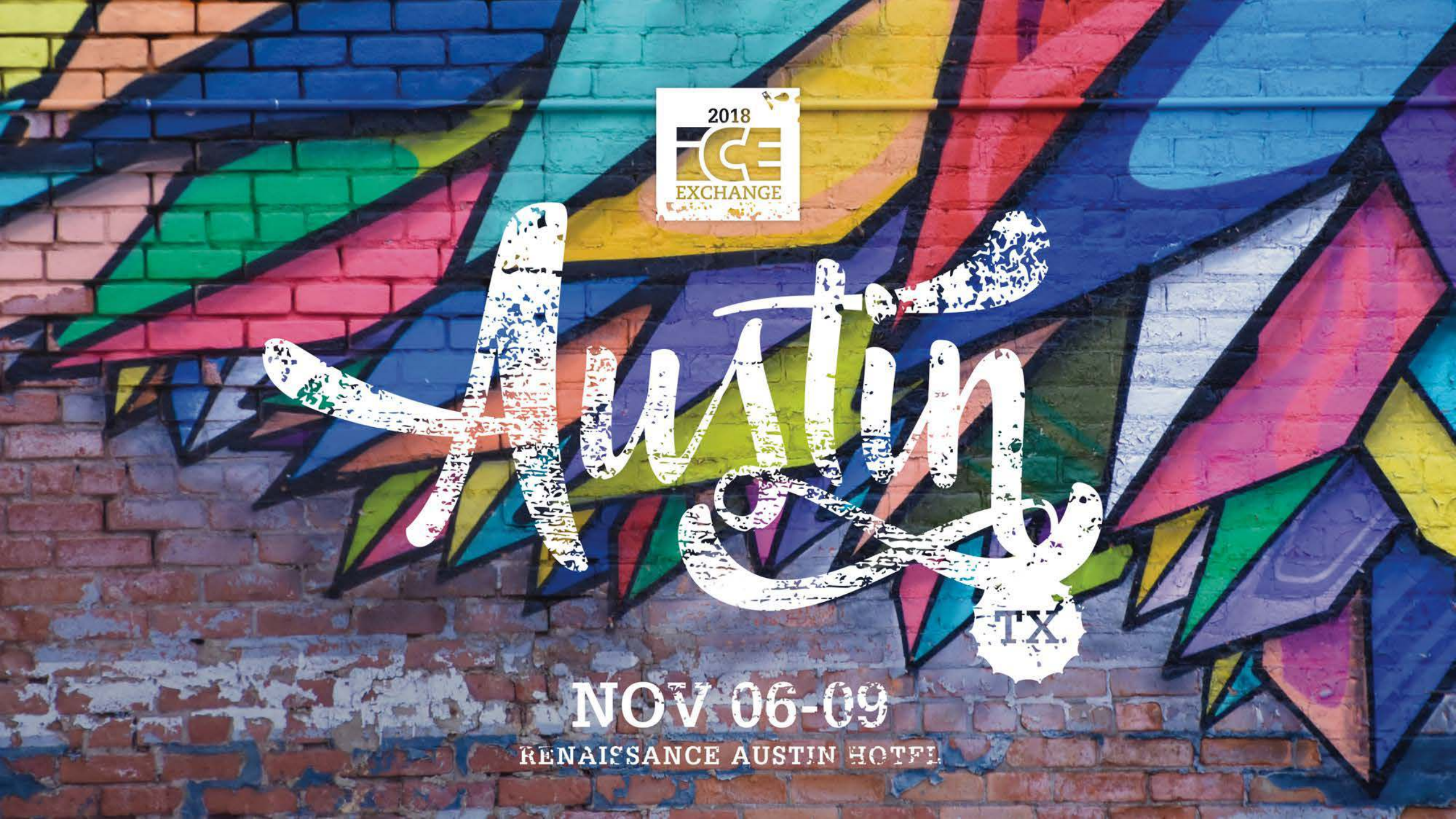

2018
ICE
EXCHANGE
Austin
TX
NOV 06-09
RENAISSANCE AUSTIN HOTEL

Do Innovative Item Types Really Work?

Amanda A. Wolkowitz, Ph.D., Alpine Testing Solutions, Inc.

Brett P. Foley, Ph.D., Alpine Testing Solutions, Inc.

Nick Respecki, AIA, NCARB, National Council of Architectural Registration Boards

What's ahead

- Introduction to Innovative Item Types
  - What are innovative items?
  - When and why should we use innovative items?
  - How do you develop innovative items?
- One Program's Experience
  - NCARB's experience with innovative items
- Results from a Study
  - Do innovative items really work?

What are innovative item types?

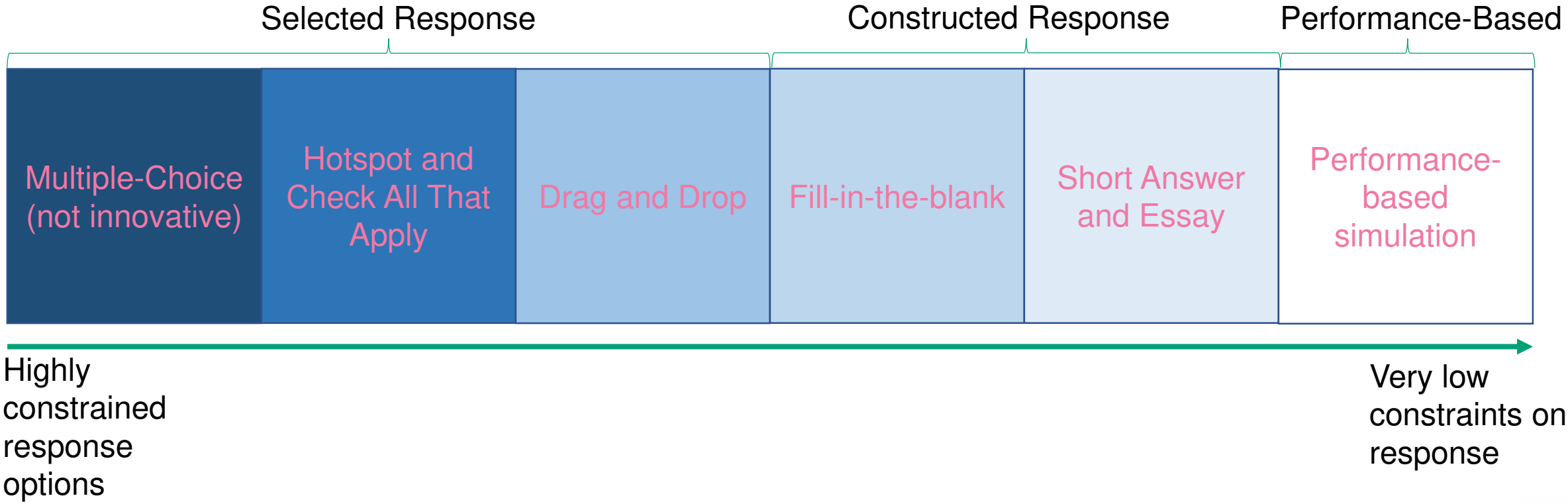
- Before talking about *innovative* item types, what's the purpose of an item?

> "A test item is a device for obtaining information about a test taker's domain of knowledge and skills or a domain of tasks that define a construct" (Haladyna & Rodriguez, 2013, p. 3)

What are innovative item types?

- Also known as *alternate item types* (AIT)
- Items that are not traditional multiple-choice items
- Many are technology enhanced
- Purpose is to increase the measurement opportunity of the construct of interest (Sireci & Zenisky, 2016).

Examples of innovative item types



(Parshall & Brunner, 2017)

Example of Selected Response: Hotspot

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

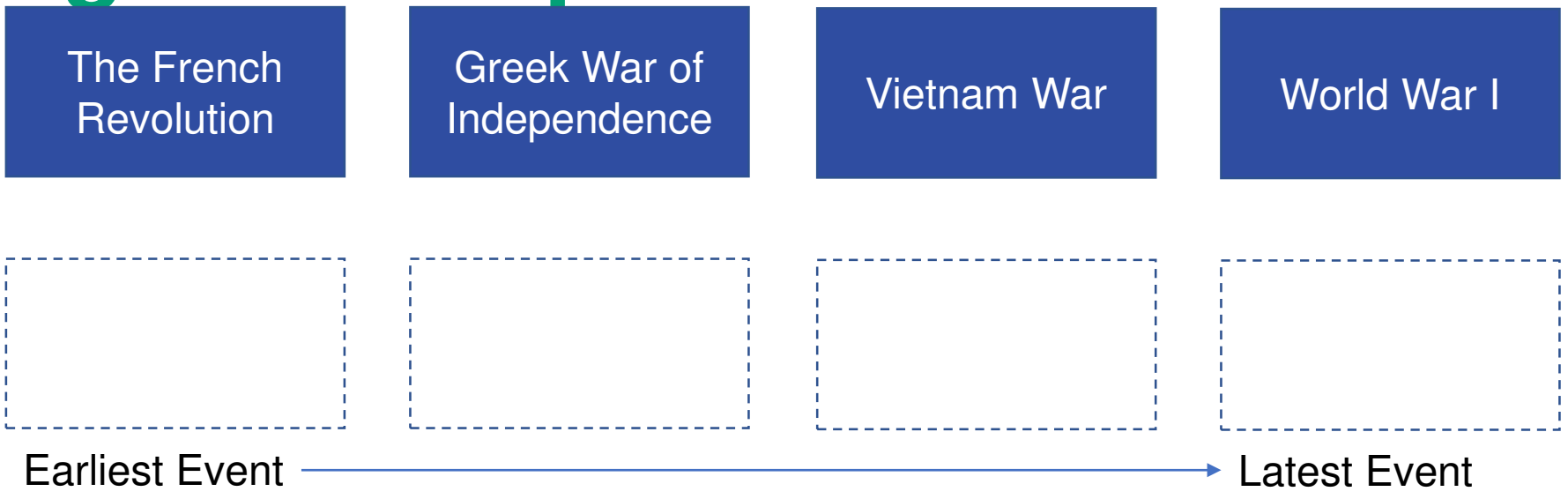
Example of Selected Response: Check-All-That-Apply

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 of Selected Response: Drag and Drop



Place the wars into chronological order in the space provided.

Example of Constructed Response: 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?

$

Example of Constructed Response: Short Answer/Essay

- Describe in 2-3 sentences the main idea of the passage.
- List 2 items from the menu that could be purchased together and the total would not exceed $10.
- What do you predict will happen next and why?

Example of Performance-Based

- Medical or dental exam:
  - A candidate is asked to perform a checkup on a “patient”.
- Physical fitness exam:
  - A candidate is asked to perform certain stretches.
- Teacher exam:
  - A candidate is asked to videotape themselves teaching a class.

Example: Constructed Response
Short Answer/Essay

- Describe in 2-3 sentences the main idea of the passage.
- List 2 errors that were made by the woman in the scenario above.

When and why should we use innovative item types?

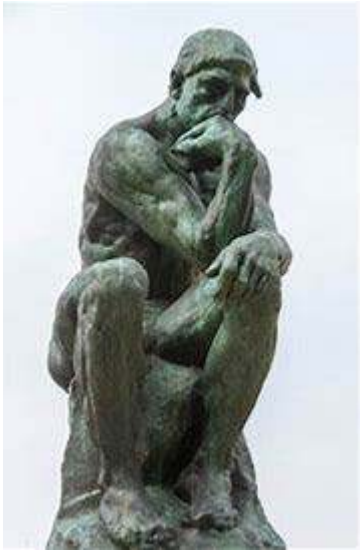
1. Is there a need? If traditional MC items do not adequately measure the knowledge, skill, or ability being assessed, which type of items will address the need?
2. Can your test delivery system handle the innovative item you want to use?

When and why should we use innovative item types?

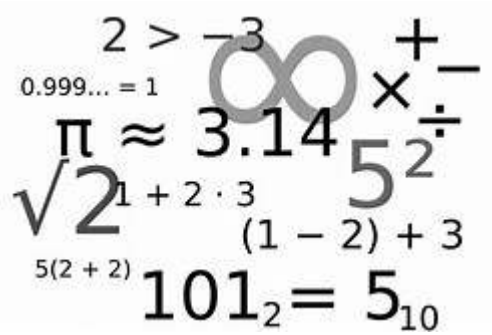
Item Type	Pros	Cons
Selected Response	• Fairly easy to understand • May assess higher level cognitive levels • May assess a variety of content • May increase candidate engagement in item	• Item development takes longer • Need for an innovative item type should be there • May increase memorability of item • Directions must be very clearly stated
Constructed Response	• Same pros as above • Less constraint on possible responses • May decrease item development time (not counting rubric development time)	• Same cons as above, • Possible rubric development (adds time and cost) • Memorability increases
Performance-Based Responses	• Same pros as above • Even less constraint on possible responses • Uses more real-life scenarios	• Same cons as above • Rubric development and raters are a necessity • Memorability increases more

How do you develop innovative item types?

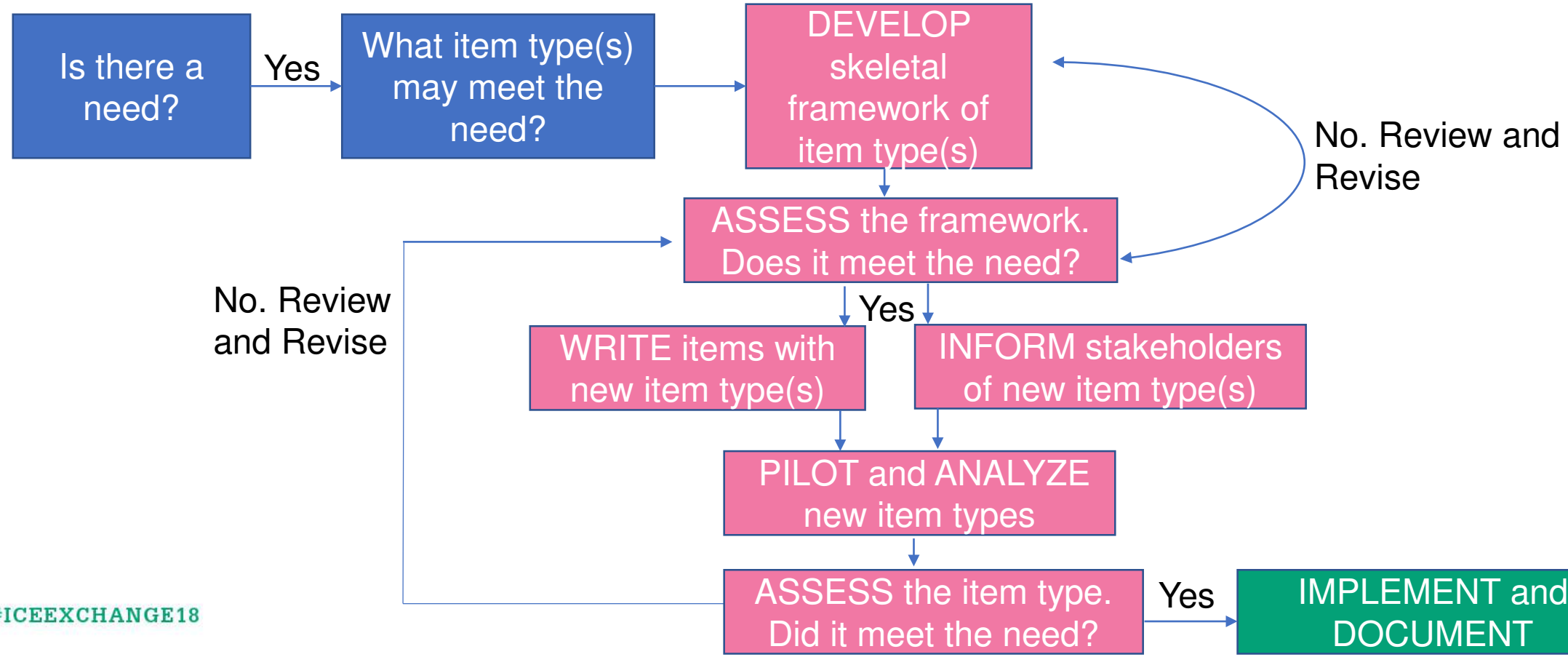
Item Writing = Creative Process



Item Analysis = Statistical Process



How do you develop innovative item types?



NCARB Experience

Architect Registration Examination® (ARE®)

- **1997** – First computer based exam, included **MC** items and **Vignettes**
- **2008** – Added **CATA** and **QFIB** items
- **2016** – Launched new exam, added **Hotspots**, **Drag and Place**, and **Case Studies**

NCARB Experience

Architect Registration Examination® (ARE®)

ARE 4.0 (2008 – 2018)	ARE 5.0 (2016 – Present)
7 divisional exams	**6 divisional exam**
MC, QFIB, CATA	MC, QFIB, CATA, DnP, HS
1-3 vignettes per exam (depends on division)	2 Case Studies per exam
Vignette section scored separately	Case studies scored as part of exam (15-20% of items on exam are case study items; at least 10 items – scored or unscored – associated with each case study)

NCARB Experience

Architect Registration Examination® (ARE®)

Item Type	Percent of Items on Forms
SMC	54-60%
CATA	10-29%
HS	2-16%
QFIB	6-13%
DnP	3-11%

Variability in percentages of any one item type across the exams is a result of the differences reflected in the test specifications for each exam.

NCARB Experience

Architect Registration Examination® (ARE®)

How do each of the innovative item types meet NCARB's need?

- Check-All-That-Apply (CATA; 6 options, 2-4 correct)
- Quantitative-Fill-in-the-Blank (QFIB; on screen calculator provided)
- Drag and Place (DnP)
- Hotspots (HS)

NCARB Experience

Architect Registration Examination® (ARE®)

Why did NCARB do away with vignettes and move to case studies?

- Cost – expensive
- Maintenance
- Old technology

- **Case Studies are a better reflection of how architects practice today**

Do innovative items really work?

- Study compared SMC, CATA, QFIB, HS, and DnP items
- Compared p-value, item time, item-total correlation, number of items flagged, Kernel density plots
- Completed distractor analysis

Do innovative items really work?

Division	# of Candidates per Item[1]	Items Analyzed
Practice Management	1,081	213
Project Management	728	218
Programming & Analysis	761	237
Project Planning & Design	1,662	349
Project Development & Documentation	1,342	324
Construction & Evaluation	419	229

[1]Because some items appeared on more than one form, the number of candidates who took each item varies. This is the median number of candidates who responded to each item within the given division.

Do innovative items really work?

P-values

	Median Difficulty[1]							
Division	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)	Case Study 1	Case Study 2	Stand Alone
PcM	0.72	0.47	0.17	0.51	0.52	0.61	0.51	0.62
PjM	0.74	0.55	0.51	0.55	0.20	0.70	0.54	0.66
PA	0.64	0.54	0.47	0.75	0.62	0.59	0.48	0.65
PPD	0.68	0.46	0.43	0.74	0.54	0.58	0.51	0.63
PDD	0.65	0.39	0.51	0.74	0.50	0.60	0.48	0.65
CE	0.75	0.56	0.54	0.69	0.43	0.60	0.81	0.70

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

Results suggest that QFIB items are among the most difficult item types and SMC and HS are among the easiest.

Results suggest that CS items are similar in difficulty, maybe slightly more difficult than standalone items.

Do innovative items really work?

Time

	Median Time[2]							
Division	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)	Case Study 1	Case Study 2	Stand Alone
PcM	62	100	245	131	197	134	108	70
PjM	58	81	201	128	118	134	124	62
PA	57	67	231	97	178	188	135	64
PPD	66	93	197	88	162	205	191	68
PDD	57	87	197	73	164	183	163	64
CE	61	76	158	81	113	168	124	71

[2]Darker shading = take candidates longer (among item types within that division)

Results suggest that QFIB items are among the most time consuming item types and SMC are among the quickest to answer.

Results suggest that CS items take more time than standalone items.

Do innovative items really work?

Item-Score Correlation (Discrimination)

	Median ISC[3]							
Division	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)	Case Study 1	Case Study 2	Stand Alone
PcM	0.24	0.24	0.30	0.13	0.32	0.30	0.27	0.23
PjM	0.22	0.24	0.33	0.15	0.19	0.22	0.28	0.22
PA	0.22	0.25	0.34	0.25	0.24	0.31	0.28	0.22
PPD	0.18	0.19	0.20	0.22	0.23	0.17	0.21	0.19
PDD	0.19	0.19	0.24	0.21	0.21	0.19	0.21	0.19
CE	0.25	0.29	0.32	0.25	0.26	0.33	0.26	0.25

[3]Darker shading = discriminates better between high/low ability candidates (among item types within that division)

Results suggest that QFIB items are among the most discriminating item types, while SMC and HS are among the least discriminating.

Results suggest that CS items discriminate better than standalone items.

Do innovative items really work?

Percent of Statistically Flagged Items

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

[4]Darker shading = more likely to be statistically flagged (among item types within that division)

Results suggest that DnP items are among the most often flagged item types, while SMC items are flagged the least.

Do innovative items really work?

Rasch Analysis (Kernel Density Plots)

- Which items type did a good job of targeting candidates at the passing score level?
  - QFIB
  - Case Study items

Do innovative items really work?

Distractor Analysis (DnP and HS)

- Candidates select areas outside of the distractor regions
- Problematic for understanding why candidates missed an item
- This can be improved upon.

Summary

Item Type	Difficulty	Time	ISC	Flags
SMC	Easy	Short	Low	Few
CATA	*	*	High	*
QFIB	Hard	Long	*	*
HS	Easy	*	Low	*
DnP	*	*	*	Many
Case Studies	*	Long	High	–
Standalone	*	Short	Low	–

*Middle rating

NCARB Experience

Architect Registration Examination® (ARE®)

What did NCARB learn from this study?

- It's still early, ARE 5.0 launched November 2016
  - Item writers continue to improve each year
- Items – Can an item type be turned off?
- Case Studies – Should the exam include more case studies?
- Should the exam contain more innovative item types?

Closing Thoughts

- SMC items are still good!
- Innovative item types work well to help make the test more real-life.
- Limitations: New item types are new to both writers and examinees. Statistics for the items may change as everyone becomes more familiar with the item types.

Closing Thoughts

- Stakeholders must have advanced noticed.

- Design considerations (e.g., may need to lengthen seat time due to time requirements)

Never be satisfied with status quo. Keep working to make the items better and address the purpose of the test.

Q & A