

The Effect of Language Proficiency on Self-Appraisal Calibration among Advanced Foreign Language Learners

LTRC June 9, 2023

Troy Cox <troyc@byu.edu>

Brigham Young University, Provo UT
troycox.xyz

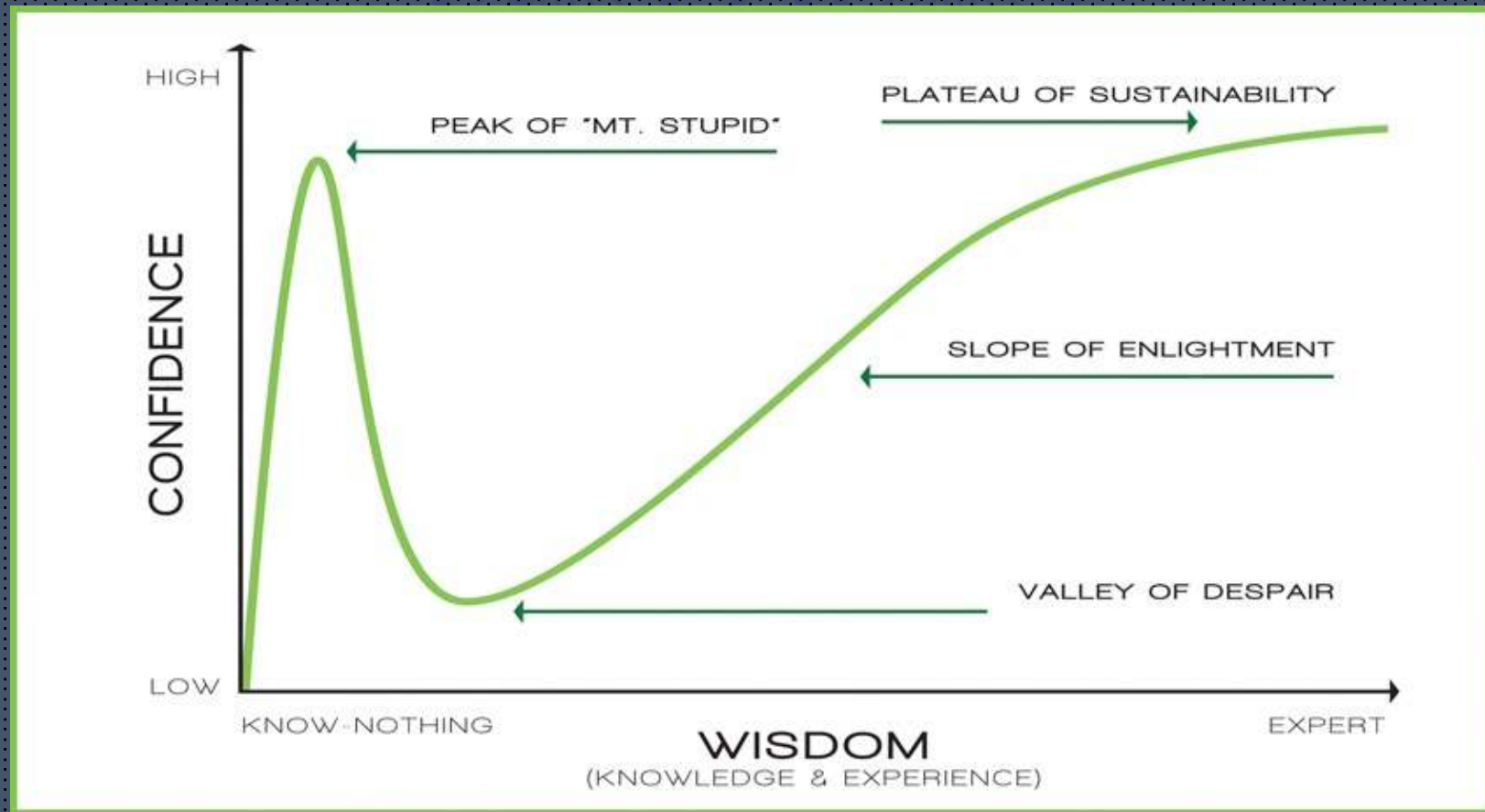
Whatever you choose
to test
becomes the
de facto learning
objectives
for a curriculum.

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

-Paraphrase of McTighe & Wiggins

The Dunning–Kruger effect is a cognitive bias whereby people with low ability, expertise, or experience regarding a certain type of a task or area of knowledge tend to overestimate their ability or knowledge.

Surprising? Not Really...



We are all subject to the
Dunning-Kruger Effect. But are
we aware of it?



© Troy L. Cox, 2022

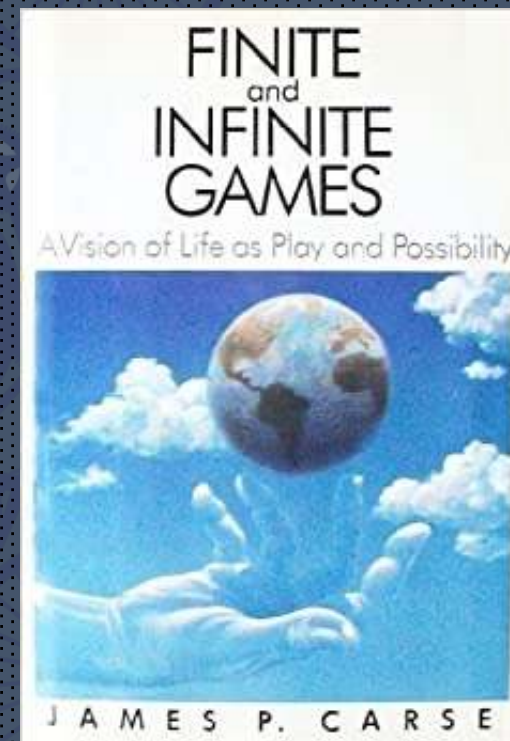




Is language learning a finite or an infinite game?

Carse, J. (2011). *Finite and infinite games*. Simon and Schuster.

Sinek, S. (2019). *The infinite game*. Penguin.



THE
INFINITE
GAME
**THE
INFINITE
GAME**
**SIMON
SINEK**
THE
INFINITE
GAME
THE
INFINITE
GAME

What's the difference between finite and infinite games?

Finite

- Played to win
- Externally defined rules
- Limits are predefined
- Players can compete to win within these limits
- Examples:
 - a football match (win/lose)
 - publishing a paper (accept/reject)
 - job interview (hired/not hired)
 - going to church (present/absent)

Infinite

- Played to keep on playing
- Internally defined rules.
- Limits are not defined
- The goal of players is to stay in the game
- Examples:
 - being an athlete for the joy of the game
 - being a scholar for the joy of discovery
 - Having a career that's fulfilling
 - Being a disciple for the purpose of fulfillment

Finite games can be within
infinite games but infinite games
can NOT be within finite games

“To be prepared against surprise is to be trained. To be prepared for surprise is to be educated.”

—James Carse

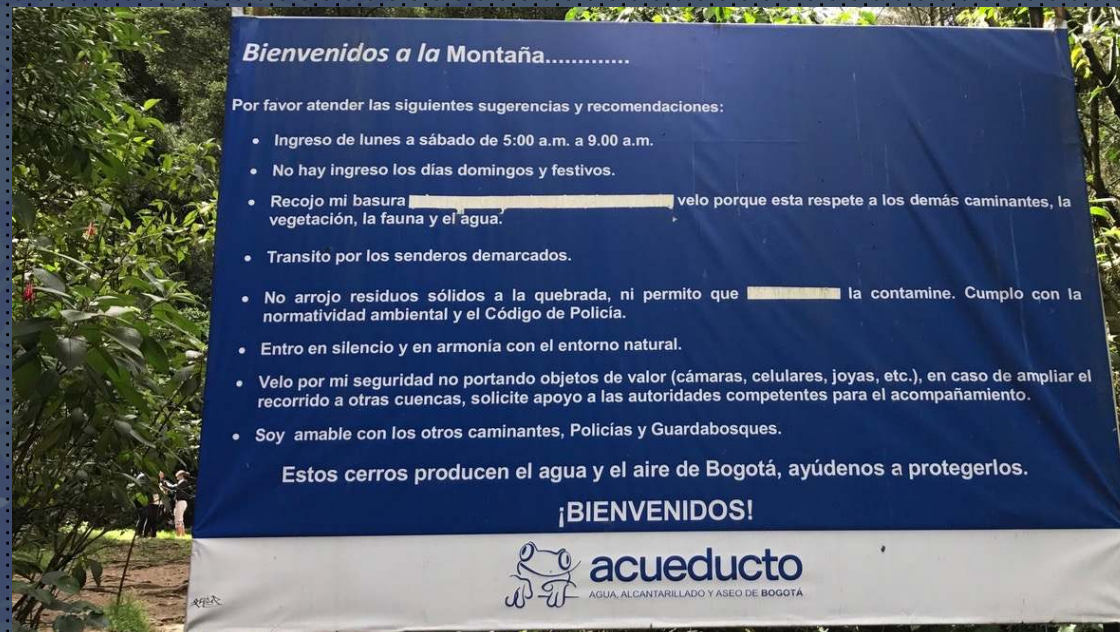
Is our goal language training or education?

What message am I sending to students through the tests I help develop?

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

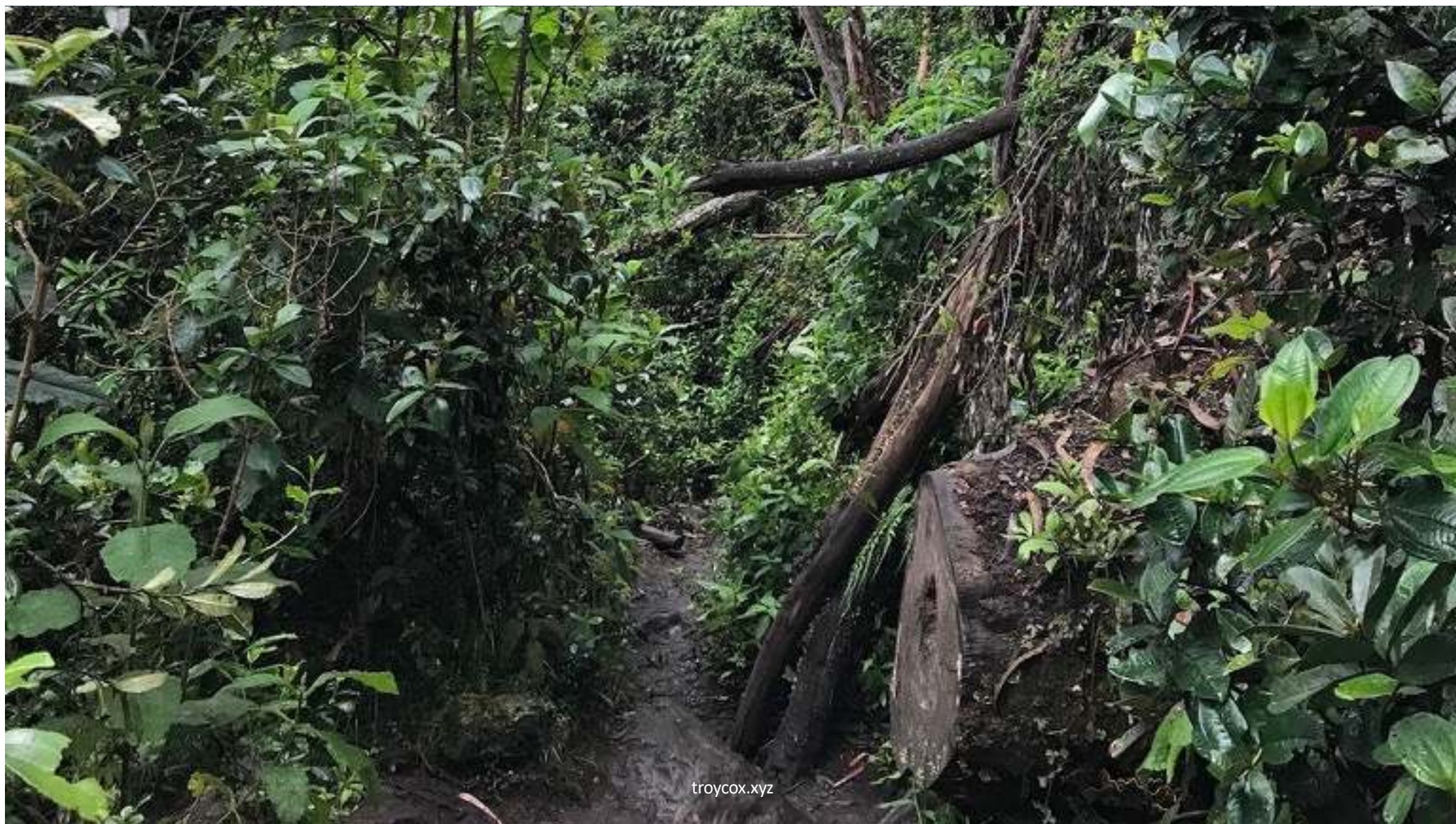
The way you choose
to assess
it becomes the
de facto teaching
and learning
philosophy.

-Paraphrase of McTighe & Wiggins









troycox.xyz





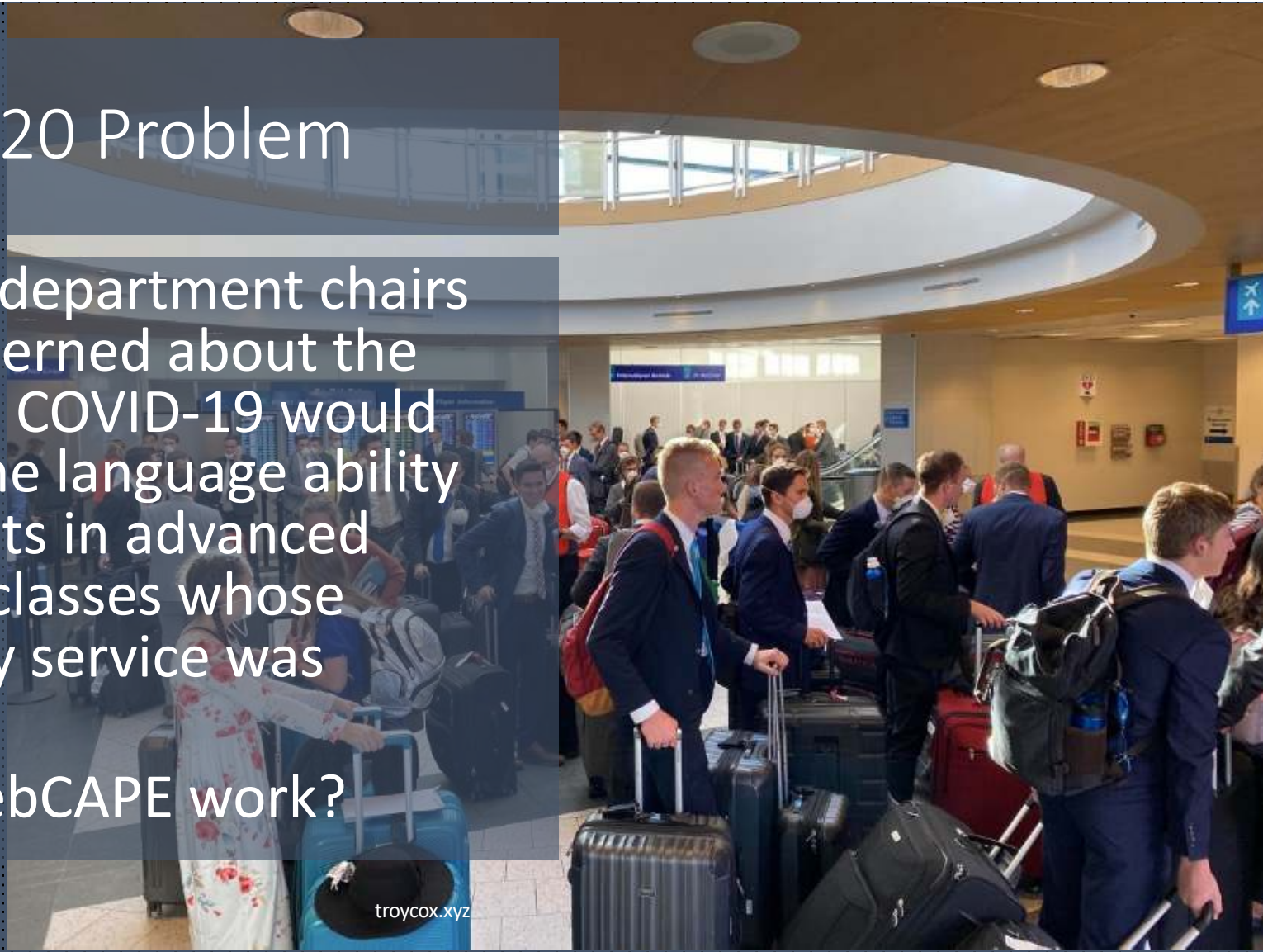
Am I self-aware and acting on my own? Do my actions align with my values?

Or am I waiting for something external to tell me what to do?



Spring 2020 Problem

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



troycox.xyz

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.



What's our goal?

- Do we want students that are automats that need to be told where to go and what to do via external checklists for a **finite game**?
- Or do we want to help students become reflective learners that can accurately self-appraise their language proficiency, choose what's best for them and enter the **infinite game** of being a lifetime language learner?

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.

-Paraphrased of McTighe & Wiggins

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

Overview

The LASER is designed to help BYU students:

- become life-long language learners as they improve their ability to self-assess and self-reflect on their language proficiency
- Provide recommendations for the appropriate language course for their ability
- verify that they are in the correct class (with instructors' feedback)
- provide information to the instructor to inform curricular decisions

Section 1: Self-report Module (30-40 minutes)

- a. Language Learning Background
- b. Language-learning Grit
- c. Willingness to communicate
- d. Productive Skills Self-assessment

Section 2: Productive Language Samples Module (20-50 minutes)

- a. Writing Test and Self-assessment
- b. Speaking Test and Self-assessment
- c. Reflection on performance (after each writing or speaking response)

Scale Design and Description

	LASER (Rasch Reliability = .94)
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 and Speaking
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) troycox.xyz

Research Questions

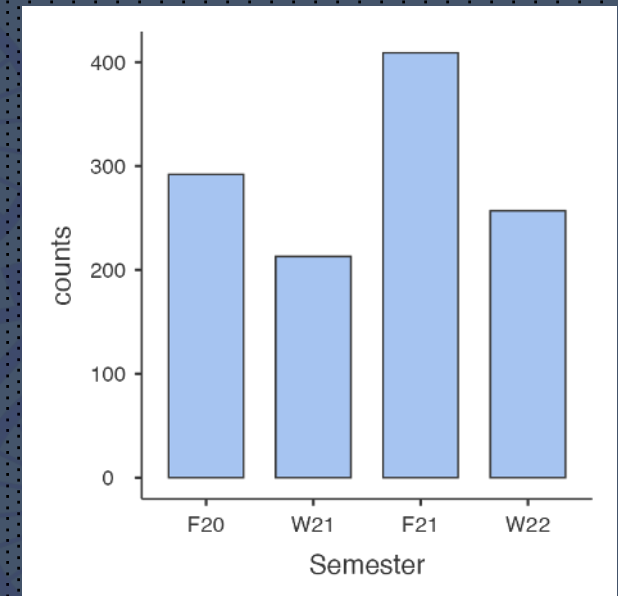
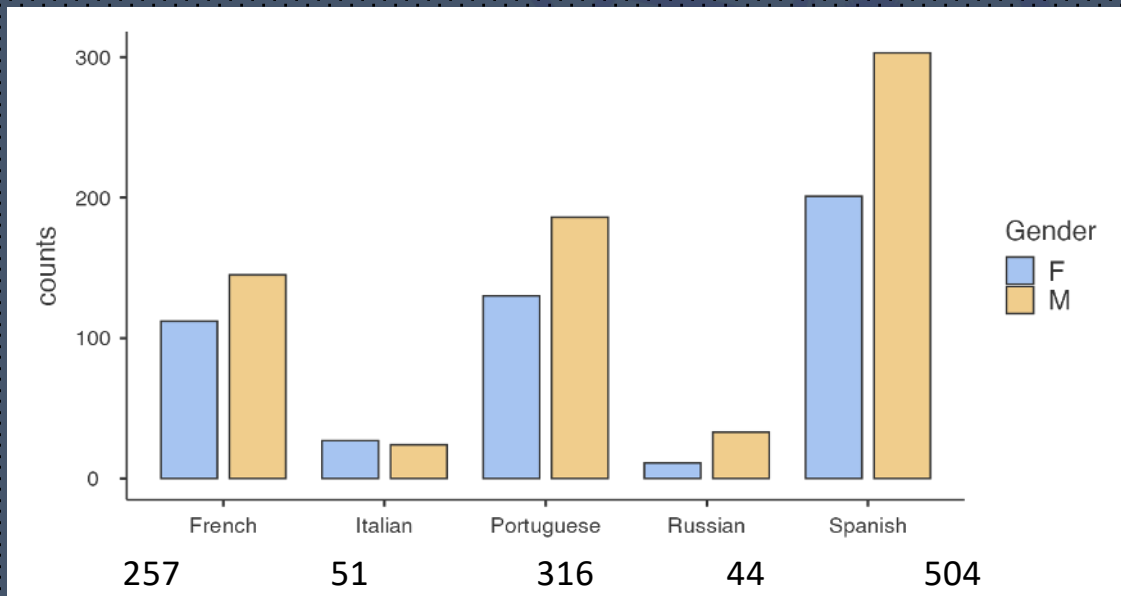
- RQ1a—How did ratings change over the course of the pandemic for self-assessed and actual proficiency in speaking?
 - RQ1b—Writing?
- RQ2a—What effect does proficiency level have on calibration in speaking?
 - RQ2b—Writing?

Materials

- LASER-Language Ability Self-assessment Evaluation Resource
- OPIc- Oral proficiency Interview-computer
- WPT-Writing Proficiency Test

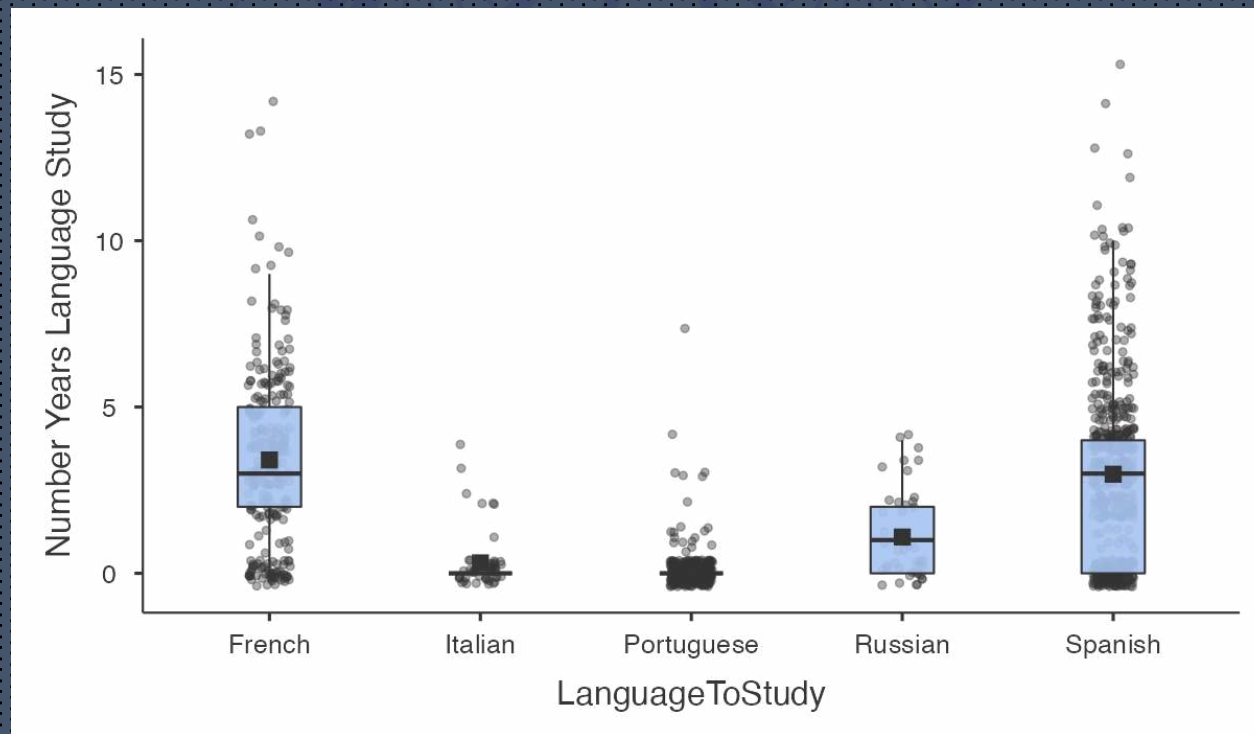
Participants

- 1,172 (691 male, 481 female) over 4 semesters

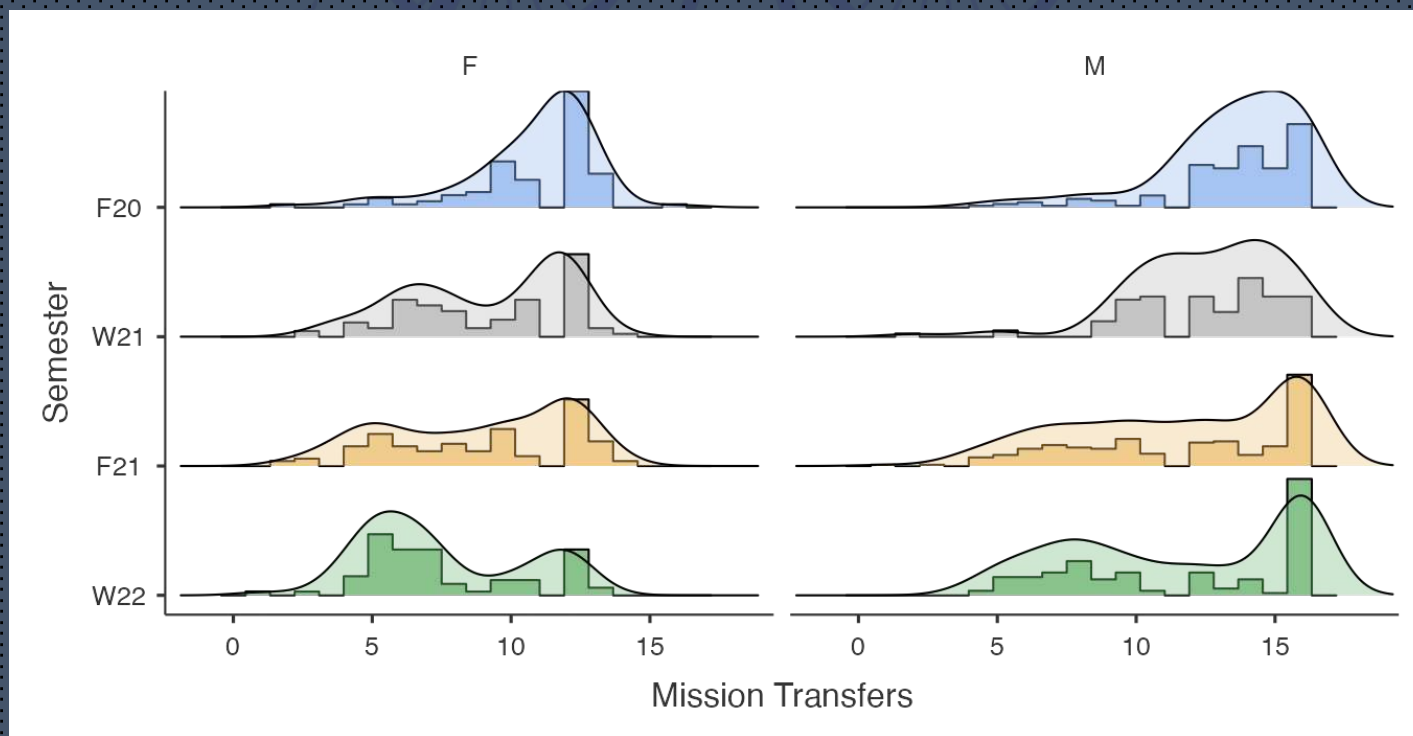


	Counts	% of Total
F20	292	24.9 %
W21	213	18.2 %
F21	409	34.9 %
W22	257	21.9 %

Years of Language Study



Missionary Service ($N = 1,036$ (out of 1,172))



$N = 375$

troycox.xyz

$N = 661$

Calculating Calibration

- Calibration = OPIc Rating – Self Assessed Rating

Value	Meaning
0	Perfectly calibrated. No difference between perceived and actual ability.
Positive Numbers	Students OVER-estimate their ability.
Negative Numbers	Students UNDER-estimate their ability

Results

RQ1a-Speaking Self-Assessments Over Course of Pandemic

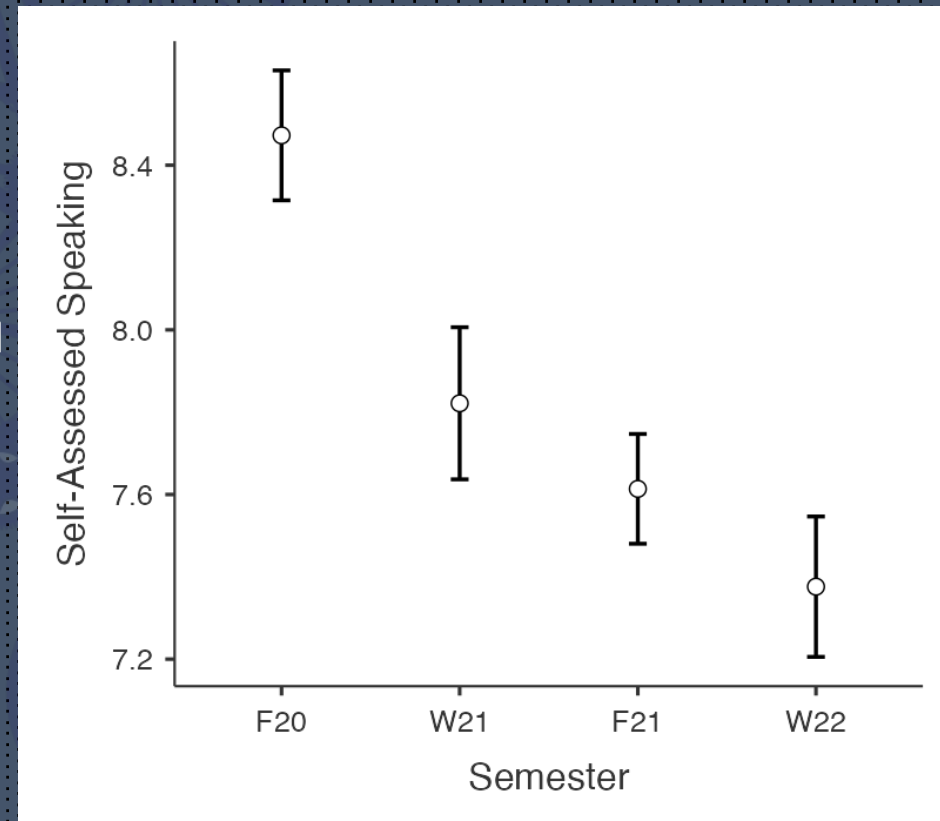
Semester	95% Confidence Interval			
	Mean	SE	Lower	Upper
F20	8.47	0.08	8.31	8.63
W21	7.82	0.09	7.64	8.01
F21	7.61	0.07	7.48	7.75
W22	7.38	0.09	7.21	7.55

ANOVA

$F(3,1160) = 33.7, p < .001$

AM

AL



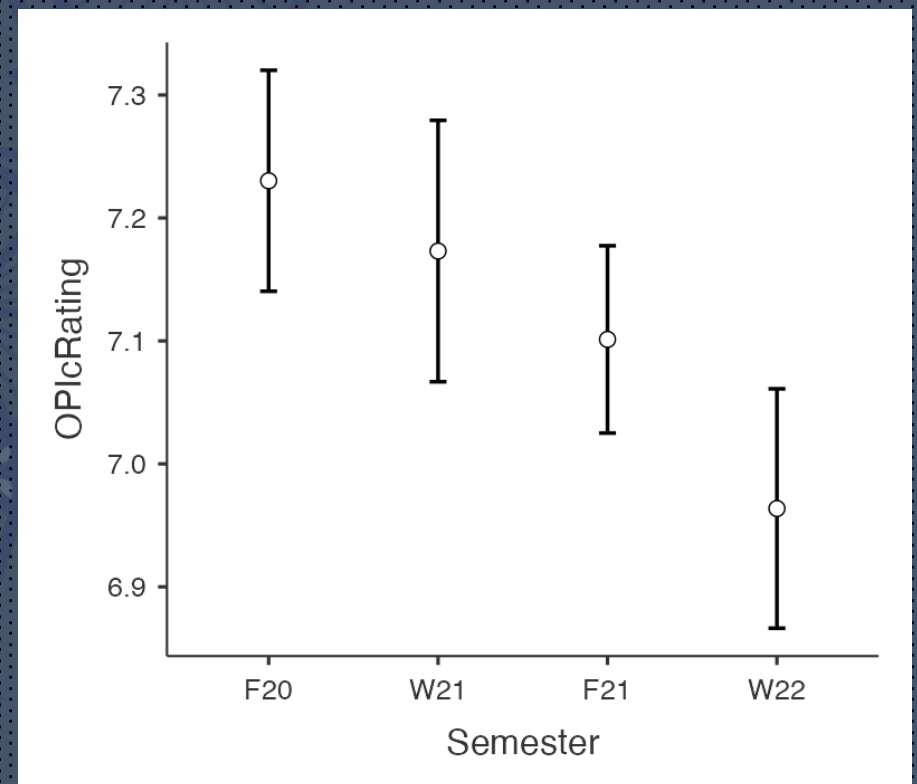
RQ1a-OPIc Ratings Over Course of Pandemic

Semester	95% Confidence Interval			
	Mean	SE	Lower	Upper
F20	7.23	0.05	7.14	7.32
W21	7.17	0.05	7.07	7.28
F21	7.10	0.04	7.03	7.18
W22	6.96	0.05	6.87	7.06

ANOVA

$$F(3,1148) = 5.63, p < .001$$

AL

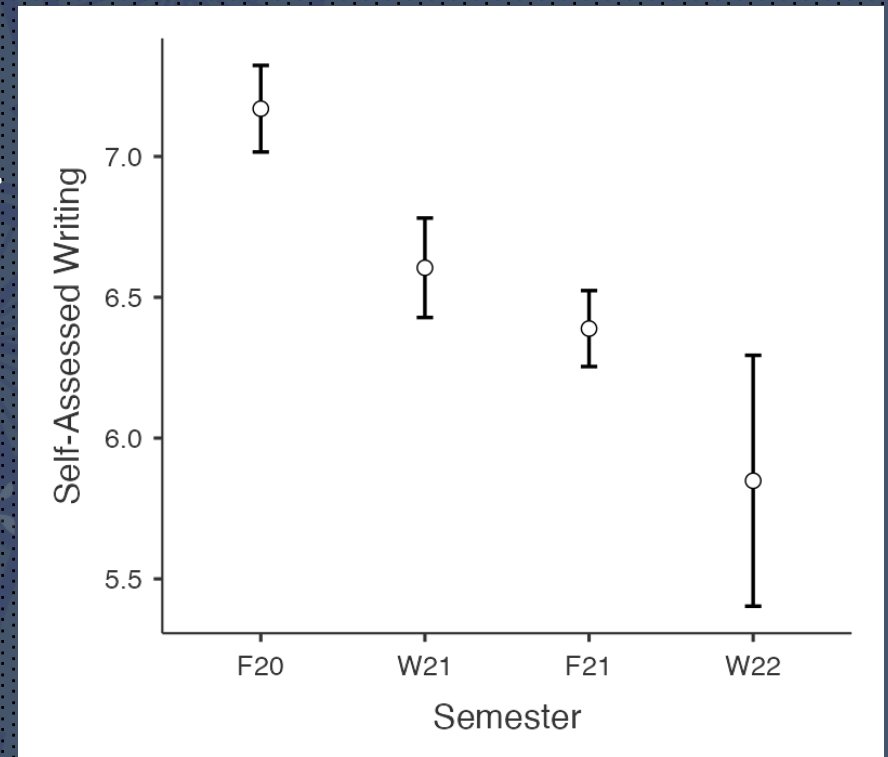


RQ1b-Writing Self-Assessment Over Course of Pandemic

Semester	95% Confidence Interval			
	Mean	SE	Lower	Upper
F20	7.17	0.07	7.02	7.32
W21	6.60	0.09	6.43	6.78
F21	6.39	0.07	6.25	6.52
W22	5.85	0.22	5.40	6.29

AL

IH



ANOVA

$$F(2, 844) = 28.5, p < .001$$

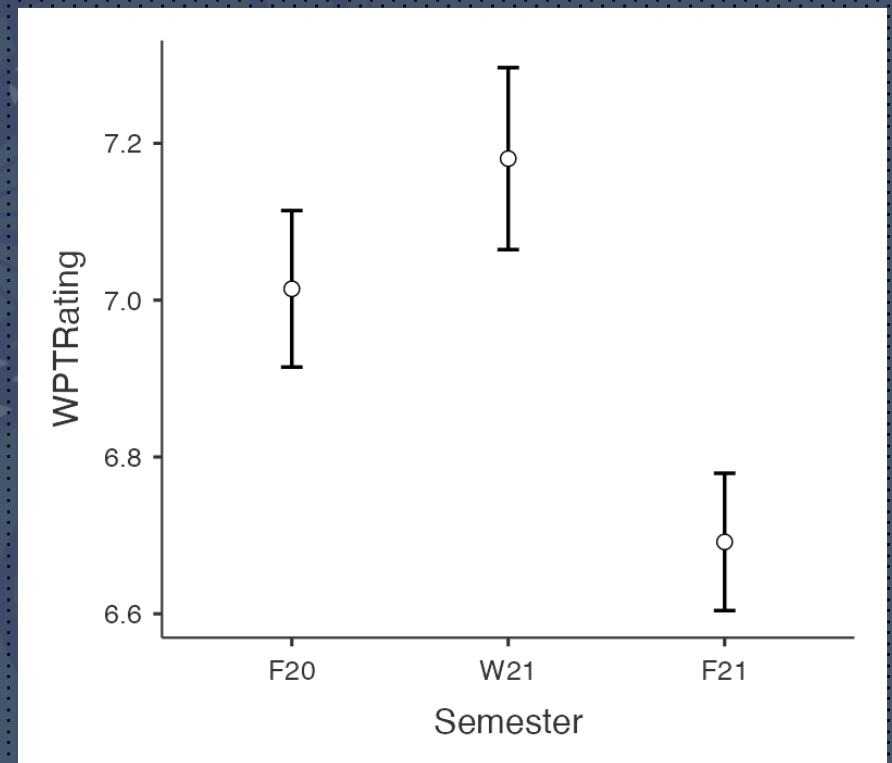
RQ1b-WPT Ratings Over Course of Pandemic

Semester	95% Confidence Interval			
	Mean	SE	Lower	Upper
F20	7.01	0.05	6.91	7.11
W21	7.18	0.06	7.06	7.30
F21	6.69	0.04	6.60	6.78
W22	No Data			

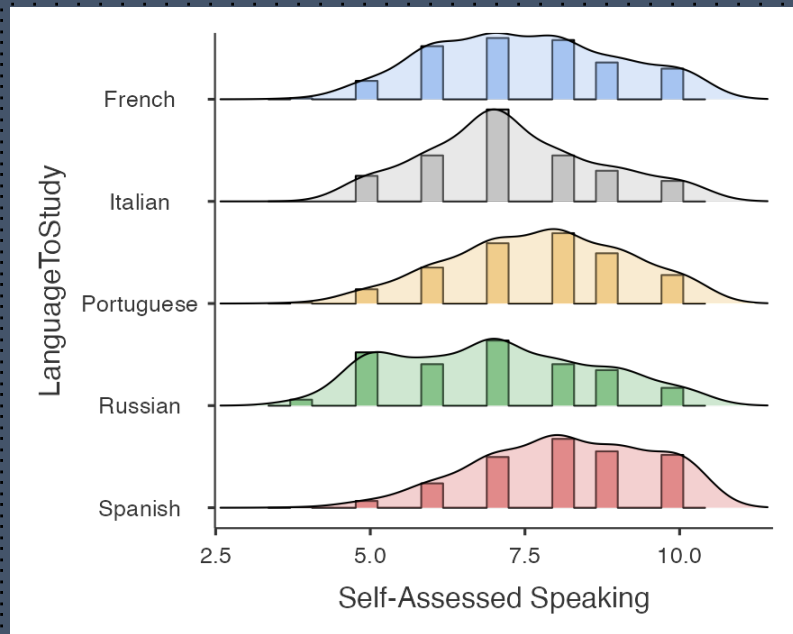
AL

ANOVA

$$F(2, 839) = 24.5, p < .001$$



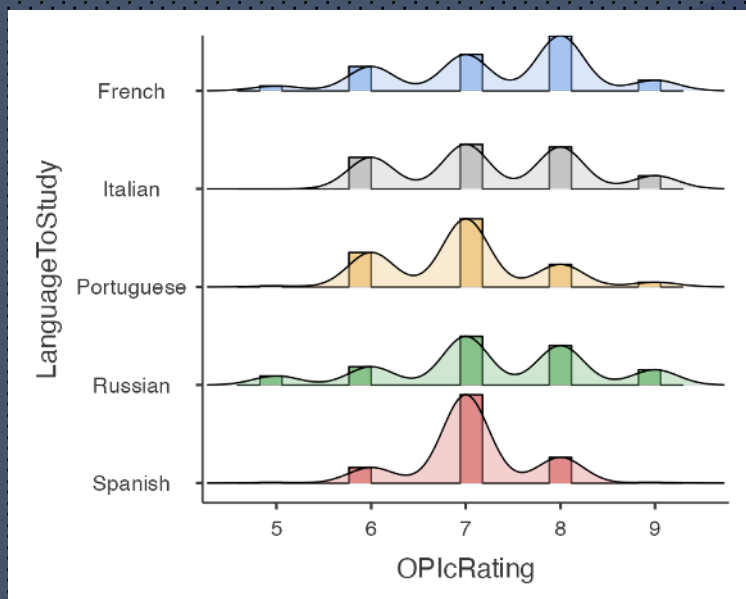
Both self-assessed and actual proficiency in speaking and writing decreased over the course of the pandemic.



RQ2a- Self-Assessed (LASER) Speaking Results

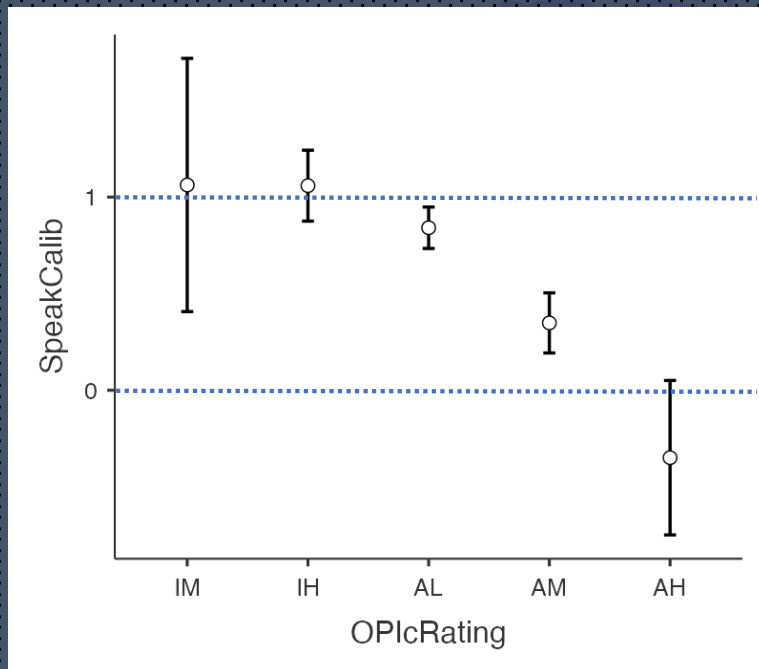
	IL	IM	IH	AL	AM	AH	S	Total
French	1	18	52	60	58	36	30	255
Italian	0	5	9	18	9	6	4	51
Portuguese	1	17	43	72	84	60	34	311
Russian	1	9	7	11	7	6	3	44
Spanish	1	13	47	98	133	109	102	503
Total	4	62	158	259	291	217	173	1164

RQ2a- OPIc Ratings



	BR/UR	NH	IL	IM	IH	AL	AM	AH	S	Total
French	6		1	9	46	69	105	20	0	249
Italian		1			12	17	16	5	0	50
Portuguese	7		2	2	81	160	53	11	0	307
Russian		1		3	6	16	13	5	0	43
Spanish	1			2	60	340	99	2	0	503
Total	14	2	3	16	205	602	286	43	0	1152

RQ2a—Speaking Calibration



Each integer represents ONE sublevel.

ANOVA

$$F(4, 1140) = 17.1, p < .001$$

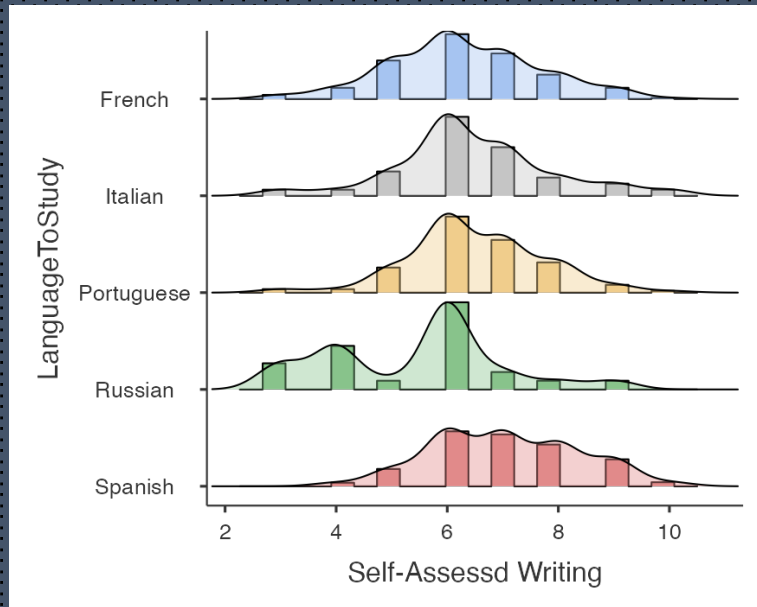
	N	Mean	SE	Lower	Upper
IM	16	1.06	0.33	0.41	1.72
IH	205	1.06	0.09	0.88	1.24
AL	602	0.84	0.05	0.73	0.95
AM	286	0.35	0.08	0.19	0.50
AH	43	-0.35	0.20	-0.75	0.05

troycox.xyz

For speaking, the higher the proficiency level of the examinee, the lower they were in their calibration.

IM & IH speakers over-assessed their ability by more than 1 sublevel.

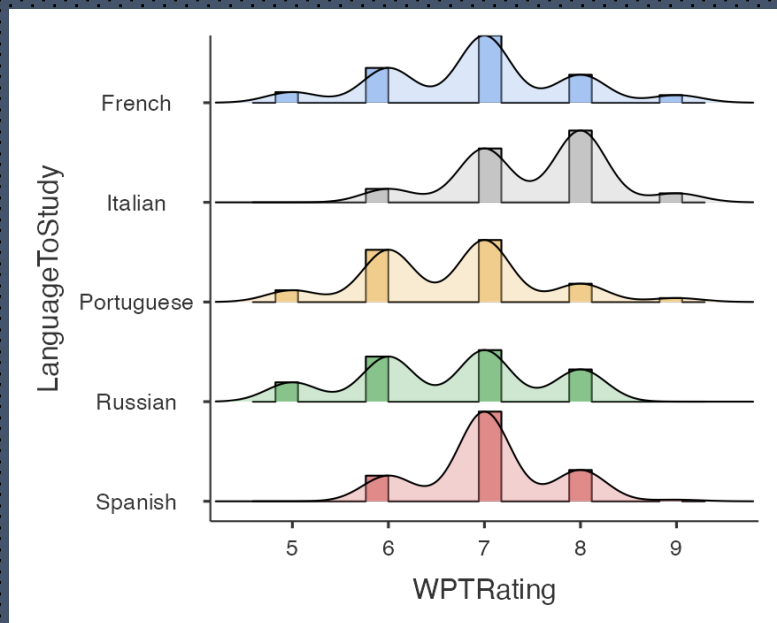
AH speakers under-assessed their ability.



RQ2b- Self-Assessed (LASER) Writing

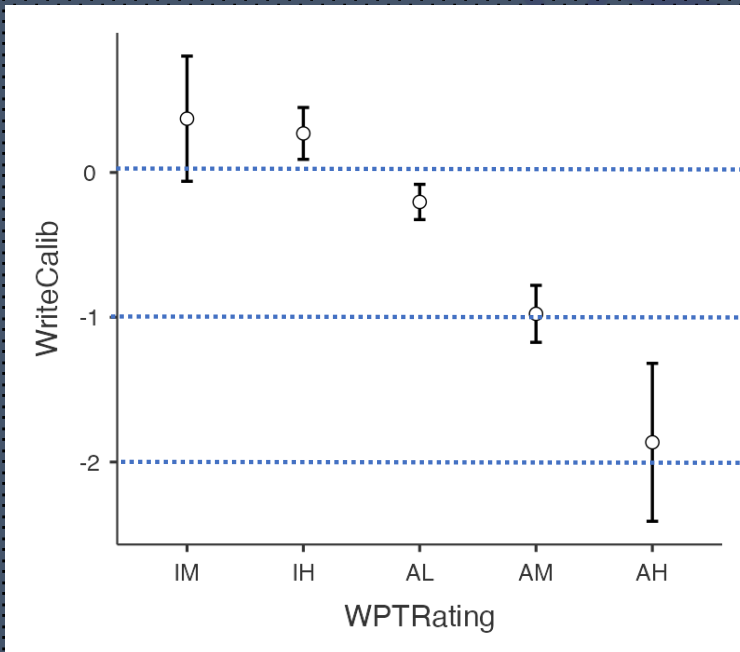
	NH	IL	IM	IH	AL	AM	AH	S	Total	
French		4	7	38	63	45	24	11	1	193
Italian		1	1	4	12	8	3	2	1	32
Portuguese		4	4	29	85	60	33	9	2	226
Russian		3	5	1	10	2	1	1	0	23
Spanish		0	7	34	108	102	80	53	8	392
Total		12	24	106	278	217	141	76	12	866

RQ2b-WPT Ratings



	BR/UR	NH	IL	IM	IH	AL	AM	AH	S	Total
French	0	1	1	14	46	89	37	10	0	198
Italian	0	0	0	0	3	12	16	2	0	33
Portuguese	0	0	0	18	80	95	28	6	0	227
Russian	0	0	0	3	7	8	5	0	0	23
Spanish	1	0	0	0	68	238	83	4	0	394
Total	1	1	1	35	204	442	169	22	0	875

RQ2b-Writing Calibration



Each integer represents ONE sublevel.

ANOVA

$$F(4, 867) = 31.8, p < .001$$

WPTRating	95% Confidence Interval			
	Mean	SE	Lower	Upper
IM	0.37	0.22	-0.06	0.80
IH	0.27	0.09	0.09	0.45
AL	-0.20	0.06	-0.33	-0.08
AM	-0.98	0.10	-1.17	-0.78
AH	-1.86	0.28	-2.41	-1.32

troyco

For writing, the higher the proficiency level of the examinee, the **lower** they were in their calibration.

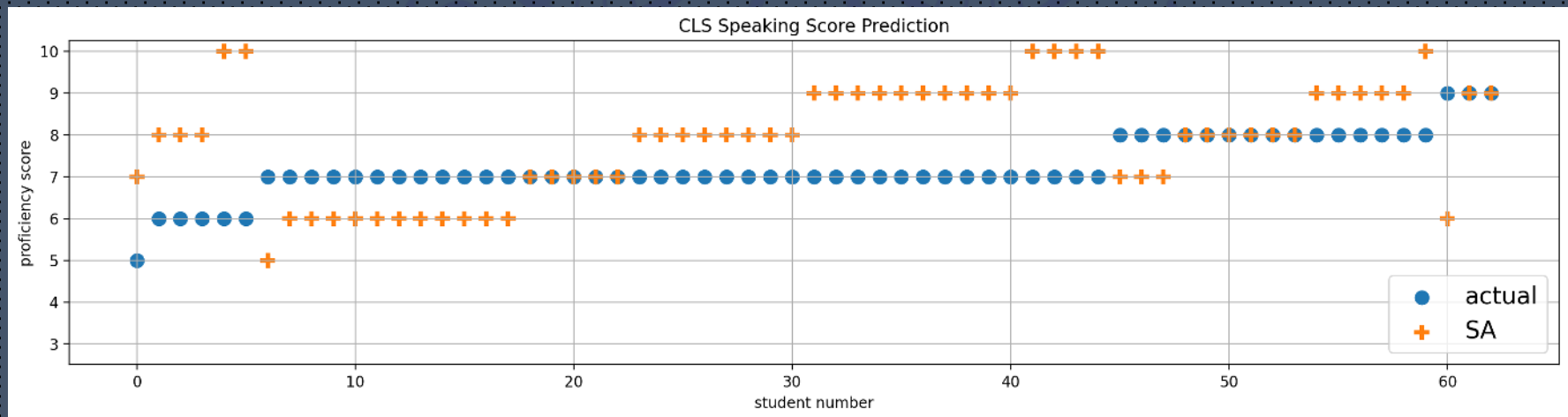
AM & AH under-assessed ability by more than 1 sublevel.

IM, IH & AL were the most accurate.

Next Steps—Combatting Dunning-Kruger

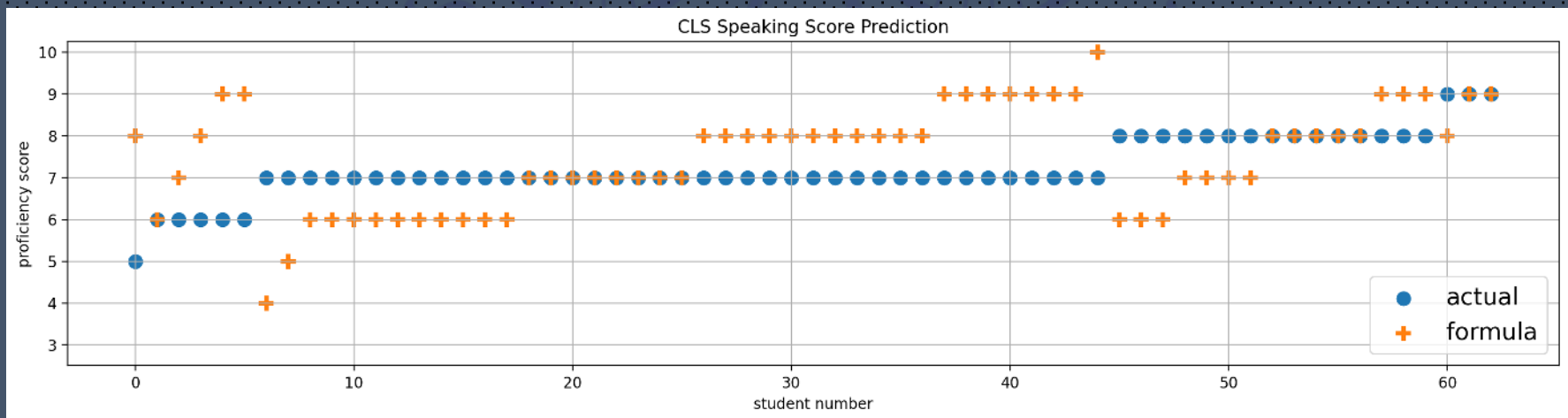
Can we pr

Self-assessment prediction of OPIc ratings



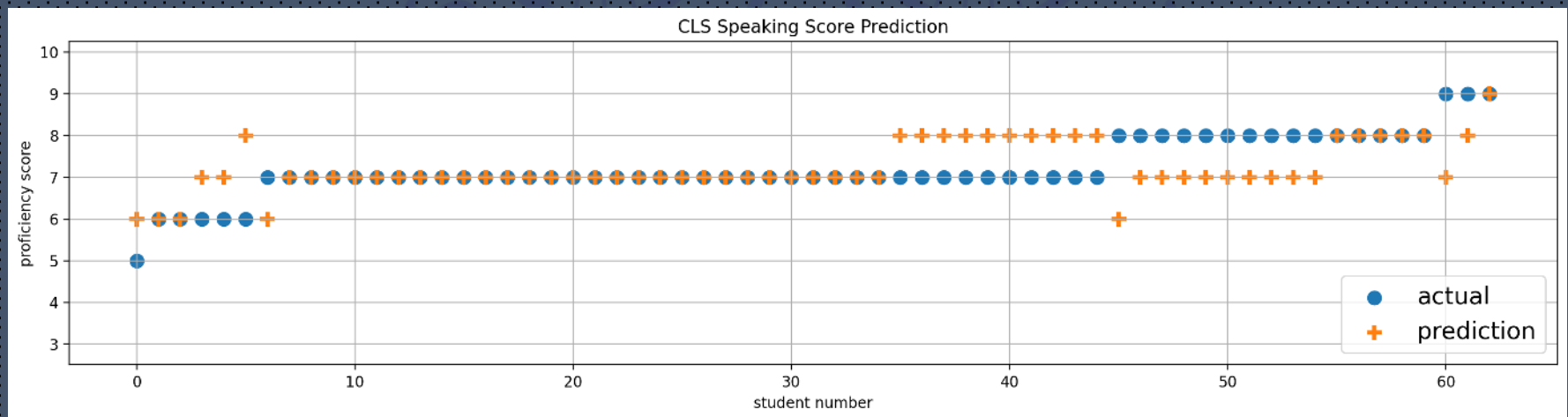
Self Assessment:
13 exact matches (21%)
27 within 1 sublevel (63%)

SA + Language learning background prediction of OPIc Ratings



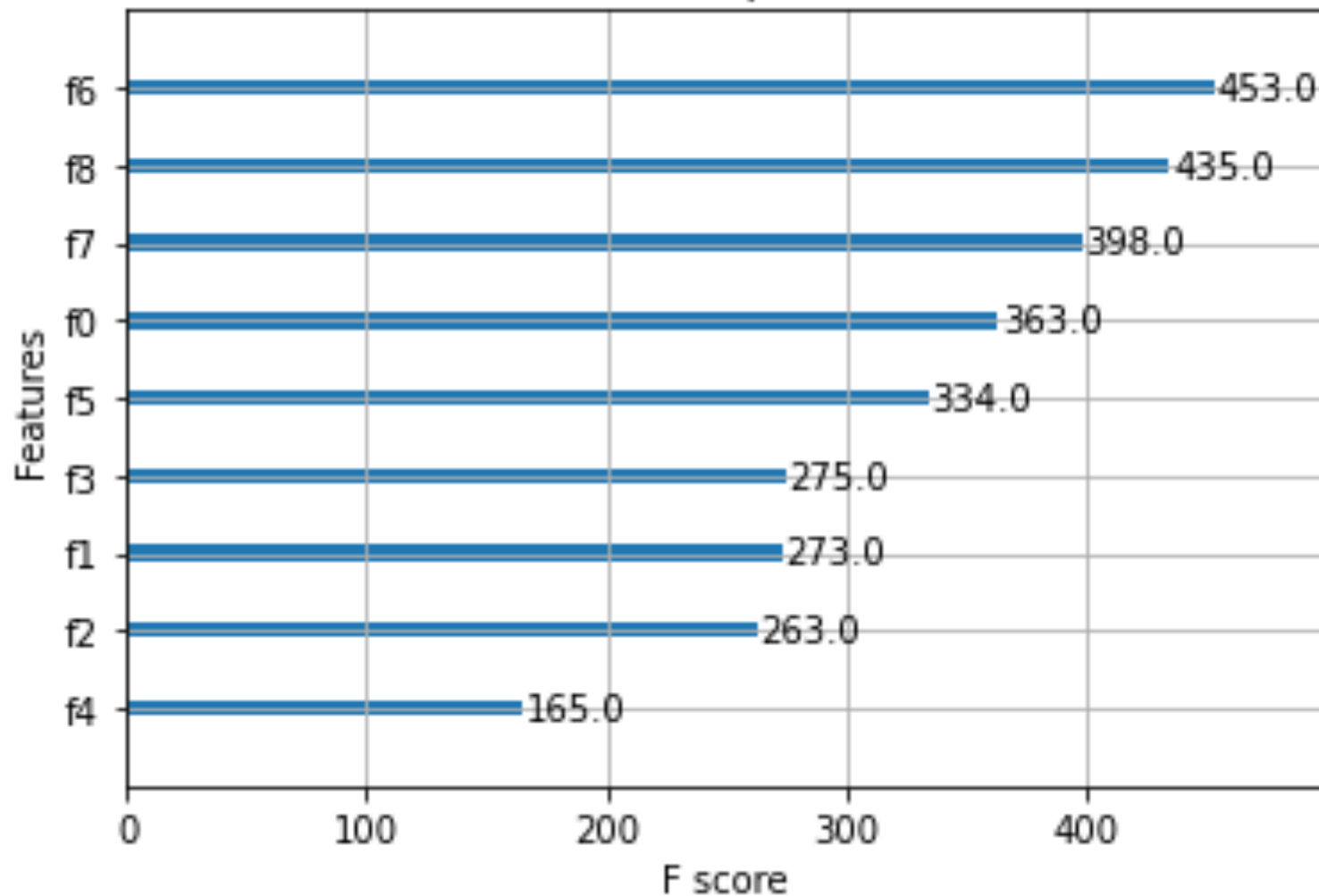
Self-assessment moderated by language learning
background variables 16 exact matches (25%)
30 withing 1 (73%)

Machine Learning Prediction of OPIc Ratings



ML model:
36 exact matches (57%)
24 within 1 (95%)

Feature importance



f6: FF Articulation Rate
f8: FF Mean Length of Utterance
f7: FF Silent Pause Ratio
f0: SA Self Rating
f5: FF Error Count
f3: Grit
f1: LB Number of Immersion Months
f2: LB Number of Months living Native Speakers
f4: LB Number of Years Studying the Language

Therefore what?

- Apple watch or fellow traveler?
- Help students RETHINK FAILURE
 - We learn more from our failures than our successes
 - You either have a good experience or you get a good story, but it can be good either way
- Intervene early to recommend students to register for the class that helps them best
- Help teachers
 - Train teachers in proficiency so they understand the scale
 - Inspire students to engage in lifelong learning
 - Use **Backward Design** to incorporate proficiency in their curriculum