Evaluating the use of 'none of the above' in multiple choice testing

Matt Pachai

McMaster University

Acknowledgements

- Dr. Joe Kim
- Dr. David DiBattista
- Yvonne Chen
- The Pedagogy Research Lab

IntroPsych
department of psychology, neuroscience & behaviour

McMaster University
Inspiring Innovation and Discovery

Outline

1) The goal of multiple choice (MC)
2) None of the above (NOTA) in MC
3) The present experiment
4) Future directions and implications

Goals of Testing

- What are your goals in testing students?
  - Assessment?
  - Discrimination?
  - Learning?

MC Guidelines

- Haladyna and Downing (1989a) examined 46 textbook passages on MC
- Produced **43** recommendations for a "good" question

Sample Guidelines

- Use Positives, not Negatives, in the Stem
- Avoid None of the Above
- Avoid complex (Type K) questions

A Bad Question

- Which of the following would not increase obedience in the Milgram experiment?
  - i. Moving the experimenter to another room
  - ii. Moving the experiment to a run down building
  - iii. Dressing the experimenter in dirty clothes
  - iv. Moving the learner closer to the teacher
    - a) i and ii
    - b) ii and iii
    - c) i, ii, and iii
    - d) iii and iv
    - e) None of the above

Empirical Support

- Only **half** of these recommendations were empirically examined
- A clear need for rigorous examination remains

Haladyna and Downing, 1989b

Measurement Tools

- How do we examine our test's ability to achieve our goals?
  - **Difficulty:** Percent Correct
  - **Discrimination:** Point-biserial correlation
  - **Learning:** Retention

Performance

- A simple way to measure knowledge at two levels
- **Students:**
  - How many questions did each student answer correctly?
- **Concepts:**
  - What percentage of students got a particular question correct?

Point-Biserial Correlation

- A measure of a question's ability to discriminate between students

- What is the correlation between the answers for a particular question and each students' final score?

Point-Biserial Correlation

Options

Grade Category	A	B	C*	D
% A	0	0	90	10
% B	5	2	83	11
% C	5	1	66	27
% D	23	5	35	37
% F	32	7	37	24

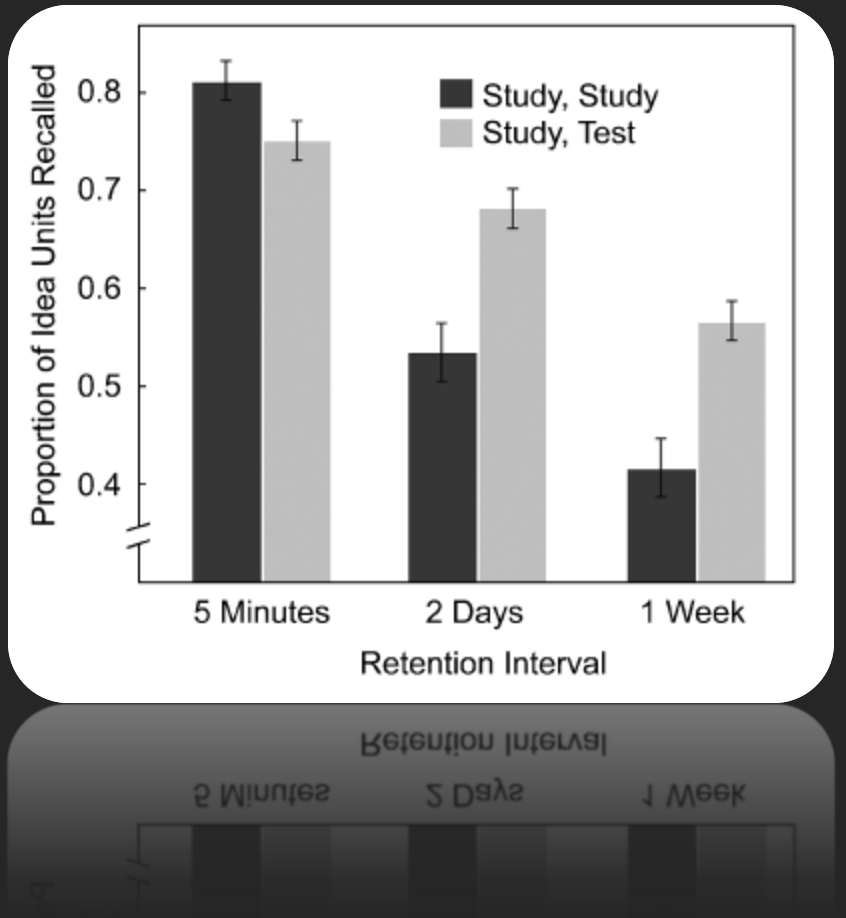
Point-biserial correlation = 0.32

Retention Experiments

- Cognitive psychologists have extensively studied retention of material
- **Basic Paradigm:**
  - Session 1: teach a concept
  - Session 2: test retention after a delay

The Positive Testing Effect

- Numerous studies suggest testing improves learning



Carpenter et al., 2008; Roediger and Karpicke (2006)

The Impact of Flaws

- Flawed questions are more difficult (Downing, 2005)
- Test flaws may hurt high achieving students more than low (Tarrant and Ware, 2008)

Specific Flaws

- Previous studies classify flawed questions based on a large number of guidelines
- Hard to decipher which specific flaws have which specific effects

The Case of NOTA

- In a recent review, **48%** of textbook authors agreed that NOTA should be avoided (Haladyna et al., 2002)

Empirical Evidence

- The few studies examining NOTA have produced mixed results
- NOTA may:
  - increase difficulty and discrimination
  - not change difficulty and discrimination
  - increase difficulty but not discrimination

Mixed Messages

- “When NOTA is correct… it rewards examinees with serious knowledge deficiencies or misinformation” … “Any stem or option format that reduces an item’s ability to distinguish between candidates with full and misinformation should not be used” (Gross, 1994)

Mixed Messages

- “NOTA should remain an option in the item-writer’s toolbox, as long as its use is appropriately considered. However, given the complexity of its effects, NOTA should generally be avoided by novice item writers.” (Haladyna et al., 2002)

General Questions

- What effect does NOTA have on:
  - Assessment?
  - Discrimination?
  - Learning? (not addressed today)

Our Study

- We examined NOTA on two of our Introductory Psychology examinations (approx **3000** students/year)
- Advantages of our population:
  - A large class
  - Highly motivated students
  - Topical questions, basic and applied

Test Design

- Five versions of each test were produced
- Each test contained 5 experimental questions, randomly distributed

Conditions

- Each test version had one question in each of the following conditions:
  - No NOTA (control)
  - NOTA as key
  - NOTA replacing distractor #1
  - NOTA replacing distractor #2
  - NOTA replacing distractor #3

Summary of Design

	FORM 1
Q1	Normal
Q2	NOTA D3
Q3	NOTA D2
Q4	NOTA D1
Q5	NOTA key

Sample Question: Normal

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  a) from which **food** was most often delivered.
  b) that provided the most contact comfort.
  c) that was present when **danger** was presented.
  d) that was present for the greatest amount of **time**.

Sample Question: NOTA Key

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  a) from which **food** was most often delivered.
  b) that was present when **danger** was presented.
  c) that was present for the greatest amount of **time**.
  d) None of the above

Sample Question: NOTA D1

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  - a) **that provided the most contact comfort.**
  - b) that was present when **danger** was presented.
  - c) that was present for the greatest amount of **time**.
  - d) None of the above

Sample Question: NOTA D2

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  a) from which **food** was most often delivered.
  b) that provided the most contact comfort.
  c) that was present for the greatest amount of **time**.
  d) None of the above

Sample Question: NOTA D3

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  a) from which **food** was most often delivered.
  b) that provided the most contact comfort.
  c) that was present when **danger** was presented.
  d) None of the above

Recoding Distractors

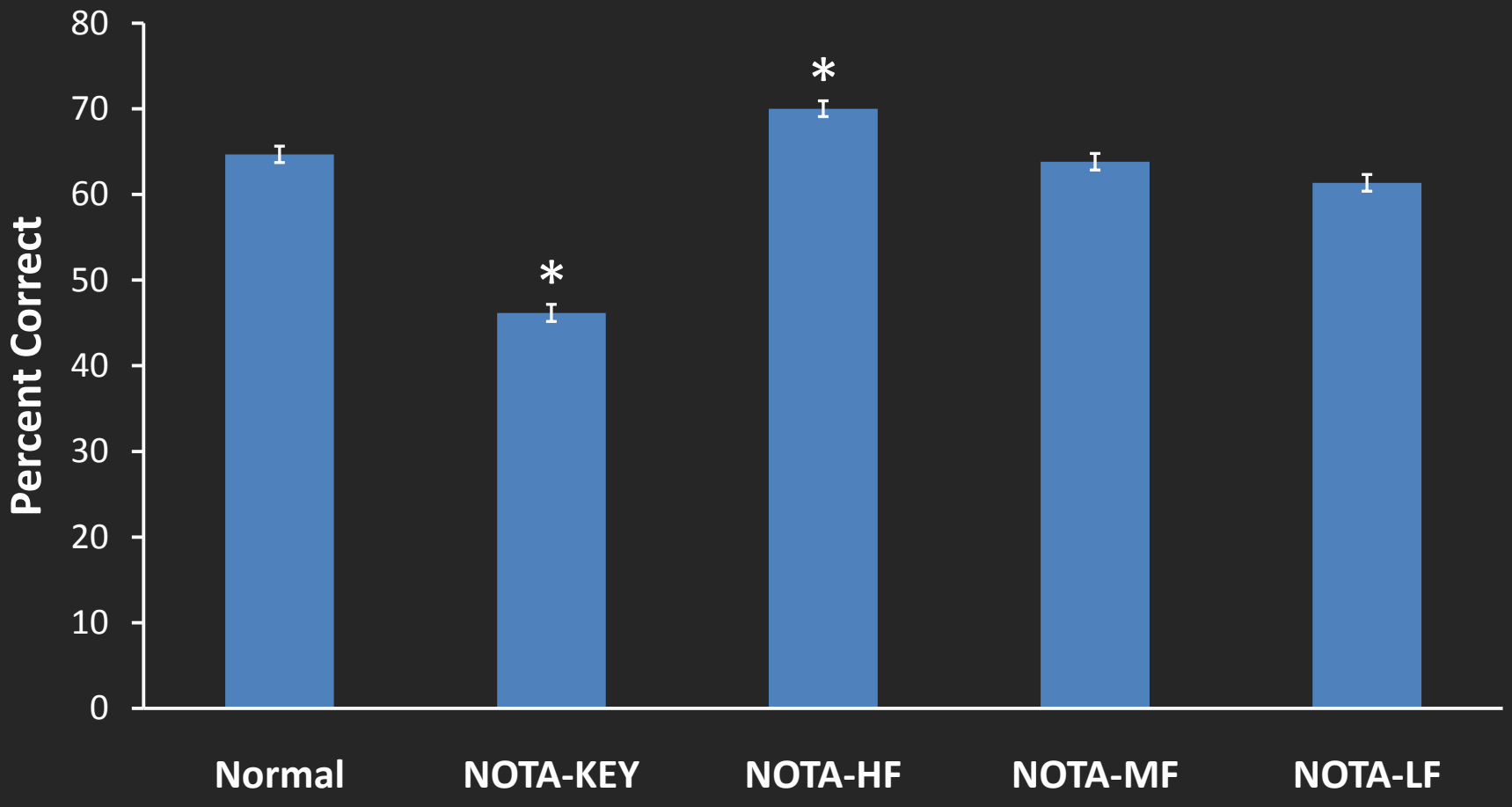
- Distractors were recoded as either high frequency, middle frequency, or low frequency selections

- **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:
  a) from which **food** was most often delivered. (**HF:** 19%)
  b) that provided the most contact comfort.
  c) that was present when **danger** was presented. (**LF:** 4%)
  d) that was present for the greatest amount of **time**. (**MF:** 17%)

Analysis

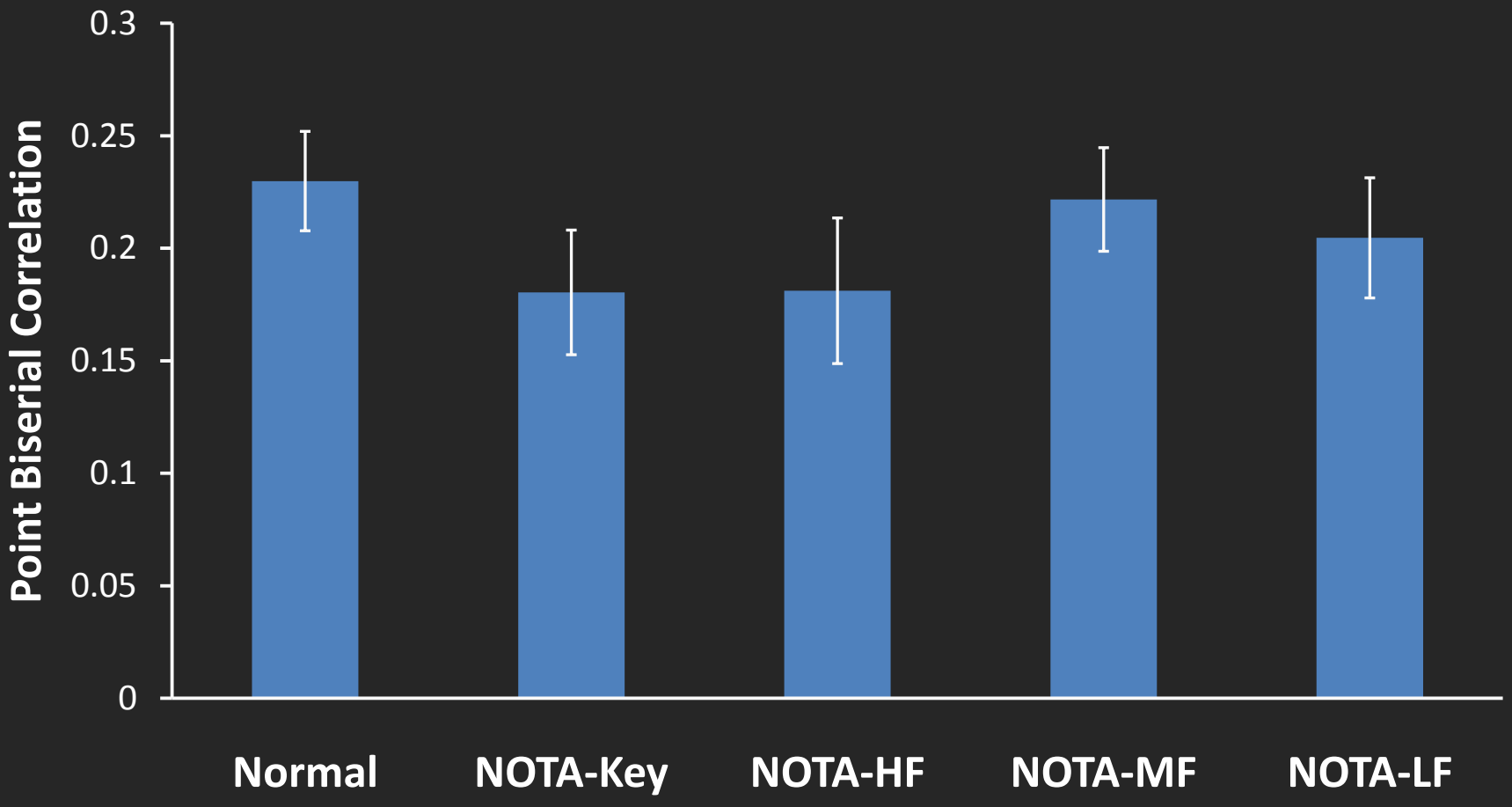
- Independent Variable: Condition
  - Normal
  - NOTA-Key
  - NOTA-HF
  - NOTA-MF
  - NOTA-LF
- Dependent Variables
  - Performance (% correct)
  - Discrimination (point-biserial correlation)

Performance



* = $p < 0.001$

Discrimination



$p > 0.05$

Implications

- What effect does NOTA have on:
  - **Assessment:**
    - **Key**: Increased difficulty
    - **Distractor:** Less effective than a good distractor
  - **Discrimination:** No effect
  - **Learning:** Negative testing effect? (Odegard and Koen, 2007)

Future Directions

- When NOTA is the correct answer, do the students selecting it know the truth?
  - Fill in the correct response for a bonus

General Conclusions

- Understanding the specific effects of writing “errors” is highly important

- Test writers should be thoughtful in question writing
  - Questions should be matched to the goals of the test

Evaluating the use of ‘none of the above’ in multiple choice testing

Questions?

References

- Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448.
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-133.
- Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of "none of the above.". *Evaluation & the Health Professions, 17*(1), 123-126.
- Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 37–50.
- Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1,* 51–78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-309.
- Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15*(8), 873-885.
- Roediger, H.L., III, & Karpicke, J.D. (2006). Test enhanced learning: Taking memory tests improves long term retention. *Psychological Science, 17 (3),* 249-255
- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206.