Worth the Squeeze?

An Investigation into the Psychometric Performance of Innovative Item Types

Amanda A. Wolkowitz, Ph.D.

Brett P. Foley, Ph.D.

Case Study

- Large Credentialing Program
- 6 related exams
- Launched a major revision
- Integrated new item types

Item Types

- Historic
  - Standard Multiple Choice (SMC)
  - Check All that Apply (CATA)
  - Quantitative Fill-in-the-Blank (QFIB)
- New
  - Hotspot (HS)
  - Drag and Place (DnP)
- All dichotomously scored

Example Check-All-That-Apply

After a bizarre turn of events, the embarassed boy returned to his home. He had commited a big mistake and felt that it was necessary to apologize to his parents immediately.

Which words are spelled incorrectly in the sentences above? Select all that apply.

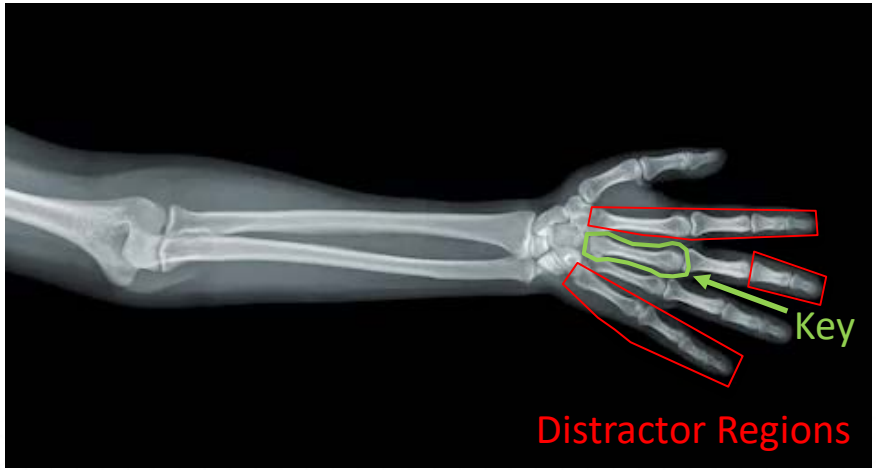
A. bizarre
B. embarassed
C. commited
D. apologize
E. immediately

Example Quantitative Fill-in-the-Blank

Apples costs $1.00 per pound at a local grocery store. The tax rate is 4.00%. If someone buys 5 pounds of apples, how much will the apples cost after tax?

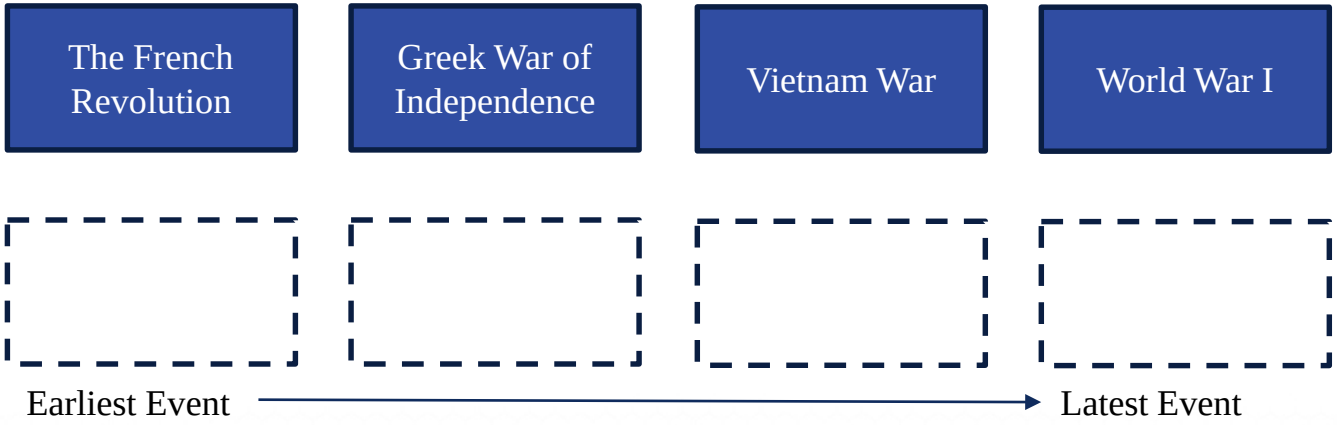
$ _____

Hotspot Example



Click on the area of the picture known as the *middle metacarpal*.

Example Drag and Drop



Place the wars into chronological order in the space provided.

Evaluate Psychometric Performance of Item Types

- Compared...
  - difficulty
  - response time
  - discrimination
  - statistical flagging
  - contribution to accuracy at the cut score
- Distractor area analysis

Exam Summary

- 6 Exams
- 4-6 forms per exam
- 185-1463 candidates per form

Item Type	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
SMC	127	129	124	183	124	211
CATA	35	47	62	33	46	37
QFIB	13	15	15	28	28	26
HS	36	24	5	45	11	47
DnP	18	22	7	35	9	28

Item Performance: P-values

	Median Difficulty[1]				
Exam	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)
1	0.72	0.47	0.17	0.51	0.52
2	0.74	0.55	0.51	0.55	0.20
3	0.64	0.54	0.47	0.75	0.62
4	0.68	0.46	0.43	0.74	0.54
5	0.65	0.39	0.51	0.74	0.50
6	0.75	0.56	0.54	0.69	0.43

[1]Darker shading = more difficult (among item types within that exam)

Item Performance: Response Times

	Median Time[2]				
Exam	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)
1	62	100	245	131	197
2	58	81	201	128	118
3	57	67	231	97	178
4	66	93	197	88	162
5	57	87	197	73	164
6	61	76	158	81	113

[2]Darker shading = candidates take more time (among item types within that exam)

Item Performance: Item-Score Correlations

	Median ISC[3]				
Exam	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)
1	0.24	0.24	0.30	0.13	0.32
2	0.22	0.24	0.33	0.15	0.19
3	0.22	0.25	0.34	0.25	0.24
4	0.18	0.19	0.20	0.22	0.23
5	0.19	0.19	0.24	0.21	0.21
6	0.25	0.29	0.32	0.25	0.26

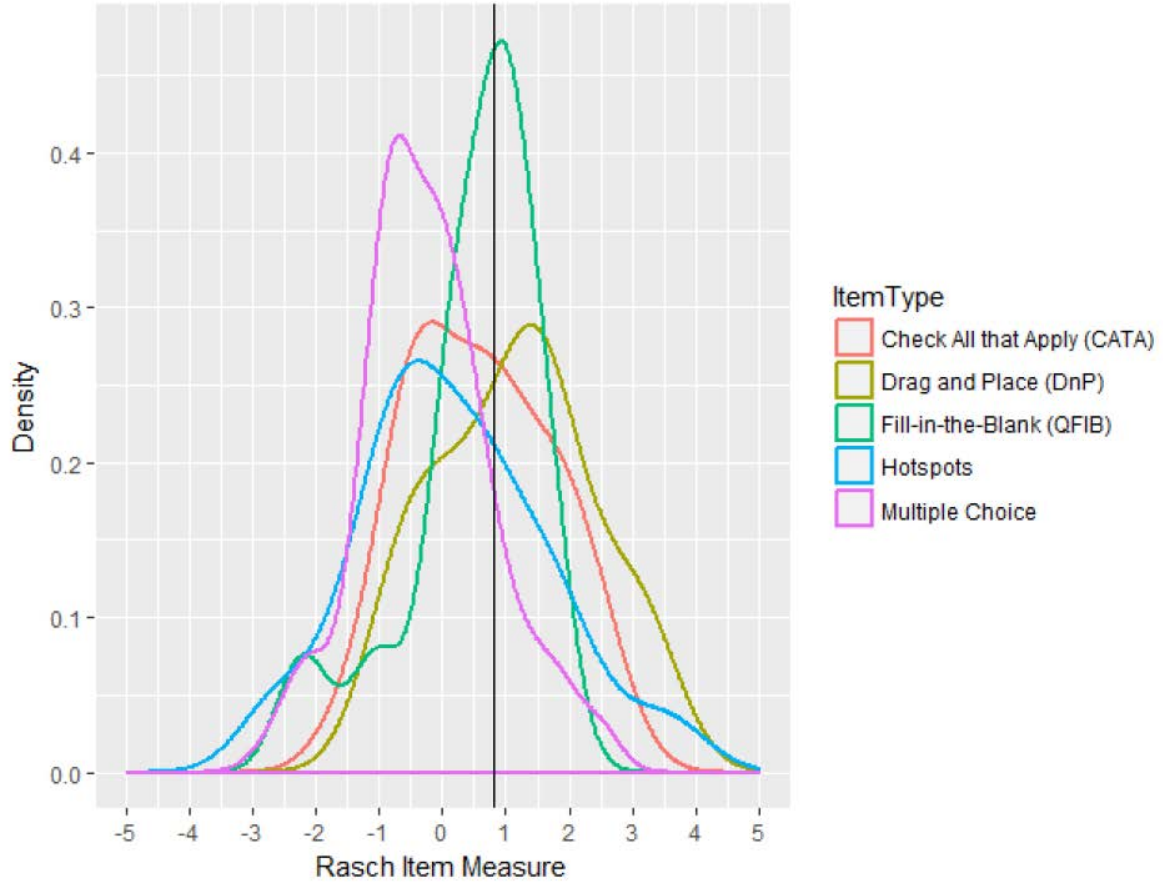
[3]Darker shading = higher discrimination (among item types within that exam)

Item Performance: Statistical Flags

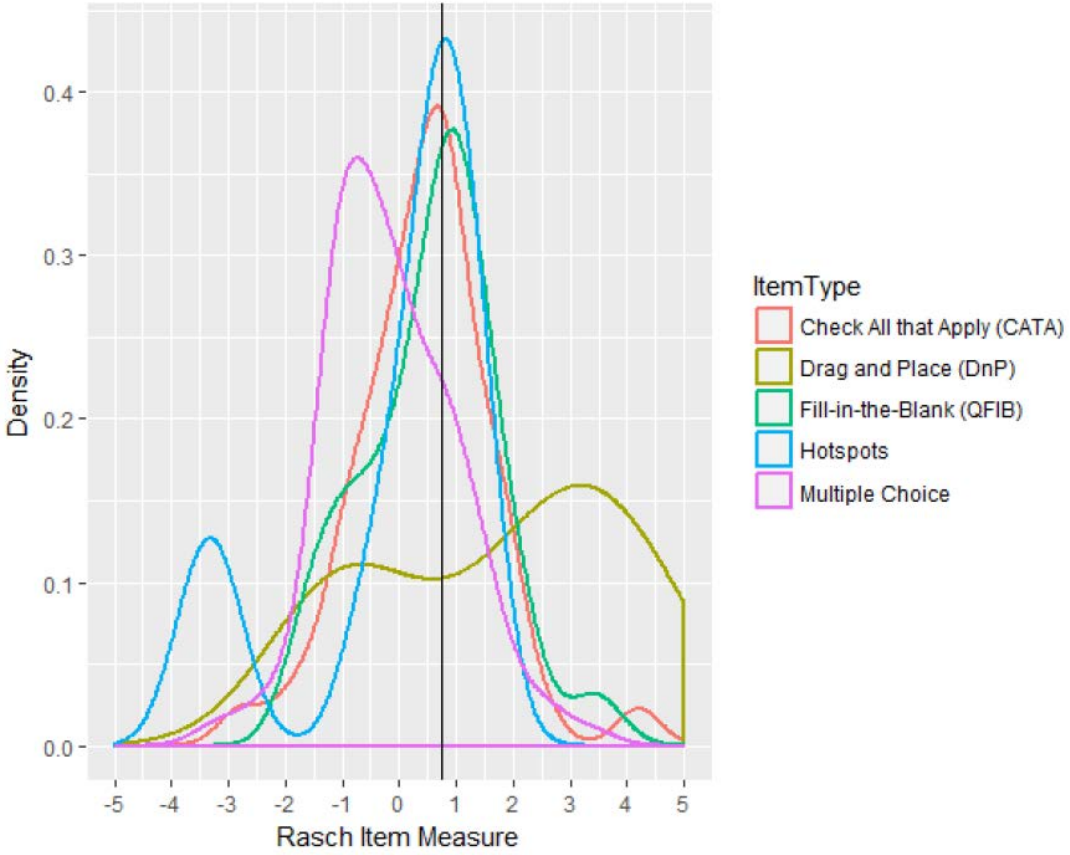
	% of Items with Statistical Flags[4]				
Exam	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)
1	11%	24%	53%	40%	43%
2	12%	22%	21%	18%	44%
3	12%	19%	40%	13%	36%
4	19%	30%	35%	17%	25%
5	10%	39%	18%	22%	34%
6	14%	11%	23%	19%	33%

[4]Darker shading = more flagged items (among item types within that exam)

Item Performance: Contribution to Precision



Item Performance: Contribution to Precision



Item Performance: Distractor Area Performance

Percent of Items in which At Least 20% of Candidates Selected a Region Outside of the Designated Distractor Regions

Item Type	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
DnP	33%	32%	0%	14%	22%	11%
HS	28%	13%	20%	20%	18%	6%

Discussion

- QFIB & DnP
  - More Difficult
  - More Time to Respond
  - More Stat Flags
  - Higher ISC
- HS
  - Easier
  - Moderate Response Time
  - Lower ISC
- CATA
  - More Difficult
  - Moderate Response Time
  - Moderate ISC
- SMC
  - Easier
  - Shortest Response Times
  - Lower ISC
  - Fewest Stat Flags
  - Furthest from Cut Score

Limitations

- Single program
- Early in exam version lifecycle
  - New content
  - Learning curve (writers and candidates)
  - Potentially atypical candidates
- Inconsistent results across exams

Worth it?

- Cons of New Types
  - Development Complexity
    - Time/expertise to write
    - More stat flags
  - Exam Time
    - Seat time
    - Content coverage
- Pros of New Types
  - Potential for strong psychometric performance
  - May better represent real expectations
  - Candidates like them