

Spanish and Beyond: AI in Language Assessment from Major to Underrepresented Languages

ECOLT 2025

Georgetown University

Troy Cox | Matthew Wilcox |

Abraham Harris | Ty Bingham



The Medici Effect



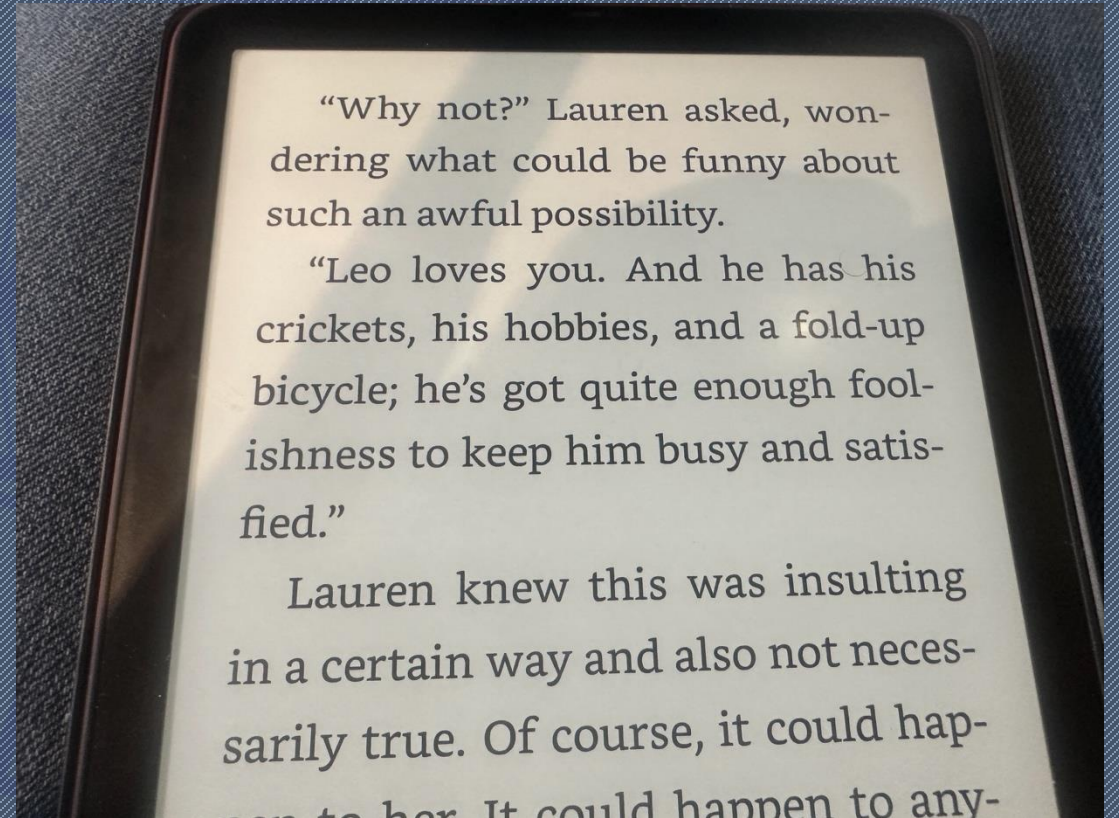
- Definition

- Innovation occurs at the intersection of diverse disciplines, cultures, and ideas.
- Named after the Medici family of Renaissance Florence, who brought together artists, scientists, and thinkers, sparking groundbreaking creativity.

- Key Principles

- Breakthrough insights emerge from cross-pollination of knowledge.
- Diversity of perspectives leads to novel connections.
- Encourages breadth of interests as a source of scholarly innovation.

- So, what do cycling, yoga, Great Britain, and an applied linguist have in common?



A faint, stylized illustration of a bicycle is centered in the background. The wheels are composed of multiple orange slices, each with a visible seed pattern. The bicycle frame, handlebars, and seat are depicted with simple, dark lines. The entire scene is set against a solid dark blue background.

A Trip Down Memory Lane

Understanding Temporal Fluency: The Influence of the ACTFL OPIc Speaking Prompts



Research Question

- What are the differences in temporal fluency features between examinees rated at different ACTFL levels?
 - articulation rate
 - mean length of run
 - silent pause ratio

Subjects

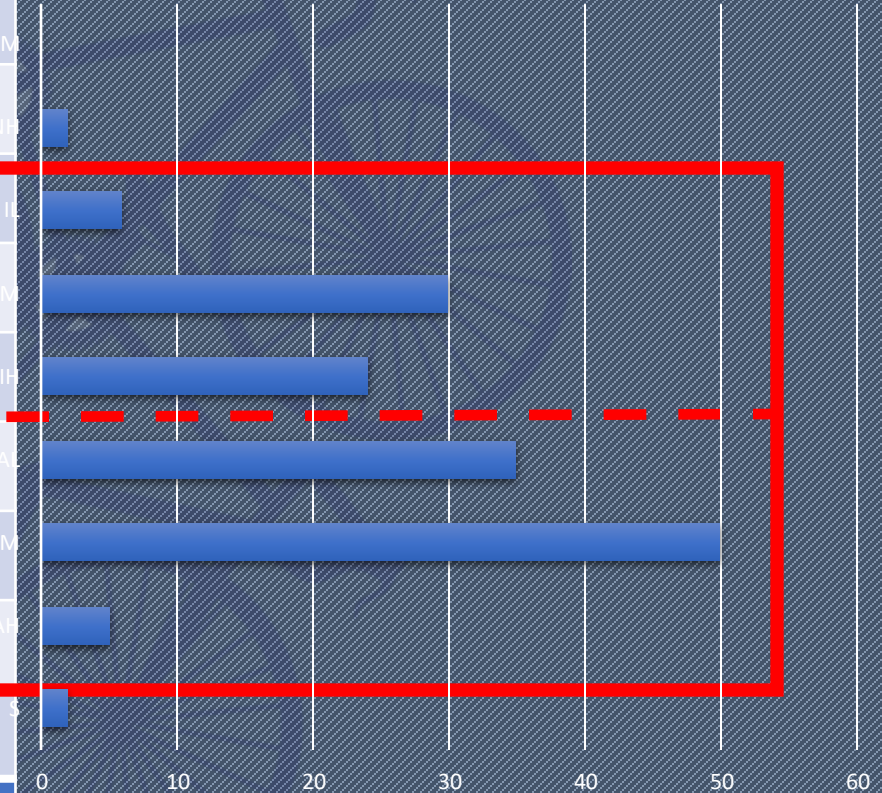
- 101 BYU students were originally recruited to participate (Fall/Winter 2013-14). Another 80 were recruited (Spring/Summer 2014) and added.
- Incentive: They could choose the higher test score for certificate.
- Completion status:
 - **Finished: 154**
 - **Incomplete: 27**

Subjects

Number of Participants	Gender	Age	Years of Study
N = 154	Male: 73 Female: 81	Mean: 23.41 SD: 5.29	Mean: 3.48 SD: 2.30

Subjects

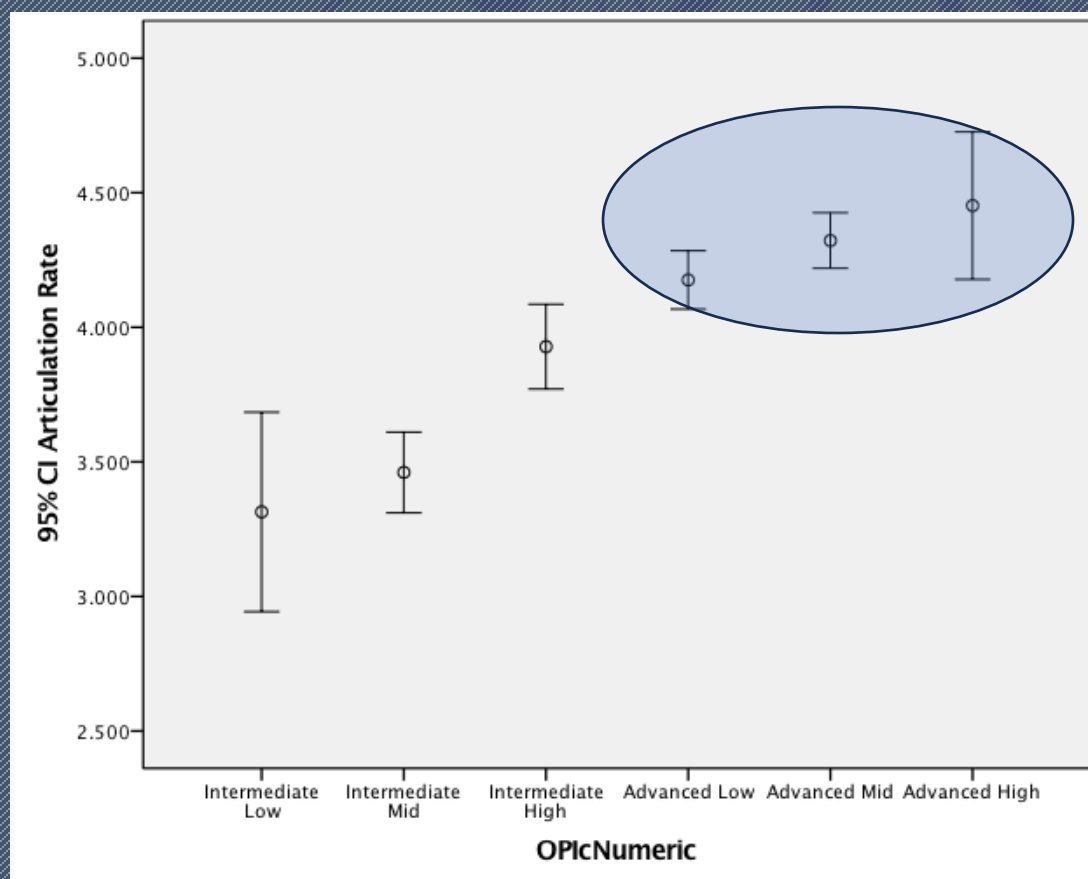
OPI Level	Number of Ratings
Novice Mid	0
Novice High	2
Intermediate Low	6
Intermediate Mid	30
Intermediate High	24
Advanced Low	35
Advanced Mid	50
Advanced High	5
Superior	2
Total	154



Acoustic Analysis (PRAAT)

- **Temporal Fluency Measures**
 - **Articulation Rate (ArtRate)**
 - Syllables per second (excluding silent pauses)
 - **Mean Length of Utterance (MLU)**
 - Average number of syllables between silent pauses
 - **Silent Pause Ratio (SPR)**
 - Number of silent pauses/ total duration

ACTFL Rating and Articulation Rate

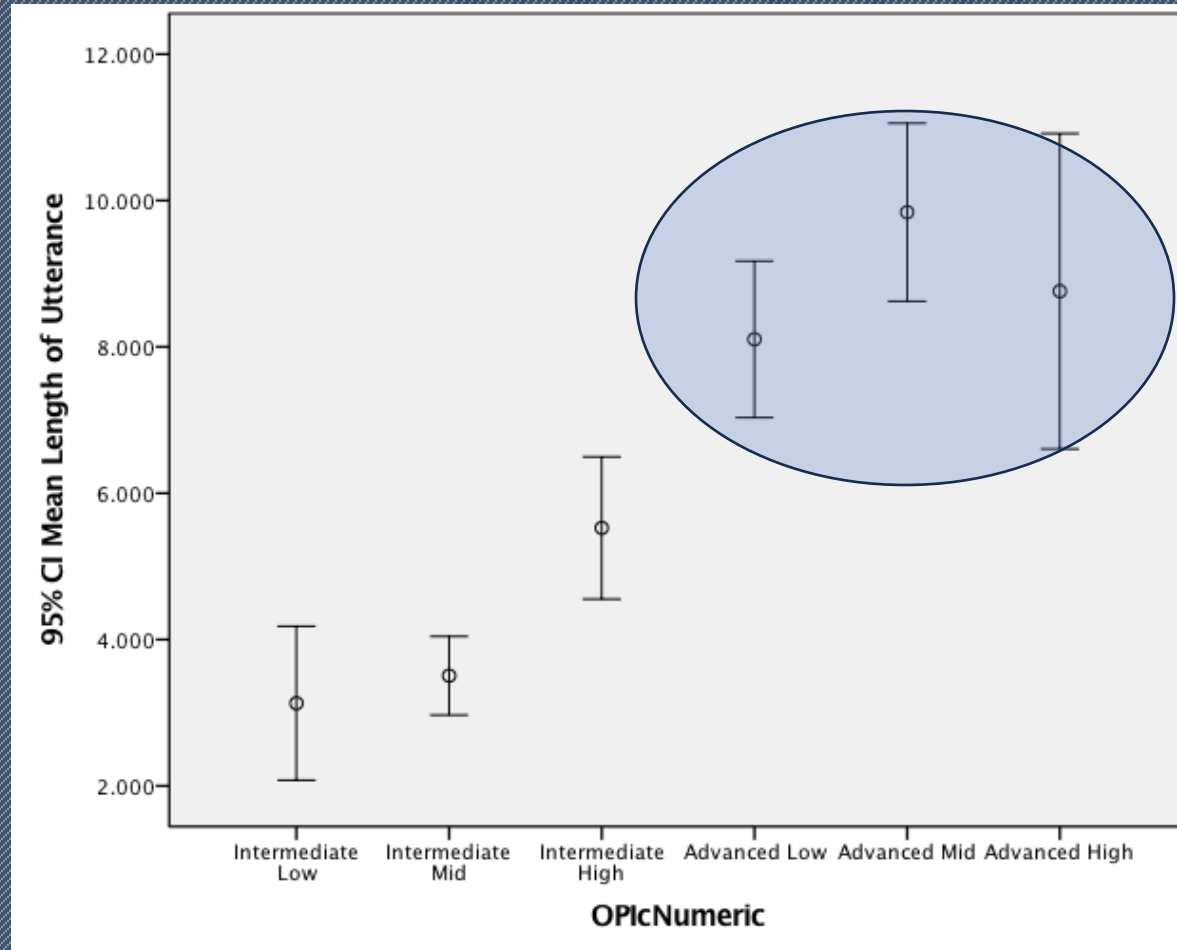


Levels	Mean Difference	Cohen's <i>d</i>
IL & IM	-0.15	0.37
IM & IH	-0.47	1.20
IH & AL	-0.25	0.73
AL & AM	-0.15	0.43
AM & AH	-0.13	0.37

Effect Size Legend	Cohen's <i>d</i>
None	.00
Small	.10
Medium	.30
Large	> .50

ANOVA $F(5, 143) = 28.37, p < .001$

ACTFL Rating and Mean Length Utterance

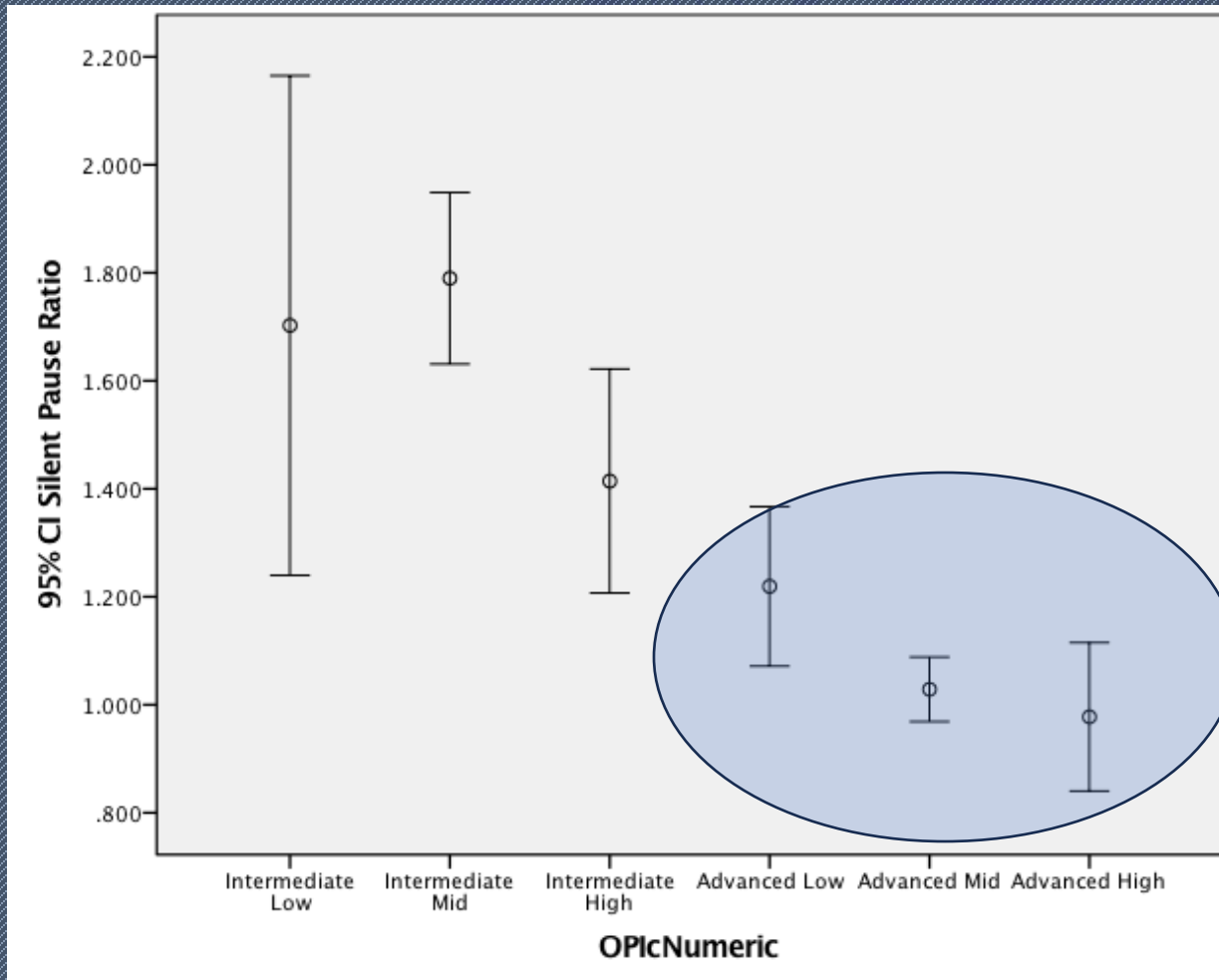


Levels	Mean Difference	Cohen's <i>d</i>
IL & IM	-0.38	0.27
IM & IH	-2.02	1.08
IH & AL	-2.58	0.92
AL & AM	-1.74	0.46
AM & AH	1.08	0.26

Effect Size Legend	Cohen's <i>d</i>
None	.00
Small	.10
Medium	.30
Large	> .50

ANOVA $F(5, 143) = 19.73, p < .001$

ACTFL rating and Silent Pause Ratio



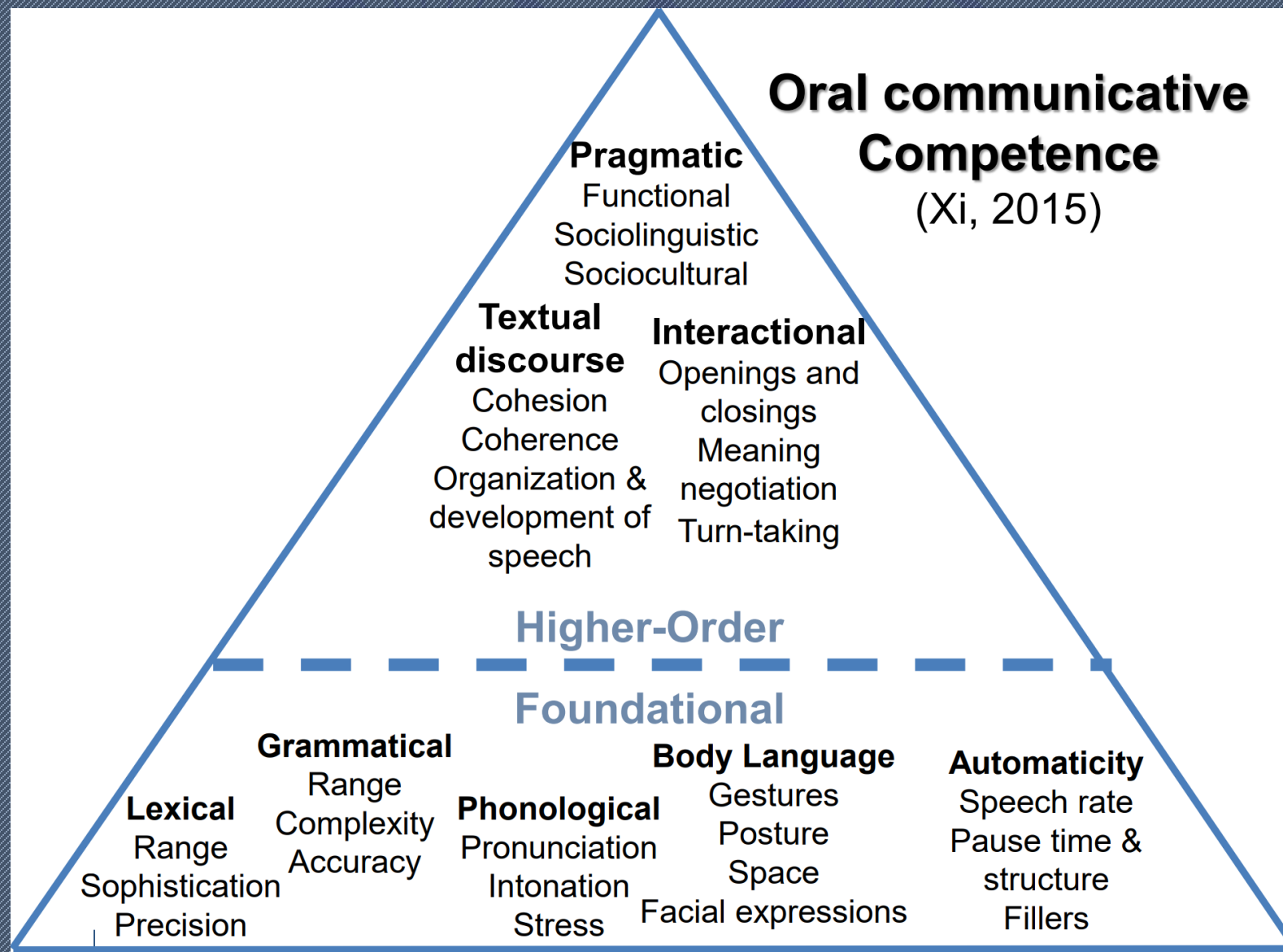
Levels	Mean Difference	Cohen's <i>d</i>
IL & IM	-0.09	0.20
IM & IH	0.38	0.82
IH & AL	0.19	0.43
AL & AM	0.19	0.60
AM & AH	0.05	0.25

Effect Size Legend	Cohen's <i>d</i>
None	.00
Small	.10
Medium	.30
Large	> .50

ANOVA $F(5, 143) = 18.15, p < .001$

Temporal fluency features were strongly related to final ACTFL Proficiency Rating but were more indicative of differences between Intermediate and Advanced

ECOLT



BYU's Language Ecosystem



- 60% students speak a 2nd language
- Majority learn language in 18-24 month missionary service
- Largest classes are the 300 level, post-mission class
- BYU teaches 84 languages
- Spanish largest language department (approx. 1,000 students every fall and winter semester)

Spring 2020 Problem

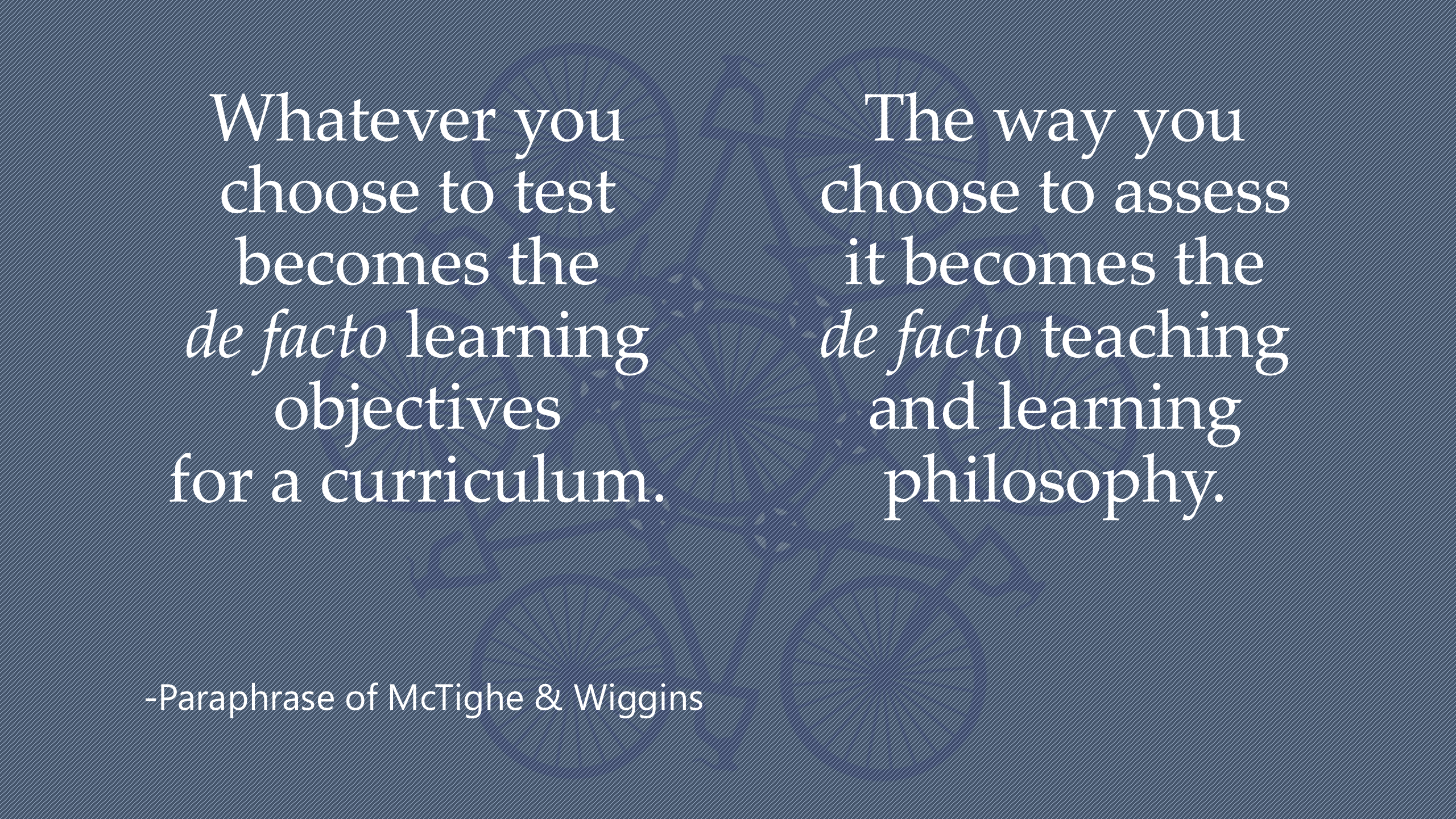
- Language department chairs were concerned about the disruption COVID-19 would have on the language ability for students whose missionary service Advanced language classes
- Would WebCAPE work?



WebCAPE (Web-Computer Adaptive Placement Exam)

- Tests students with MC questions.
- Adaptive algorithm gives each student a unique set of questions at varied difficulty levels to determine score.
- Results are low stake.





Whatever you
choose to test
becomes the
de facto learning
objectives
for a curriculum.

The way you
choose to assess
it becomes the
de facto teaching
and learning
philosophy.

-Paraphrase of McTighe & Wiggins

Bloom's Taxonomy

WebCAPE

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual	– Factual knowledge of terminology. – Factual knowledge of specific details and elements.					
Conceptual	– Knowledge of classifications and categories. – Knowledge of principles and generalizations. – Knowledge of theories, models, and structures.					
Procedural	– Knowledge of subject-specific skills and algorithms. – Knowledge of subject-specific techniques and methods. – Knowledge of criteria for determining when to use appropriate procedures.					
Metacognitive	– Strategic knowledge. – Knowledge about cognitive tasks, including appropriate contextual and conditional knowledge. – Self knowledge.					

How can we create an instrument that assesses different types of knowledge and higher order skills?

A faint, repeating pattern of bicycles is visible in the background, rendered in a light blue color against a darker blue textured background.

Back when we would do high speed dubbing of interviews, I could usually guess the rating within one sublevel.

—ACTFL Facilitator

What is the LASER?

- The Language Ability Self-Evaluation Resource is a self-assessment diagnostic tool designed to help students evaluate the strengths and weaknesses of their language proficiency through nationally recognized standards.
- As students improve in their ability to self-assess, they will improve in their ability to grow and develop their language proficiency.



laser.byu.edu

Scale Design and Description

Construct	Indirect measure Speaking and Writing proficiency
Format	Section 1 (Fixed form) • Language Learning Background • Language Learning Grit (9 items, 5 pt. scale) • Willingness to Communicate (26 items (random subset, 100% time scale)) • Self-Assessment of Speaking and Writing (24 items, 5 pt. Scale) Section 2 (items targeted at the self assessed level from section 1) • Writing & Speaking Self-assessed Passages (10 passages (subset), 5 pt. scale)
Languages	Can be adapted for most languages
Results	ACTFL scale (10 Point) from NL to S (not official). Self-assess results moderated by language learning background, length of immersion, and distance from English. Sent to both student and instructors who have (hopefully) been trained in the proficiency scale
Delivery	Qualtrics (not proctored)

Case, M (2022) Temporal Fluency in L2 Self-Assessments: A Cross-Linguistic Study of Spanish, Portuguese, and French, BYU MA Thesis

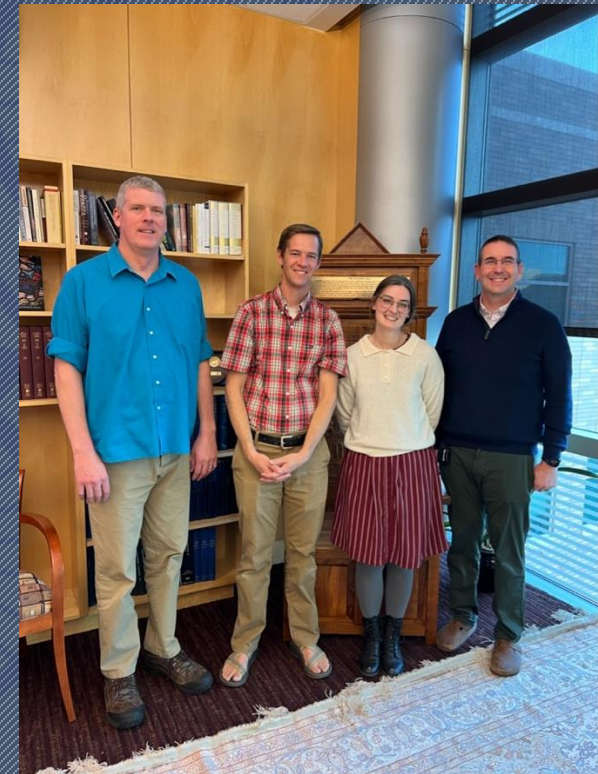
Participants

$n = 283$; male = 187, female = 96

French ($n = 62$, 21.9%)

Portuguese ($n = 106$, 37.5%)

Spanish ($n = 115$, 40.6%).



Higher OPIc levels are associated with faster speech, longer utterances, and fewer pauses,

Figure 7

Articulation Rate by OPIc Assessed Proficiency Level

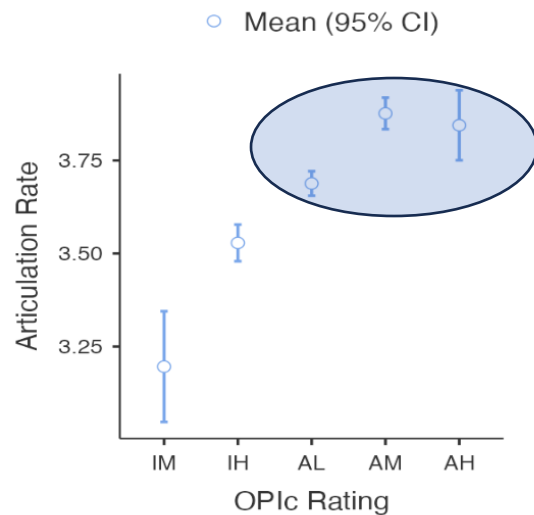


Figure 8

MLU by OPIc Assessed Proficiency Level

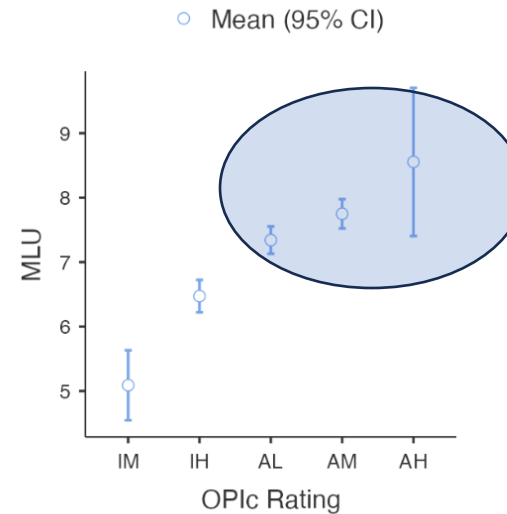
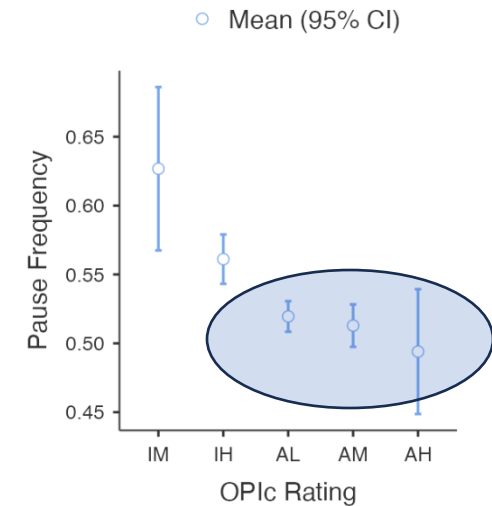


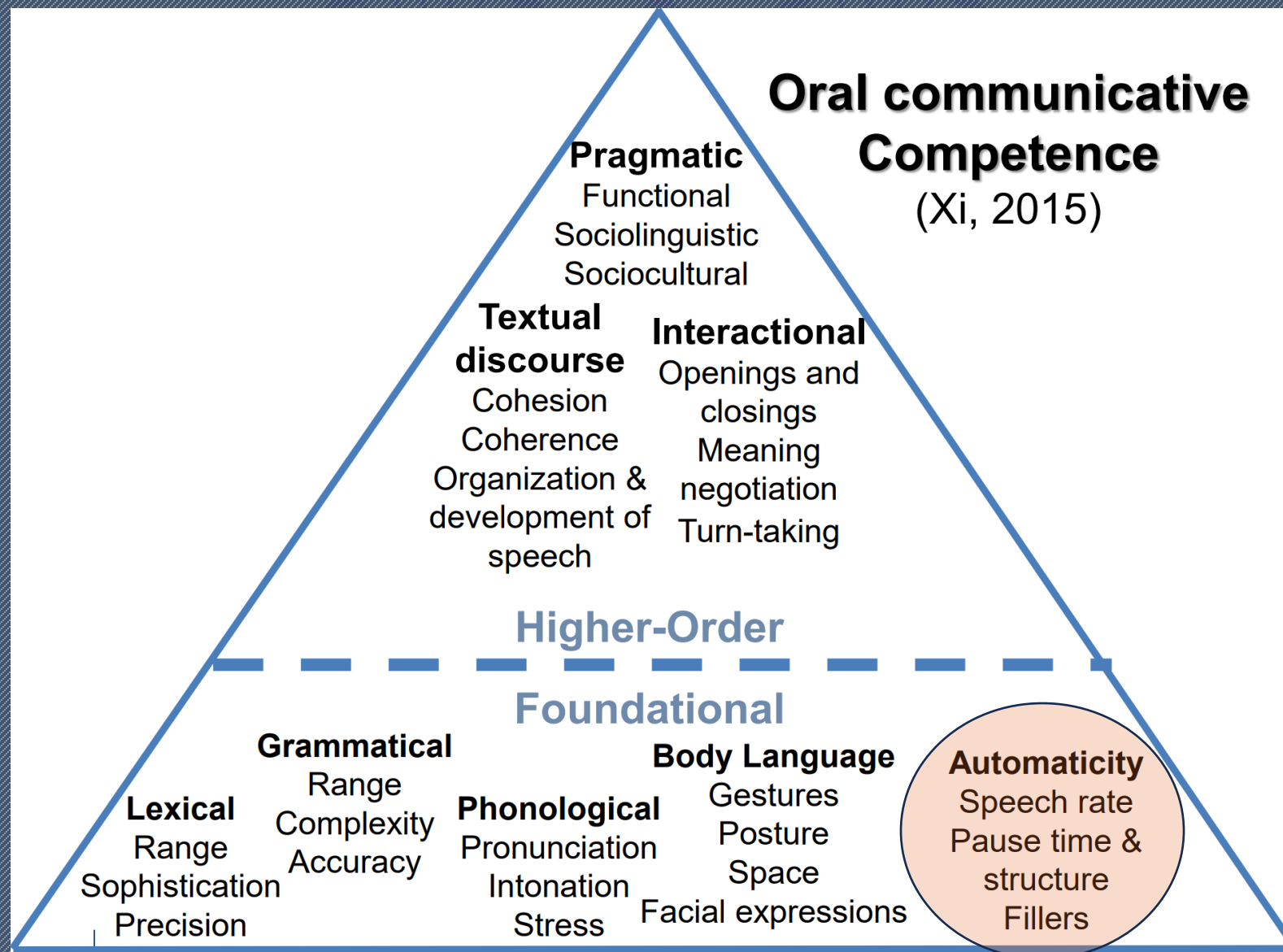
Figure 9

Pause Frequency by OPIc-Assessed Proficiency Level



Weakness: Not many Superior or Intermediate speakers in the LASER subject pool

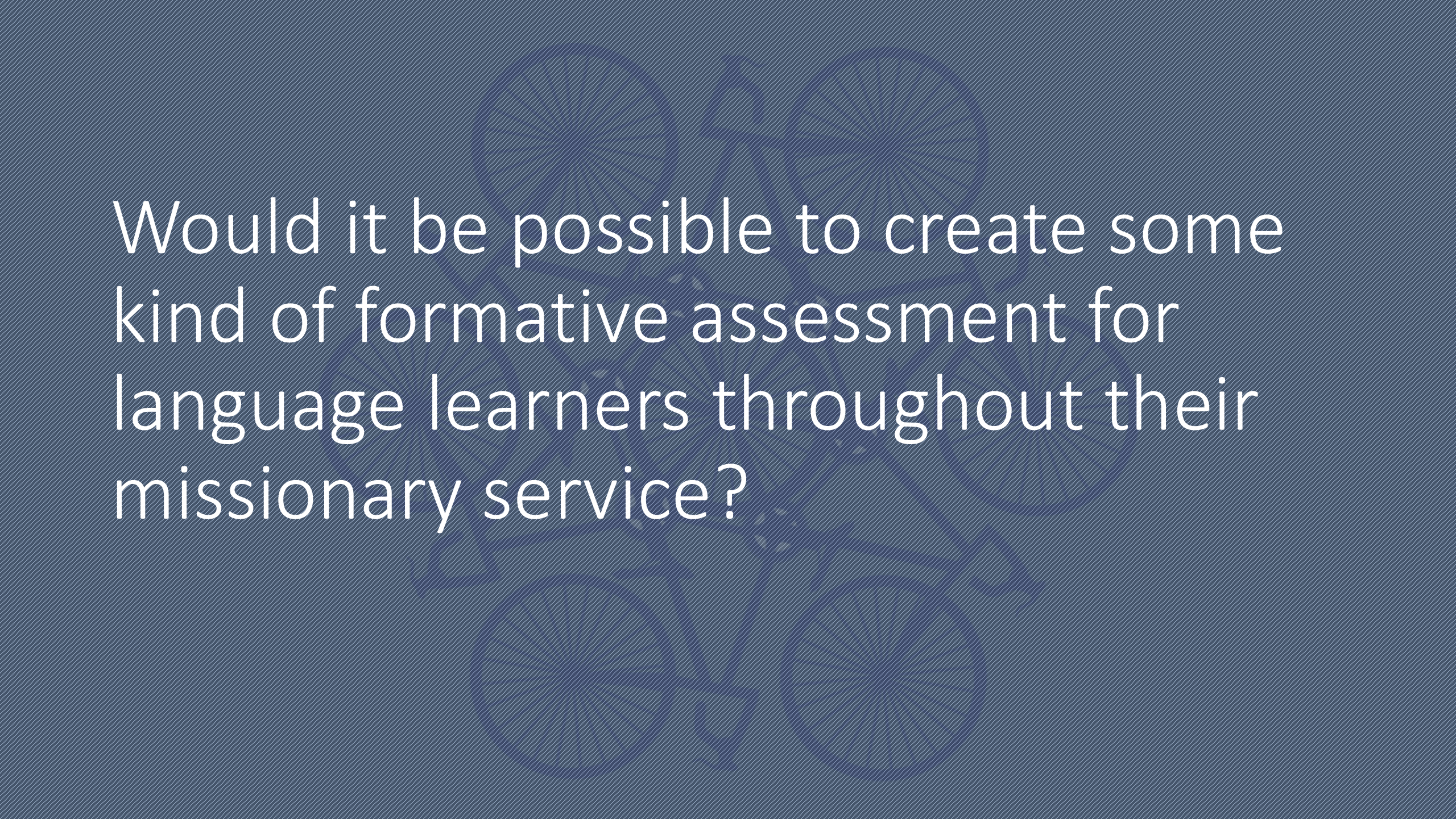
ECOLT



A background image showing a close-up of two hands shaking in a firm grip, symbolizing a partnership or agreement. The hands are wearing business suits. In the blurred background, several other people in professional attire are visible, suggesting a corporate or business setting.

Partnership with Missionary Training Center and BYU

In other words, they poached our CS major student programmer

The background of the slide features a repeating pattern of stylized bicycles and wheels in a light blue color, set against a darker blue textured background.

Would it be possible to create some kind of formative assessment for language learners throughout their missionary service?

Automated Proficiency Assessment

Goal 1:

- Train a predictive model that assigns proficiency ratings as well as trained human raters for high resource languages
 - As a practical goal, we shoot for **60% exact accuracy** and **95% within-one accuracy**

Goal 2:

- Train a predictive model that assigns proficiency ratings well across languages (especially low-resource languages)

ACTFL Agreement Rates

Study Background

- ACTFL statistical analysis by Human Resources Research Organization for ACTFL¹
- Published August 18, 2023
- Contains OPI data for three languages from 2020-2022

Scoring Process:

- A language-learner is interviewed by a rater and assigned a proficiency score
- A second interviewer watches the interview and assigns a proficiency score
- If there is a discrepancy, a third rater provides a rating

Table 2.3. Summary of Exact and Adjacent Agreement of OPI by Language

Language	Rating	1	Exact	Adjacent	Adjacent	Exact + Adjacent
		<i>n</i>	%	<i>n</i>	%	%
Mandarin Chinese	Rating 1 - Rating 2	1,673	68.29	708	28.9	97.18
	Rating 1 - Rating 3	326	42.01	391	50.39	92.4
	Rating 2 - Rating 3	327	42.14	375	48.33	90.46
	Rating 1/2 - Rating 3	653	84.15	116	14.95	99.1
French	Rating 1 - Rating 2	792	69.78	334	29.43	99.21
	Rating 1 - Rating 3	187	54.52	148	43.15	97.67
	Rating 2 - Rating 3	140	40.82	190	55.39	96.21
	Rating 1/2 - Rating 3	326	95.04	14	4.08	99.13
Spanish	Rating 1 - Rating 2	5,479	66.23	2,666	32.24	98.46
	Rating 1 - Rating 3	1,539	55.08	1,171	41.91	96.99
	Rating 2 - Rating 3	1,045	37.4	1,614	57.83	95.23
	Rating 1/2 - Rating 3	2,578	92.27	214	7.66	99.93

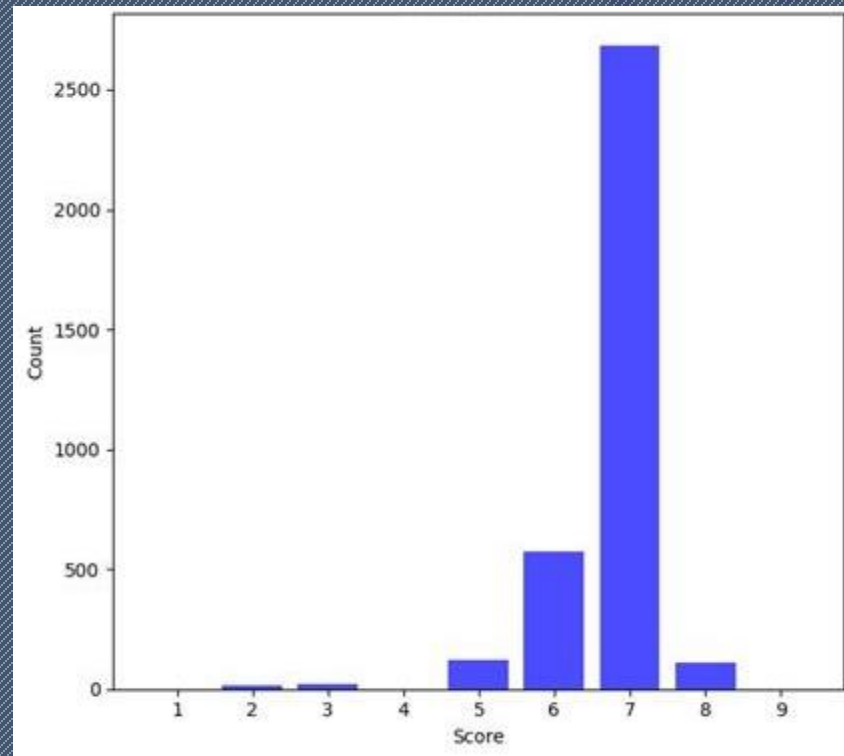
¹ Examination Evaluation of the ACTFL Oral Proficiency Interview (OPI) in Mandarin Chinese, French, and Spanish for the ACTFL Review, ACTFL, August 18, 2023. https://www.actfl.org/uploads/files/general/Documents/assessments/assessments/OPI_Report_Part_B_Statistical_Analysis_2023.pdf

Elicited Imitation (EI) Test

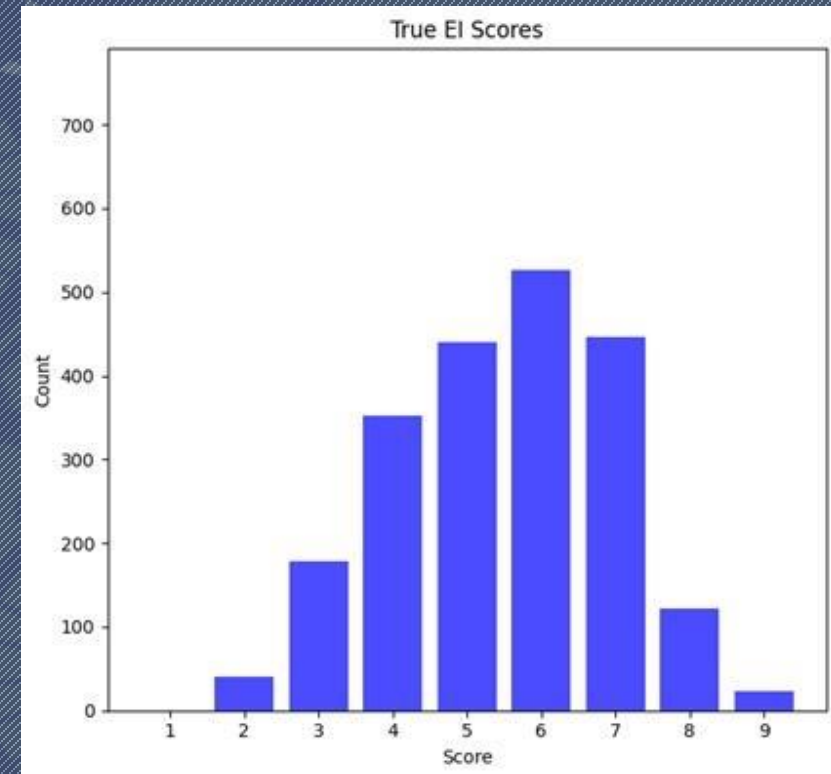
- Published by TrueNorth (now RosettaStone)
- Users heard audio in their target language and repeat it
 - They were scored based on the faithfulness of repetition (0-10 scale)
- At the end, open-ended speaking prompts were recorded
 - Open-ended responses were NOT scored, however,
 - They were willing to give the open-ended responses to the MTC
- Supported in nine languages
 - English, Spanish, Portuguese, French, Japanese, Mandarin, Italian, German, Tagalog

Why use the open-responses from the EI Test?

Distribution of OPIc Scores based on LASER participants



Distribution of EI Scores based on missionaries



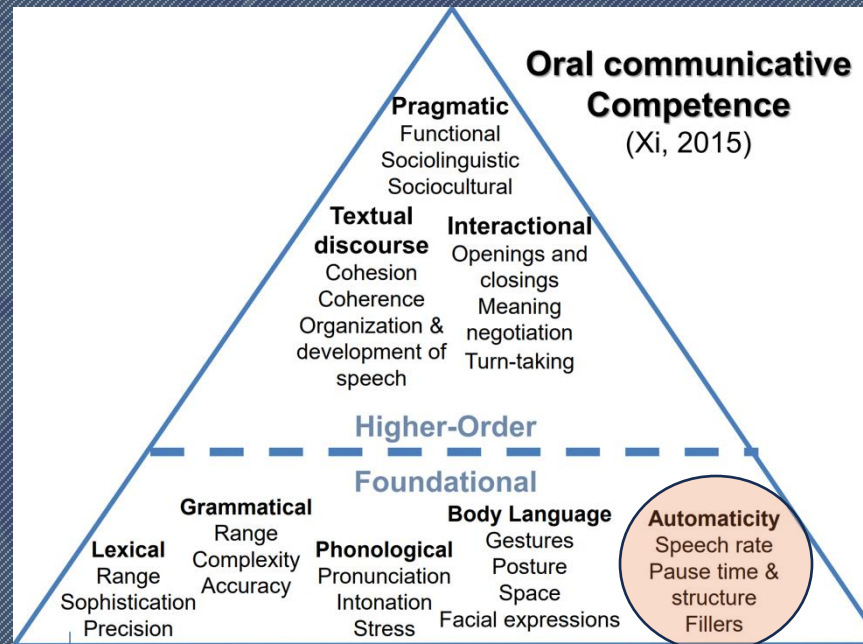
EI Data Breakdown

- We analyze the open responses from the EI test (NOT the elicited imitation audio) and their associated scores
- Three audio files per respondent
 - Average 49 s duration per audio file
 - 2 ½ minutes of recorded audio
- Demographics
 - Roughly 18-25 years old
 - Missionaries learning a language, often in fully immersive environments

Language	Entries
Portuguese	5,008
Spanish (Sample)	3,328
French	1,760
German	951
English	747
Italian	639
Japanese	627
Mandarin	423
Tagalog	157

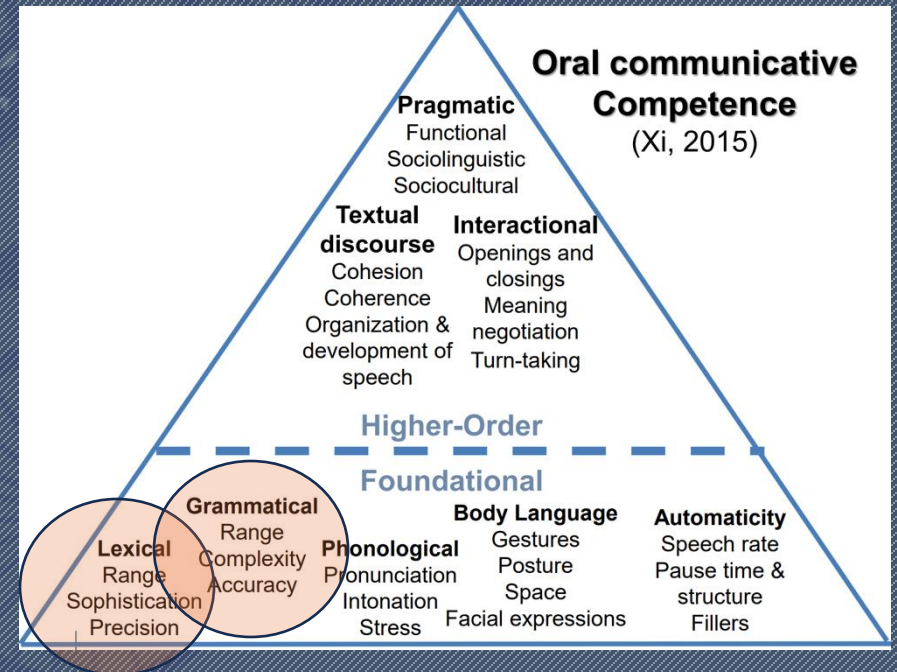
Feature Extraction Part 1

- Audio features extracted with Praat
 - Articulation rate, speech rate, average syllable duration, mean length of utterance, pause frequency



Feature Extraction Part 2

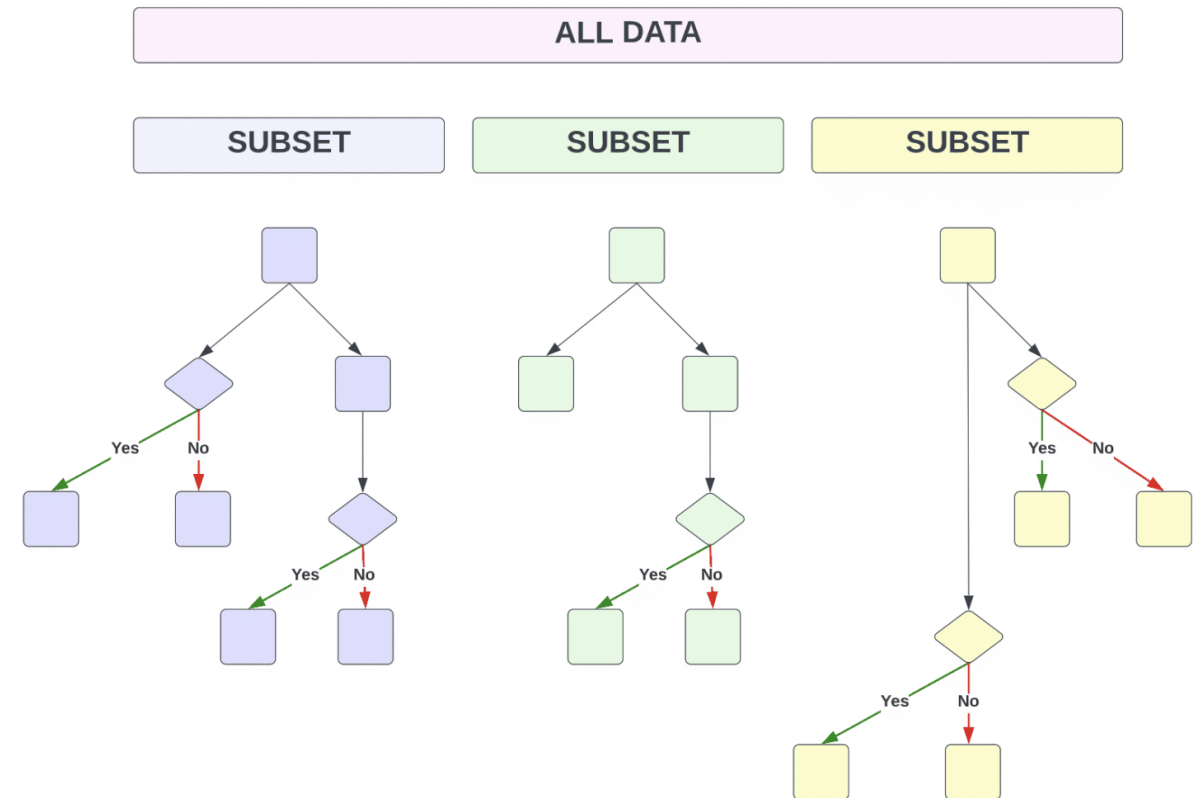
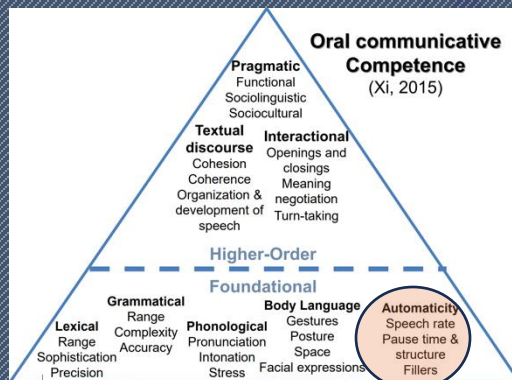
- Each audio file is transcribed using OpenAI's Whisper (large version) which supports 98 languages
- From transcriptions, we extract additional features:
 - Lexical diversity measures
 - Error frequencies
 - POS, tense, and mood frequencies



Modeling Approach 1: XGBoost



- Boosted decision tree
- Trained on audio features
- Tested prediction accuracy against true EI score



Sankshep Mahendra, "Introduction to XGBoost and Its Uses in Machine Learning," *AI Plus Info*, June 22, 2022, <https://www.aiplusinfo.com/files/introduction-to-xgboost-and-its-uses-in-machine-learning/>.

XGBoost Results on EI Open Responses

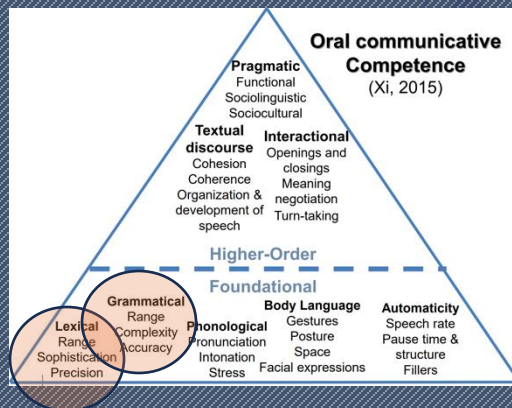
Scoring Metric	Score	Target Score
Accuracy	0.43	0.60
Within-One Accuracy	0.88	0.95



Modeling Approach 2: mDeBERTa



- Multilingual transformer-based language model
- ChatGPT is also a transformer, but it is built for generative tasks while mDeBERTa is built for understanding tasks
- Fine-tuned on transcriptions of open-response audio and true EI score



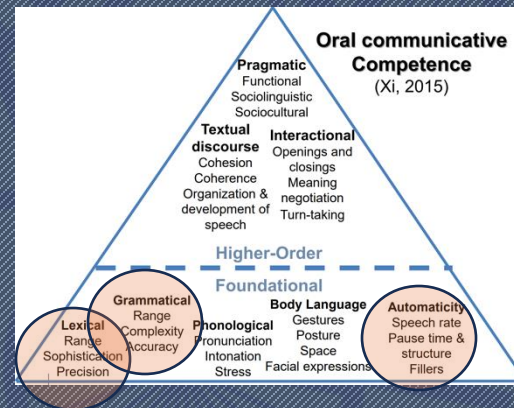
mDeBERTa Results

Scoring Metric	mDeBERTa Score	XGBoost Score	Target Score
Accuracy	0.53	0.43	0.60
Within-One Accuracy	0.96	0.88	0.95
Spearman Correlation	0.90	--	--



Modeling Approach 3: Combined Model

1. Get a prediction from mDeBERTa based on the transcription
2. Use XGBoost on the most important features PLUS the mDeBERTa prediction
3. Predict an EI rating



Combined Model Results

Scoring Metric	Combined Model Score	mDeBERTa Score	XGBoost Score
Accuracy	0.50	0.53	0.43
Within-One Accuracy	0.94	0.96	0.88



Target
Score

0.60

0.95

Testing on an Unseen Language

- Multi-lingual model fine-tuned on EI open response data from 8 languages
- Tested on a held-out Tagalog EI dataset of 157 respondents

Language	Entries
Portuguese	5,008
Spanish (Sample)	3,328
French	1,760
German	951
English	747
Italian	639
Japanese	627
Mandarin	423
Tagalog	157

Testing on an Unseen Language Results

Scoring Metric	Score
Accuracy	0.47
Within-One Accuracy	0.90
Spearman's Correlation	0.79

Target Score

0.60

0.95



Confusion Matrix										
1	0	0	0	0	0	0	0	0	0	
2	0	0	0	1	0	0	0	0	0	
3	0	0	12	17	2	0	0	0	0	
4	0	0	4	8	7	1	0	0	0	
5	0	0	3	14	28	1	0	0	0	
6	0	0	0	6	13	15	2	0	0	
7	0	0	0	0	1	5	3	1	0	
8	0	0	0	0	0	1	1	3	2	
9	0	0	0	0	0	0	0	1	4	
10	0	0	0	0	0	0	0	0	1	
	1	2	3	4	5	6	7	8	9	10

Back to the LASER and OPIC Scores

- We used our best EI model to predict proficiency levels on the LASER data

Language	Entries
Spanish	1,968
French	581
Portuguese	524
Italian	129
German	128
Russian	126
Japanese	63
Swedish	1
Albanian	1

LASER and Proficiency (ILR/ACTFL) Prediction Results



Scoring Metric	mDeBERTa + XGBoost mapping	Target Score
Accuracy	0.49	0.60
Within-One Accuracy	0.94	0.95

To What Extent Did We Reach Our Goals?

Goal 1:

- Train a predictive model that assigns proficiency ratings as well as trained human raters for high resource languages
 - As a practical goal, we shoot for 60% exact accuracy and 95% within-one accuracy

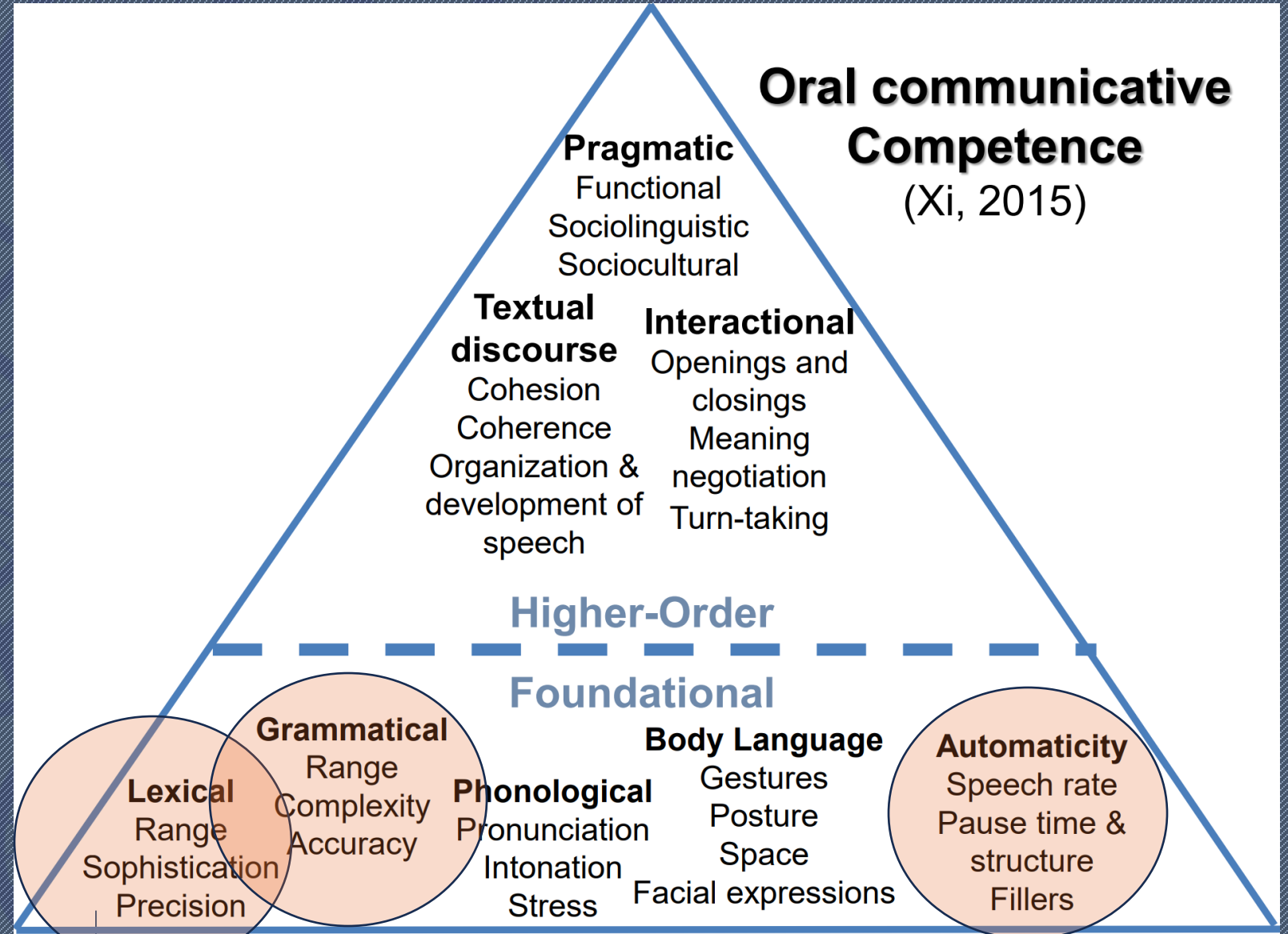
Goal 2:

- Train a predictive model that assigns proficiency ratings well across languages (especially low-resource languages)

Key Takeaways

- **Foundational skills such as temporal fluency, lexical range, and grammatical accuracy might work across languages**
- **Practical applications**
 - Scalable alternative to human raters
 - Time, cost, and effort
- **Limitations**
 - Good for low-stakes (limited for high-stakes)
 - Combining different datasets

ECOLT 2015



Future Work

- Expand to more languages
- Moving from Advanced to Superior likely requires language specific higher order skills in addition to the foundational

The background of the slide is a solid dark blue color. Overlaid on this background is a repeating pattern of light blue line art. The pattern consists of bicycles and wheels arranged in a circular, radial fashion, creating a mandala-like effect. The bicycles are shown from a side profile, and the wheels are depicted with many thin spokes. The pattern is centered and covers most of the slide area.

Questions?

- So, what do cycling, yoga, Great Britain, and an applied linguist have in common?



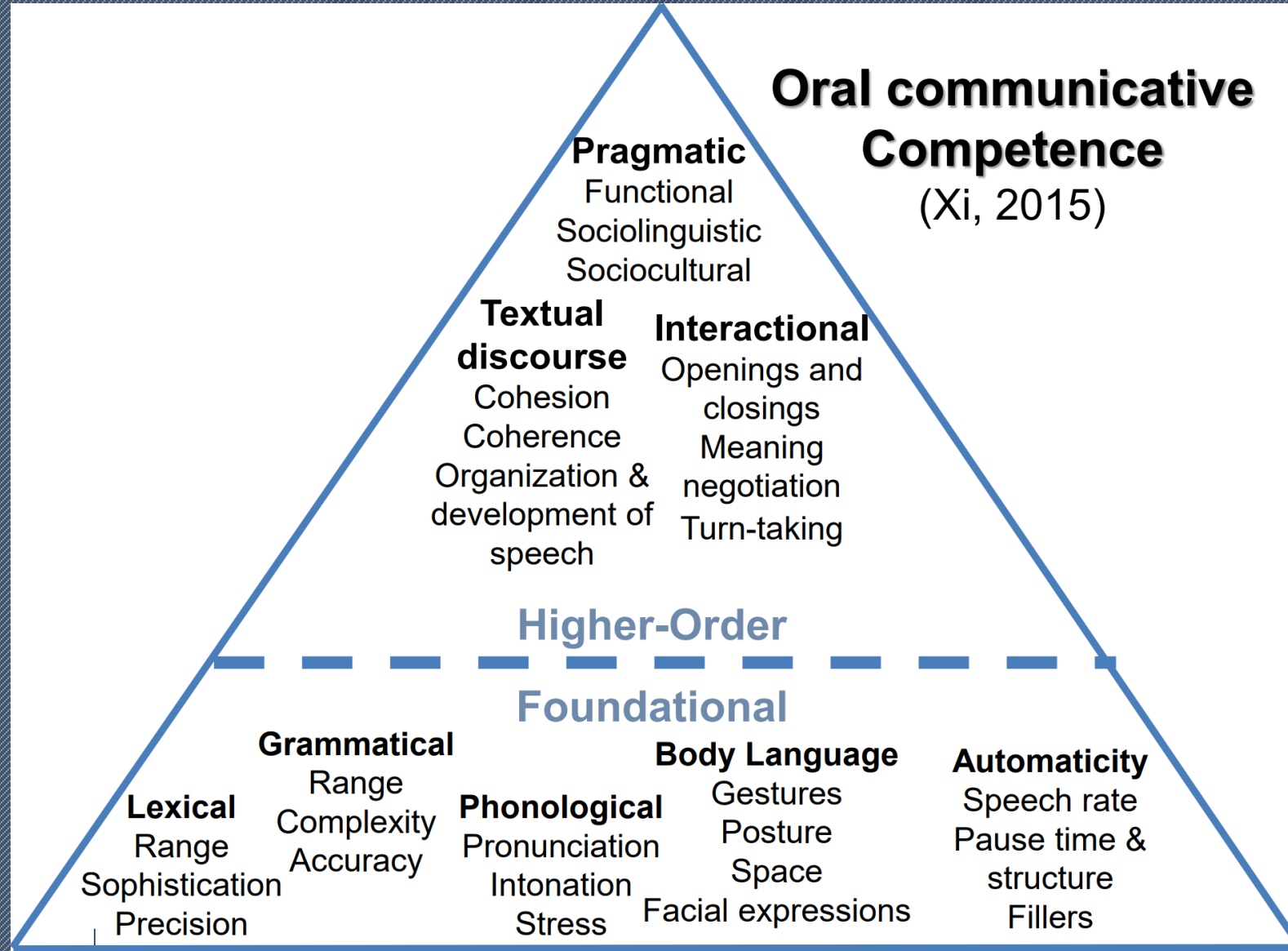
Aggregation of Marginal Gains



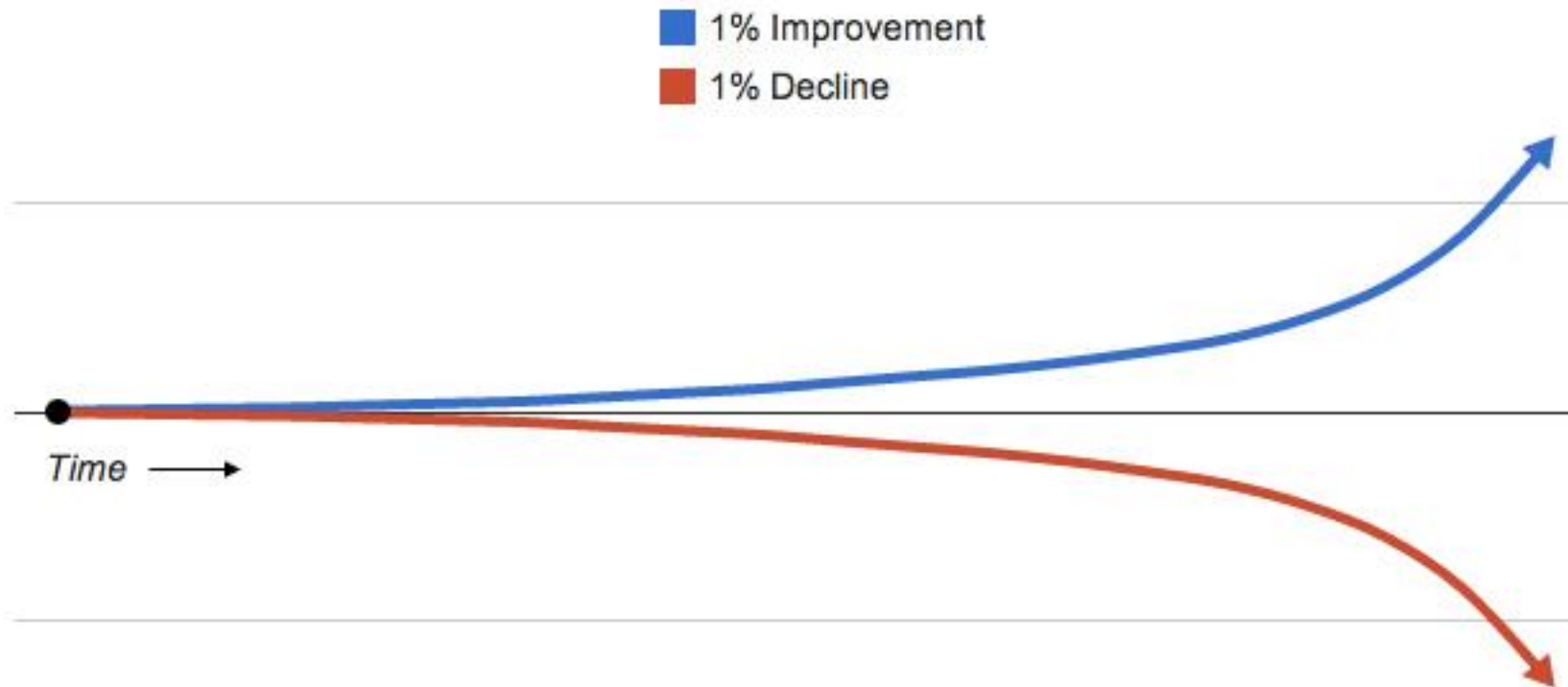
Marginal Gains

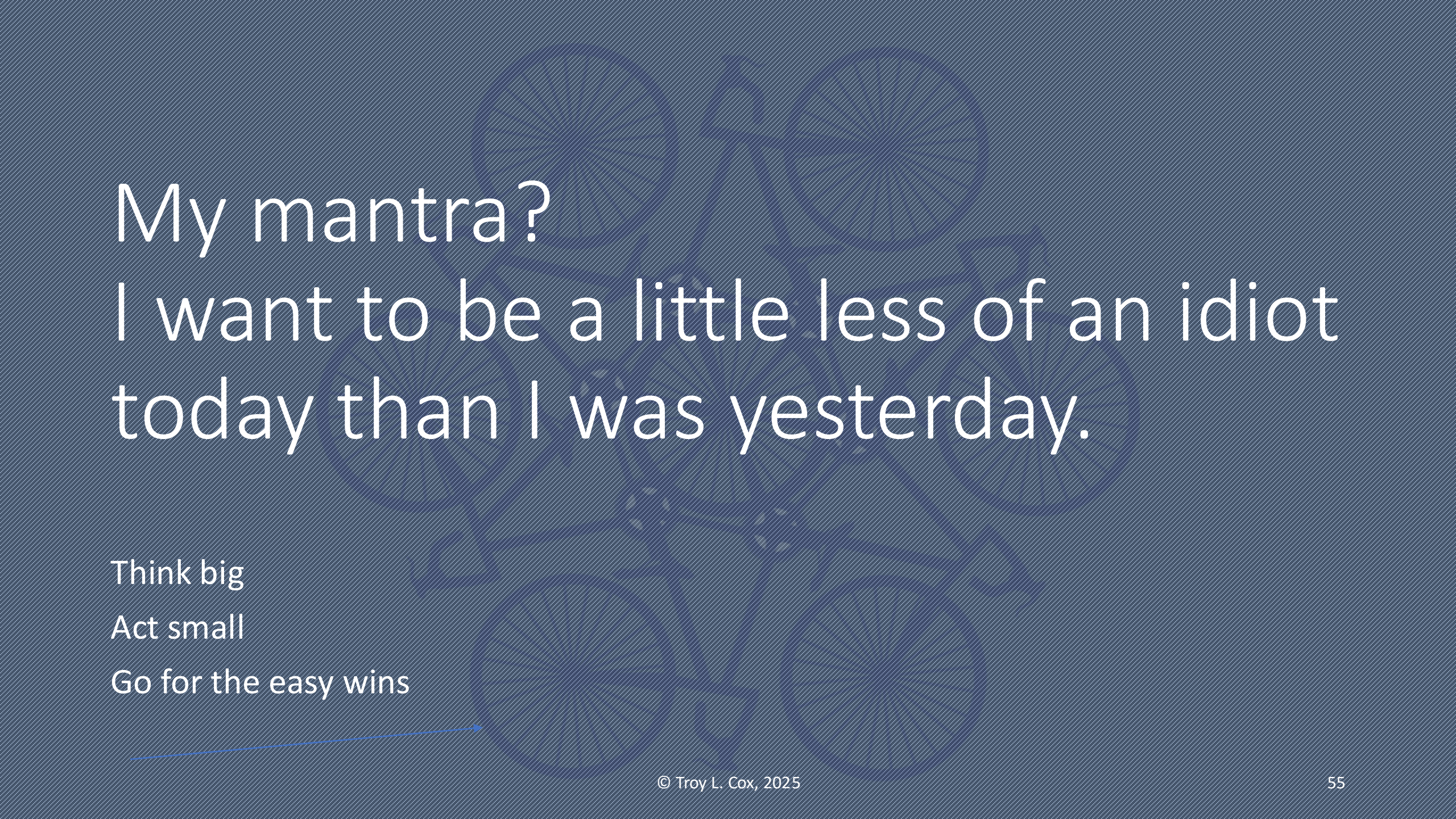
A method of reaching high-performance levels through constantly making small, incremental improvements.

Oral communicative Competence (Xi, 2015)



Aggregation of Marginal Gains



The background of the slide features a repeating pattern of stylized bicycles and wheels in a light blue color, set against a dark blue background.

My mantra?

I want to be a little less of an idiot
today than I was yesterday.

Think big

Act small

Go for the easy wins

