FSI

FOREIGN SERVICE INSTITUTE

WE SPARK LEARNING

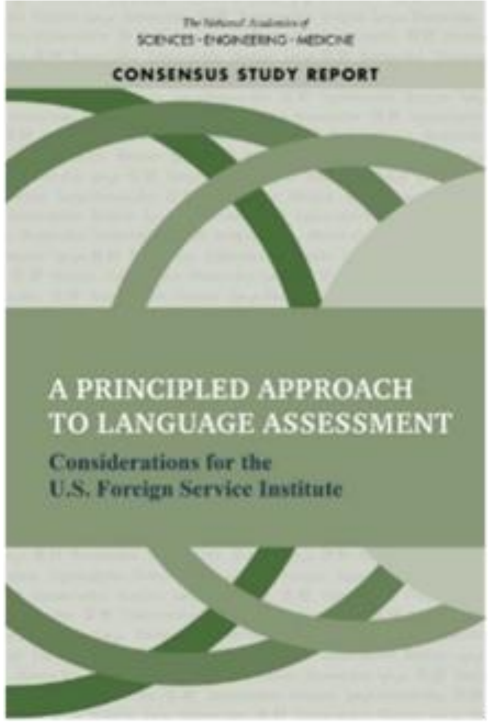
# An Overview of FSI's Speaking/Listening Rubric Pilot

Division of Language Testing and Assessment

School of Language Studies

January 24, 2025

# FSI Rubric Pilot – Introduction



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

A PRINCIPLED APPROACH TO LANGUAGE ASSESSMENT

Considerations for the U.S. Foreign Service Institute

*National Academies of Sciences, Engineering, and Medicine. 2020.* A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute. *The National Academies Press. https://doi.org/10.17226/25748.*

**Dr. David Sawyer – Intro, Rasch, and Conclusion**

**Dr. Shannon Grippando- Overview**

**Dr. Catherine Pulupa – Factor Analysis**

**Dr. Kristin Rock – G- and D-Studies**

**Dr. Will Fox – Use of AI**

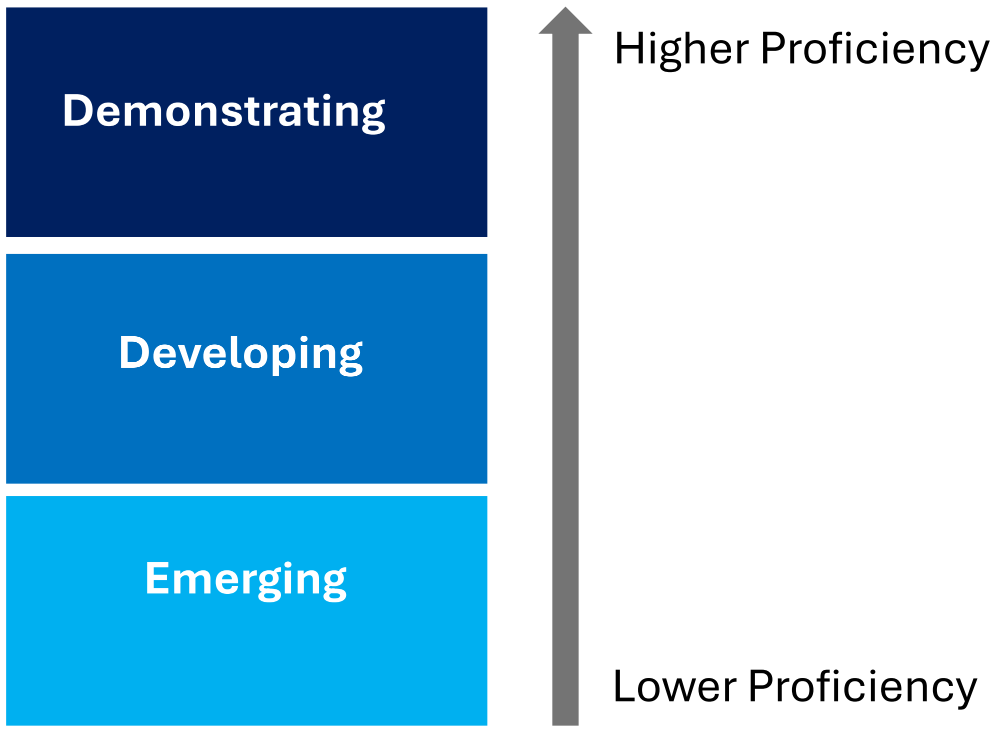
# FSI Rubric Pilot – Revised Speaking/Listening Test

- Launched January 2023
- Scenario-based
  - **Part 1**: Social Conversation
  - **Part 2:** Q&A and Topical Conversation
  - **Part 3:** Gathering and Reporting Information

# FSI Rubric Pilot – Draft Constructs

| Construct | Definition |
|---|---|
| **Conversational Fluency** | The ability to communicate with clarity across interactions and topics through articulation, pronunciation, and flow. |
| **Listening Comprehension** | The ability to process speech in real time, evidenced by responding relevantly and reporting accurately. |
| **Interactional Management** | The ability to participate and collaborate in conversation, negotiate meaning, and adjust speech to context and subject matter. |
| **Production Quality** | The ability to combine structure and vocabulary to convey meaning. |

# FSI Rubric Pilot – Bands and Scoring



Demonstrating

Developing

Emerging

Higher Proficiency

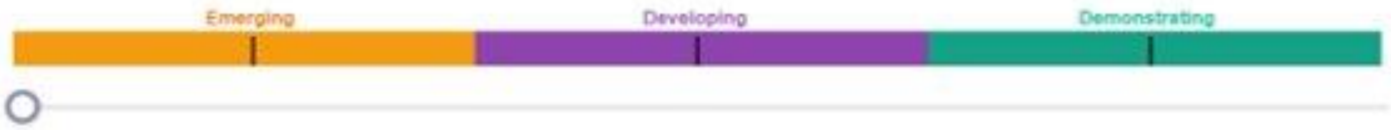
Lower Proficiency

# FSI Rubric Pilot – Pilot Structure

|  | RATERS | TESTS |  |  |  |
|---|---|---|---|---|---|
|  |  | ILR 0 to 1+ | ILR 2/2+ | ILR 3 | ILR 3+ to AP |
| Arabic | 4 | 15 | 15 | 15 | 15 |
| French | 4 | 15 | 15 | 15 | 15 |
| Mandarin | 2 | 14 | 15 | 15 | 15 |
| Portuguese | 4 | 15 | 15 | 15 | 15 |
| Russian | 4 | 15 | 15 | 15 | 8 |
| Spanish | 4 | 15 | 15 | 15 | 15 |

# FSI Rubric Pilot – Scoring

| | Conversational Fluency | Listening Comprehension | Interactional Management | Production Quality |
|---|---|---|---|---|
| **Part I** | X | X | X | X |
| **Part II** | X | X | X | X |
| **Part III** | X | X | X | X |



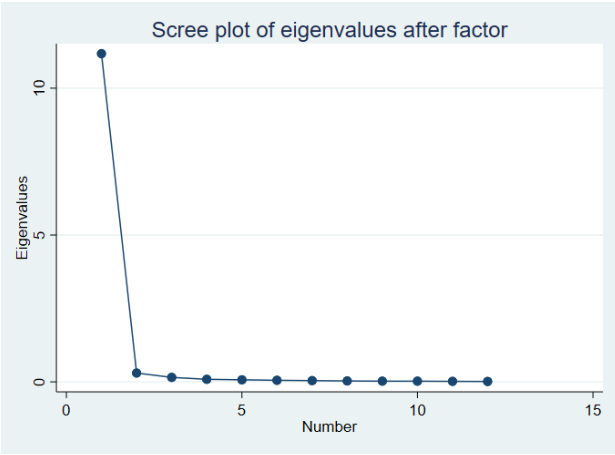
# FSI Rubric Pilot – Data and Procedures

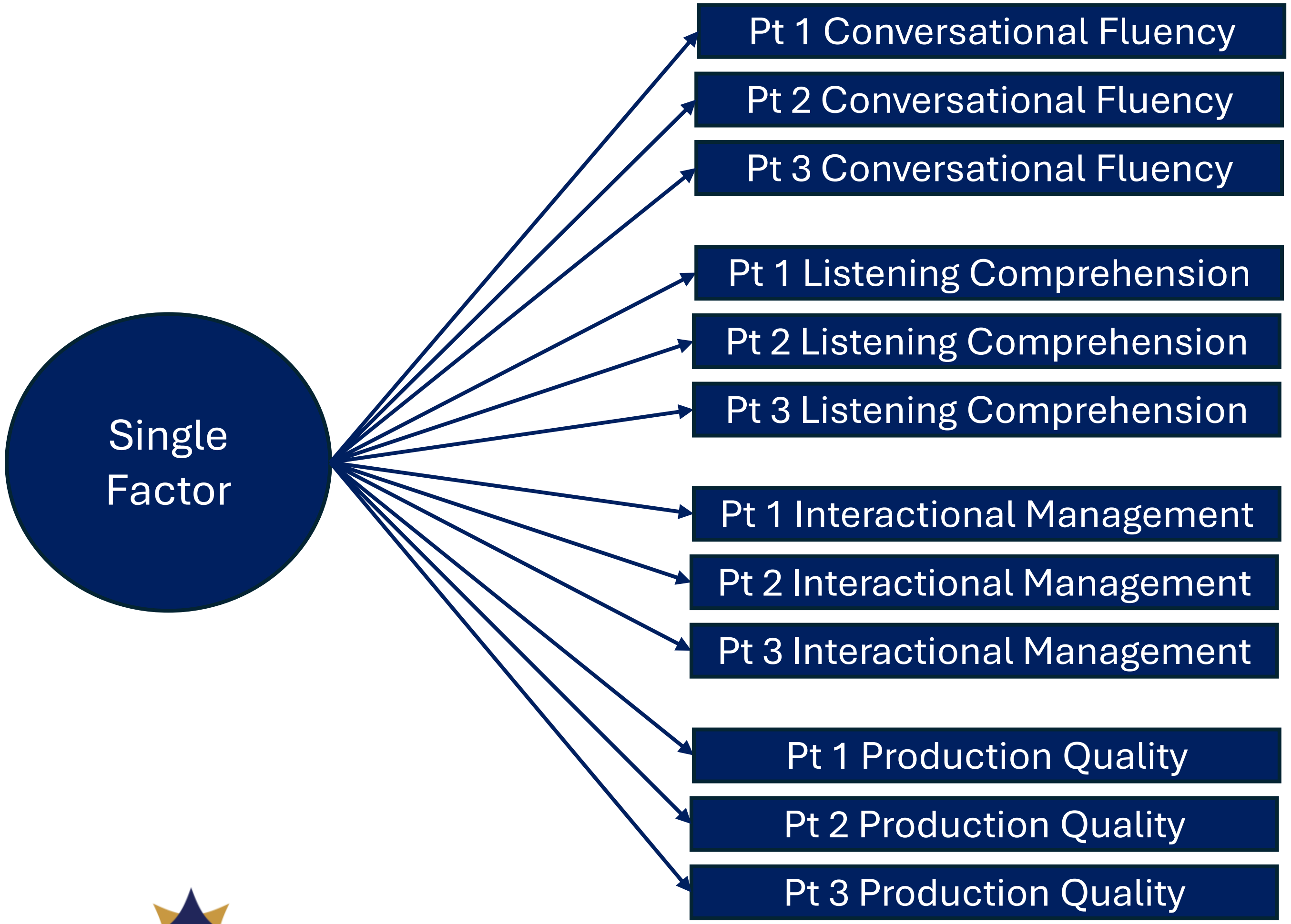
| Data Collected | Timing / Sample | Procedures |
|---|---|---|
| **Scores** | During each individual pilot test | Factor Analysis<br>Rasch<br>G- & D- Studies |
| **Scaled Rater Confidence** | | Descriptive Statistics |
| **Qualitative Rater Comments** | | Qualitative Analysis – AI and manual |
| **Scaled Rater Opinions of Ease and Quality of Rubric** | End of each individual pilot test | Descriptive Statistics |
| **Rater Opinions on Independence of Constructs** | | Qualitative Analysis – AI and manual |
| **Think Alouds** | Stratified sample of individual pilot tests | Qualitative Analysis – AI and manual |
| **Rater Identification of Useful and Problematic Rubric Features** | Post piloting period with each rater | Qualitative Analysis – AI and manual |

# FSI Rubric Pilot – Exploratory Factor Analysis

- Performed using data from all languages together, and individual languages independently

- Found no support for a 2, 3, or 4 factor model, support for a 1 factor model - unidimensional




Single Factor

Pt 1 Conversational Fluency

Pt 2 Conversational Fluency

Pt 3 Conversational Fluency

Pt 1 Listening Comprehension

Pt 2 Listening Comprehension

Pt 3 Listening Comprehension

Pt 1 Interactional Management

Pt 2 Interactional Management

Pt 3 Interactional Management

Pt 1 Production Quality

Pt 2 Production Quality

Pt 3 Production Quality

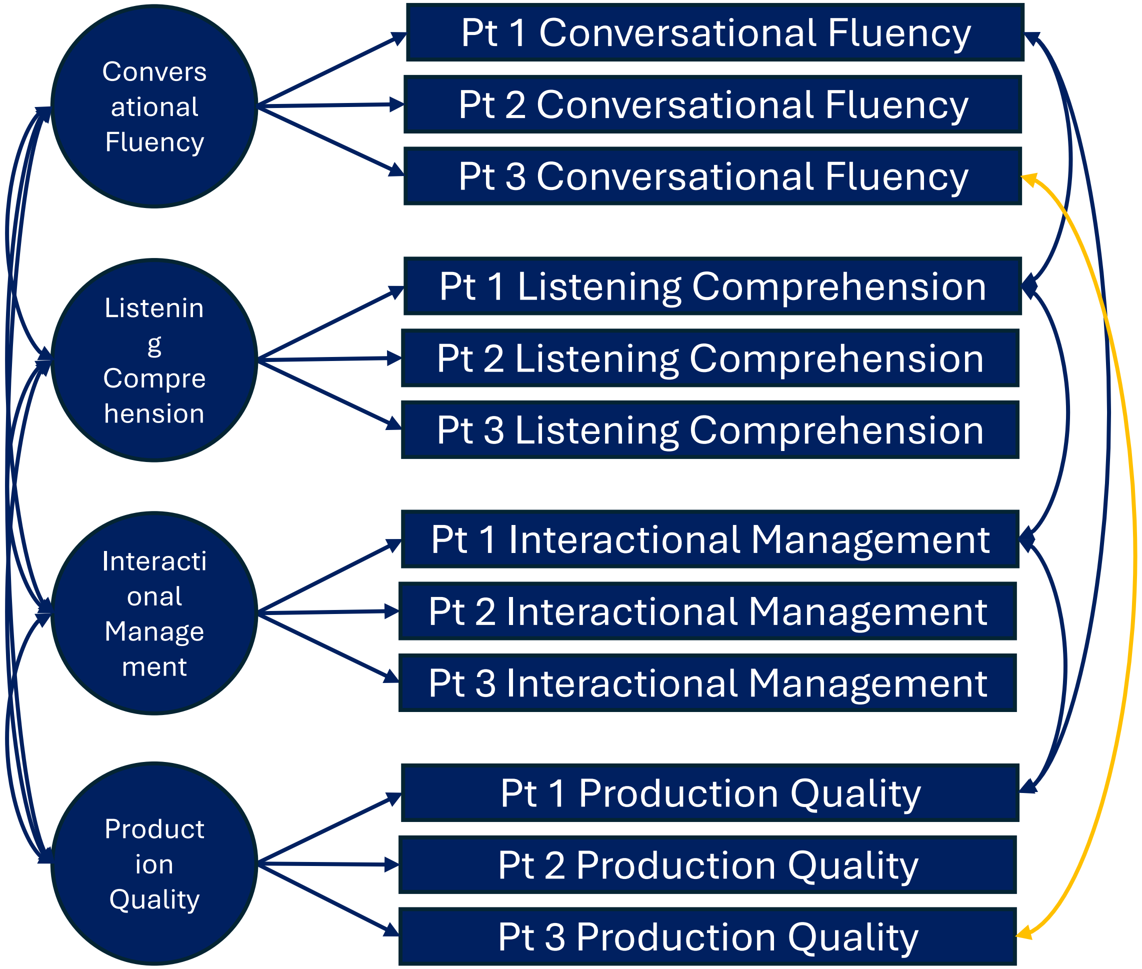
# FSI Rubric Pilot – Confirmatory Factor Analysis

- Performed using data from all languages together, and individual languages independently

- Similar results across languages
    - Most did not fit without some additional suggested modifications

    - Pt 1 items consistently interrelated, especially conversational fluency and production quality

    - Strong reliability of constructs (factor determinacies)

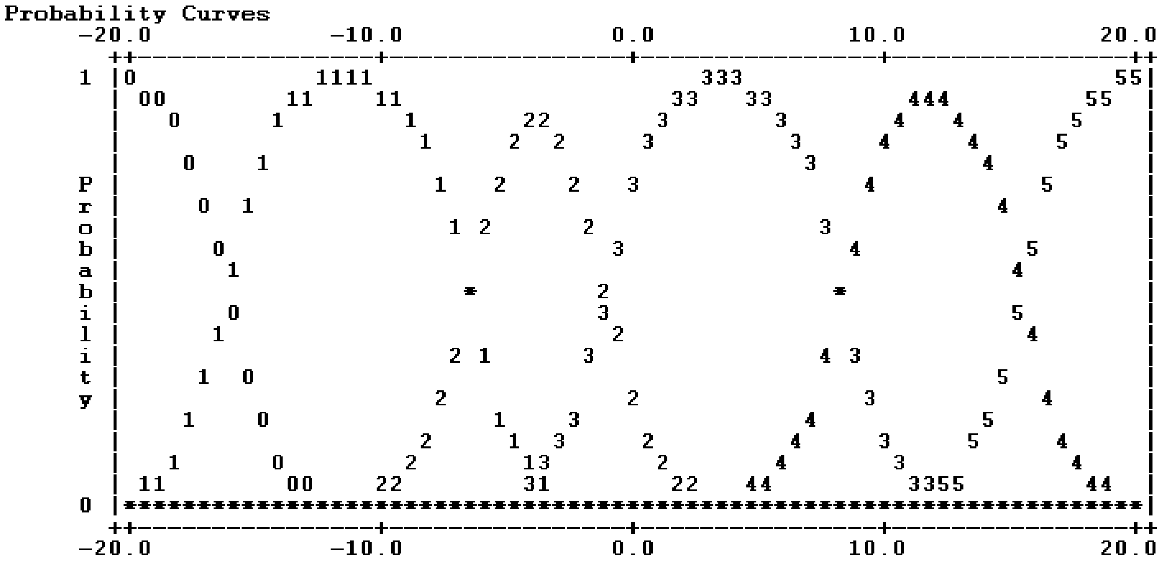


Convers
ational
Fluency
Listenin
g
Compre
hension
Interacti
onal
Manage
ment
Product
ion
Quality
Pt 1 Conversational Fluency
Pt 2 Conversational Fluency
Pt 3 Conversational Fluency
Pt 1 Listening Comprehension
Pt 2 Listening Comprehension
Pt 3 Listening Comprehension
Pt 1 Interactional Management
Pt 2 Interactional Management
Pt 3 Interactional Management
Pt 1 Production Quality
Pt 2 Production Quality
Pt 3 Production Quality

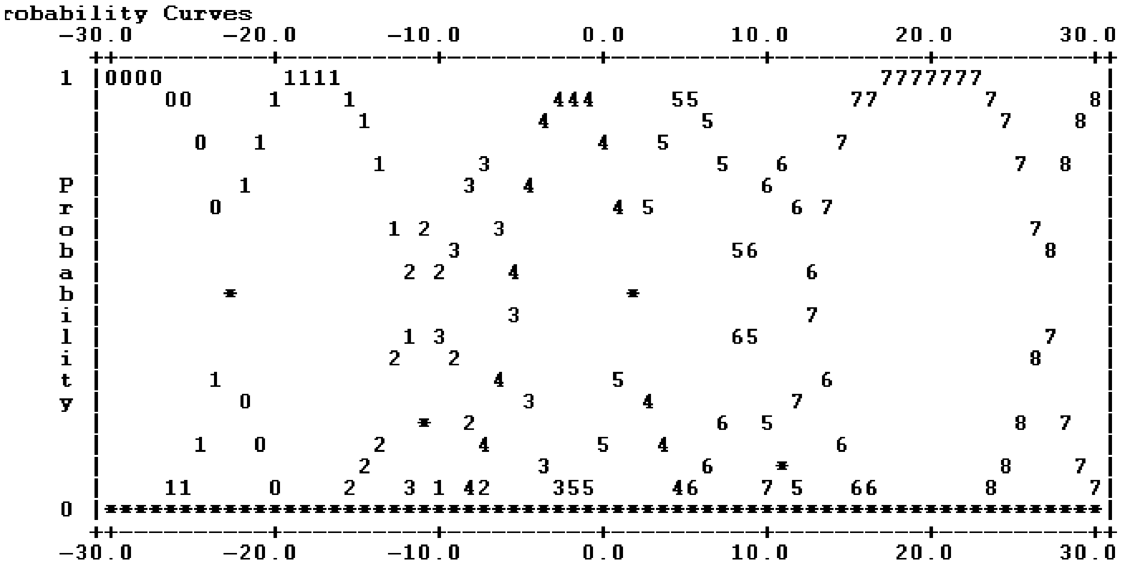
# FSI Rubric Pilot – Rasch Analysis

## French Data – Category Probability Curves

### 6 Levels



### 9 Levels



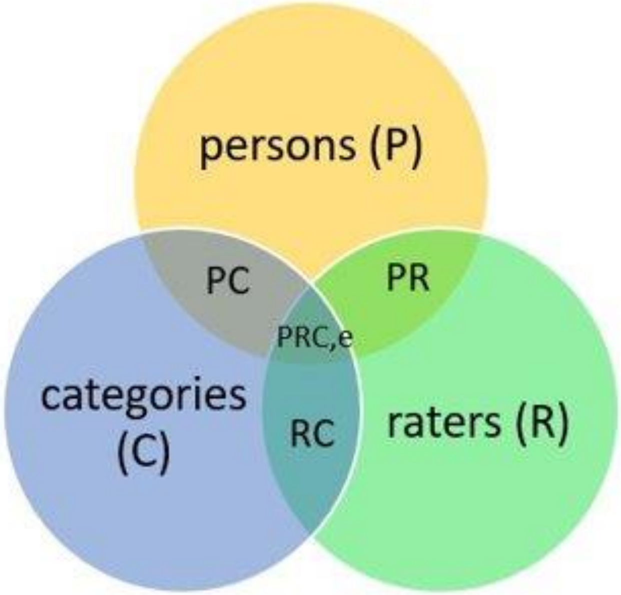
# Rubric Pilot: Cut Scores

- Microsoft Excel spreadsheets created (both language-specific and all together)
  - Personally identifiable information removed
  - Examinee number—Original ILR score—Average rater score linked
  - Entries organized from smallest to largest average score
  - Step increases inspected visually
  - Cut scores identified
    - In cases where divisions between ILR score levels were unclear, cut scores that pushed examinees upward (instead of downward) were selected

| Average Score | ILR Score | New "Maps to" |
|---|---|---|
| 14.5 | 0.5 | 0+ |
| 15.75 | 0.5 | |
| 15.75 | 0.5 | |
| 21.75 | 1.5 | 1 |
| 23..25 | 1 | |
| 23.75 | 1 | |
| 24 | 1.5 | 1.5 |
| 25 | 1.5 | |
| 26.25 | 1.5 | |
| 28.25 | 1 | |
| 29.5 | 1.5 | |
| 32.5 | 1.5 | |
| 39 | 2 | 2 |
| 40 | 2.5 | |
| 42.5 | 2 | |

# Rubric Pilot: Generalizability Theory

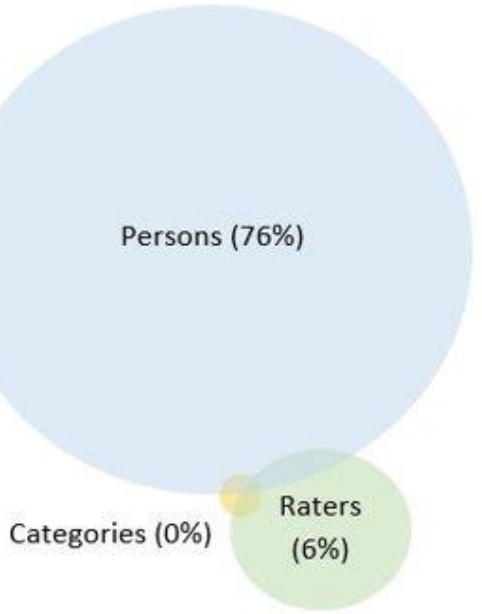
- **Generalizability Theory (G-Theory)**
  - Generalizability theory equivalent to reliability in Classical Test Theory (CTT)
  - Addresses reliability on multiple dimensions
    - i.e., rater variability, examinee ability, and their interaction
- **Step 1: G-Study**
  - Identifies sources of variance
  - GENOVA Software (Brennan, 2001) calculated variance components (VCs) for:
    - Persons
    - Categories
    - Raters
    - Persons x categories
    - Persons x raters
    - Categories x raters
    - Persons x raters x categories (error)



# Rubric Pilot: Generalizability Theory

- **G-Study Analysis**
  - Across languages, largest source of variance: Examinees
  - Examinees counted for 66% to 84% of total variance
- Three weighted models were also considered, but largest source of variance was categories

| | Arabic | Chinese | French | Portuguese | Russian | Spanish |
|---|---|---|---|---|---|---|
| P | 0.77 | 0.84 | 0.78 | 0.79 | 0.73 | 0.66 |
| R | 0.05 | 0.00 | 0.06 | 0.02 | 0.07 | 0.16 |
| C | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| PR | 0.11 | 0.25 | 0.09 | 0.14 | 0.09 | 0.09 |
| PC | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| RC | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
| PRC | 0.05 | 0.06 | 0.07 | 0.05 | 0.08 | 0.07 |



# Rubric Pilot: Generalizability Study

- Step 2: The Decision Study (D-Study)
  - Variance components from G-Study used to calculate:
    - Generalizability coefficients for different combinations of raters and categories
    - Dependability estimates for proposed cut scores
- For this project, the phi lambda dependability index was used
  - Values for lambda (a cut-score expressed as a proportion) were used to calculate reliability estimates (phi lambda)
- Across languages, dependability estimates for all cut scores were 0.94 or higher, including at the ILR 2+ to 3 threshold

$$\Phi(\lambda)=\frac{\sigma_p^2+(\mu-\lambda)^2}{\sigma_p^2+(\mu-\lambda)^2+\sigma_e^2(\Delta)}$$

# AI ANALYSIS

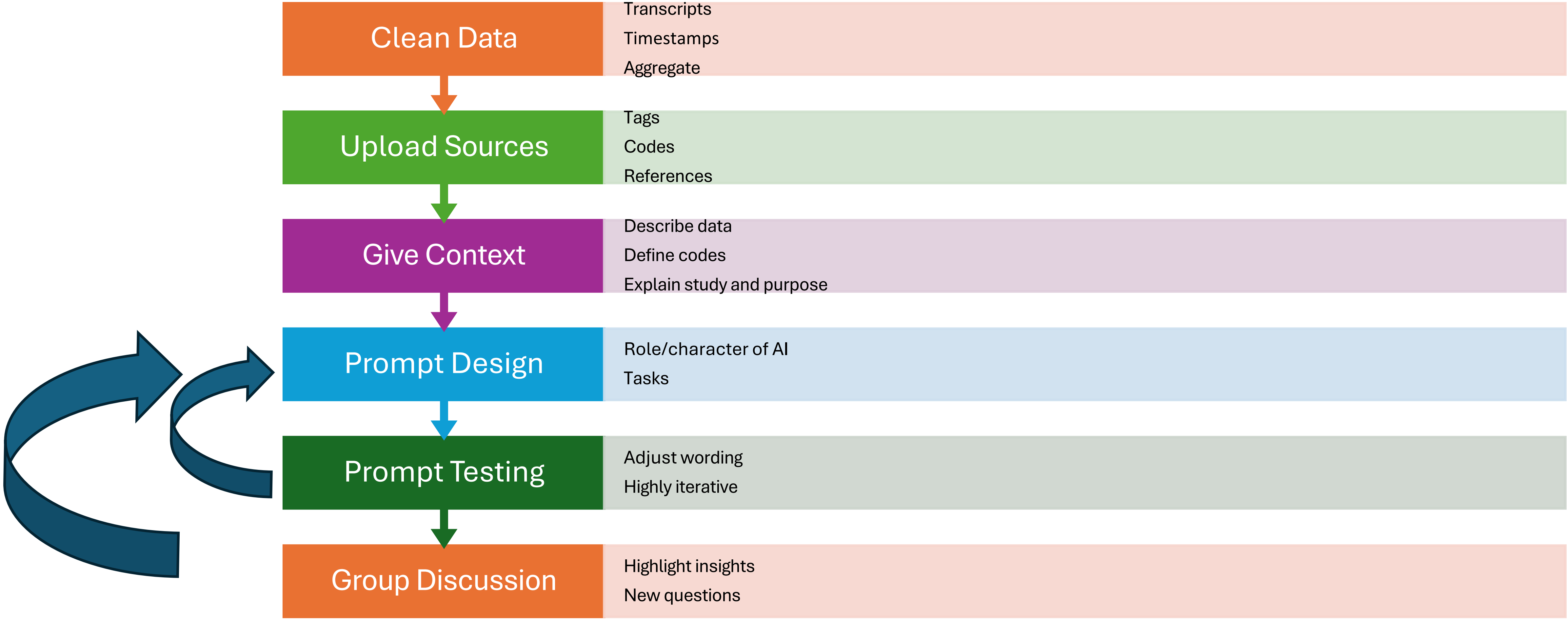
*StateChat*

## 17 total documents

- Test context and pilot study background documents
- Think aloud protocol transcripts
- Feedback session (focus group) notes
- Scoring form comments
- Final survey questions and responses

## ~500 pages of qualitative data

# AI Analysis Process Flow



Clean Data
- Transcripts
- Timestamps
- Aggregate

Upload Sources
- Tags
- Codes
- References

Give Context
- Describe data
- Define codes
- Explain study and purpose

Prompt Design
- Role/character of AI
- Tasks

Prompt Testing
- Adjust wording
- Highly iterative

Group Discussion
- Highlight insights
- New questions

# AI Prompt Engineering

## Team of 3 Prompt Engineers

- Background, explanation of study
- Document descriptions
  - Rater IDs, Language, Coding
- Task explanation
- Foll
  - P
  - B

You are a Ph.D. trained evaluation specialist with a focus on qualitative analysis. You are evaluating a rubric to assign scores for a foreign language proficiency test for professional diplomats who interact in that foreign language for job-related purposes.

The thirteenth c
Comments. This
responses
test using
e English sp
ter ID (e.g
age (e.g., S

Using all the background information about the study, the test, and the participants, conduct qualitative analyses on the documents separately. Identify the main themes in each document. Also focus on identifying themes or patterns across raters and languages. Provide a count of the statements that contributed to each theme.

Provide five representative quotes to support the themes and patterns from the texts that do not restate the main themes, and which are unique and are not repeated from other themes.

Identify concrete ways to improve the rubric, supported with examples from the documents. Provide a concise summary of the suggested improvements along with a more detailed discussion.

# AI Analysis Tasks

General qualitative analysis of data

Research questions

Two alternative formulations of the rubric

Summary Document & Action Items

# AI Analysis Conclusions

- Distinct points on the slider
  - "We might want a same standard for what the middle point is and what it takes to move above the middle point." (General) - Feedback Session
- Tailored rubric for Part III
  - "It always feels to me that, in order to correctly rate an EE's performance, each construct should have statements that closely reflected what we are looking for in each part of the test; instead of having constructs with the same descriptors/statements throughout the 3 parts." (1PYLTA, Portuguese) - Score Form
- Clarity in wording, descriptions
  - "I still struggle to understand and use the descriptions/notes in 'Interactional Management' cells of the rubric." (2RuLTA, Russian) - Think Aloud Protocol

# FSI Rubric Pilot – Conclusions

| Analysis | Finding | Interpretation / Decision | Complementary Analysis |
|---|---|---|---|
| **Factor Analysis of Scores** | Support for unidimensionality<br>Part 1 scores interrelated, not Parts 2 and 3 | Examine why scores within Parts 2 and 3 vary from each other | Rasch<br>Qualitative analysis |
| **Rasch Analysis of Scores** | Some misfit<br>Probability curves break down with more than 6 levels | Revise rubric<br>Include 6 distinct levels | Factor analysis<br>Qualitative analysis |
| **G- and D-Theory Analyses** | Test takers were primary source of variance in scores; cut scores were dependable | Confident that rubric scores are separating examinees of different ability levels | |
| **Qualitative Comments** | • Slider needs more precision<br>• Not all descriptions match test parts<br>• Some definitions and descriptions are unclear | • Created 6 distinct levels<br>• Part III-specific constructs<br>• Final edits of rubric language | Factor analysis<br>Rasch |

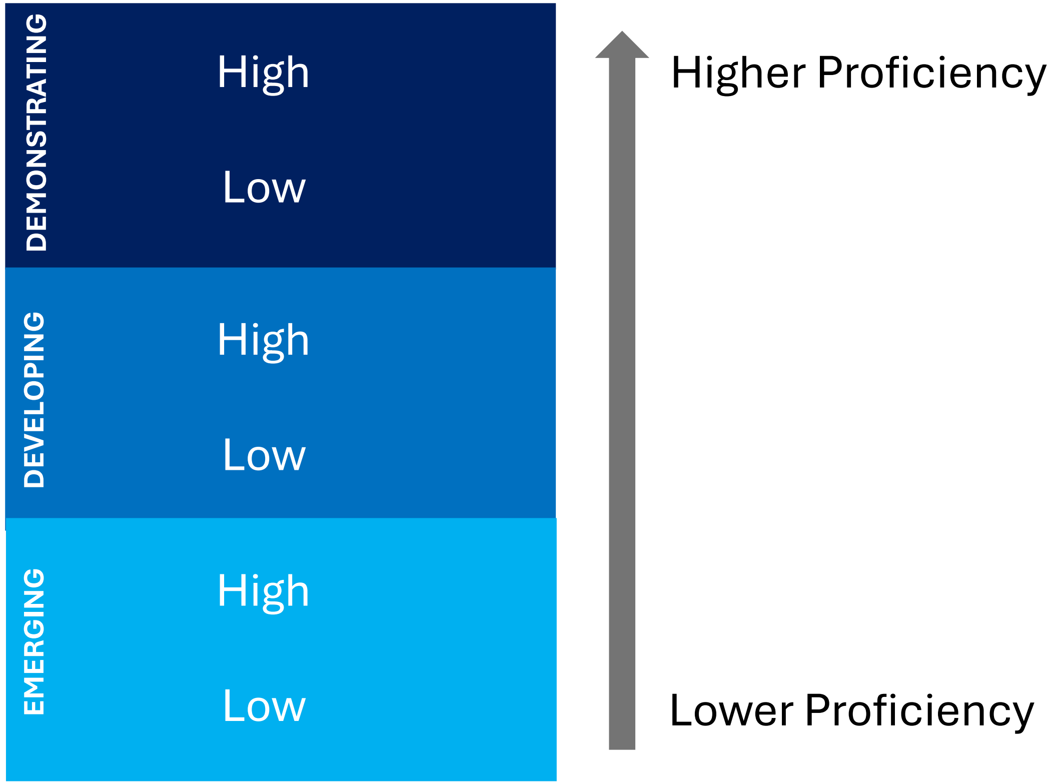
# FSI Rubric Pilot – Constructs

| CONSTRUCT | DEFINITION |
|---|---|
| **PART 1 & 2** | |
| **Conversational Fluency** | The ability to communicate in conversations across topics and interactions. |
| **Interactional Management** | The ability to take turns, transition across topics and interactions, ask follow-up questions, and clarify understanding. |
| **Listening Comprehension** | The ability to understand and respond on topic. |
| **Production Quality** | The ability to use grammatical structure and vocabulary to convey meaning. |
| **Part 3** | |
| **Reporting Ability** | The ability to convey in English the information gathered. |
| **Question Formulation** | The ability to inquire, elicit, and gather information. |

# Thank You

Stay connected with FSI

# FSI Rubric Pilot – Proficiency Bands



DEMONSTRATING

High

Low

DEVELOPING

High

Low

EMERGING

High

Low

Higher Proficiency

Lower Proficiency

# FSI Rubric Pilot – Scoring

| | Conversational Fluency | Interactional Management | Listening Comprehension | Production Quality | Reporting Ability | Question Formulation |
|---|---|---|---|---|---|---|
| **Parts I & II** | X | X | X | X | ■ | ■ |
| **Part III** | ■ | ■ | ■ | ■ | X | X |