Memorizing your ABCD's: Detection of item preknowledge and application for future item development

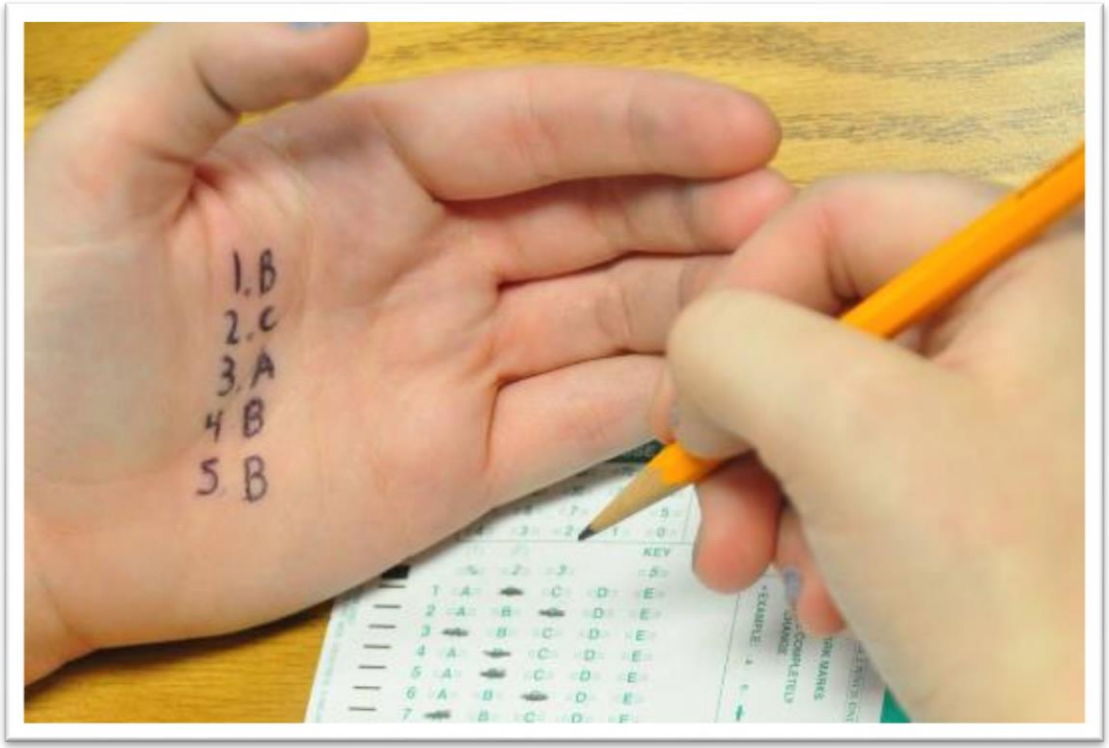
Brooke Dresden and Nicole Tucker

October 11, 2018

Operational Definition

Item Preknowledge

Occurs when a candidate begins a test with knowledge of questions and/or answers obtained from an outside source, such as a teacher, website, or past examinee (Sinharay, 2017).



Purpose of Licensure Examinations

- To **assess professional competence** in terms of the knowledge, skills, and abilities required to successfully perform the important tasks that practitioners must be able to carry out at the time of initial licensure

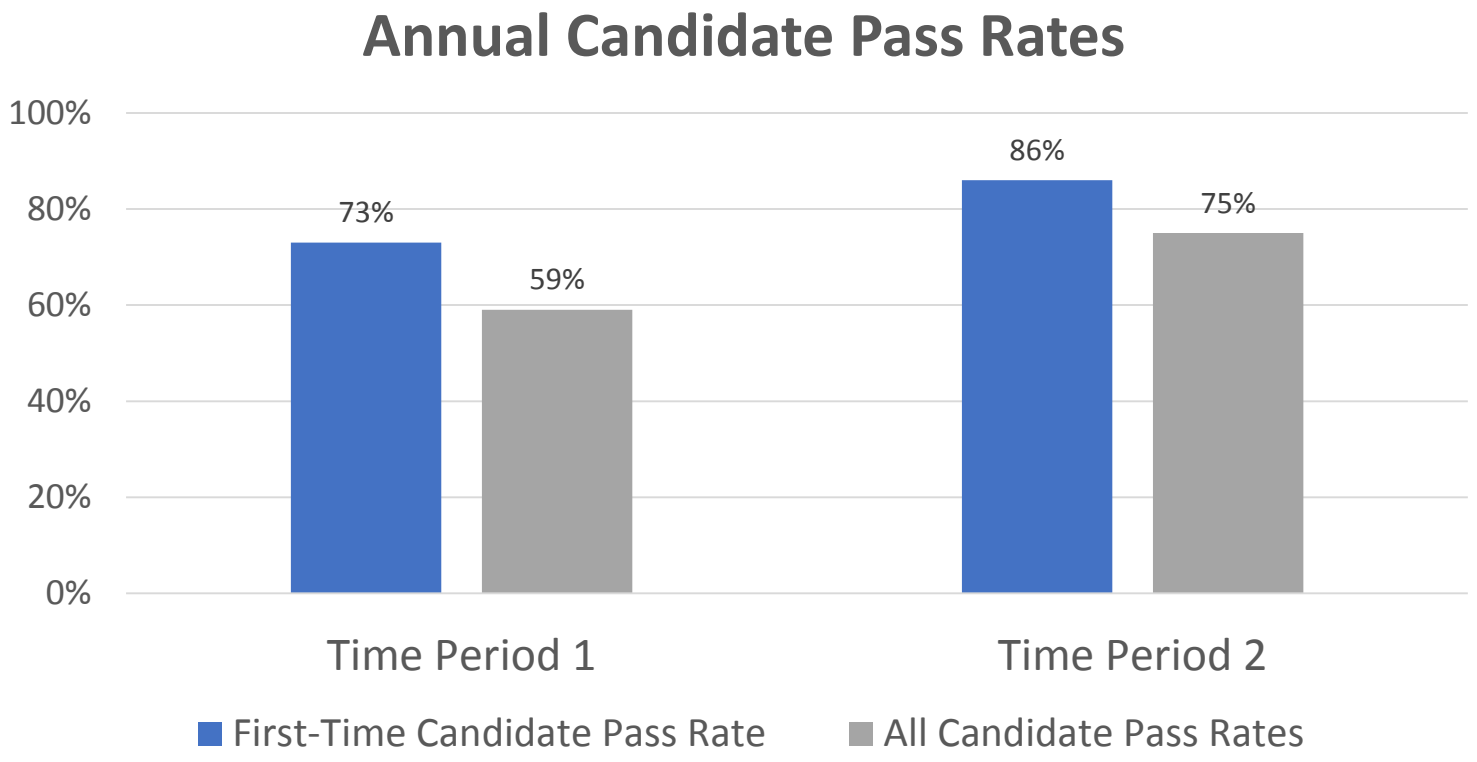
- To provide **a valid and reliable** means of identifying candidates who are at least **minimally competent** to practice in the profession

- To **protect the public** from harm

Background – Exam One

High-stakes state licensure examination

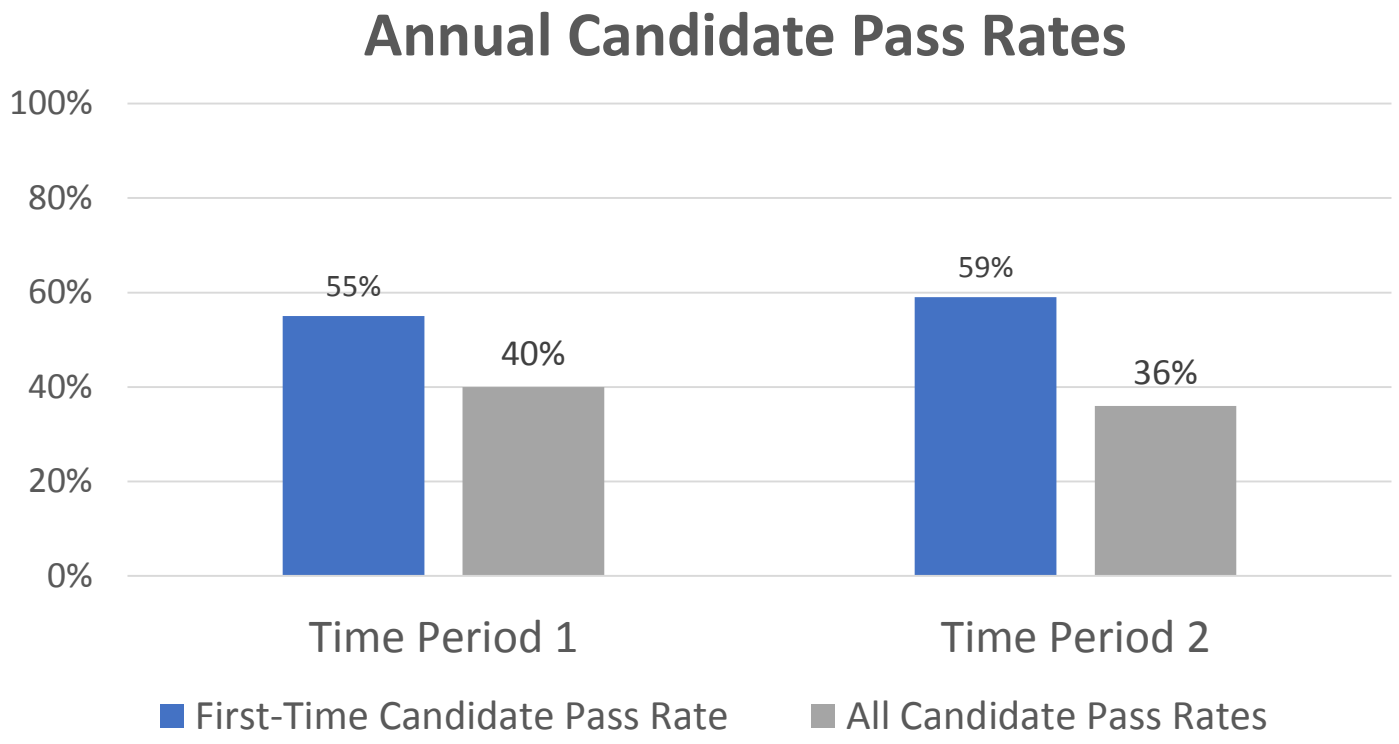
- 542 annual candidates
- Noticed a sharp increase in candidate pass rates
- 70 item test; 100 minutes allowed test time



Background – Exam Two (Control)

High-stakes state licensure examination

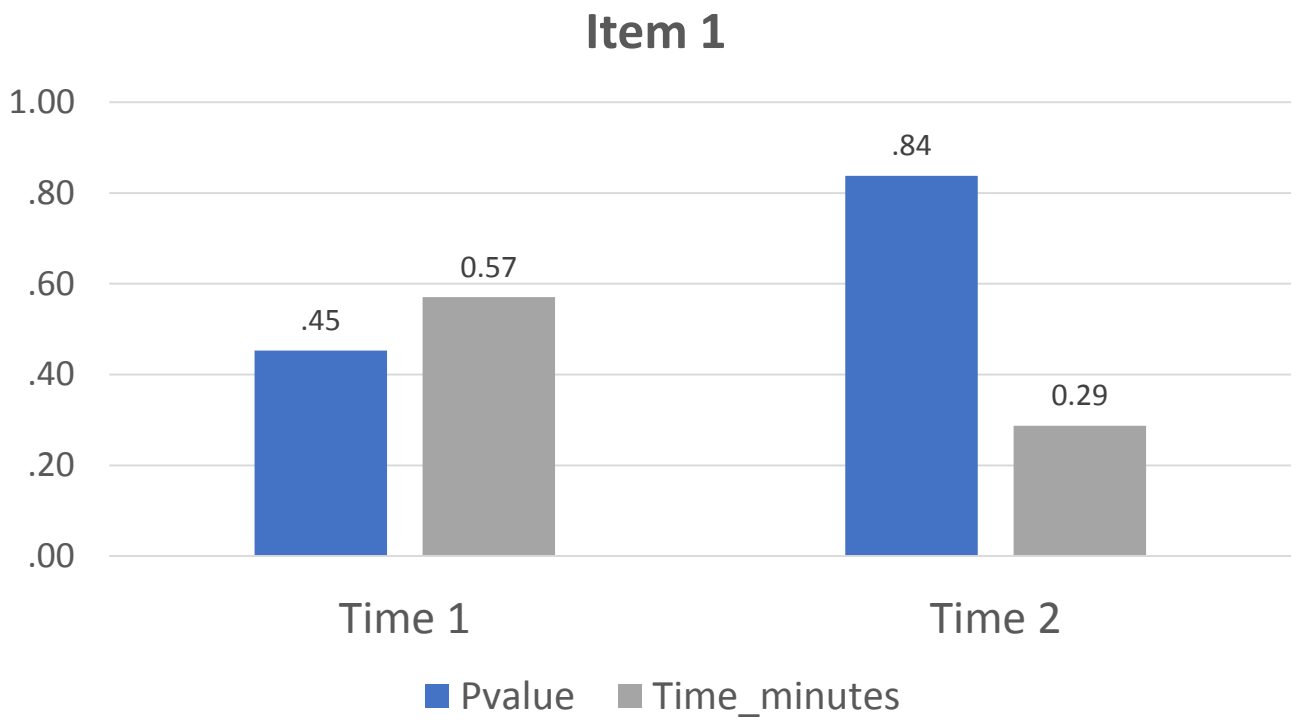
- 403 annual candidates
- Consistent candidate pass rates
- 105 item test; 120 minutes allowed test time



Present Study – Example Item

Compare Item Parameters between Time 1 and Time 2

- Item Difficulty
- Item Time



Additional Background Information

- Both Exam One and Exam Two first launched in 2015.

- ***Time Period 1***
  - October 1 2015 through September 30 2015
  - Total Candidate Volume: Exam One - 586, Exam Two - 591

- ***Time Period 2***
  - October 1 2017 through September 20 2018
  - Total Candidate Volume: Exam One – 591, Exam Two - 422

- Slight modification to test forms (turned on or turned off items) between Time 1 and Time 2, but overall high degree of item overlap between two time periods for both exams.

- No change in cut score or passing standard between time periods.

- Exams share source items there were slightly modified for specific jurisdictions.

Item-Focused Approach

- The focus of the study is to explore simple but meaningful methods for detecting item pre-knowledge, with a goal of informing test development staff of the best items for retirement.

- Following identification of items which suggested potential pre-knowledge, items were examined by test developers and content experts to identify characteristics which made the items easier to remember.

Parameters for Pre-knowledge Detection

Item Difficulty

Pvalue: Change in pvalue for first time test takers between Time 1 and Time 2.

Differential Item Functioning: Rasch-based DIF Analysis with Time 1 as the baseline and Time 2 candidate data as Group 2 using.

Time

Time Spent to Answer Item Correctly

For the purposes of this study, we compared mean time spent answering the item correctly between Time 1 and Time 2. We also explored using the natural log transformation of candidate time spent (Van der Linden, 2006).

Flagging Items and SME Review: Exam One

Identifying items

Items were flagged if there was a 20% increase in pvalue between Time 1 and Time 2.
- 30 items were identified. All had decreases in mean time spent to answer correctly.

Items were flagged if DIF contrast was less than -1.0 logits and the difference was significant.
- 3 items were flagged. All had decreases in mean time spent and were included in the above pvalue flag.

SME Review

Three SMEs independently reviewed the identified items and a small number of items with no change in pvalue. They were asked the following questions:

1. *Is there anything about the item that may make it easier for candidates to memorize?*
2. *Is there anything that can explain the increase in pvalue between Time 1 and Time 2?*
3. *For items with no change in performance, is there anything about the item that can explain why its performance stayed consistent?*

Results – Item Characteristics

- Stem Length did not matter
- Some evidence that response option length did
  - Average response length of identified items: 13.88 characters
  - Average response length of remaining items: 22.90 characters
- Items with little to no change in pvalue tended to have stronger distractors.
- SMEs indicated that there has been an industry push to increase candidate preparedness for certain topics.

Comparison: Exam Two

Reminder

Exams One and Two cover similar content, but are administered in different jurisdictions.

Exam One had a pass rate that changed from 73% at time one to 86% at time two.

Exam two had a pass rate that changed from 55% at time one to 59% at time two.

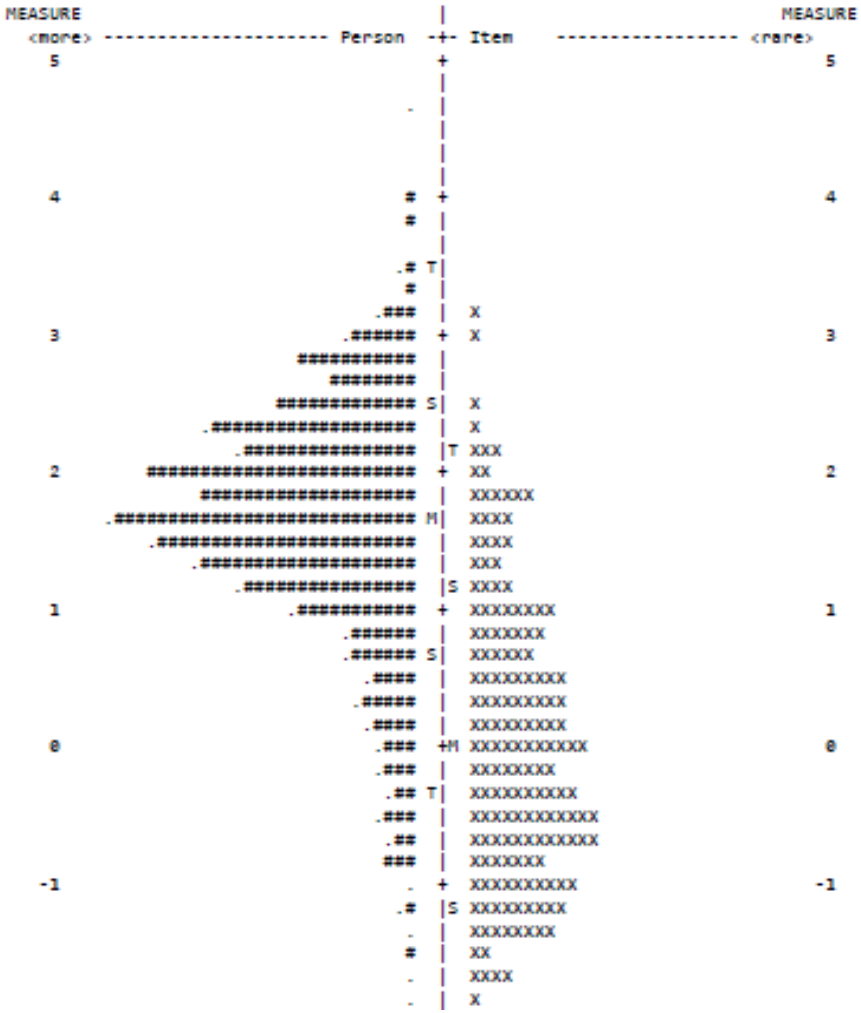
Identified Items

Compared to 30 items flagged on Exam 1, 8 items were flagged for Exam 2 using a criteria of a 20% increase in pvalue between Time 1 and Time 2.

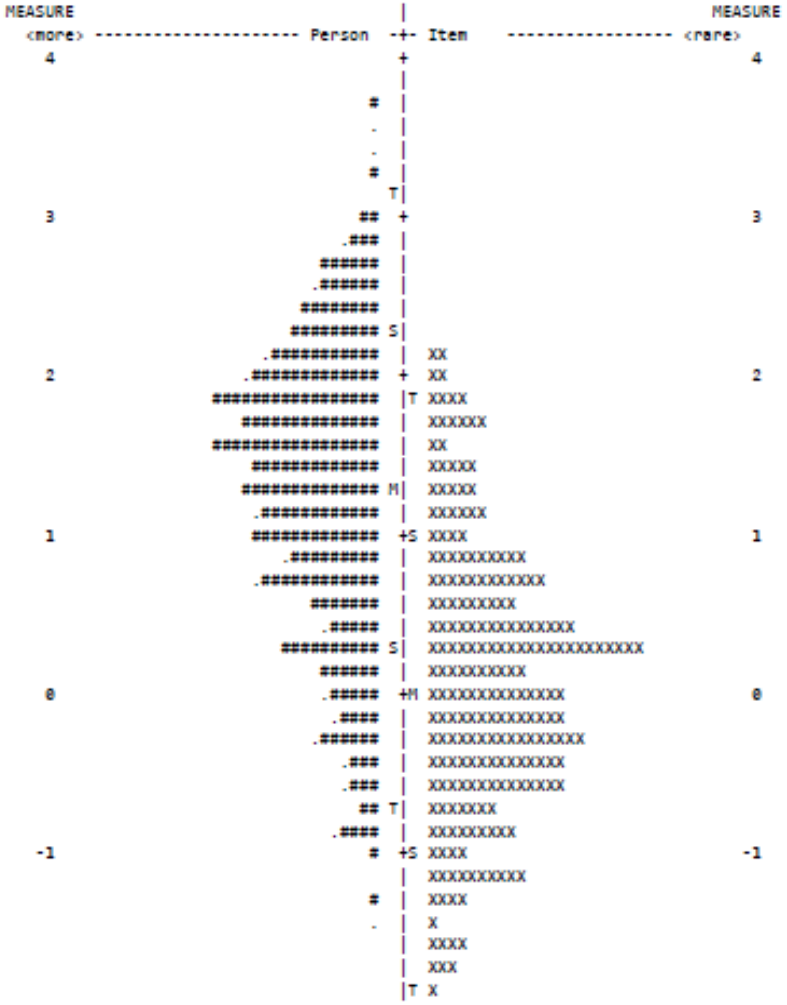
There was no consistent pattern on time spent for items flagged by pvalue.

No items were flagged by using the DIF contrast between Time 1 and Time 2.

Exam 1: Item Person Map



Exam 2: Item Person Map



Discussion – Future Directions

- Recommendations to test developers for Exam One.
  - ✓ Retire easy items.
  - ✓ Revise flagged items, consider updating distractors.
  - ✓ Continue to right-size item bank by introducing harder items.
- Continue analyses with other exams with larger candidate volume
- Implement SME feedback and analyze results

References

- Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.
- Meijer, R. R., & Sotaridona, L. (2006). Detection of advance item knowledge using response times in computer adaptive testing.
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38-47.
- Kingston, N., & Clark, A. (Eds.). (2014). *Test fraud: Statistical detection and methodology*. Routledge.
- Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*(1), 46-68.
- Van der Linden (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181-204
- Verbic, S., & Tomic, B. (2009). Test item response time and the response likelihood. *arXiv preprint arXiv:0901.4356*
- Wollack, J. A. & Maynes, D. D. (2017). Detection of test collusion using cluster analysis. In. G. J. Cizek and J. A. Wollack (Eds.) Handbook of Quantitative Methods for Detecting Cheating on Tests (pp. 124-150), New York, NY: Routledge.

Thank You!