



The Effect of Brief Proficiency Training

on Self-Assessment Calibration in Second Language Learners

Rachel Reynold • Troy Cox | Brigham Young University

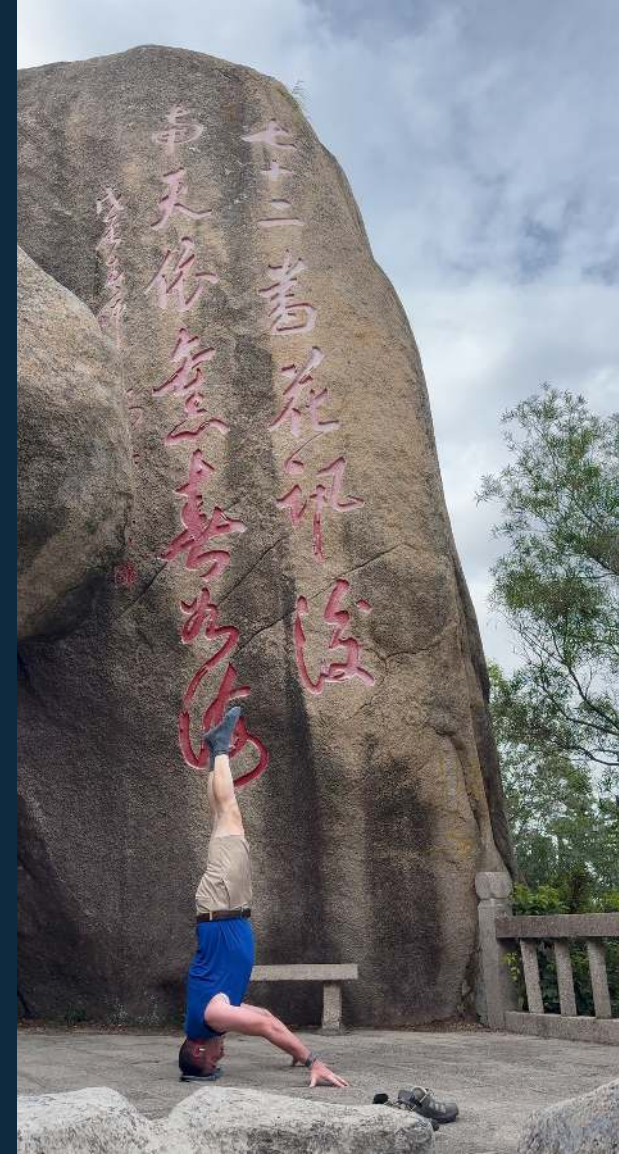
Psychology of Language Learning Conference (PLL6) • 2026





A little learning is a dangerous thing;

*Drink deep, or taste not the Pierian spring:
There shallow draughts intoxicate the brain,
And drinking largely sobers us again.*
—Alexander Pope, An Essay on Criticism



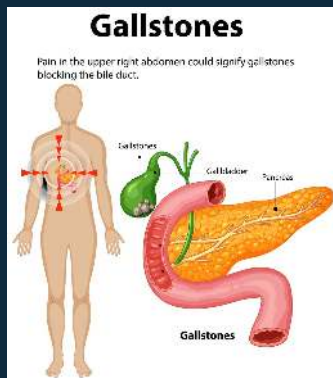
Story Time: My 30th Wedding Anniversary Celebration



5/2/2026



“It’s
probably
gallstones”
-Dr. Cox, Phd
not to be confused with MD



WHY THIS MATTERS

Self-assessment only works if learners can see themselves clearly.

Self-assessment shows up everywhere in language education — placement, goal-setting, reflection, program evaluation. But its usefulness rests on a single assumption: that learners can accurately judge what they can do.

Across studies, that assumption is shaky. Learner self-ratings often diverge from expert ratings — sometimes substantially.

The calibration problem

$$\text{perceived} - \text{actual} = \text{calibration}$$

+ over-assessment

0 perfect calibration

— under-assessment

WHAT PRIOR WORK SAYS

Training is the field's standard answer — but findings are mixed.



What training is supposed to do

Build internal criteria · teach learners to apply them · align judgments through feedback · transfer skills to new tasks.



Where the evidence is messy

Studies vary widely in length (a single session to whole semesters), scope, and learner population. Effects are inconsistent.



Proficiency level matters

Intermediate learners are often the least calibrated; lower- and higher-proficiency learners tend to do better.



Gap in the literature

Few studies test brief, targeted training in authentic placement contexts — and fewer still examine how training interacts with proficiency level.

RESEARCH QUESTIONS

RQ1

Does a brief (~10-minute) training module on the ACTFL proficiency guidelines improve the calibration of learners' perceived and actual proficiency?

RQ2

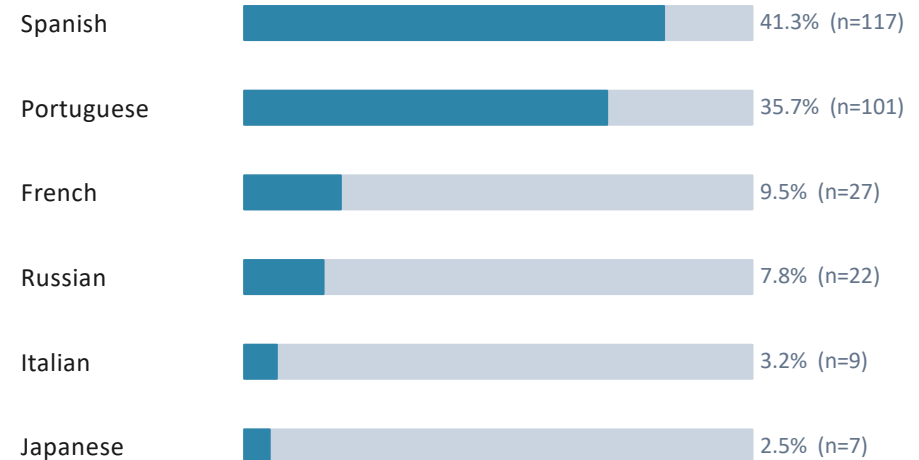
Does the effect of training on calibration vary across learner proficiency levels?

METHODS • PARTICIPANTS & DESIGN

283 third-year language learners, six languages, random assignment.

- N = 283 university students (Fall 2021)
- Third-year world language courses
- Random assignment: control (n = 139) vs. experimental (n = 144)
- Within-30-day window between LASER and OPIc

Languages studied

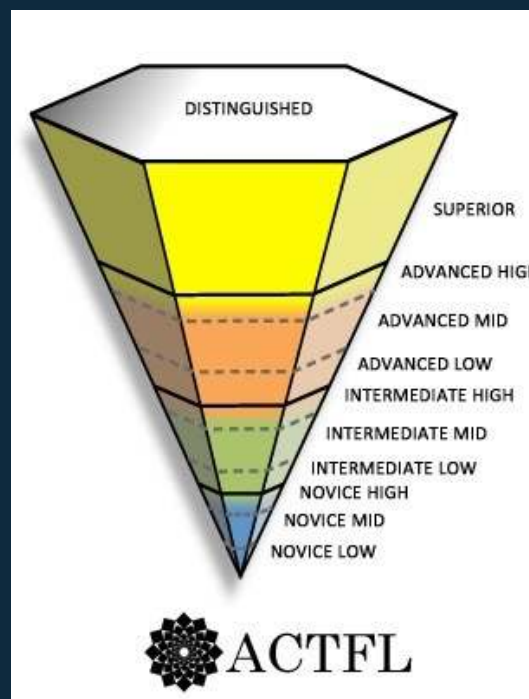


Three pieces: a perceived measure, an actual measure, a 10-minute module.

PERCEIVED	ACTUAL	INTERVENTION
LASER <i>Language Ability Self-Evaluation Resource</i> Online placement test. 24 can-do statements (12 speaking, 12 writing) of increasing complexity; 5-point confidence scale. Items weighted by function, accuracy, content, and text type.	OPIc <i>Oral Proficiency Interview — Computer</i> ACTFL's adaptive computer-delivered oral interview, rated by certified human raters. Converted to a 10-point scale (NL = 1 ... Superior = 10).	Training Module <i>~10 minutes, delivered before the LASER</i> Two parts: (1) explicit instruction on each ACTFL level via short video clips, written samples, and L1-use reflection; (2) practice with immediate feedback on video/audio clips sourced from ACTFL.

The ACTFL Proficiency Guidelines

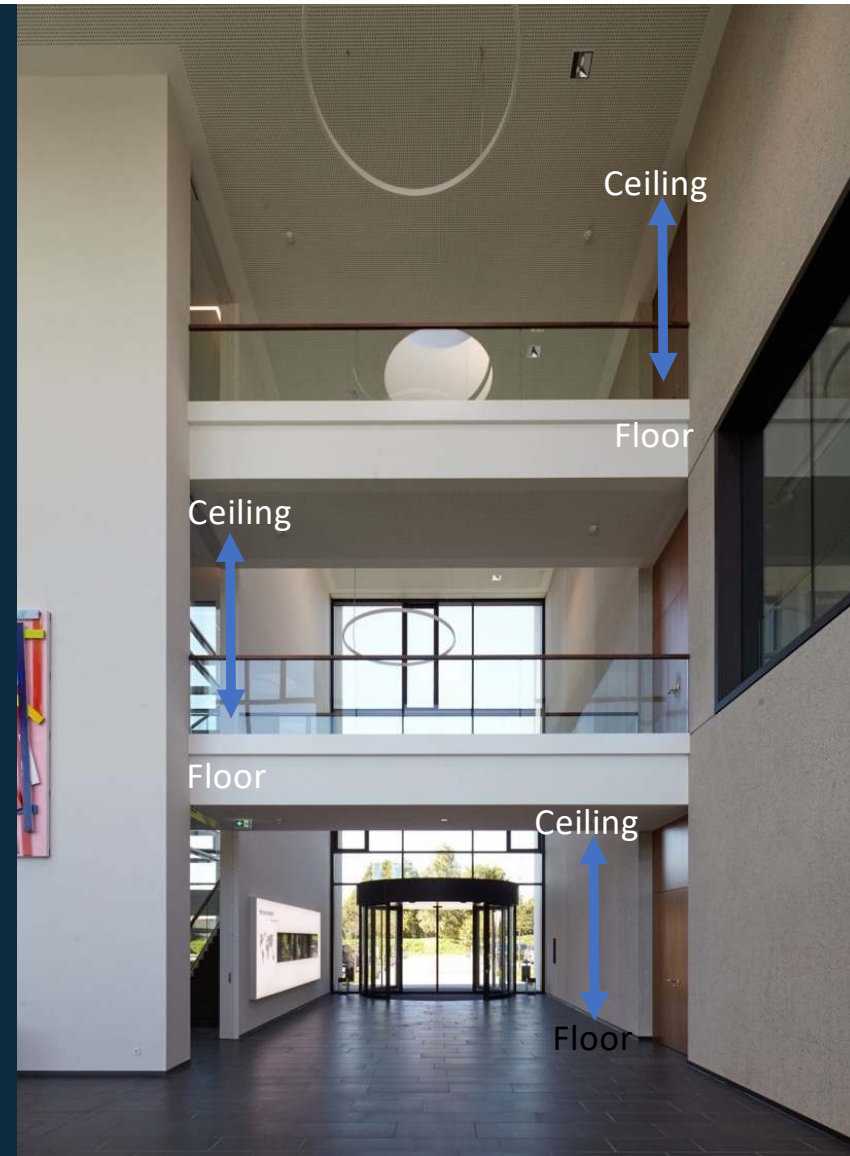
- Not what you KNOW about the language, but what you can DO SPONTANEOUSLY in real world situations.
- Five major levels (each subsumes all lower levels) that describe a confluence of:
 - **F**unction
 - **A**ccuracy
 - **C**ontent/context
 - **T**ext Type



*Keep your
FACTs straight
and you can
determine the
major level!*

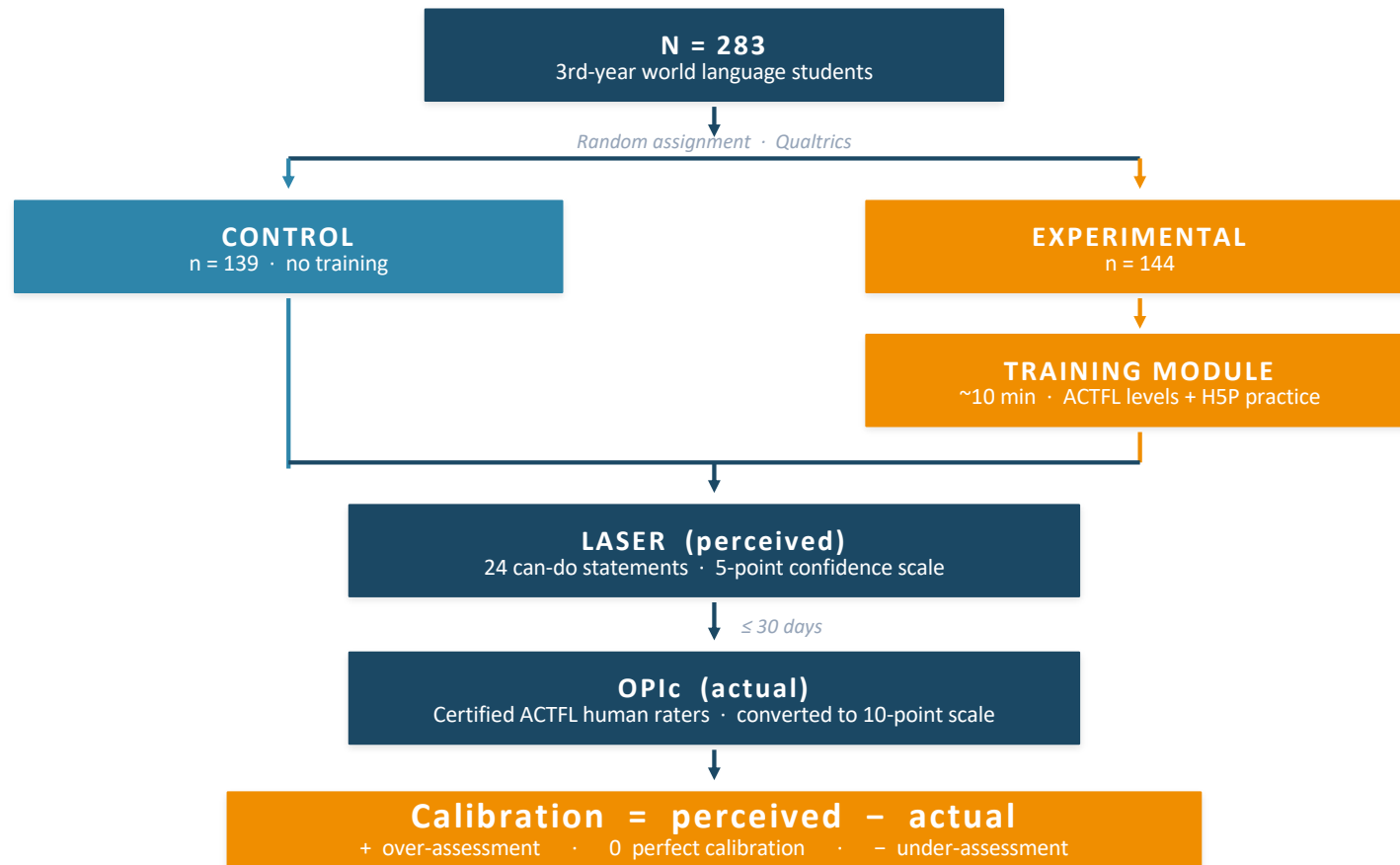
- Proficiency Guidelines describe major levels of language development.
- **Floor** describes highest level that can be maintained consistently.
- **Ceiling** describes lowest level in which patterns of breakdown appear.

ThePhoto by PhotoAuthor is licensed under CCYSA.



METHODS · PROCEDURE

From random assignment to calibration score.

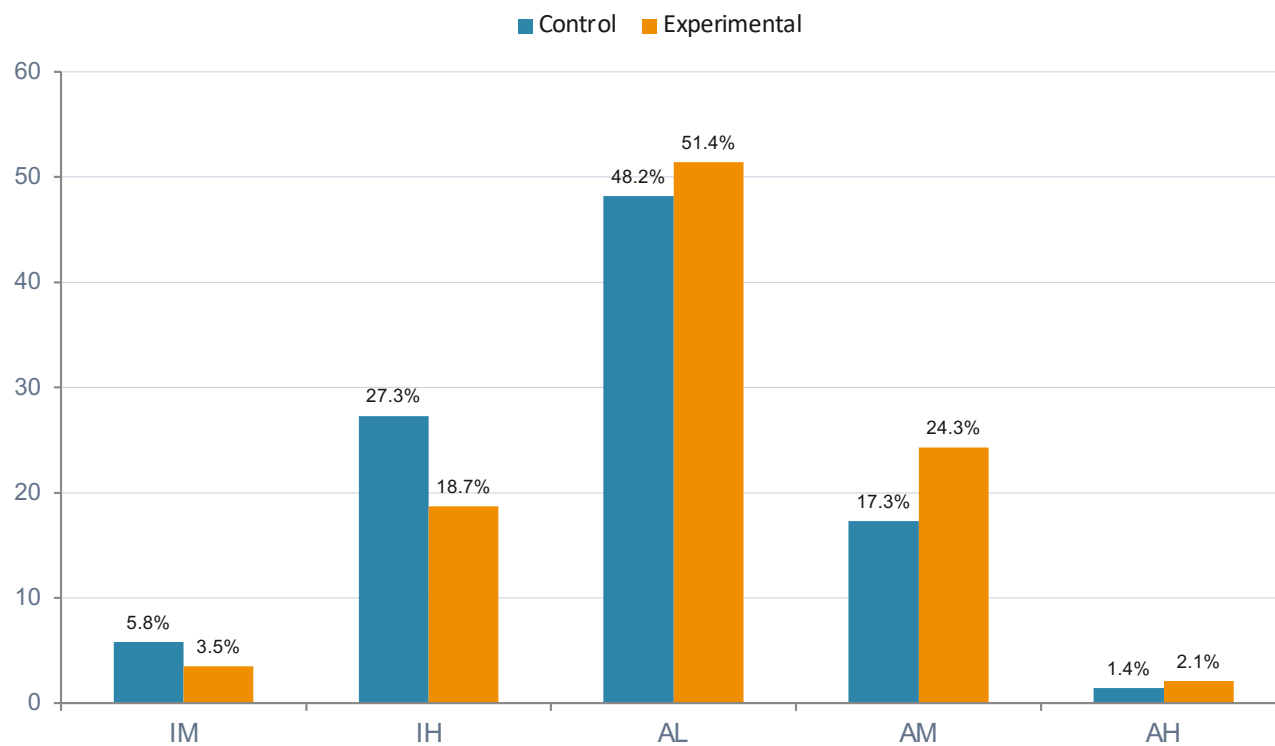


Pause Point

- Turn to someone, introduce yourself, and predict what you think the findings will be

RESULTS • ACTUAL PROFICIENCY

OPic ratings clustered at Advanced Low — and were comparable across groups.



Headline

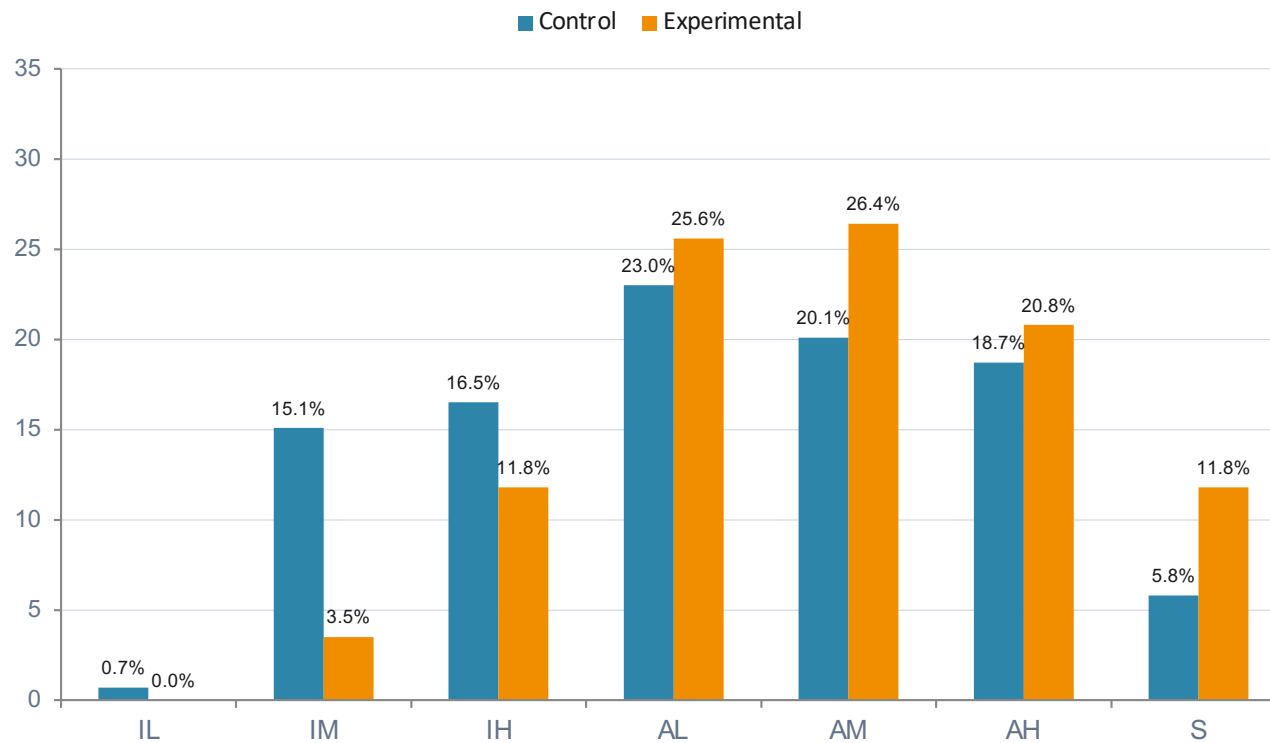
49.8%

of participants rated at Advanced Low — the modal level.

Groups were statistically comparable (χ^2 n.s., Cramér's V small).

RESULTS • PERCEIVED PROFICIENCY

Self-ratings spread wider — and the experimental group skewed higher.



Headline

59% vs 45%

self-rated AM or above — experimental vs. control.

Self-ratings even reached IL and Superior — levels that didn't appear in the OPIc data at all.

RESULTS • CALIBRATION

Training shifted calibration — in the wrong direction.

26.9%

perfectly calibrated

53%+

over-assessed

67%

within one sublevel

The unexpected result

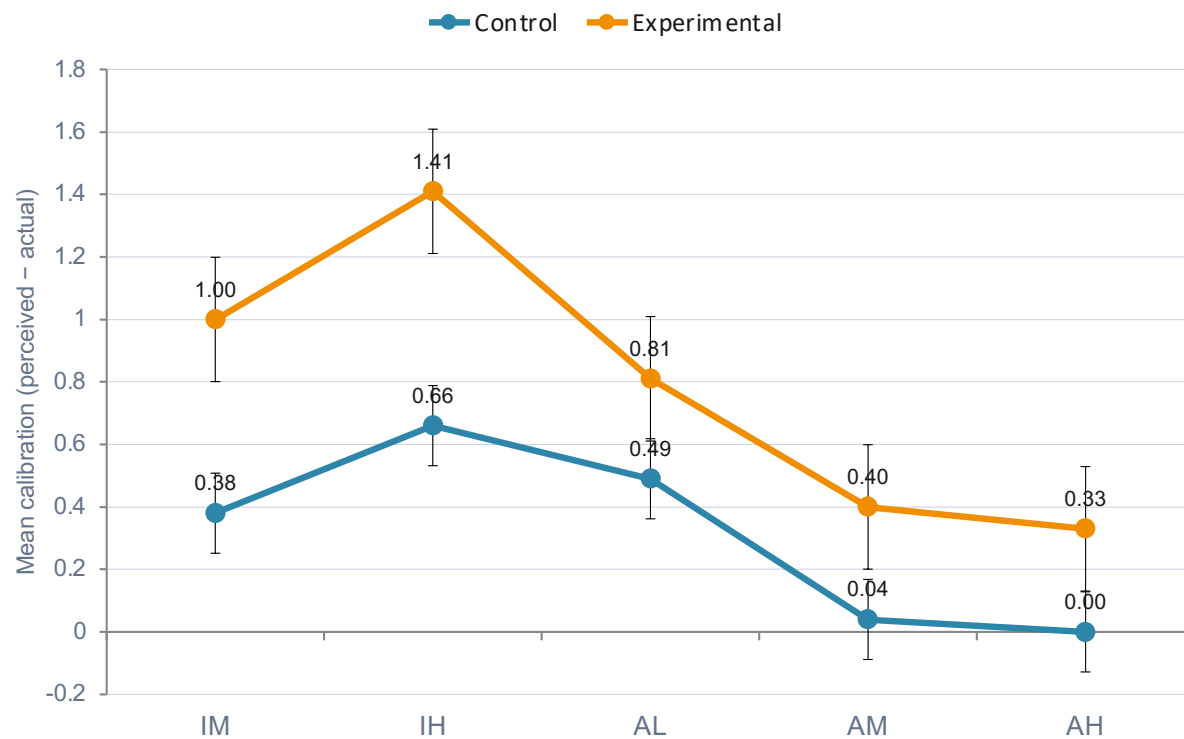
The experimental group over-assessed more than the control group — not less. Trained learners rated themselves higher relative to their actual OPIc scores (independent-samples t-test, $p < .05$; small effect).

Perfect calibration: Control 30.9% · Experimental 22.9%

Over-assessed by ≥ 1 sublevel: Control 46% · Experimental 60%

Two-way ANOVA confirmed both main effects: training condition ($p < .05$) and OPIc proficiency level ($p < .05$). The only reliable pairwise difference: Intermediate High vs. Advanced Mid ($p = .018$).

Mean calibration drops monotonically from IH to AH as proficiency rises.



Reading the line

Lower-proficiency learners systematically over-assessed; higher-proficiency learners were close to accurate.

The experimental line sits above the control line at every observed level — a consistent training-induced shift.

Pattern is consistent with Dunning–Kruger and prior L2 self-assessment findings (Boud & Falchikov, 1989; Ma & Winke, 2019).

Pause Point

- Why might a 10-minute module make things worse?

DISCUSSION

Why might a 10-minute module make things worse?

1

Incomplete training, confident learners

Brief exposure may give learners just enough familiarity with proficiency descriptors to justify pre-existing self-perceptions — without prompting the critical reflection real calibration requires.

2

A textbook Dunning–Kruger pattern

Brief instruction can paradoxically increase overconfidence in low-to-mid performers. The monotonic drop from IH to AH fits the 'unskilled and unaware' literature directly.

3

Levels read as exclusive, not cumulative

Post-training responses showed many learners thought reaching Advanced meant they no longer use Intermediate functions — a misreading of the ACTFL framework itself.

4

The label problem

Language majors may resist identifying as 'Intermediate' regardless of their communicative ability. Naming the levels may itself bias the rating.

LIMITATIONS & IMPLICATIONS

What we'd caveat — and what teachers can do with this.

LIMITATIONS

- Time-on-module was not directly measured — only total LASER duration.
- Some learners misread ACTFL levels as exclusive rather than cumulative.
- A technical glitch briefly affected one practice activity.
- Module phrasing implied that language majors are 'Advanced' speakers.

PEDAGOGICAL IMPLICATIONS

- For classroom reflection, ballpark accuracy (within one sublevel) may be good enough — most learners had this.
- For placement or other high-stakes uses, brief broad-scope training is not enough.
- Teaching the criteria (function, accuracy, content, text type) without leaning on ACTFL labels may reduce label bias.

TAKEAWAY

Brief, broad-scope training on the ACTFL guidelines didn't sharpen self-assessment.

If anything, it nudged learners further into overconfidence — especially at the Intermediate–Advanced boundary.

Calibration is a skill —
not a side effect of proficiency.

Brief training isn't free —
it can reinforce the wrong intuitions.

Design matters more than dose —
what's taught matters more than how long.

Thank you · Questions?

troyc@byu.edu · Brigham Young University



“So you have gallstones, but they aren’t causing the pain. You have a pulmonary embolism. There’s a blood clot that has made its way into your lungs.”

A little learning is a dangerous thing;

*Drink deep, or taste not the Pierian spring:
There shallow draughts intoxicate the brain,
And drinking largely sobers us again.*
—Alexander Pope, An Essay on Criticism

