ILR Skill Level Descriptions for Proficiency Revisions

Hanan Agag, Rachel Brooks, Beth Mackey, David Sawyer

ILR SLD Revisions Committee

May 14, 2021

ILR SLD Revisions Initiative

The ILR SLDs from 1985 to today

- Much has been learned about the ILR SLDs since they were approved in 1985 through trainings, collaboration, experience, and research, including:
  - Feedback from users (examinees, test developers, raters, trainers, managers)
  - ILR Speaking and Listening Summits
  - Interagency Speaking Comparability Study
  - Testing and Assessment Expert Group-sponsored Interagency Speaking Summits
  - Defense Language Testing Advisory Project (DELTAP)
  - Discussions with ACTFL and CEFR communities

The Case for Revisions

Why are we revising the Proficiency SLDs?

- They are over 30 years old. Language within the descriptions needed to be updated to reflect current use and needs.
- Different agency missions have evolved (intelligence, diplomatic, defense, judicial).
- A very different test taker population exists, including native and heritage speakers as well as learners.
- Interagency collaboration and resources have advanced and these should be reflected.
- Experience and feedback from testers/raters/test developers has shown us the SLDs strengths and weaknesses.
- Applied linguistics and language testing research have evolved.

ILR Revisions Committee

- Members include representatives from CIA, DLIELC, DLIFLC, DLI Washington, DLNSEO, FBI, FSI, NSA, Canadian FSI
  - Committee members are ILR SMEs for their organizations
- Members reviewed draft versions with their testing specialists and other stakeholders and brought valuable comments back to the ILR Revision Working Group
- Meetings occurred regularly over the past 7 years, both in person and virtually
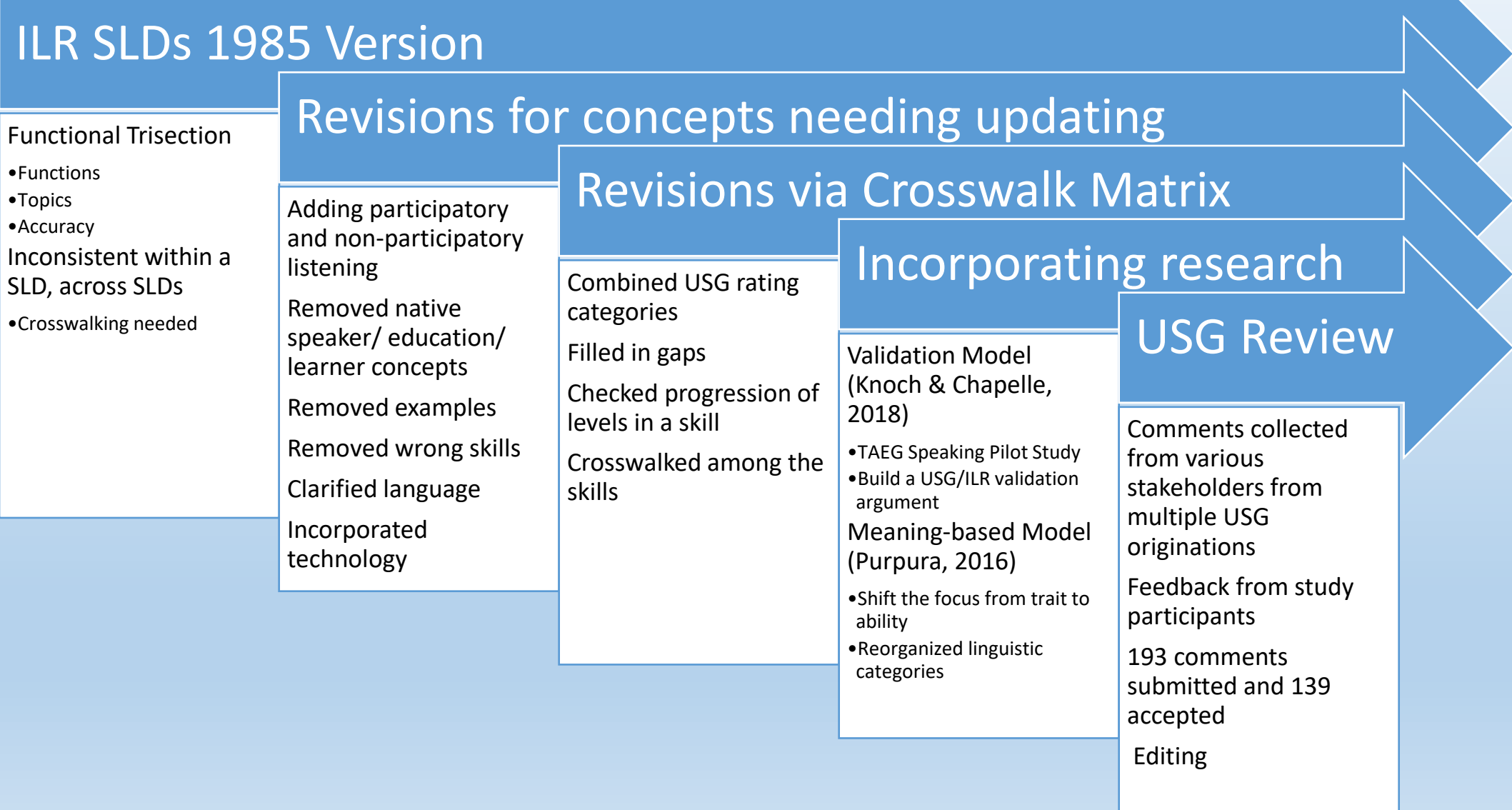
Revisions Process

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

Process

ILR SLDs 1985 Version

Functional Trisection

- Functions
- Topics
- Accuracy

Inconsistent within a SLD, across SLDs

- Crosswalking needed

Revisions for concepts needing updating

Adding participatory and non-participatory listening

Removed native speaker/ education/ learner concepts

Removed examples

Removed wrong skills

Clarified language

Incorporated technology

Revisions via Crosswalk Matrix

Combined USG rating categories

Filled in gaps

Checked progression of levels in a skill

Crosswalked among the skills

Incorporating research

Validation Model (Knoch & Chapelle, 2018)

- TAEG Speaking Pilot Study
- Build a USG/ILR validation argument

Meaning-based Model (Purpura, 2016)

- Shift the focus from trait to ability
- Reorganized linguistic categories

USG Review

Comments collected from various stakeholders from multiple USG originations

Feedback from study participants

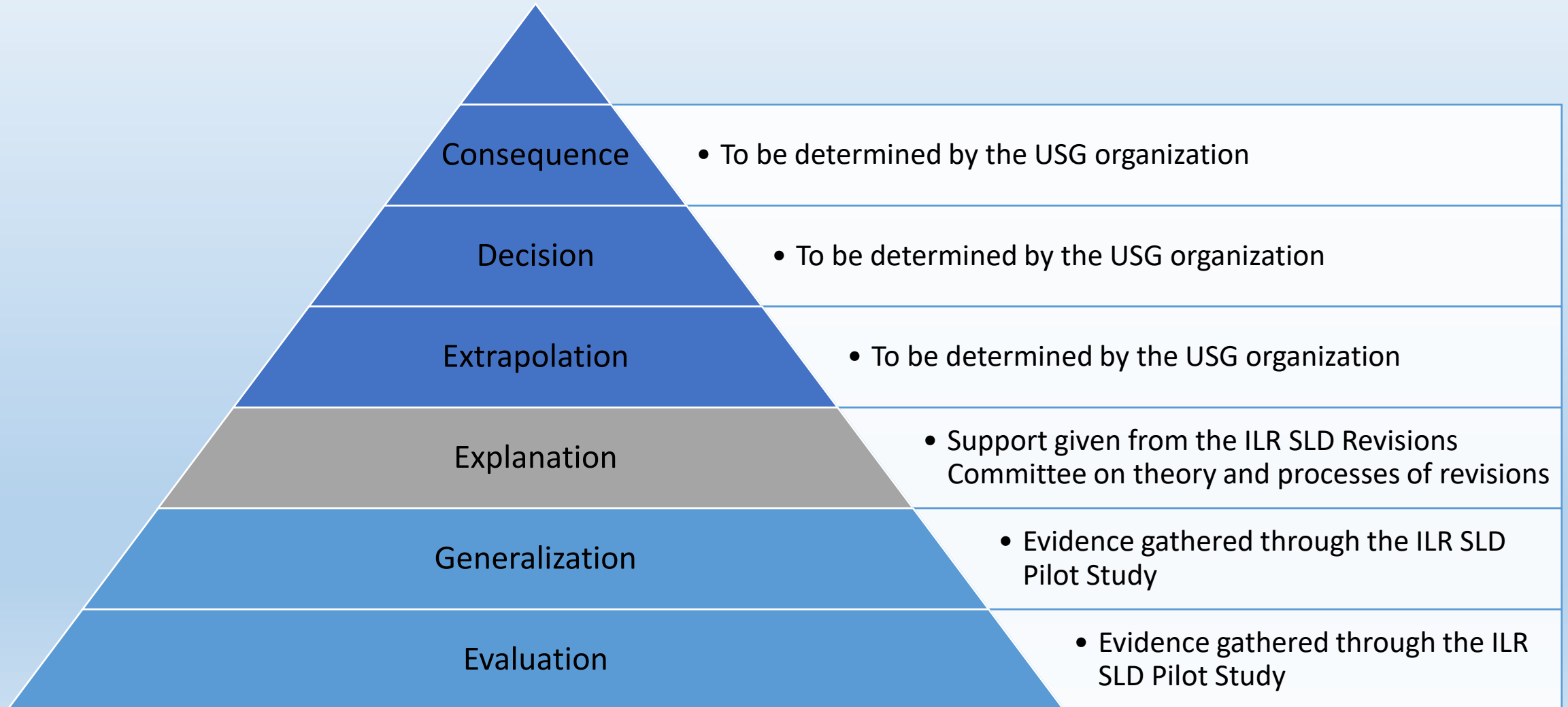193 comments submitted and 139 accepted

Editing

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.

- Knoch, U., Deygers, B, & Khamboonruang, A. (2021). Re-visiting rating scale development – modelling construct choices made by scale developers. *Language Testing.* 1-25.

- 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.
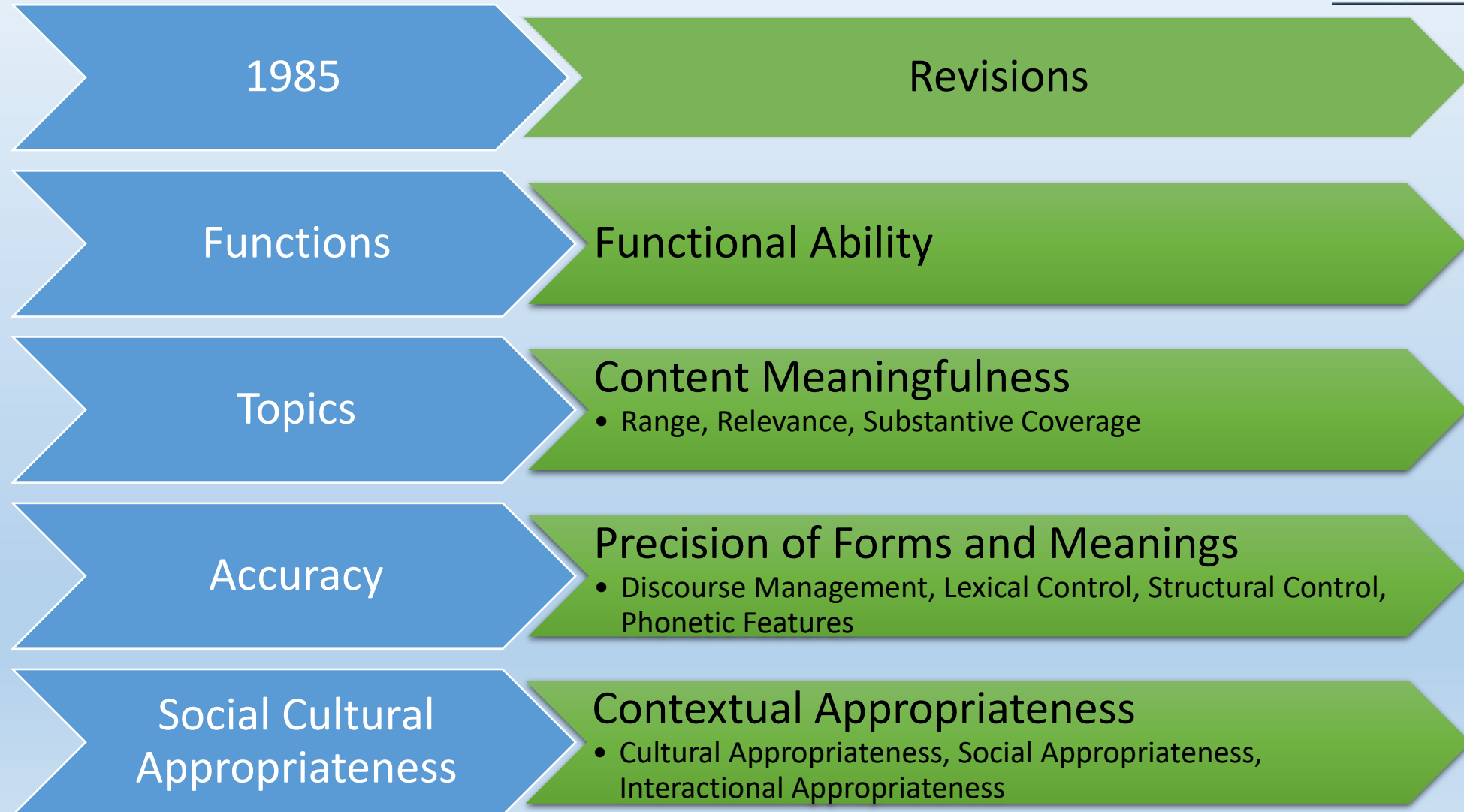
Building a Validity Argument for the ILR SLDs

Sample Revisions

Category Evolution

1985	Revisions
Functions	Functional Ability
Topics	**Content Meaningfulness** • Range, Relevance, Substantive Coverage
Accuracy	**Precision of Forms and Meanings** • Discourse Management, Lexical Control, Structural Control, Phonetic Features
Social Cultural Appropriateness	**Contextual Appropriateness** • Cultural Appropriateness, Social Appropriateness, Interactional Appropriateness

Change: Main statements

Speaking 3 (General Professional Proficiency) Able to speak the ... [covered] ...ent accuracy to participate effectively in formal and informal structural accuracy and vocabulary to participate effectively in ... [covered] ... professional topics and tasks. Can perform a range of tasks, such as conversations in practical, social and professional topics. Never... [covered] ... issues and their implications, state and defend a position or policy, limitations generally restrict the professional contexts of language use to matters of shared support opinions, hypothesize, elicit information and informed opinion, resolve unexpected knowledge and/... [covered] ...al uses the situations, or p... [covered] ...nships of ideas.

language accep... [covered] ... never May employ s... [covered] ...frequency
interfere with u... [covered] ...can vocabulary tha... [covered] ...nd abstract or
effectively com... [covered] ...ly. The societal issues... [covered] ...atic
individual spea... [covered] ...n natives expressions. U... [covered] ...In general,
speaking the st... [covered] ...complete. basic structure... [covered] ...nally interfere
Although cultu... [covered] ...may not with communi... [covered] ...arely impede
be fully underst... [covered] ...ion may be communication. Can discuss own interests and fields of competence. Otherwise, linguistic limitations
obviously foreign. Individual sounds are accurate: but stress, intonation and pitch control in depth and precision generally restrict language use to matters of general interest. Information
may be faulty. **Examples:** Can typically discuss particular interests and special fields of conveyed is on topic. Supports ideas in a clear manner without much unnecessary information. Able
competence with reasonable ease. Can use the language as part of normal professional to fulfill common cultural norms and expectations in routine and professional interactions. Speech
duties such as answering objections, clarifying points, justifying decisions, understanding the may be culturally awkward at times. Can usually control formal and informal registers in various
essence of challenges, stating and defending policy, conducting meetings, delivering settings and has command of most social conventions of conversations. Able to manage turn-taking
briefings, or other extended and elaborate informative monologues. Can reliably elicit smoothly and naturally. Can frequently use turn-taking devices such as rapid speaker changes and
information and informed opinion from native speakers. Structural inaccuracy is rarely the invitations to participate in turn closings. Assumes joint responsibility in maintaining the interaction.
major cause of misunderstanding. Use of structural devices is flexible and elaborate. Can convey some mood, feeling, emotion, or position appropriate to the context. Speaks readily, at a
Without searching for words or phrases, the individual uses the language clearly and rate of speech that does not impede comprehension.
relatively naturally to elaborate concepts freely and make ideas easily understandable to
native speakers. Errors occur in low-frequency and highly complex structures. (Has been
coded S-3 in some nonautomated applications.) [Data Code 30]

Main Statements: QUATRASECTION

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations in practical, social and professional topics.

Able to speak with sufficient accuracy to participate effectively in formal and informal contexts on general social and professional topics and tasks.

13

Change: Native speaker removed

1985

Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country where the language is natively spoken. (Level 5 Speaking)

Revised

Able to speak with clarity of expression and precision, using diverse linguistic resources to produce articulate, nuanced, and engaging language in almost any context. (Level 5 Speaking)

Change: Clarified

1985

Has broad enough vocabulary that rarely has to ask for paraphrasing or explanation. (Level 3 Listening)

Revised

Readily understands high frequency vocabulary and shows some inconsistency in mid frequency vocabulary. May understand some widely used idiomatic language and cultural references, though not fully.

(Level 3 Listening)

Change: Contexts modernized

1985

Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. (Level 2 Reading)

Revised

Can understand factual texts, including online or printed materials such as news items about frequently occurring events; routine business correspondence, email, and text messages; and clearly written descriptions and comments. (Level 2 Reading)

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.

Participants

- (n = 32) testers from CIA, DLIELC, DLIFLC, FBI, FSI

Data

- (n = 860), tests from ILR Levels 0+ - 5

Languages

- English, Mandarin Chinese, Spanish

ILR Speaking Pilot Study Results

Can USG speaking raters reliably identify examinees' levels across the entire ILR SLDs?

Yes! Inter-rater reliability: Weighted Kappa = 0.832

Are USG speaking raters more confident in giving ratings using the revised ILR SLDs for Speaking?

Yes! The rater participants preferred the revisions, saying that they were clearer (92%), more complete (88%) and easier to use (84%).

Do the ILR SLD Speaking abilities and sub-abilities, as outlined in the ILR Crosswalk Matrix, assess separate features of speaking?

Yes! A confirmatory factor analysis showed that the four main abilities fit nicely into model with no need to rate at the sub-ability level

Is there any patterned shift of scores on tests rated on the current scale vs the revised scale?

No! A paired samples t-test showed no significant difference between 1985 SLD scores and the revised SLD scores (t(859) = -1.66, p= .097).
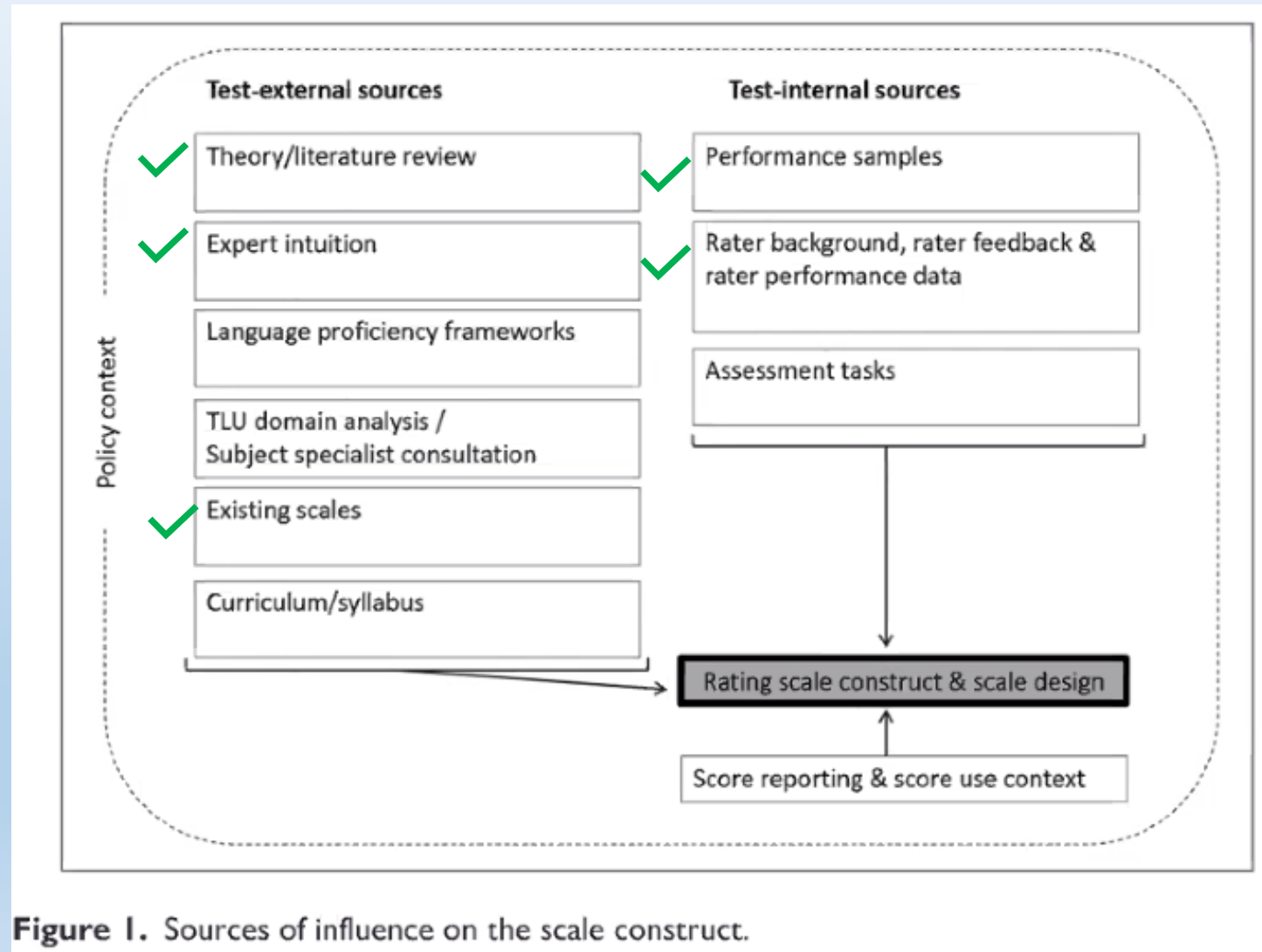
Sources informing SLD revisions

Figure 1. Sources of influence on the scale construct.

Our Product: ILR Documents

- ILR Skill Level Descriptions Prose Versions (4 documents)
  - Official ILR SLDs revised for Listening, Reading, Speaking, and Writing
  - Each include a preface that gives general description for the documents
  - Includes descriptions for Levels 0 – 5, including plus levels
- ILR SLD Crosswalk Matrix
  - Abilities tab shows definitions of the ILR Abilities and Sub-abilities for each skill
  - Speaking, Listening, Reading, and Writing tabs show the ILR descriptions for each skill subdivided by the ability and sub-ability categories

Next steps

- The revised SLDs are with the individual USG organizations who are conducting their own review and approval processes.

- The timeline for adopting the SLDs is determined by individual organizations and will vary by organization.

- The revised SLDs will be posted at the ILR website when the organizations indicate that they are ready.

- If you would like to see a copy of the revised SLDs, please work through your point of contact at your organization.

Questions?   Thank you!

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.