Answer-changing Behavior in Multiple-choice Questions: Looking beyond the Impact of Changes

Zhi Li, Michelle Y. Chen, & Jayanti Banerjee

International Testing Commission (ITC), July 4th, 2018

Montréal , Québec

Overview of Answer Changes

- A common phenomenon in objective tests
  - Most test takers made some changes (Balance, 2006; Bath, 1967; Jacobs, 1972; Mathews, 1929)
- Effects of answer changes vs. common beliefs
  - First instinct fallacy vs. It-pays-to-switch (Foote & Belinky, 1972; Di Milla, 2007)
- Factors related to answer-changing behaviors
  - Test takers' characteristics (proficiency, gender, personality)
  - Item characteristics (difficulty, discrimination, etc.)

Answer Changes in Language Tests

- Relatively few studies in language-related testing
  - The Michigan English Language Institute College English Test - Grammar, Cloze, Vocabulary, Reading (Al-Halmly & Coombe, 2005)
  - The Graduate Record Examinations (Liu et al., 2015)

- No studies on listening tests

TESTING ENTERPRISES Paragon

Picture source: https://goo.gl/eKk7fm, https://goo.gl/Gya13t

Listening Comprehension Tests

- Listening comprehension
  - As a complex process of meaning making
  - Goal setting, decoding aural/visual input, … monitoring comprehension (Taylor & Geranpayeh, 2011)

- Two types of listening performance tests
  - While-listening performance tests
  - Post-listening performance tests

- Three stages in a while-listening-performance tests
  - Question preview, Question responding, Answer review

Answer-changing Behaviors and Test Validation (I)

- Validity: “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (Standards; AERA, APA, & NCME, 2014: P11)

- Response-process based validity evidence
  - One of the five major sources of validity evidence (AERA, APA, & NCME, 2014)
  - Contributes to the construct validity (Anderson et al., 1991; Cohen, 2006)
  - Often missing in validation studies (Zumbo & Chan, 2014)
    - ➢ Response process is difficult to capture

Definition of Response Process

The *Standards* (AERA et al., 2014, p.15)

- "Cognitive process engaged in by test-takers"

A broader definition

- Response processes include test-takers' cognitive processes, and processes related to their behaviors and emotions during a test (e.g., Hubley & Zumbo, 2017).

Answer-changing Behaviors and Test Validation (II)

- Answer-changing behaviors as part of response process
  - They represent test takers’ behaviors
  - They may reflect test takers’ strategies

    e.g., make predictions, monitoring
  - They can be recorded through *timestamped log data* in computer delivered tests
  - The outcome of answer-changing behaviors is directly related to test performance/scores

Research Questions

With an eye towards test score validation, this study investigates:

1. Who made the changes?

2. When did the changes take place throughout the listening test?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers’ proficiency levels?

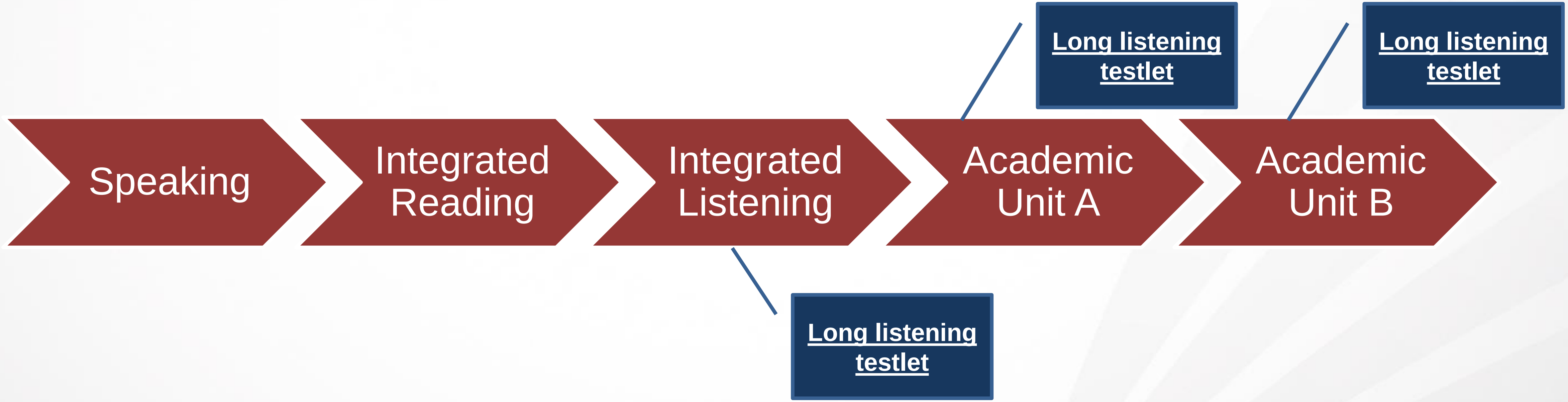
3. What were the outcomes of the changes?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers’ proficiency levels?

The CAEL CE Listening Test

The Canadian Academic English Language (CAEL) Test, Computer Edition (CE)

–An integrated and topic-based test of English for academic purposes

–Reporting scale: 10-90 band score



Speaking

Integrated Reading

Integrated Listening

Academic Unit A

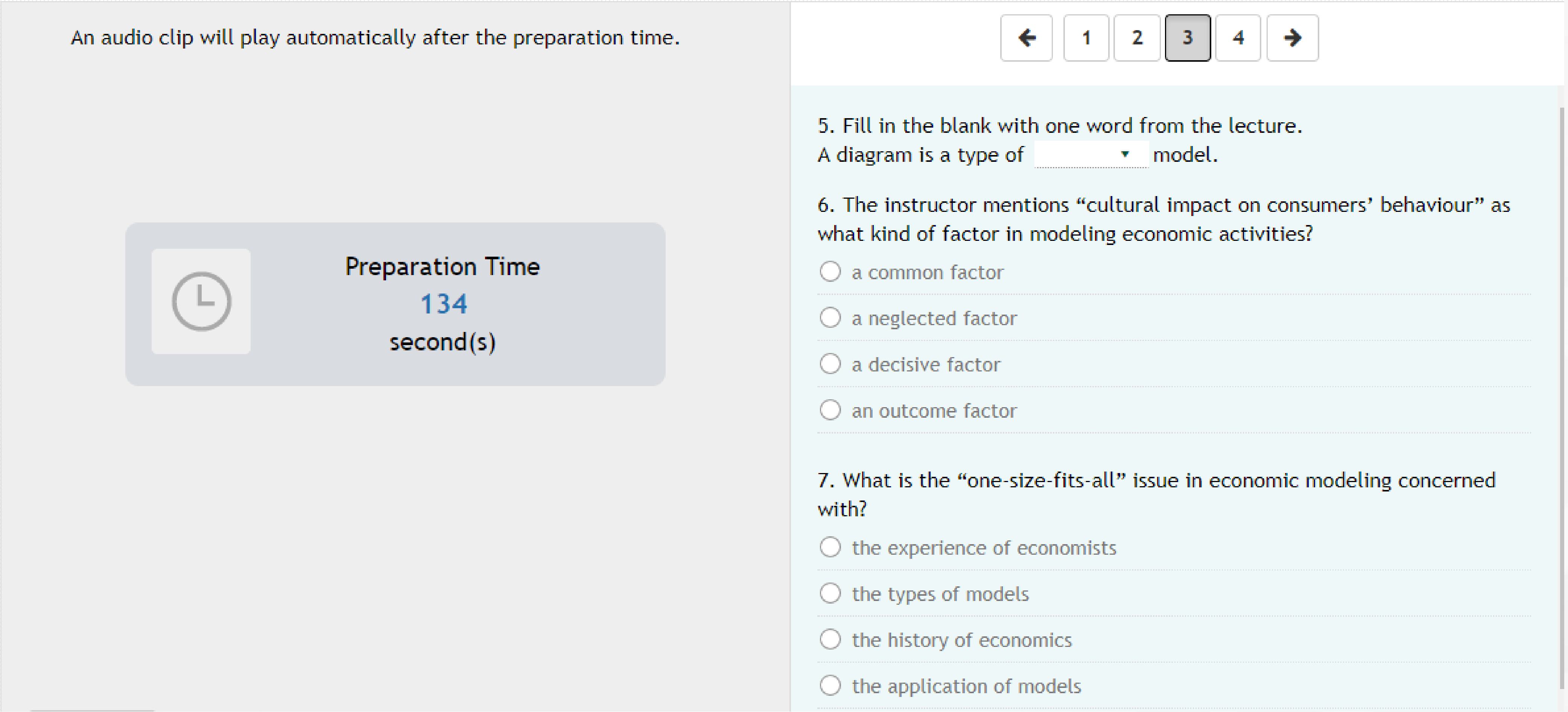
Academic Unit B

Long listening testlet

Long listening testlet

Long listening testlet

The CAEL CE Listening Test – Sample Interface



An audio clip will play automatically after the preparation time.

Preparation Time
134
second(s)

5. Fill in the blank with one word from the lecture.
A diagram is a type of ______ ▾ model.

6. The instructor mentions “cultural impact on consumers’ behaviour” as what kind of factor in modeling economic activities?

○ a common factor
○ a neglected factor
○ a decisive factor
○ an outcome factor

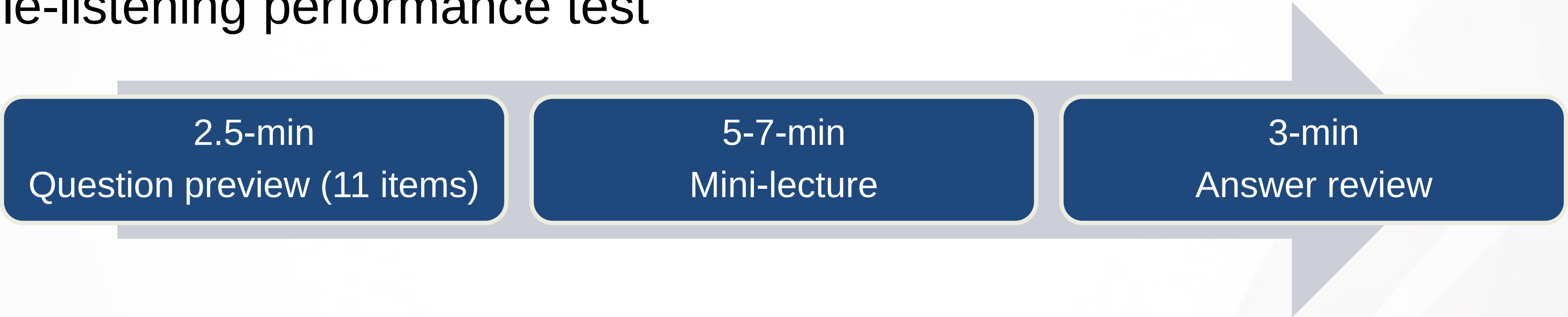
7. What is the “one-size-fits-all” issue in economic modeling concerned with?

○ the experience of economists
○ the types of models
○ the history of economics
○ the application of models

Note: This is a screenshot of an example listening test.

The CAEL CE Listening Test

- We focused on the multiple choice questions in the three long listening testlets
- Features of these listening testlets
  - While-listening performance test



  - Academic topics: Two topics in arts & One topic in science
  - Subskills: Comprehending *local* information, Comprehending *global* information, & Making *inferences*

Participants

88 participants recruited for a pilot test

- Gender: 48 females and 40 males
- Major first language (L1) groups:
  - Chinese, Farsi, Arabic, Spanish, & Korean
- Proficiency levels (CAEL CE listening band score):
  - Low: Band score 20-40 (16)
  - Median: Band score 50-60 (41)
  - High: Band score 70-90 (31)

Data Collection & Analysis

- Data
  - Timestamped log data: answer-changing behaviors
  - Test performance and item score

- Preliminary Analysis
  - Mostly based on descriptive statistics to look for the patterns and possible relationships

Overview of Answer-changing Behaviors

Lecture topic	# of TTs[a]	Total # of changes	Min, Max	Average # of changes per TT	SD
Topic 1	67	304	(1, 28)	4.5	5.0
Topic 2	61	208	(1, 36)	3.4	2.7
Topic 3	66	245	(1, 8)	3.7	3.5
TOTAL	87[b]	757	(1, 47)	8.7	9.1

Note: TT = Test taker

a. number of test takers who made at least one change (N = 88 TTs, k= 28 MCQs).

Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers' proficiency levels?

3. What were the outcomes of the changes?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers' proficiency levels?

Who Made the Changes (by Listening Proficiency Levels)

Proficiency (Listening)	# of TTs	Total # of changes	(Min, Max)	Average # of changes per TT	SD
Low	16	117	(1, 22)	7.3	5.9
Mid	41	448	(1, 47)	10.9	12.2
High	31	192	(1, 15)	6.2	4.2

TT = Test taker

Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers' proficiency levels?

3. What were the outcomes of the changes?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers' proficiency levels?

When Did the Changes Take Place

Pre-listening	During-listening	Post-listening	Total
43	370	344	757

Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

How Were the Changes Related to the Types of Subskill

Subskills	Pre-listening	During-listening	Post-listening	Total # of changes	Average # of changes	Average of TTs[a] per item
Global (k=8)	12	128	135	275	1.9	17
Local (k=11)	15	144	69	228	1.4	15
Inference (k=9)	16	101	147	264	1.3	19

Note: a. number of test takers who made at least one change on one item (N = 88)

Research Questions

1. Who made the changes?

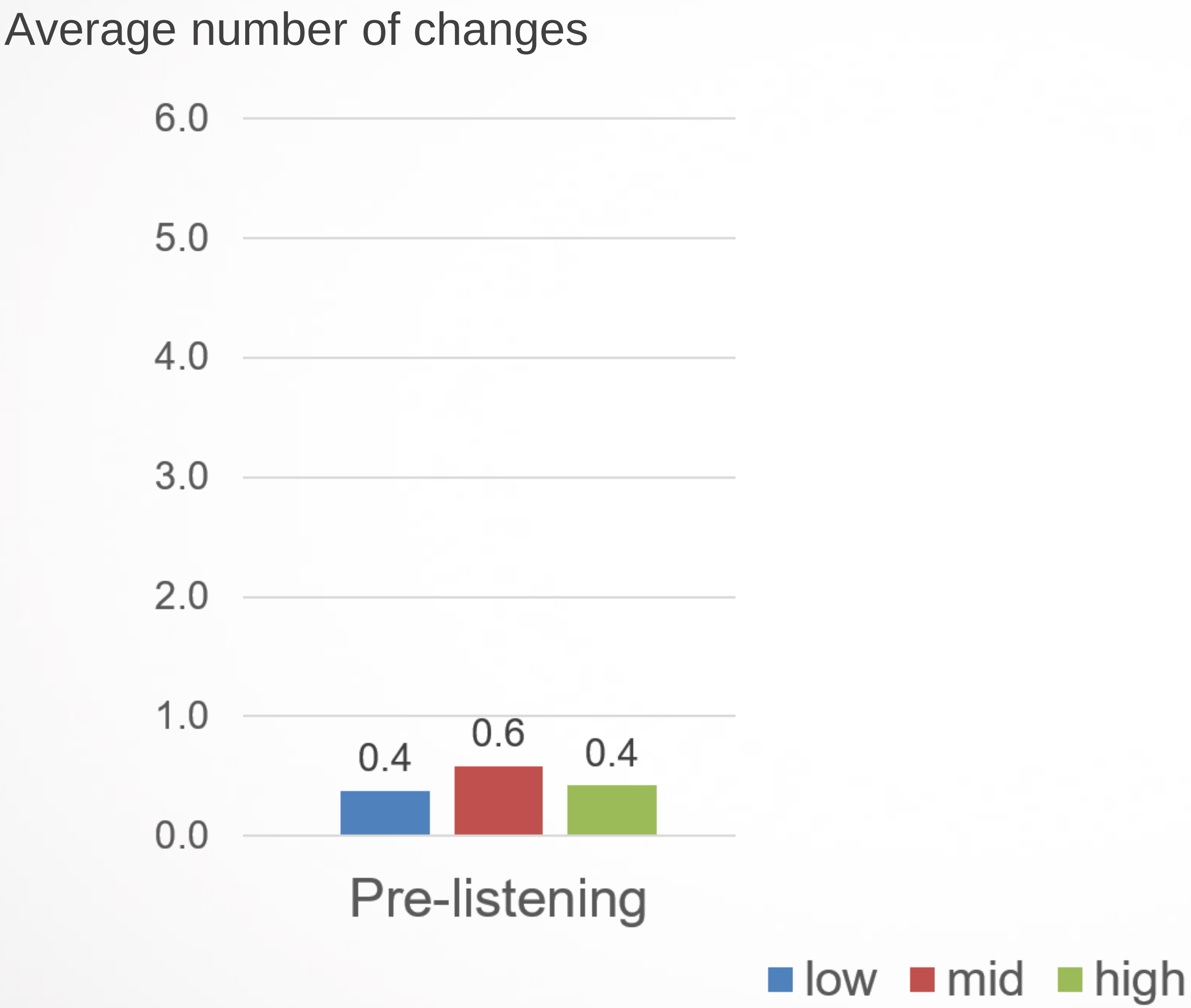
2. When did the changes take place throughout the listening test?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   i. Was it dependent on the subskills measured?
   ii. Was it dependent on test takers’ proficiency levels?

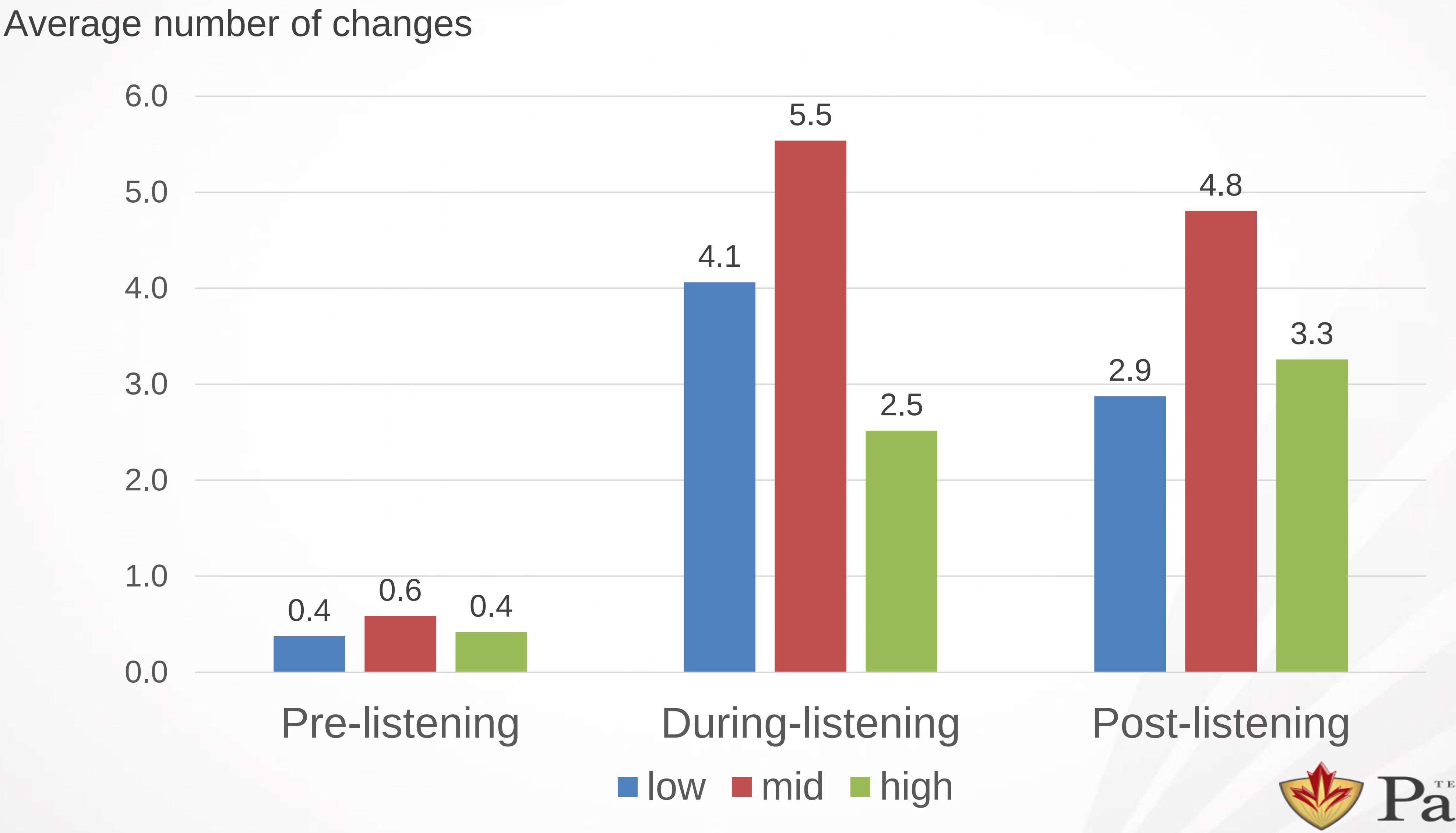
When Did the Changes Take Place

Proficiency (Listening)	**Pre-listening**	**During-listening**	**Post-listening**	**Total**
Low	6	65	46	117
Mid	24	227	197	448
High	13	78	101	192
Total	43	370	344	757

When Did the Changes Take Place



When Did the Changes Take Place



Paragon TESTING ENTERPRISES

Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers’ proficiency levels?

RQ3: What Were the Outcomes of the Changes

Outcomes	Topic1	Topic2	Topic3	Total
W → W	67	56	64	187
R → W	32	31	25	88
W → R	56	43	56	155
R → R	9	16	13	38

RQ3: What Were the Outcomes of the Changes

Outcomes	Topic1	Topic2	Topic3	Total
W → W	67	56	64	187
R → W	32	31	25	88
W → R	56	43	56	155
R → R	9	16	13	38
Correct rate (WR)	0.34	0.29	0.35	**0.33**
Loss rate (RW)	0.20	0.21	0.16	**0.19**

RQ3: What Were the Outcomes of the Changes

Outcomes	Topic1	Topic2	Topic3	Total
W → W	67	56	64	187
R → W	32	31	25	88
W → R	56	43	56	155
R → R	9	16	13	38
Correct rate (WR)	0.34	0.29	0.35	**0.33**
Loss rate (RW)	0.20	0.21	0.16	**0.19**

Research Questions

1. Who made the changes?

2. When did the changes take place throughout the listening test?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
    i. Was it dependent on the subskills measured?
    ii. Was it dependent on test takers’ proficiency levels?

How Were the Changes Related to the Types of Subskill

Subskills	**Most changes happened at**	**Average correct rate (W→R)**
Global (k=8)	Post-listening, During-listening	0.41
Local (k=11)	During-listening	0.37
Inference (k=9)	Post-listening	0.44

Note: number of test takers who made at least one change on one item (N = 88)

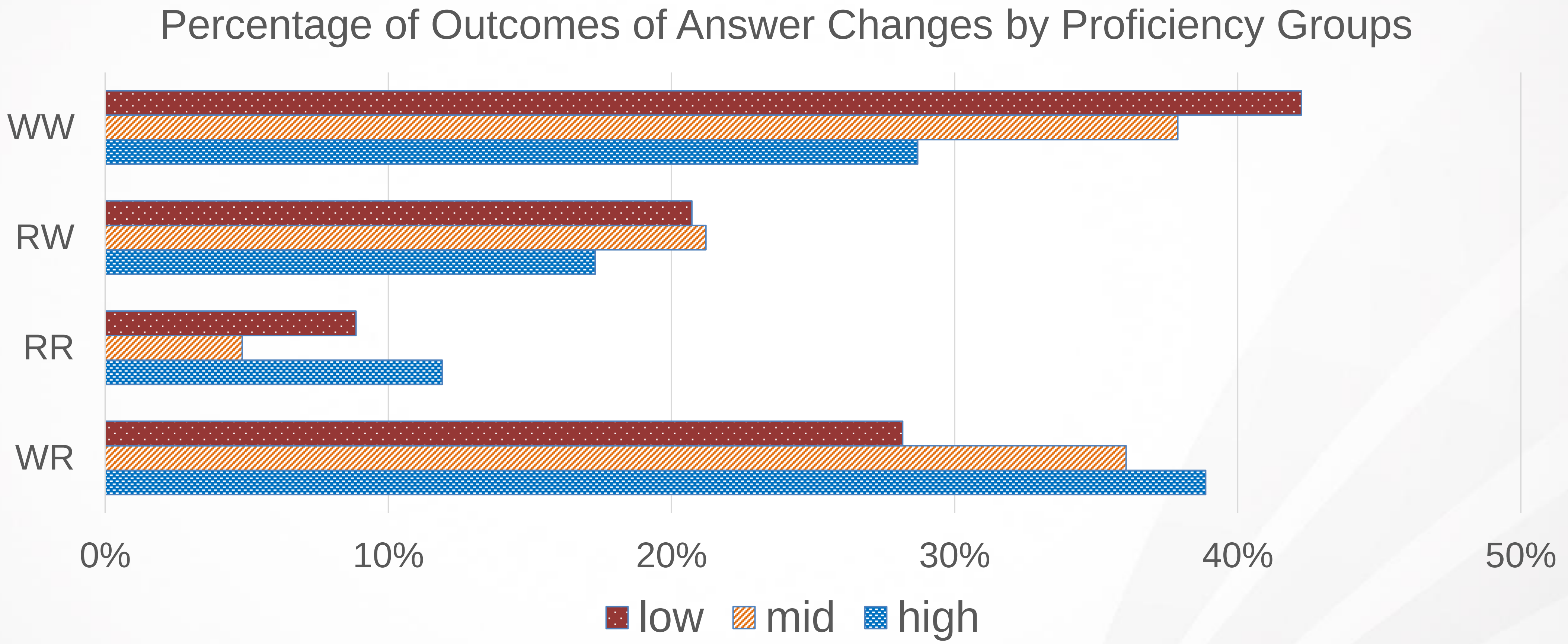
Research Questions

1. Who made the changes?

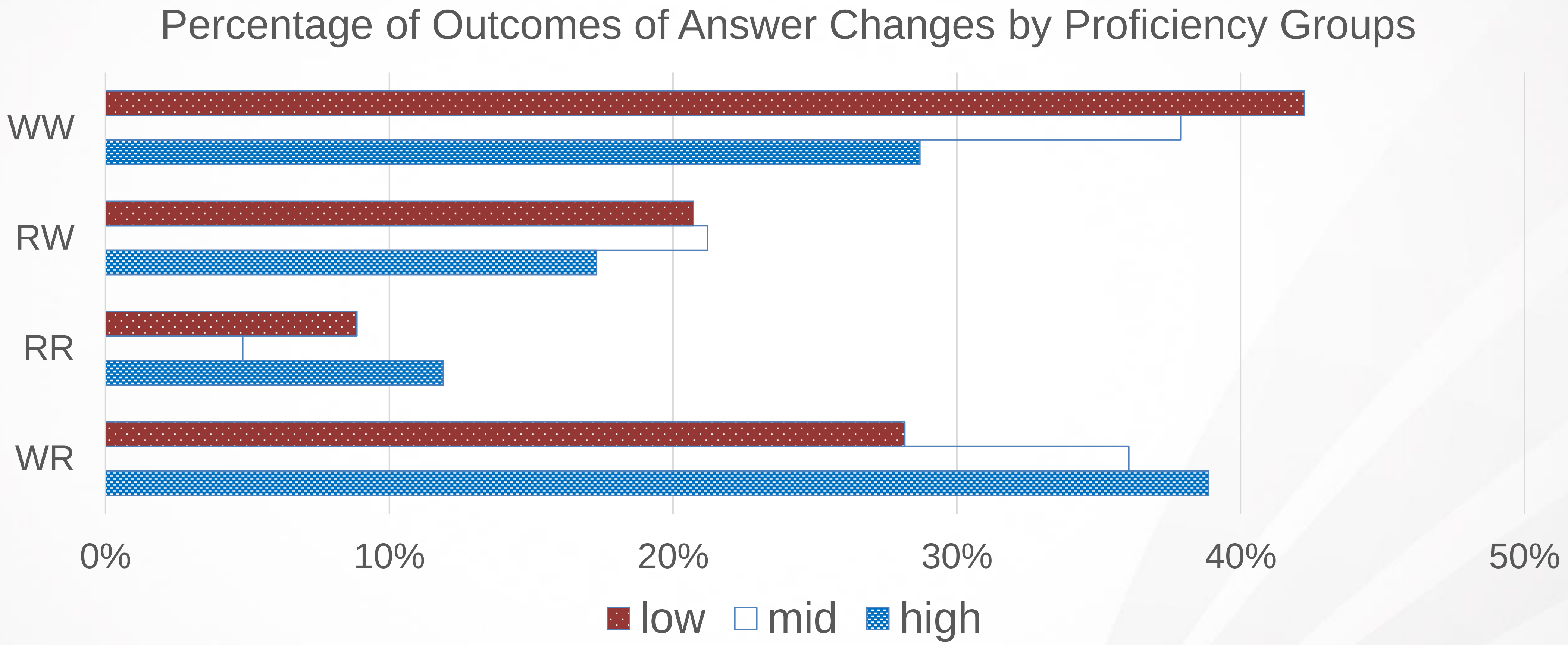
2. When did the changes take place throughout the listening test?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers’ proficiency levels?

3. What were the outcomes of the changes?
   1. Was it dependent on the subskills measured?
   2. Was it dependent on test takers’ proficiency levels?

What Were the Outcomes of the Changes



What Were the Outcomes of the Changes



Answer-changing Behaviors and Test Validation

Understanding answer-changing behaviors

- Who
- Effectiveness of the changes

- When
  - Related to the target skills/constructs
  - Reflect test taking strategies? Metacognitive strategies?

Future Studies

Timestamped responding data + other data types

- – The findings can be triangulated with an analysis of other behavioral data (e.g., eye-tracking) and/or think-aloud data

→ Better understanding of test-taking processes and their relationships with the measured construct

Thank You

research@paragontesting.ca

Selected References

- Al-Hamly, M., & Coombe, C. (2005). To change or not to change: investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing*, *22*(4), 509–531. http://doi.org/10.1191/0265532205lt317oa
- Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a Foreign Language Listening Tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing*, 15(3), 1–23.
- Ballance, C. T. (2006). Item characteristics and answer-changing Behaviors. *Psychological Reports*, *98*(1), 205–208.
- Couchman, J. J., Miller, N. E., Zmuda, S. J., Feather, K., & Schwartzmeyer, T. (2016). The instinct fallacy: The metacognition of answering and revising during college exams. *Metacognition and Learning*, *11*(2), 171–185.
- Foote, R., & Belinky, C. (1972). It pays to switch? Consequences of changing answers on multiple-choice examinations. *Psychological Reports*, *31*(2), 667–673. http://doi.org/10.2466/pr0.1972.31.2.667
- Jacobs, S. S. (1972). Answer Changing on objective tests: Some implications for test validity. *Educational and Psychological Measurement*, *32*(4), 1039–1044. http://doi.org/10.1177/001316447203200420
- Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of Response Changes in the GRE Revised General Test. *Educational and Psychological Measurement*, *75*(6), 1002–1020. http://doi.org/10.1177/0013164415573988
- Qualls, A. L. (2005). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice*, *20*(1), 9–16. http://doi.org/10.1111/j.1745-3992.2001.tb00053.x
- Stylianou-Georgiou, A., & Papanastasiou, E. C. (2017). Answer changing in testing situations: the role of metacognition in deciding which answers to review. *Educational Research and Evaluation*, *23*(3–4), 102–118.
- Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire (MALQ): Development and validation. Language Learning, 56(3), 431–462.