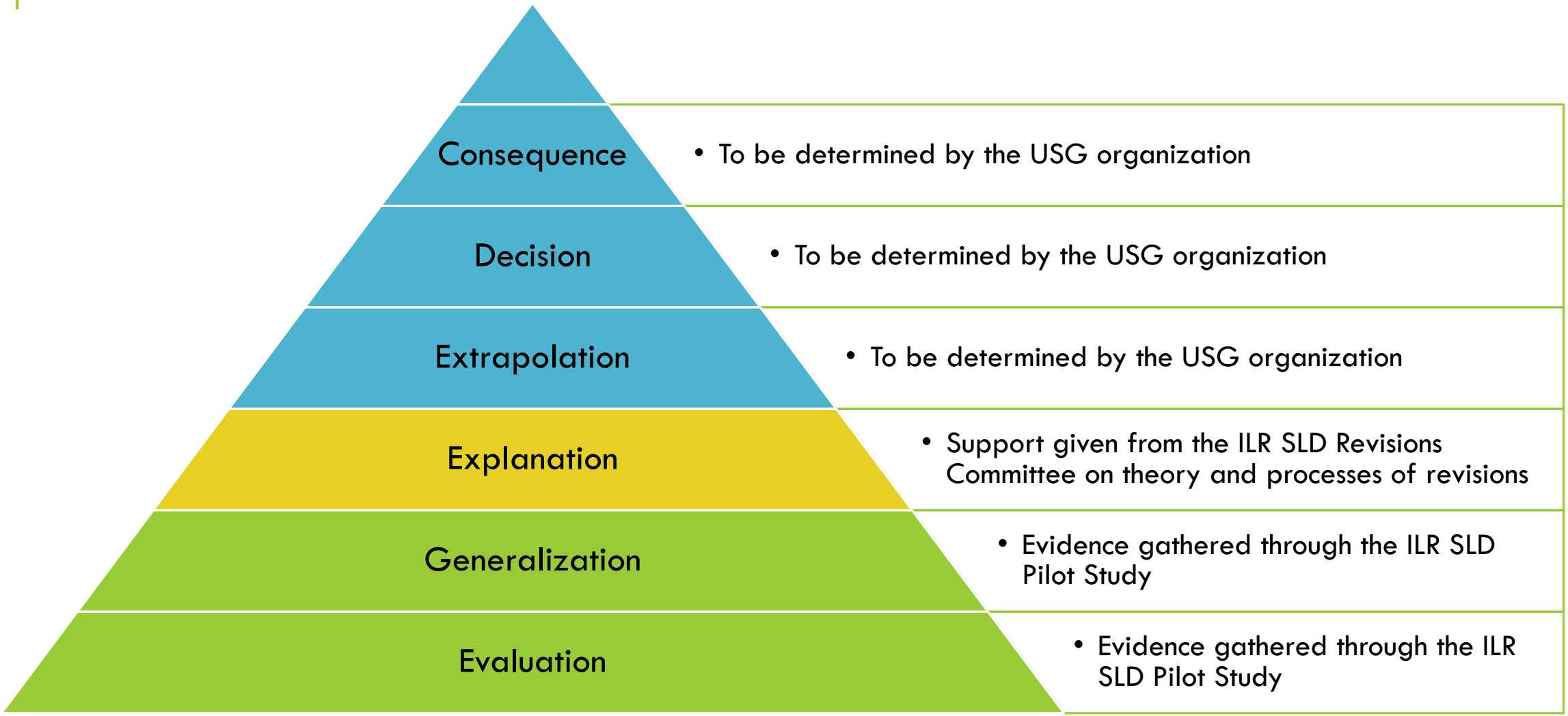
Scoring inference

- Appropriate administration conditions
- Effective rater training
- Sound rating criteria

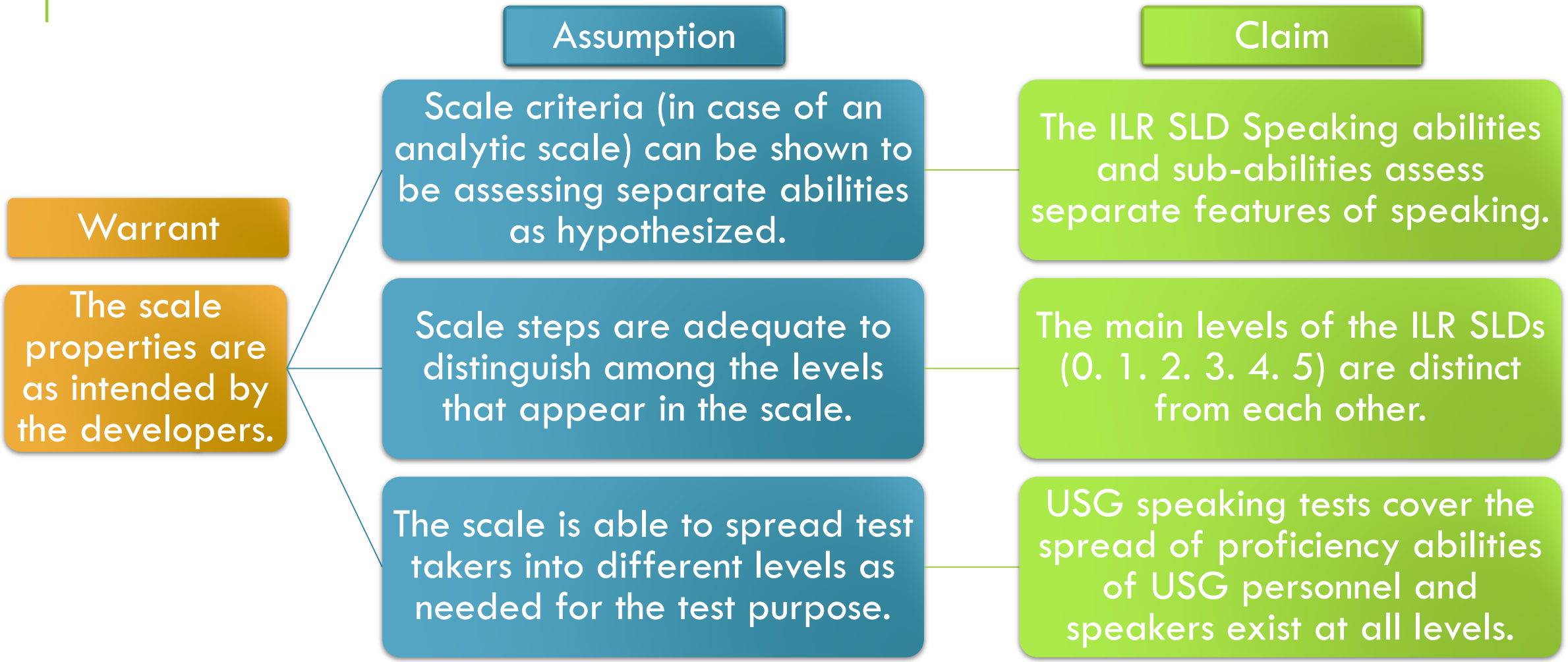
INFERENCES AND THEIR ASSOCIATED CLAIMS EXPRESSING THEIR MEANINGS (KNOCH & CHAPELLE, 2018, P.35)

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

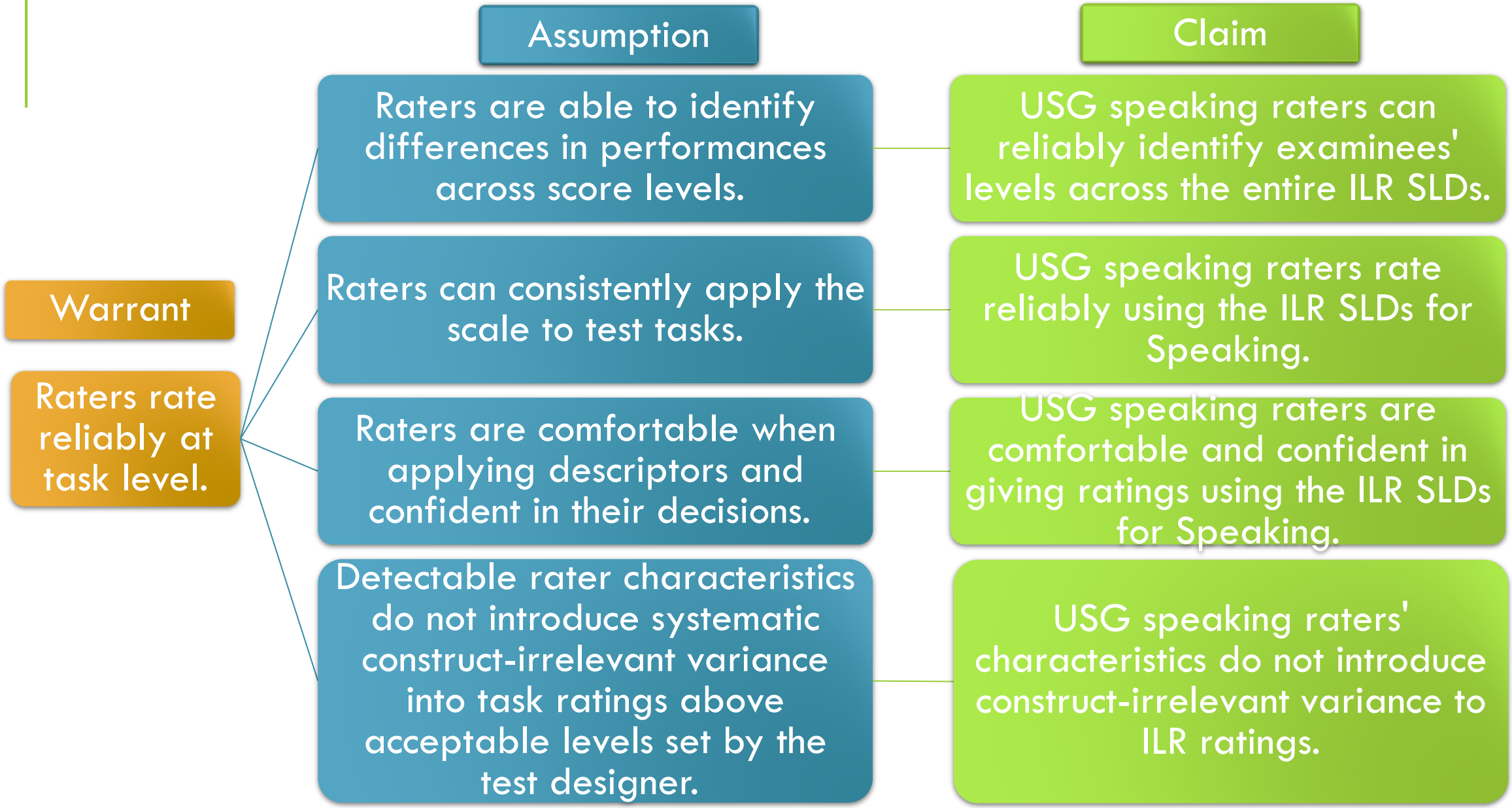
BUILDING A VALIDITY ARGUMENT FOR THE ILR SLDS



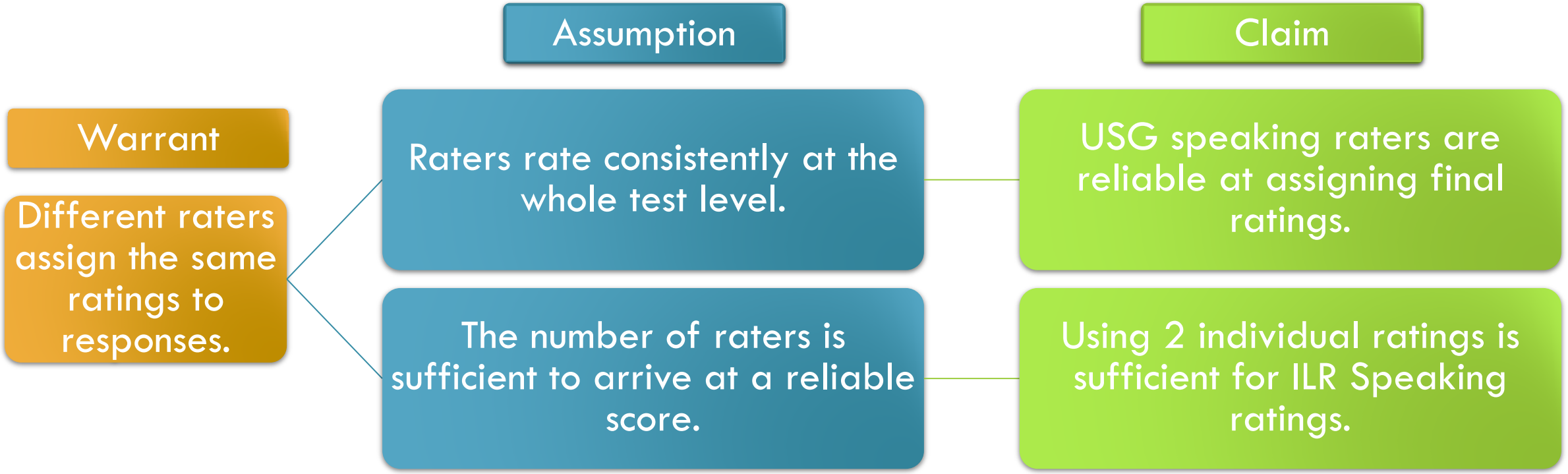
EVALUATION INFERENCE



EVALUATION INFERENCE



GENERALIZATION INFERENCE



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.

PARTICIPANTS

N = 32 experienced speaking test raters

Language	ILI	DLIELC	DLIFLC	FBI	FSI	Total
Chinese – Mandarin	3	0	3	3	3	12
Spanish	3	0	3	3	3	12
English	0	4	0	4	0	8
Total	6	4	6	10	6	32

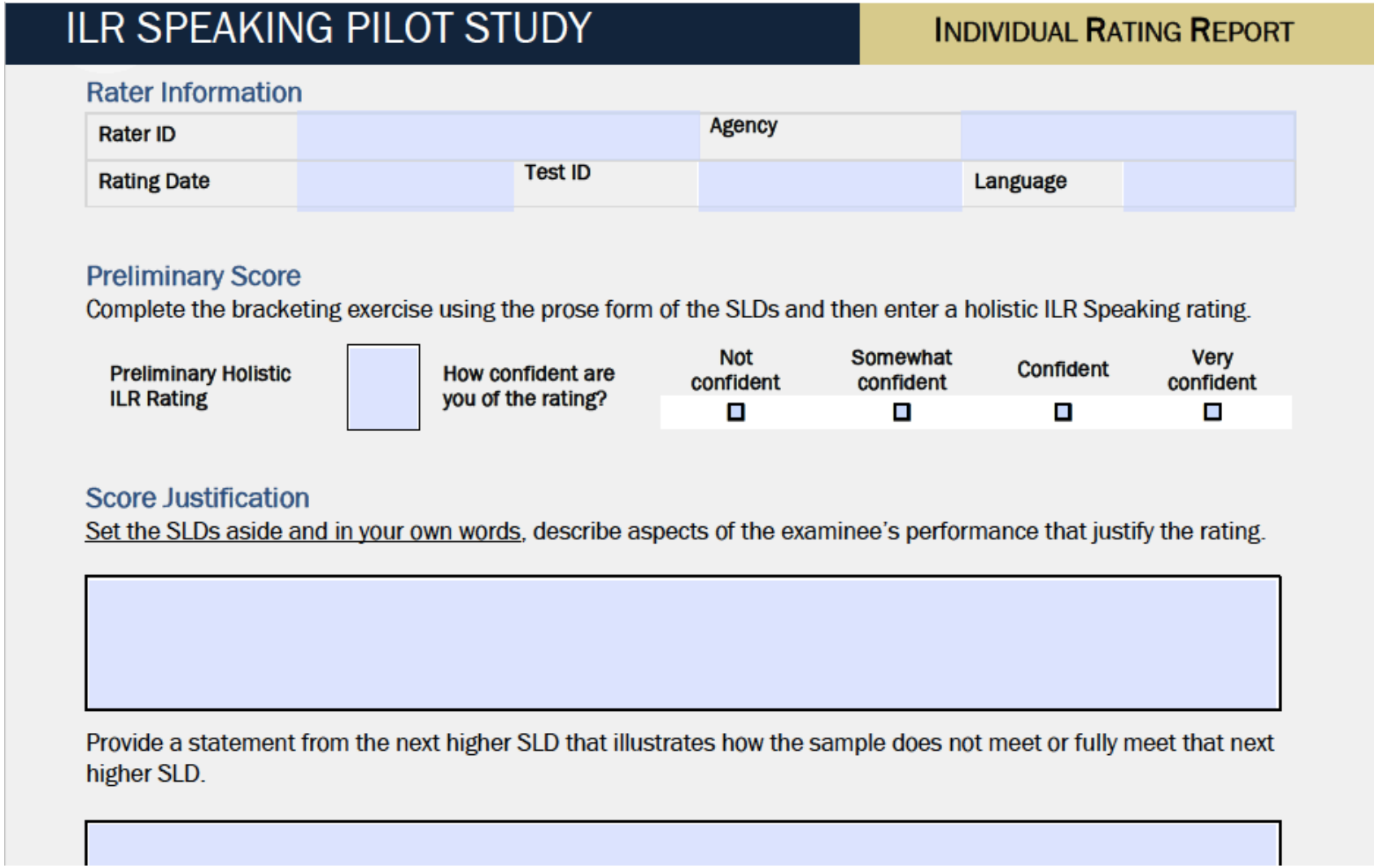
TESTS

N = 30 for Mandarin Chinese and Spanish raters, n = 40 for English raters

Mandarin Chinese Spanish	DLI FLC	FBI	FSI	ILI	Total Tests	Total Ratings
Lower-Level (0+-2+)	15	15	15	15	60	180
Upper-Level (3-5)	15	15	15	15	60	180
Total	30	30	30	30	120	360

English	DLI ELC	FBI	Total Tests	Total Ratings
Lower-Level (0+-2+)	20	20	40	160
Upper-Level (3-5)	20	20	40	160
Total	40	40	80	320

INDIVIDUAL RATER REPORT



ILR SPEAKING PILOT STUDY

INDIVIDUAL RATING REPORT

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.

Preliminary Holistic ILR 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.

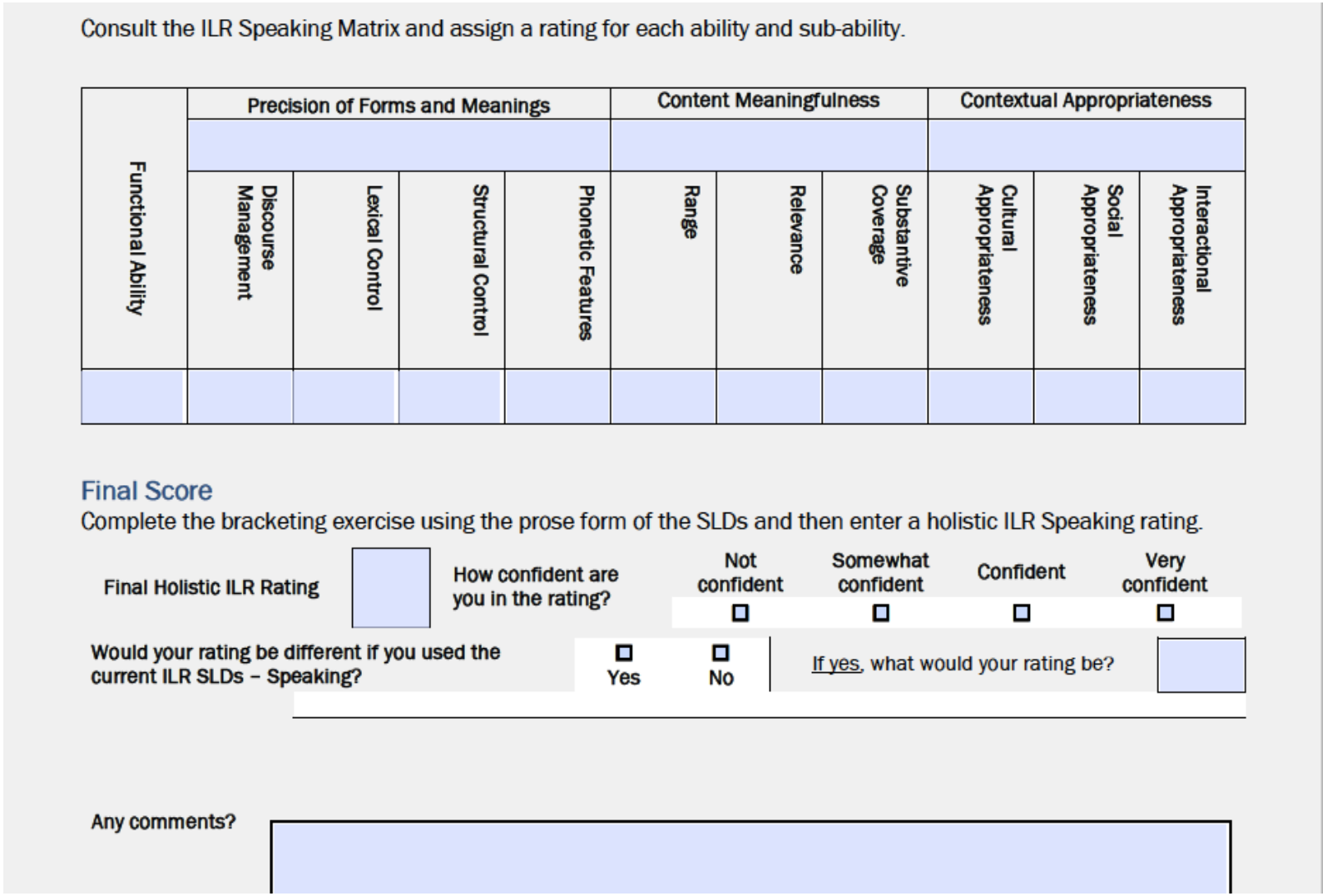
INDIVIDUAL RATER REPORT

Ability Review

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

Functional Ability		
Precision of Forms and Meanings	Discourse Management	
	Lexical Control	
	Structural Control	
	Phonetic Features	
Content Meaningfulness	Range	
	Relevance	
	Substantive Coverage	
Contextual Appropriateness	Cultural Appropriateness	
	Social Appropriateness	
	Interactional Appropriateness	

INDIVIDUAL RATER REPORT



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

Functional Ability	Precision of Forms and Meanings				Content Meaningfulness			Contextual Appropriateness		
	Discourse Management	Lexical Control	Structural Control	Phonetic Features	Range	Relevance	Substantive Coverage	Cultural Appropriateness	Social Appropriateness	Interactional Appropriateness

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?

LINKING THE DATA TO THE CLAIM

USG speaking raters' characteristics do not introduce construct-irrelevant variance to ILR ratings.	1. Rater characteristics and IRR: final ILR ratings;	2. IRR: Rating justifications
USG speaking raters' characteristics do not introduce bias to ILR ratings.	Rater characteristics and IRR: final ILR ratings	
USG speaking raters are reliable at ILR base levels.	IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; rater reliability statistics	
IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5	IRR: Final ILR ratings from tests across the range of ILR Levels, 0-5; G-theory	

ILR SPEAKING VALIDATION STUDY TIMELINE

Dates	Objective
October 2020	Raters attend the two-day, six-hour training and rate one practice test..
November 2020 – January 2021	Raters rate all tests individually and submit reports. Then, questionnaires are completed.
February – March 2021	Data is compiled and analyzed.
April 2021	Results and conclusions completed.

QUESTIONS? | Thank you!