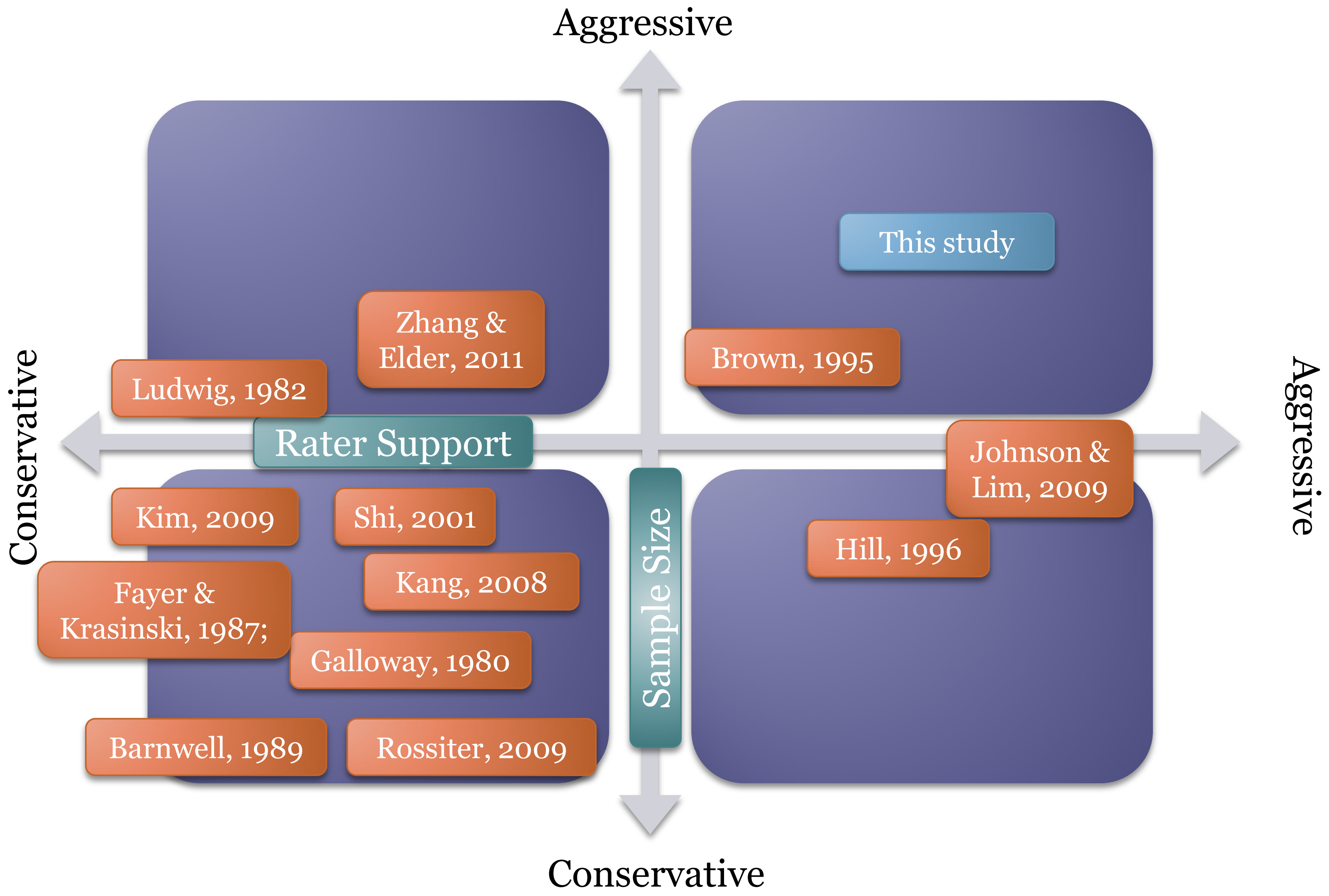
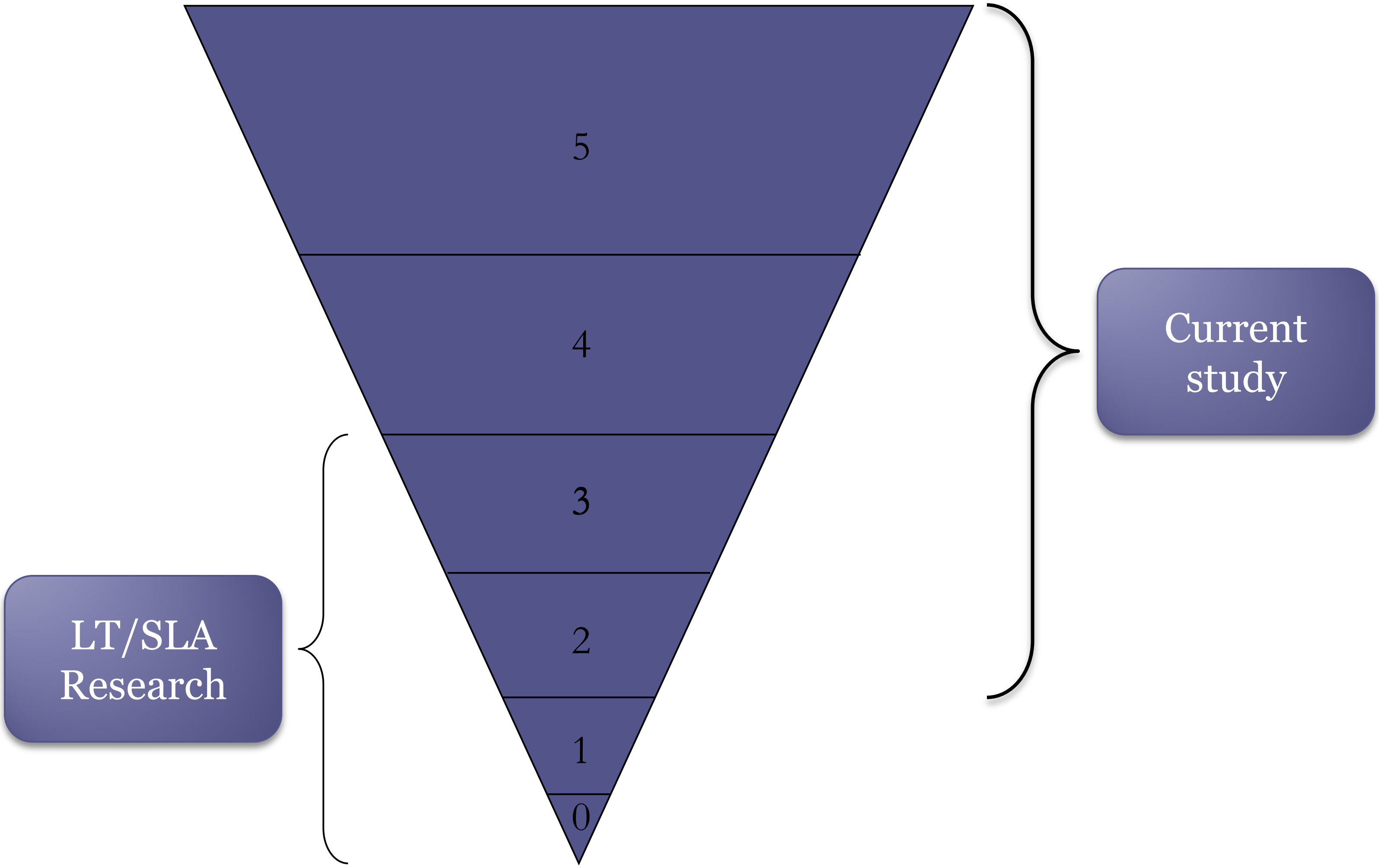
Comparing native and non-native raters of US Federal Government speaking tests

Rachel L. Brooks, PhD
Federal Bureau of Investigation
ALTE 2014

Native Speaker in Language Testing



ILR Skill Level Descriptions



What is a n

If L1/mother tongue is relevant, then a non-native rater with an L1 similar to the language tested might rate more accurately.

Paikeday, 1985

1. A person who has a specified language as the mother tongue or first learned language
   - hav[ing] at least a bachelor's degree from a reputable college or university
2. A competent speaker of a specified language
   - who
     or i
     incl
     synt

If competence/ability is relevant, then a non-native rater with a higher speaking proficiency might rate more accurately.

1. Acquired L1/native language in childhood
2. Has intuitions (acceptability/ productiveness) about his idiolectal grammar
3. Has intuitions about standard grammar
4. Is widely fluent, spontaneous, with huge vocabulary and communicative competence

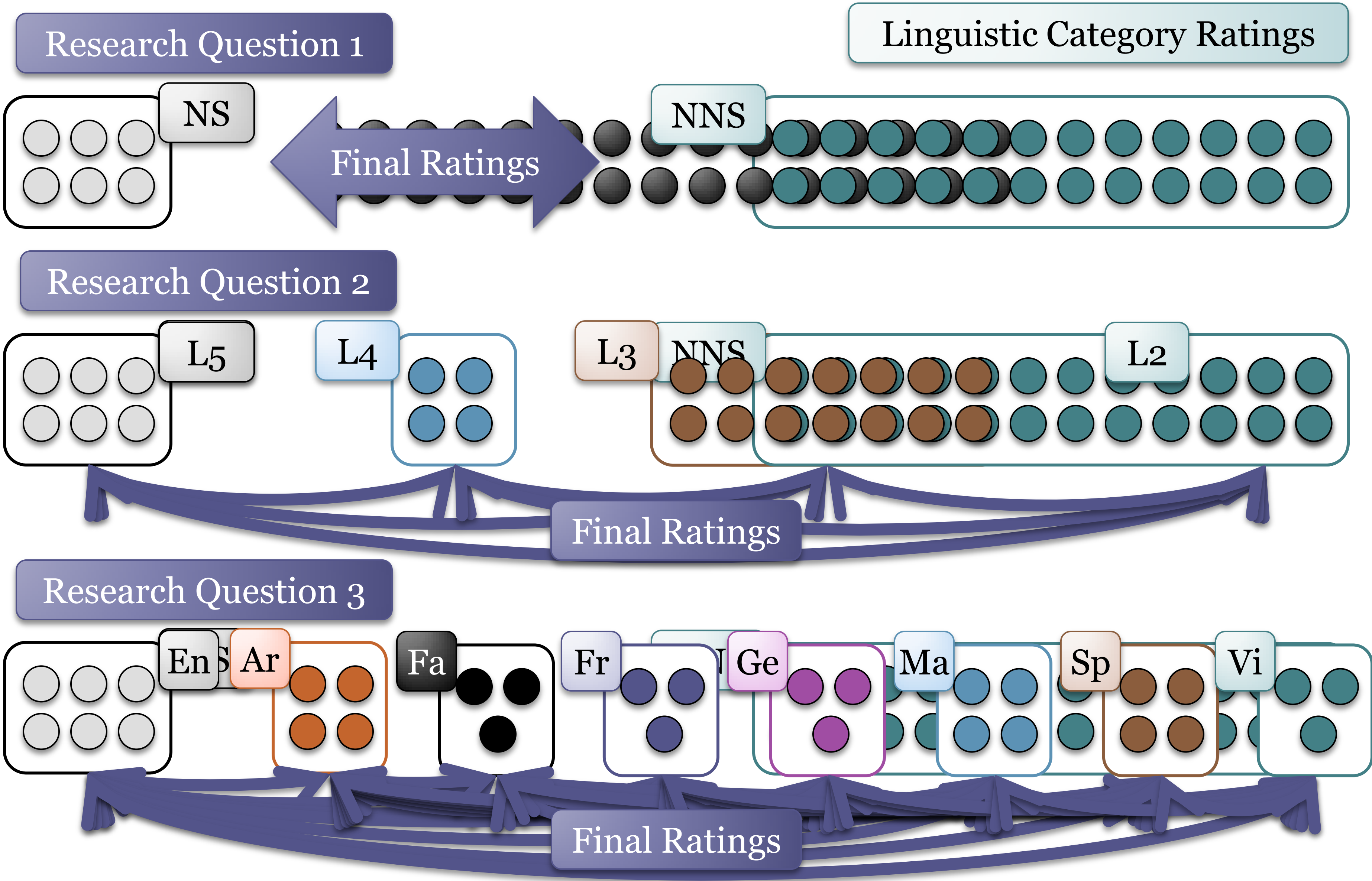
Writes creatively

Has a unique capacity to interpret or translate into L1

Rater Distribution (n=30)

Research Question 4

Linguistic Category Ratings



Research Question 1

Research Question 2

Research Question 3

Research Questions

1. Do native and non-native speaker raters assign comparable ratings on speaking tests?
2. Does speaking proficiency level affect a rater’s ability to reliably evaluate speaking proficiency?
3. Does the first language learned affect a rater’s ability to reliably evaluate speaking proficiency?
4. Do native and non-native raters assess the specific linguistic features of the speaking samples comparably?

Raters/ Samples Evaluated

Raters

Language	Total
English (NS)	6
Arabic (NNS)	4
Farsi (NNS)	3
French (NNS)	3
German (NNS)	3
Mandarin (NNS)	4
Spanish (NNS)	4
Vietnamese (NNS)	3
Total	30

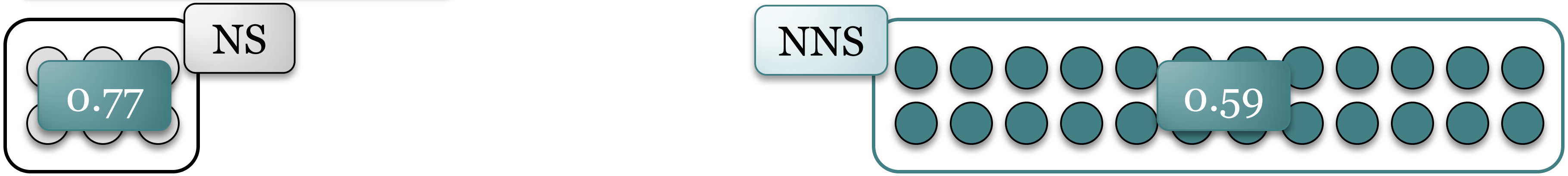
Exams Rated

ILR Level	NS	NNS	Total
4/4+/5	5	0	5
3/3+	7	7	14
2/2+	1	5	6
Total	13	12	25

= 750 evaluations

Inter-rater Reliability (Krippendorf's alpha)

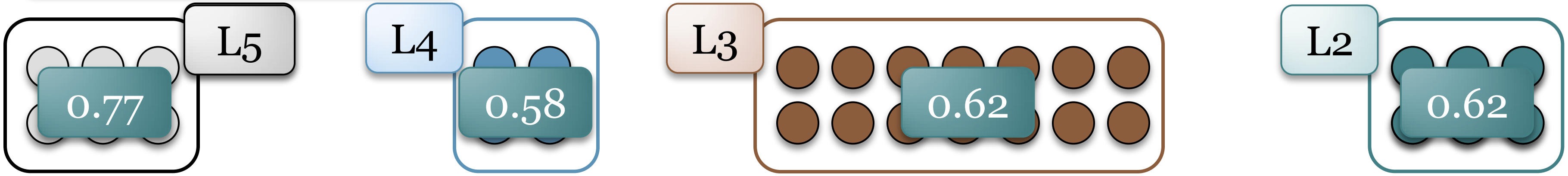
Research Question 1



NS 0.77

NNS 0.59

Research Question 2



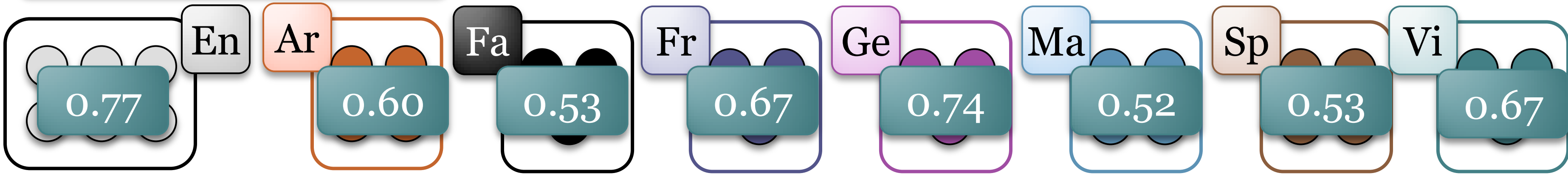
L5 0.77

L4 0.58

L3 0.62

L2 0.62

Research Question 3



En 0.77

Ar 0.60

Fa 0.53

Fr 0.67

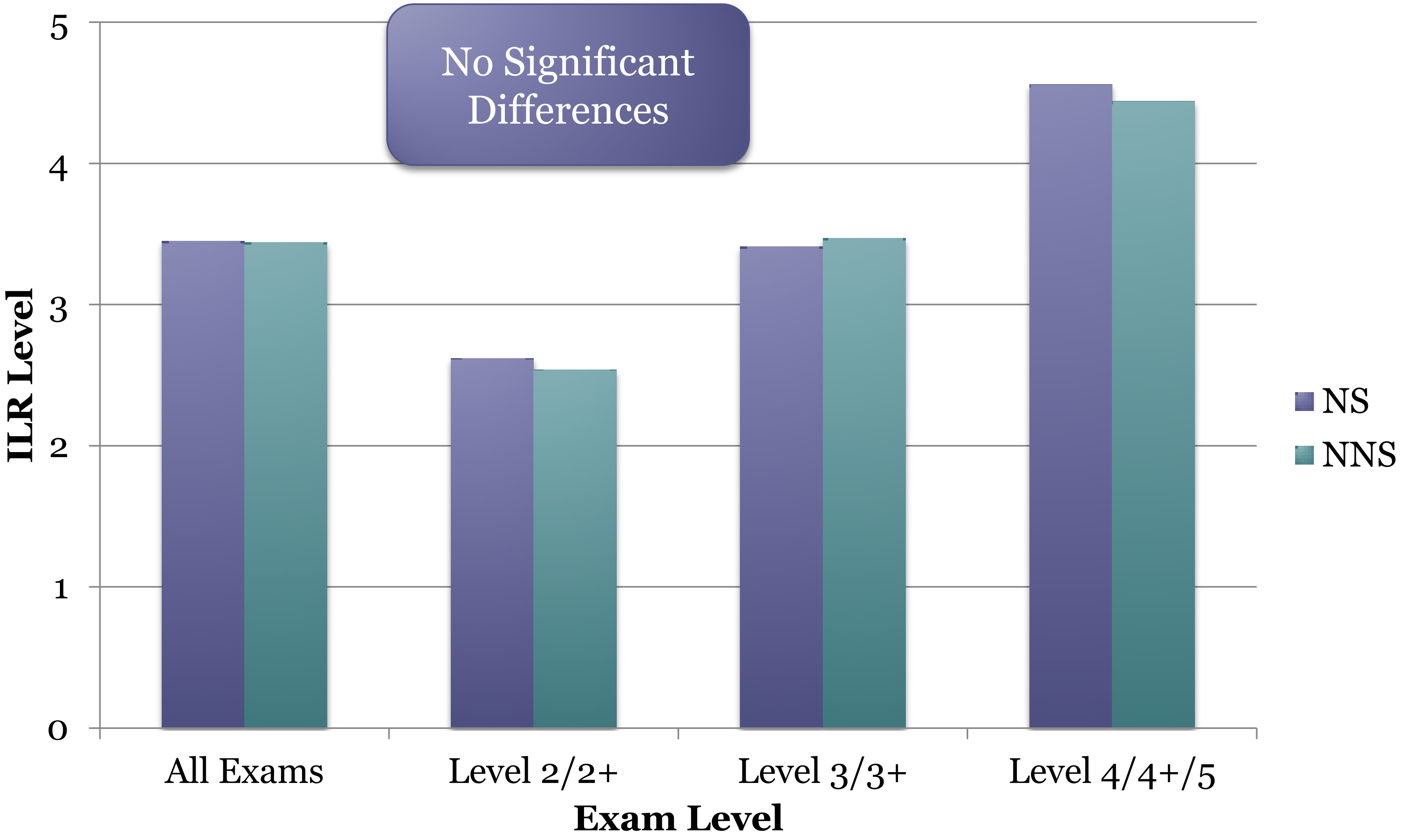
Ge 0.74

Ma 0.52

Sp 0.53

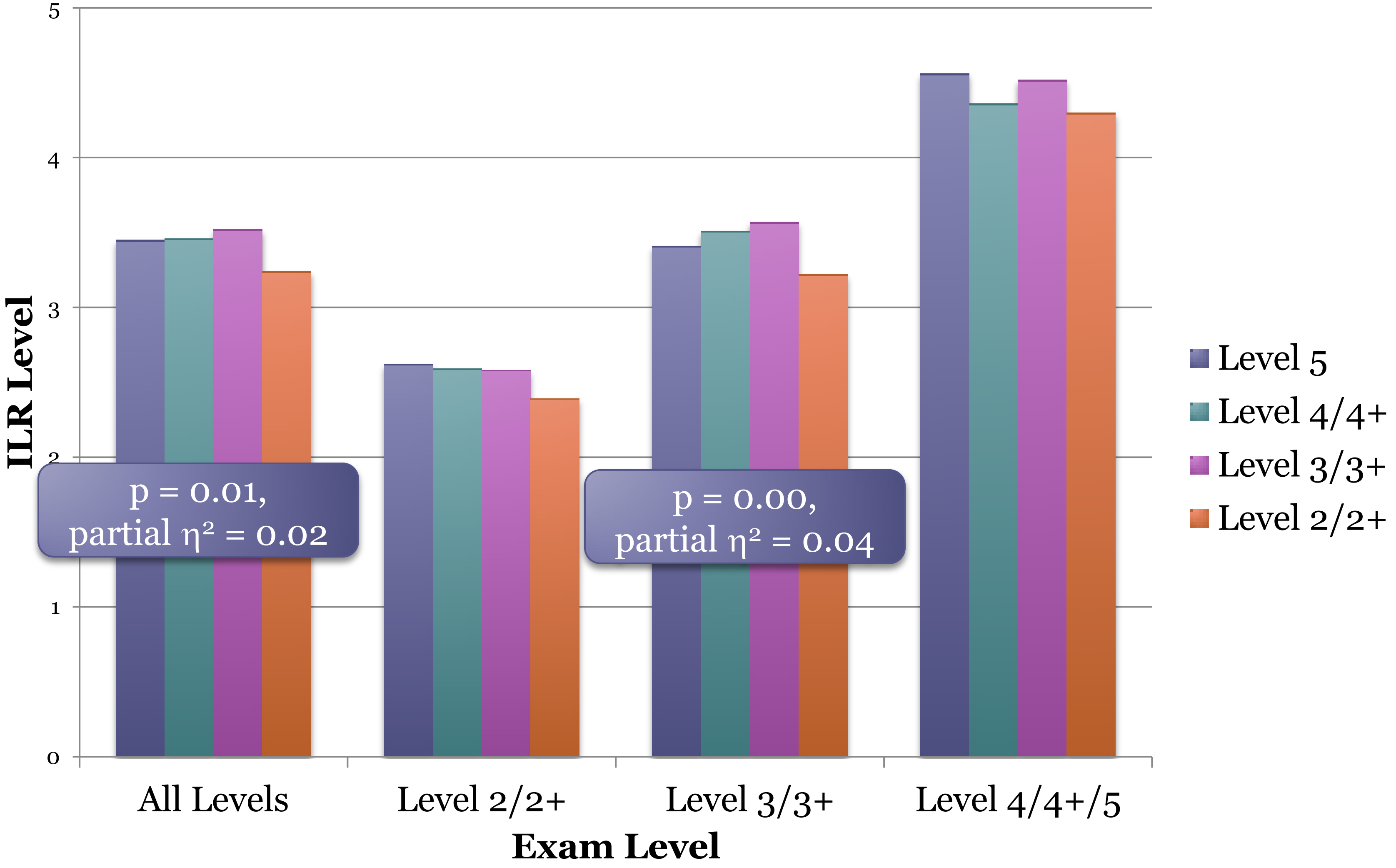
Vi 0.67

RQ 1: NS and NNS Group Mean Ratings



No Significant Differences

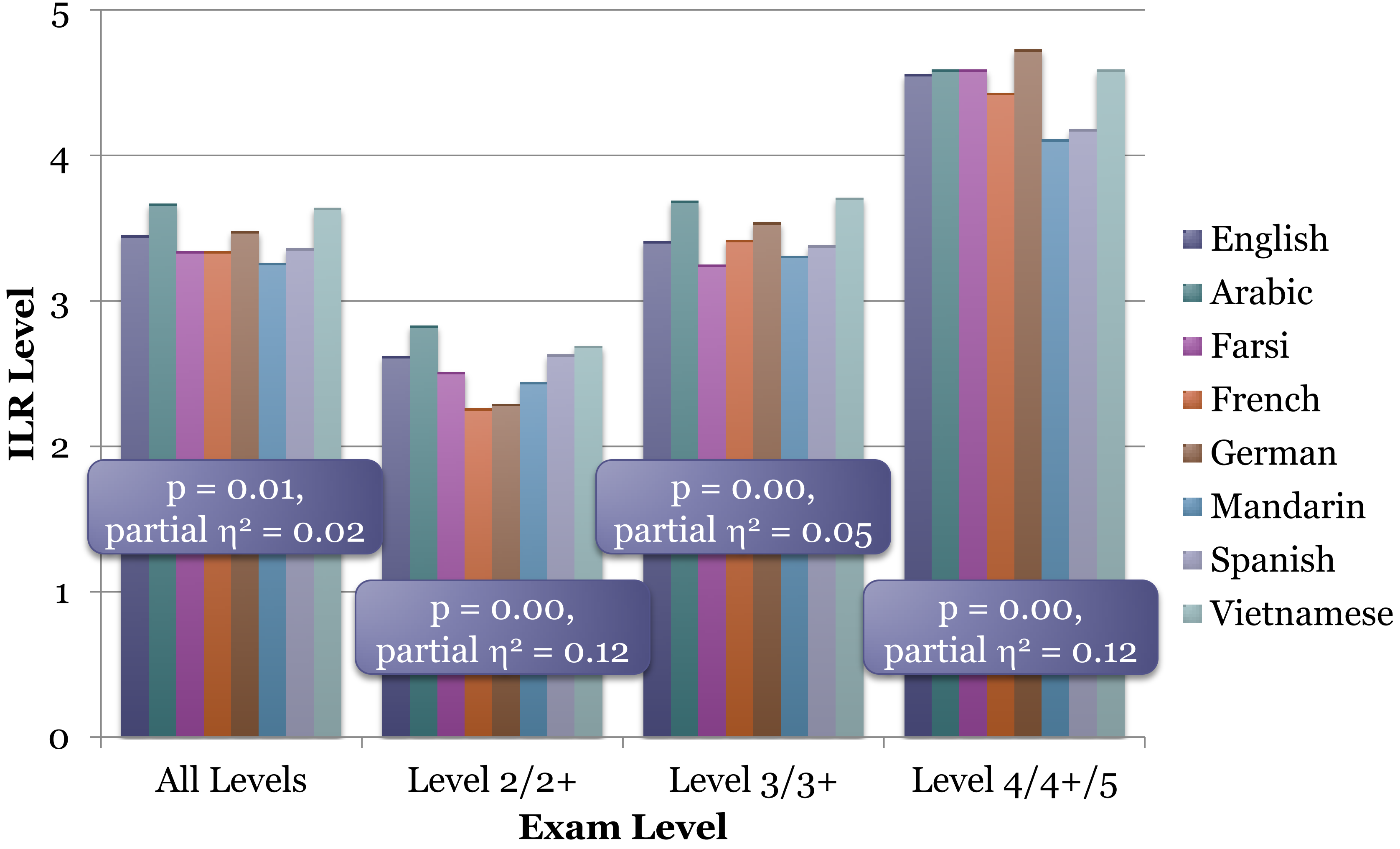
RQ 2: English Proficiency Level Group Mean Ratings



p = 0.01, partial η² = 0.02

p = 0.00, partial η² = 0.04

RQ 3: First Language Mean Ratings



p = 0.01, partial $\eta^2 = 0.02$

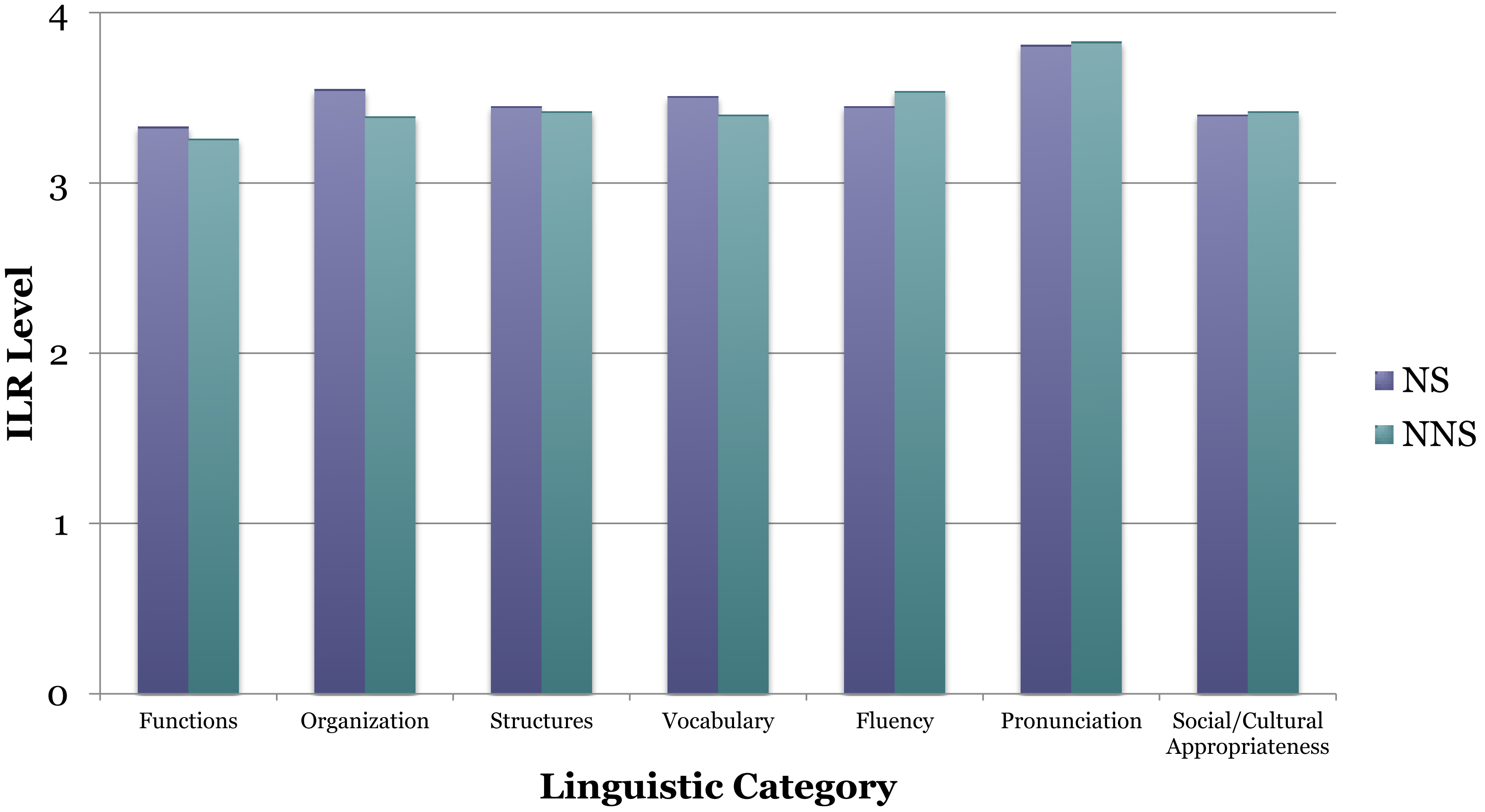
p = 0.00, partial $\eta^2 = 0.12$

p = 0.00, partial $\eta^2 = 0.05$

p = 0.00, partial $\eta^2 = 0.12$

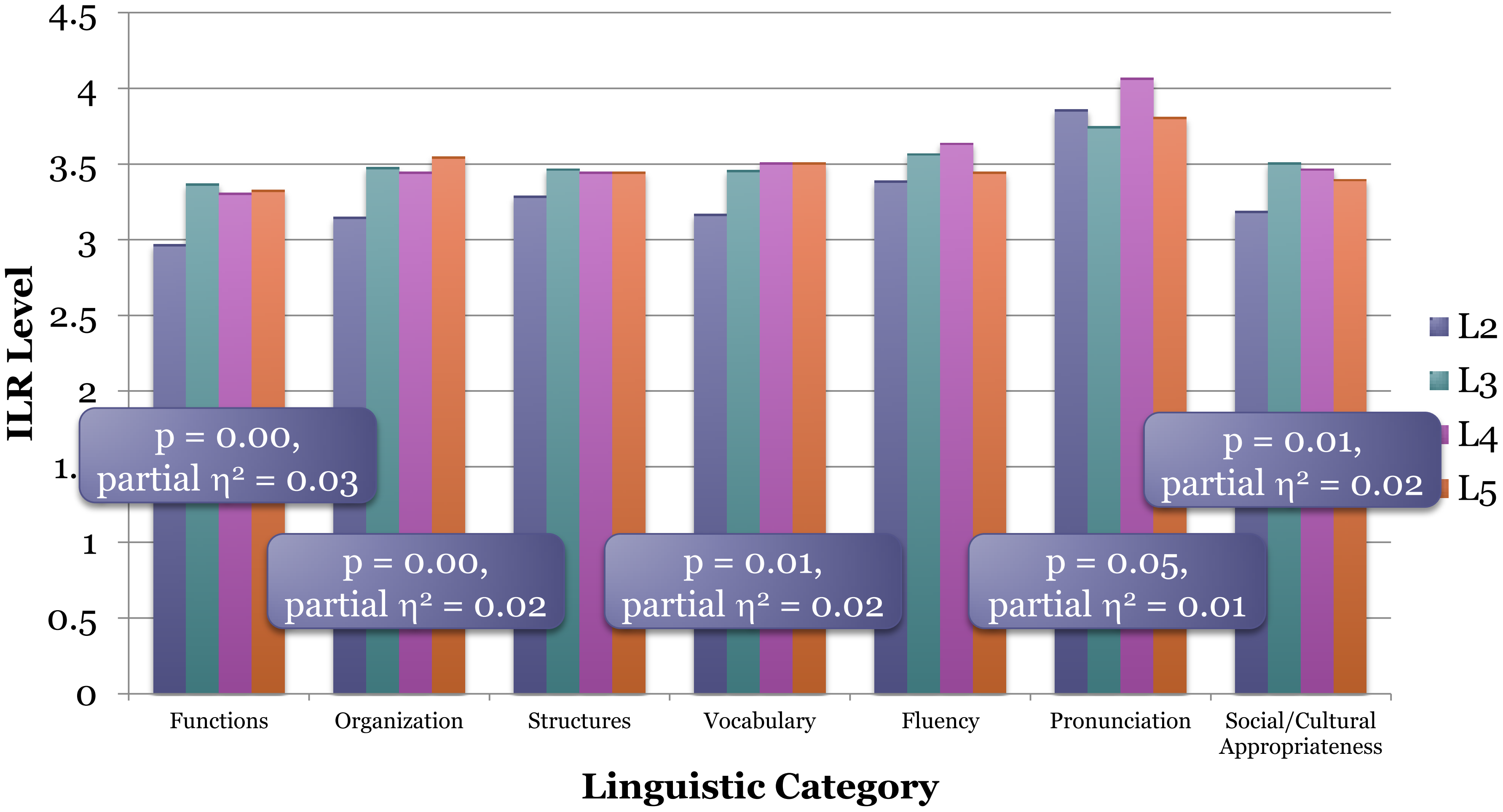
Overall:
$p = 0.00$,
partial $\eta^2 = 0.04$

RQ4: NS and NNS Raters: Mean Linguistic Category Ratings



Overall:
$p = 0.00$,
partial $\eta^2 = 0.04$

RQ4: English Proficiency: Mean Linguistic Category Ratings



$p = 0.00$,
partial $\eta^2 = 0.03$

$p = 0.00$,
partial $\eta^2 = 0.02$

$p = 0.01$,
partial $\eta^2 = 0.02$

$p = 0.05$,
partial $\eta^2 = 0.01$

$p = 0.01$,
partial $\eta^2 = 0.02$

Conclusions

1. No significant difference between NS and NNS raters
   - Any differences can be overcome by training
   - FBI SPT raters are not typical people
   - Inter-rater reliability impact?
2. Proficiency should be considered over NS
   - Level 2+ raters should be excluded
3. L1 has an impact on rating
   - But not compared to English raters
   - Language distance matters
4. Ratings of specific features show more group differences
   - Rater proficiency and L1 groups
   - Differences never occur in “structures”

The native speaker

- (Re)defined
  - Need for clear definition
  - Native speaker assumptions
  - Native speaker is a social construct, not a measurement construct
    - It is associated with acquisition method, culture, identity, confidence
  - Call the ideal speaker something else, specify what it is
- Justification
  - Appropriateness for use: is it fair?
    - Decisions: standard variety, correctness
    - Consequences of misuse go beyond test itself

Qualifying speaking raters by speaking proficiency ability

- It takes one to know one?
  - ▫ Much of impact seen in Level 3 tests
    - Level 2 is below rater's proficiency level
    - Level 4-5 is limited by ceiling effect
- Competence vs. performance
- Training: the great equalizer

Holistic versus analytic rating

- Trend:
  - No differences are found in overall ratings
  - Differences found in linguistic features, except Structures
- What construct are raters using to rate?
- Does construct matter if final ratings are not significantly different?

Limitations and future research

- Current study deals only with rating, not test administration
- Replicate with NS raters < ILR 5
- Replicate in a language other than English
- Analyze the rater comments
- Further investigate rater competencies: linguistic, cognitive, cultural, and evaluative competencies

Thank you

Rachel Lunde Brooks
rachel.brooks@ic.fbi.gov