SmartItems and their Surprising Effects on Test Security

Sarah Toton
Tara Williams
Chris Foster
David Foster
Alison Foster Green

Conference on Test Security

October 11, 2018

Outline

- Introduction to SmartItems and their Security Benefits
- Developing SmartItems
- Psychometric Quality of Game of Thrones SmartItems
- SmartItem Simulations
- Conclusions

Introduction to SmartItems

Purpose of the Session: Comparing SmartItems and Traditional MC

Definition of a SmartItem:

1. A SmartItem covers a skill/objective/competency completely.
2. A SmartItem looks different to each examinee.
3. A SmartItem may vary in thousands, even millions of ways.
4. A SmartItem can be any item type.

Security Benefit: Stealing SmartItems is not feasible or profitable. The most harmful forms of cheating are impossible.

Worth repeating…

SmartItems CANNOT BE functionally:

stolen,
harvested,
pirated,
purloined,
borrowed, or
pilfered!

Imagine that outcome for a minute…

Also, almost all types of cheating are prevented!

Typical Response to SmartItems

I would love to worry very little about security, but do SmartItems really work?

In other words: Do they support valid interpretations of test scores?

In other words: What is their psychometric quality?

That is the purpose of this research, to compare the psychometric properties of SmartItems and traditional items, on the same content.

We used Game of Thrones content to access research participant volunteers from GoT fan forums.

Experimental Design

21 Items	Test Form A	Test Form B	Test Form C
7 items	SmartItem DOMC	MC	SmartItem MC
7 items	SmartItem MC	SmartItem DOMC	MC
7 items	MC	SmartItem MC	SmartItem DOMC

Uses of Randomization:
Items were randomized to forms.
Item order was randomized on the test.
The order of MC and DOMC options was randomized.
Test forms were randomly assigned to participants.

Content Development:
The 21 items were developed by a GoT expert to completely cover the breadth of 21 different skills or objectives.

Developing SmartItems

Types / Possibilities

For our research:
Multiple Choice
DOMC

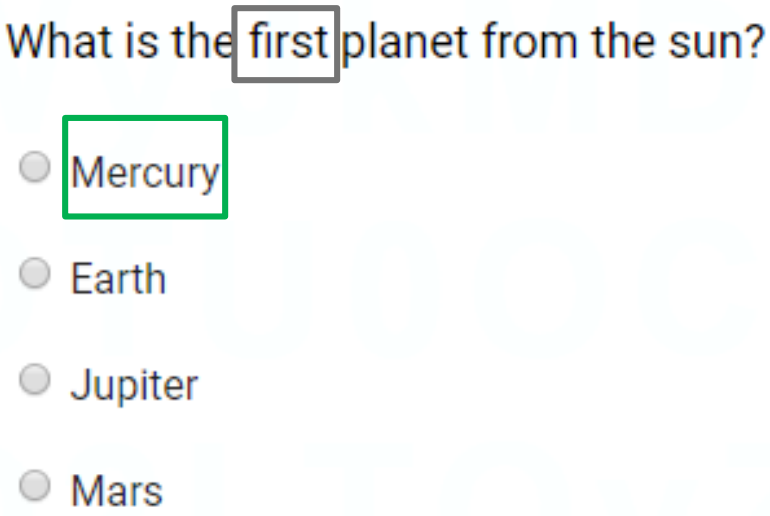
Beyond our research:
Matching
Build List
Short Answer

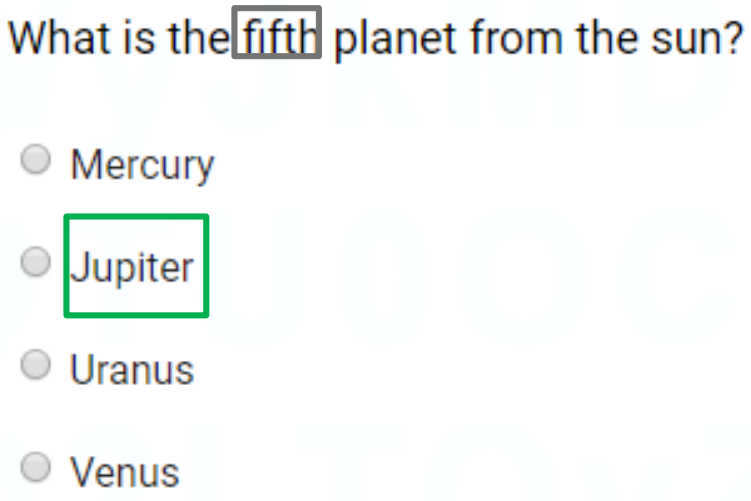
Additionally:

-Scenarios
-Images, videos
-Code snippets

Sample SmartItem (MC)

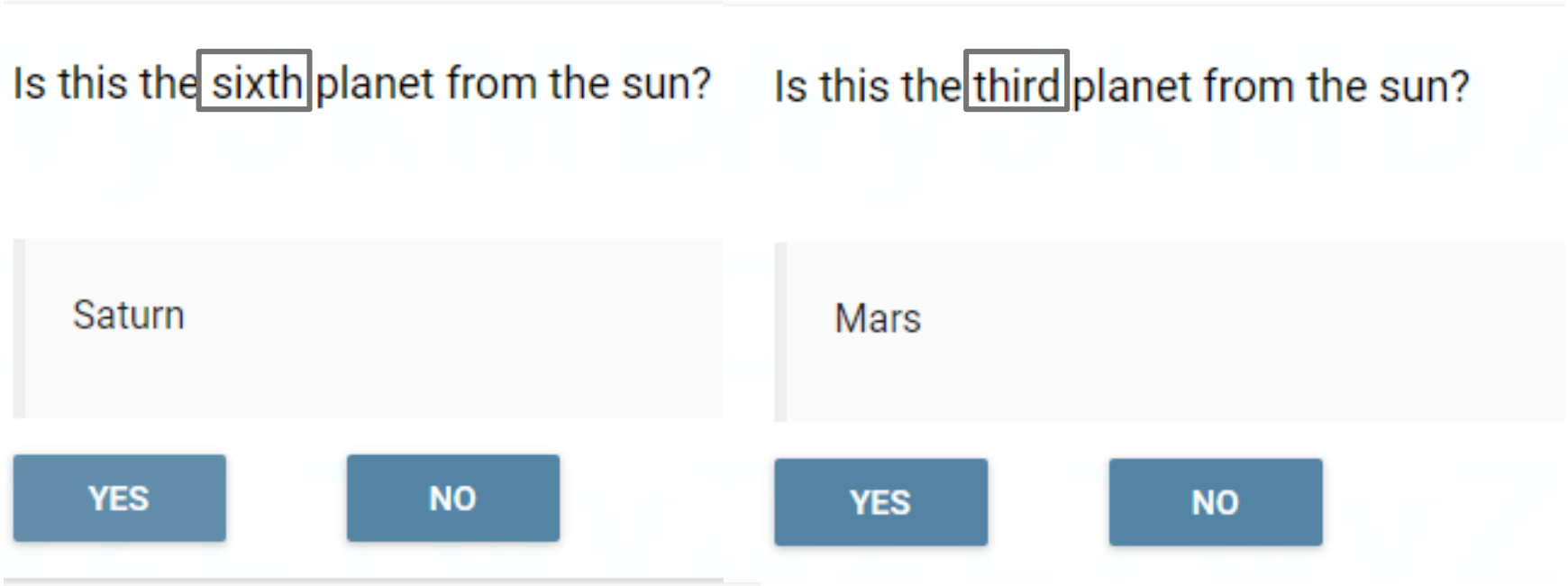
Objective:
Know the order of the planets in our solar system from the sun





All possibilities for stem: Which planet is **[first, second, third, fourth, fifth, sixth, seventh, eighth]** from the sun?

Sample SmartItem (DOMC)



Development Process for GoT

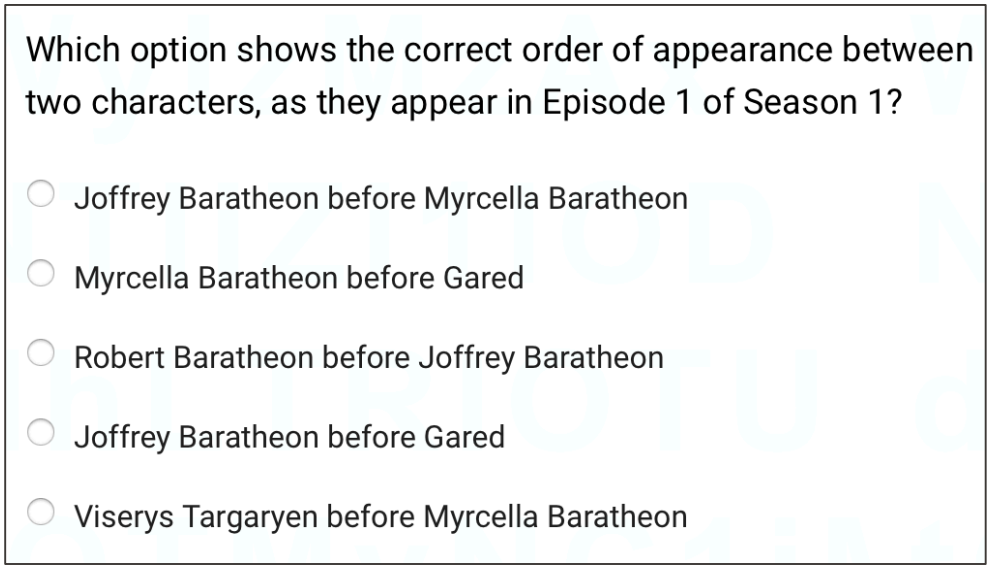
1. Development time: 3 weeks
2. Beta tested items
3. Hard launch

SmartItem and Traditional Item Variability

(Options for MC items were randomized, providing 120 different views of each.)

Item	MC	SmartItem MC	SmartItem DOMC
1C	120	158400	254760
10C	120	480	772
21C	120	1200	1930
2C	120	600	965
11C	120	277200	445830
3C	120	3720960	5984544
4C	120	18720	30108
17C	120	25200	40530
12C	120	1681680	2704702
5C	120	16800	27020
13C	120	360	579
19C	120	840	1351
16C	120	7200	11580
14C	120	600	965
15C	120	480	772
6C	120	4080	6562
7C	120	360	579
8C	120	50400	81060
9C	120	153600	247040
18C	120	480	772
20C	120	480	772

Anatomy of a SmartItem (Item 12)



Objective: Recall the correct order of appearance of characters in Season 1.

It is **important** to note that SmartItems can be designed and programmed in many ways, limited only by the wording of the competency description, creativity of the SME and the skill of the coder.

35 GoT characters available

14,014 different sets of options

1,681,680 different options x option order for MC

2,704,702 different options x option order for DOMC

How to Create a SmartItem

Three key steps:

1. Review the objective
2. Map the item
3. Code the item

How to create a SmartItem

1. Review objective

Objective: Know name, owner, and status of direwolves

Ask yourself: What are *the pieces of knowledge or sub-tasks* that make up this objective? Are there any that are especially critical? Which ones, if any, are NOT relevant to our audience or exam purpose? With this exercise, we are determining what should be covered, what should not be, and why.

SEASON ONE
SPOILERS AHEAD!

How to create a SmartItem

2a. Map the SmartItem: Set up content

All owners, wolves, and wolf statuses:

Jon Snow	Arya	Rickon	Sansa	Robb	Bran
Ghost	Nymeria	Shaggydog	Lady	Grey Wind	Summer
Alive	Unknown	Alive	Deceased	Alive	Alive

How to create a SmartItem

2b. Map the SmartItem: Build scaffolding

Is this the name and status at the conclusion of Season 1 of {{ StarkName's }} direwolf?

How to create a SmartItem

2c. Map the SmartItem: Confirm keys/distractors

Jon Snow	Arya	Rickon	Sansa	Robb	Bran
Ghost	Nymeria	Shaggydog	Lady	Grey Wind	Summer
Alive	Unknown	Alive	Deceased	Alive	Alive

Use incorrect matches as distractors
EX: If stem says “Rickon,” “Shaggydog, Deceased” is incorrect. “Lady, Deceased” is incorrect.

How to create a SmartItem

3. Code the SmartItem

```
{% let StarkName = choice(['Jon Snow\'s', 'Arya\'s', 'Sansa\'s', 'Bran\'s']) %}

{% if StarkName in ['Jon Snow\'s'] %}

{% let options = ['Ghost, Alive', 'Ghost, Deceased', 'Shaggydog, Alive', 'Shaggydog, Deceased', 'Lady, Deceased', 'Lady, Alive', 'Grey Wind, Alive', 'Grey Wind, Deceased', 'Summer, Alive', 'Summer, Deceased', 'Nymeria, Unknown', 'Lady, Unknown'] %}

{% elif StarkName in ['Arya\'s'] %}

{% let options = ['Nymeria, Unknown', 'Shaggydog, Alive', 'Shaggydog, Deceased', 'Lady, Deceased', 'Lady, Unknown', 'Lady, Alive', 'Grey Wind, Deceased', 'Grey Wind, Alive', 'Summer, Deceased', 'Summer, Alive', 'Nymeria, Deceased', 'Ghost, Unknown'] %}

{% elif StarkName in ['Sansa\'s'] %}

{% let options = ['Lady, Deceased', 'Summer, Deceased', 'Shaggydog, Deceased', 'Nymeria, Alive', 'Nymeria, Deceased', 'Shaggydog, Alive', 'Lady, Alive', 'Grey Wind, Deceased', 'Grey Wind, Alive', 'Summer, Alive', 'Nymeria, Unknown', 'Ghost, Unknown'] %}

{% elif StarkName in ['Bran\'s'] %}

{% let options = ['Summer, Alive', 'Lady, Unknown', 'Nymeria, Unknown', 'Shaggydog, Deceased', 'Nymeria, Alive', 'Nymeria, Deceased', 'Shaggydog, Alive', 'Lady, Deceased', 'Lady, Alive', 'Grey Wind, Deceased', 'Grey Wind, Alive', 'Summer, Deceased'] %}

{% endif %}


Is this the name and status at the conclusion of Season 1 of {{ StarkName }} direwolf?
```

What is the name and status at the conclusion of Season 1 of Sansa's dire wolf?

- Summer, Alive
- Lady, Deceased
- Lady, Alive
- Nymeria, Deceased
- Shaggydog, Alive

Psychometric Quality of SmartItems

Research Participants

- Participants (1,156) were recruited from various GoT fan forums, *N* after exclusions =1,031
  - Only first time-takers (>95%) who did not use “help” (>98.5%), finished more than half the items, did not respond extremely quickly, and did not take more time than the time allowed
- 445 Female, 535 Male, 51 Other/Prefer Not to Answer/Missing
- Caucasian (492), Asian (251), Black (54), Hispanic (31), Native American (12), Pacific Islander (4), Mixed/Prefer Not to Answer/Other/Missing (187)

Results- Validity

- Self-reported GoT knowledge is just as related to SmartItem scores as it is to multiple choice scores

GoT Knowledge	*N*	Mean Score	SD of Score
Beginner	160	8.06	2.96
Intermediate	652	9.99	3.05
Expert	219	11.91	2.50

For test scores (21 items)
$r(1,029) = 0.37, p < .001$

For MC scores (7 items)
$r(1,029) = 0.30, p < .001$

For SmartItemDOMC (7 items)
$r(1,029) = 0.21, p < .001$

For SmartItemMC (7 items)
$r(1,029) = 0.29, p < .001$

Results- Form and Demographic Differences

1. SmartItems showed similar relationships between demographics and scores as multiple choice items. Thus, SmartItems did not introduce additional bias for any demographic
2. There were no differences in the three forms based on scores, demographics, reliabilities, or validity coefficients

Mystery Item Formats Exercise

Can you figure out which items are multiple choice and which are SmartItems by looking at the reliability values? What about the item statistics?

Results- Reliability by Item Format

7-Item Subtests by Form	*N*	Predicted Alpha with 70 Items	7-Item Alpha	SEM
Format X (Form A)	360	.48	.08	1.21
Format X (Form B)	319	.81	.29	1.22
Format X (Form C)	354	.85	.36	1.17
Format Y (Form A)	360	.85	.37	1.19
Format Y (Form B)	319	.88	.42	1.16
Format Y (Form C)	354	.74	.23	1.20
Format Z (Form A)	360	.80	.28	1.11
Format Z (Form B)	319	.63	.14	1.20
Format Z (Form C)	354	.88	.43	1.15

Item Formats To Assign
Multiple choice
SmartItemMC
SmartItemDOMC

Results- Reliability by Item Format

7-Item Subtests by Form	*N*	Predicted Alpha with 70 Items	7-Item Alpha	SEM
SmartItemDOMC (Form A)	360	.48	.08	1.21
SmartItemDOMC (Form B)	319	.81	.29	1.22
SmartItemDOMC (Form C)	354	.85	.36	1.17
SmartItemMC (Form A)	360	.85	.37	1.19
SmartItemMC (Form B)	319	.88	.42	1.16
SmartItemMC (Form C)	354	.74	.23	1.20
Multiple choice (Form A)	360	.80	.28	1.11
Multiple choice (Form B)	319	.63	.14	1.20
Multiple choice (Form C)	354	.88	.43	1.15

Item Formats To Assign
Multiple choice
SmartItemMC
SmartItemDOMC

Results- Item Statistics

- Which format is which?

Item Formats To Assign
Multiple choice
SmartItemMC
SmartItemDOMC

	Average *P*-Value (SD)	Average CITC (SD)	Average RT (SD)
Format J	0.48 (.22)	0.19 (.15)	26.10 (11.21)
Format K	0.43 (.18)	0.16 (.13)	25.20 (9.34)
Format L	0.53 (.19)	0.21 (.11)	26.25 (10.36)

Results- Item Statistics

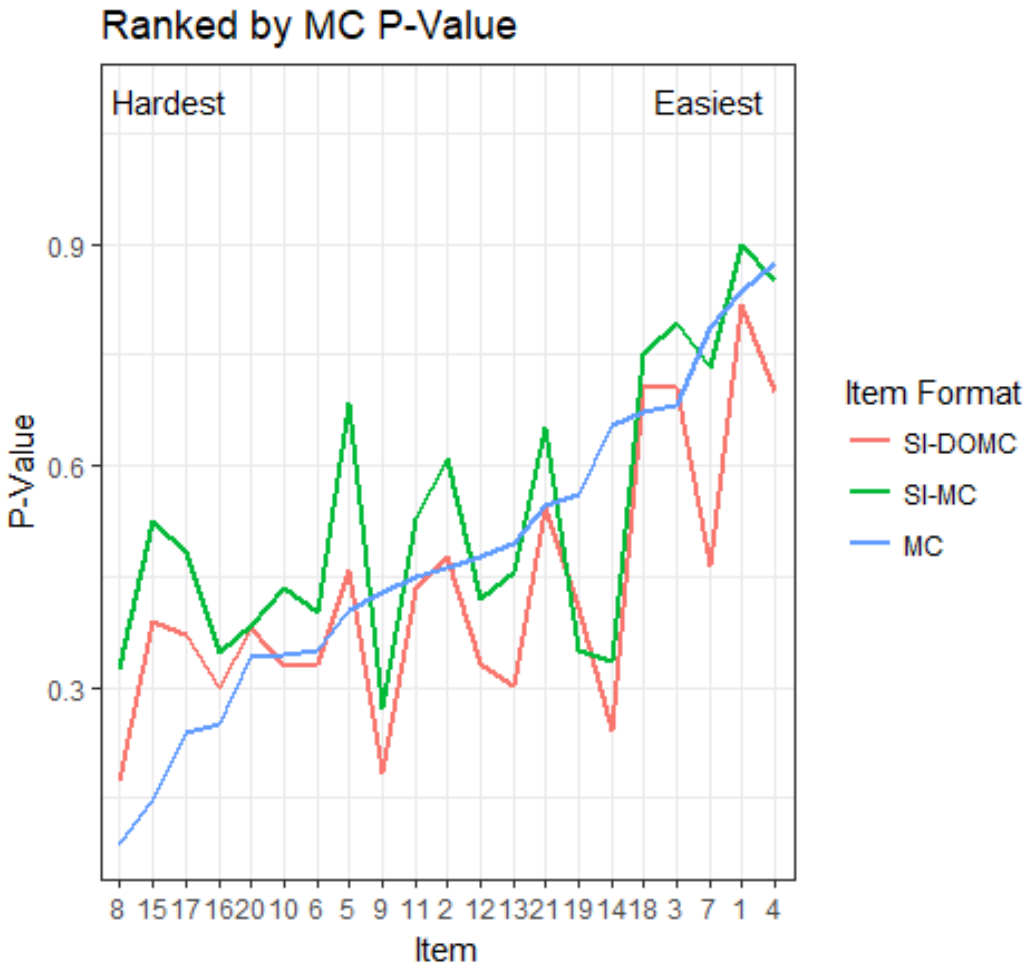
- Which format is which?
- No significant differences

Item Formats To Assign
Multiple choice
SmartItemMC
SmartItemDOMC

	Average *P*-Value (SD)	Average CITC (SD)	Average RT (SD)
Multiple choice	0.48 (.22)	0.19 (.15)	26.10 (11.21)
SmartItemDOMC	0.43 (.18)	0.16 (.13)	25.20 (9.34)
SmartItemMC	0.53 (.19)	0.21 (.11)	26.25 (10.36)

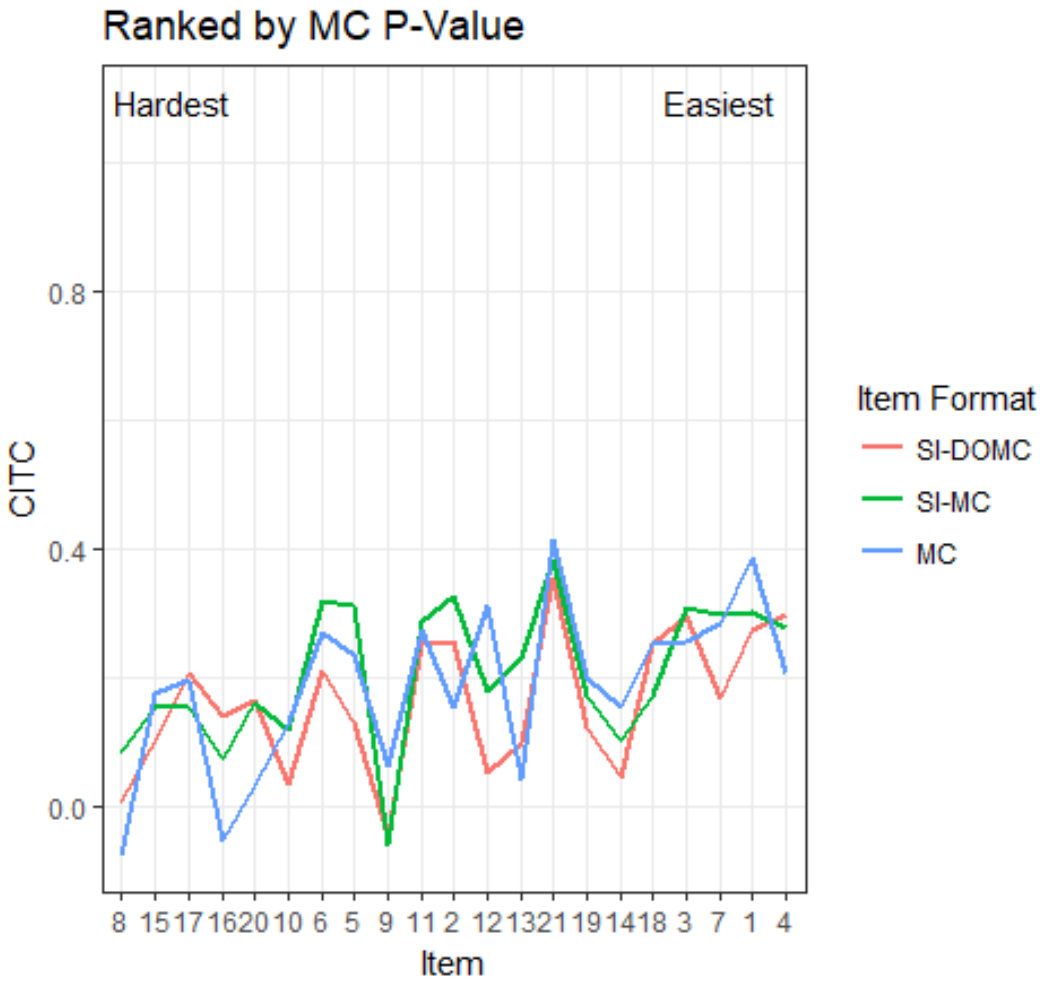
Plots of Item Statistics by Item Format and Item: **P-values**

Inter-Format Correlations
MC and SI-MC: .68
MC and SI-DOMC: .70
SI-MC and SI-DOMC: .93



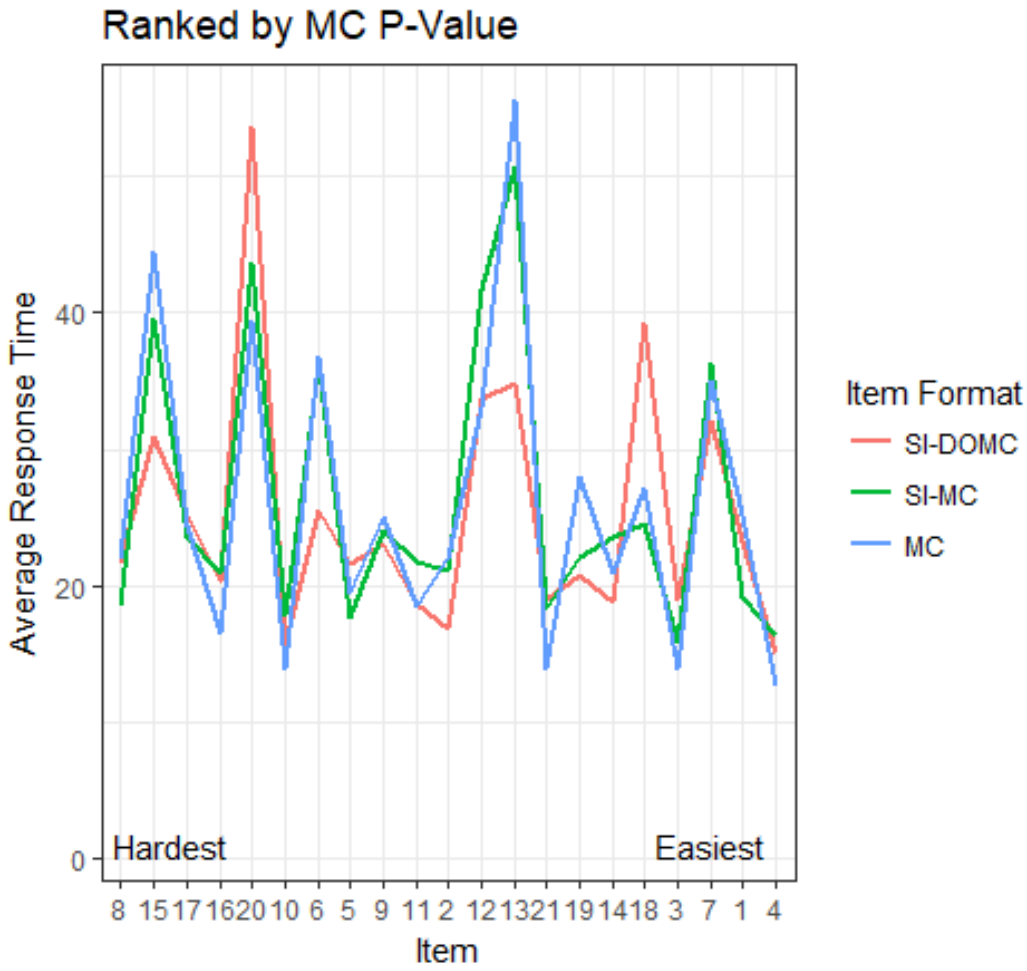
Plots of Item Statistics by Item Format and Item: **CITCs**

Inter-Format Correlations
MC and SI-MC: .69
MC and SI-DOMC: .60
SI-MC and SI-DOMC: .79



Plots of Item Statistics by Item Format and Item: **Average RTs**

Inter-Format Correlations
MC and SI-MC: .93
MC and SI-DOMC: .72
SI-MC and SI-DOMC: .77



Response Time Differences

- For MC
  - Faster for correct
- For SmartItemMC
  - Faster for correct
- For SmartItemDOMC
  - Faster for incorrect
    - On average
    - Medians are similar

Item score	*n*	Mean	SD	Med
0	3,731	27.06	28.57	21.00
1	3,474	25.09	29.43	18.80

Item score	*n*	Mean	SD	Median
0	3,349	28.20	38.74	21.54
1	3,861	24.89	33.48	18.63

Item score	*n*	Mean	SD	Median
0	4,127	24.12	22.39	19.11
1	3,076	26.02	23.15	19.92

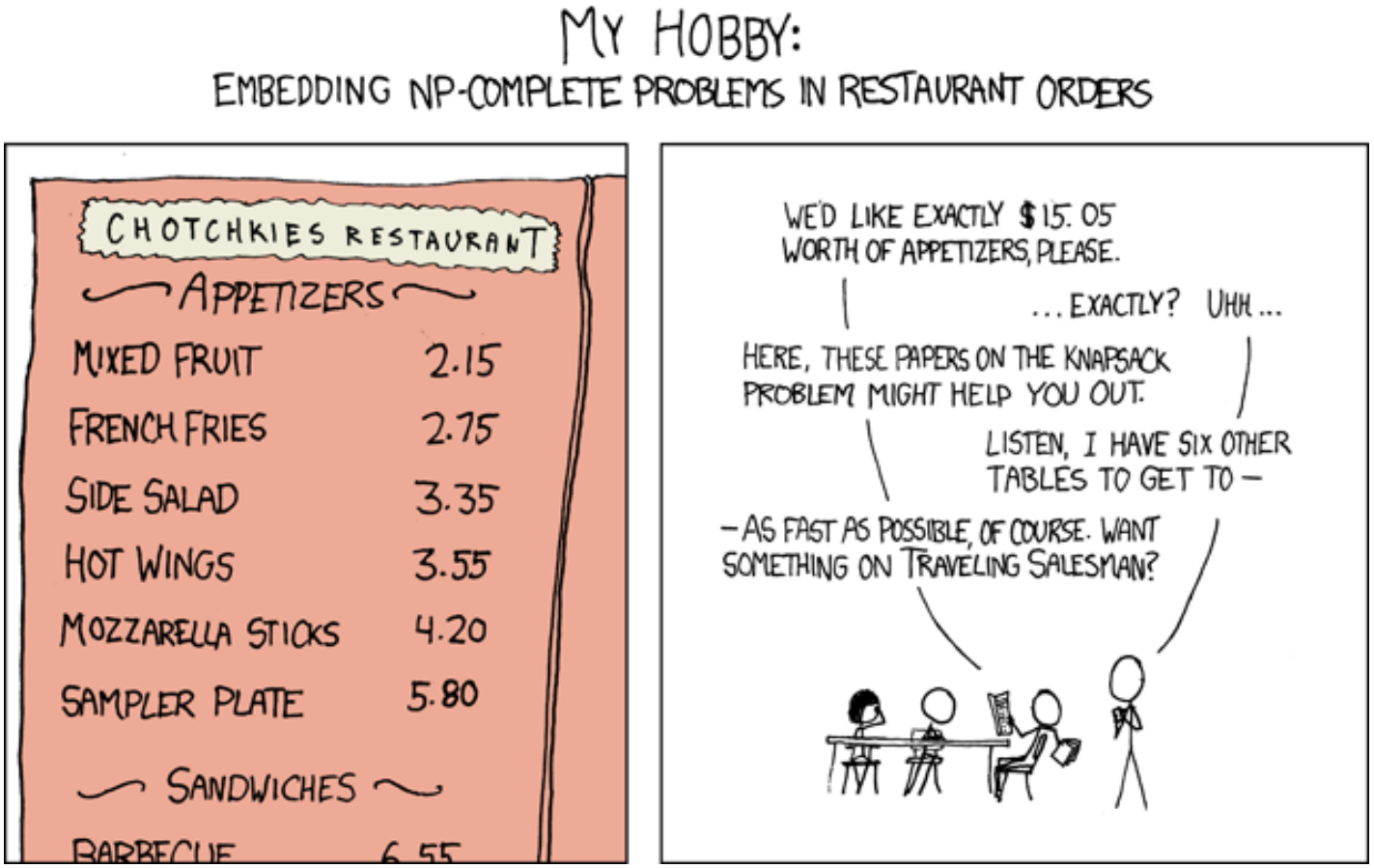
Simulations on SmartItems

Form Building

- With smart items no two forms would look completely identical
- As a psychometrician I want forms to be even.
- I am fine with different forms, as long as they are equivalent

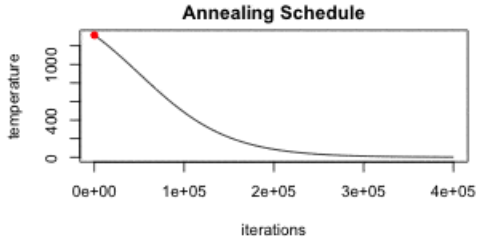
- Building equivalent forms
- Simulation of random form creation

The Traveling Salesman



Traveling Salesman

Distance: 43,499 miles
Temperature: 1,316
Iterations: 0



Traveling Salesman and Psychometrics

- What does the traveling salesman have to do with testing?
- Superficially, not much
- The traveling salesman is what is called an NP-hard problem
  - People do not know how to “solve” it without testing every possible solution
- Form Building is also an NP-hard problem
  - Choose a selection of items so that forms are equivalent
- What does equivalent mean?
  - Cronbach: Equal average p-value, equal average standard deviation, same sampling of content area, item correlations
  - Implied equivalent factor structure
  - Similar response times? Similar option difficulties? Similar word count?

Current Form Building

Matching	
Difficulty	Most of the time
Variance of difficulty	Maybe
Discrimination	Maybe
Variance of Discrimation	Probably not
Content Area	Most of the time
Difficulty*content area	Probably not
Discrimnation*content area	Probably not
Response Times	No
Distractors	No
….	No

Measurement Invariance

- Generally investigates measurement invariance across forms (or time).
  - Similar Fit indices
  - Metric invariance (similar loadings)
  - Similar intercepts
  - Constrain means and intercepts to be equal
  - Constrain residual variances to be equal across groups

  - It is hard to match equivalent forms on all parameters. Forms are never truly equal. Random forms may not be that bad.
  - But at least effort is put into building equivalent forms and they are not constructed randomly.

caveon™

Construction of Random Forms

- What happens when we construct forms randomly instead of trying to match?
- Perfect opportunity for a simulation

Data Generation

- Generated an item bank of 10,000 items using the 2 parameter logistic model:
  - Discrimination mean: 1 standard deviation .1
  - Difficulty mean:0 standard deviation 1
- Generated 10,000 examinee ability parameters
  - Mean: 0 Standard Deviation: 1
- Item parameters and examinee abilities were used to generate response vectors
- Using the item bank, we developed a 40 item fixed-length test using the items above:
  - Average P-value: .501
  - 95% confidence interval: 0.45 to 0.55
  - Reliability: .89

Equivalent Forms

- We talked quite a bit about equivalent forms. For the purpose of the simulation, we are just going to match on equivalent difficulty.
- For the simulation, a random form is equivalent to our 40 item test if the average difficulty of items falls within the 95% confidence interval of the fixed form.
- There are other statistical ways to test equivalency.

Administration

- Gave each of the 10,000 examinees tests of length 10-60

caveon™

Results: Percent Similar

Item Count	Percent Similar
10	.70
15	.81
20	.89
25	.85
30	.84
35	.90
40	.94
45	.98
50	.98
55	.98
60	1.00

Results

- For similar length tests (40 items) only 5% of examinees received a different difficulty exam
- Random forms are similar to smart item forms in that each examinee receives a variation pulled from a item universe (bank)
- How does receiving a different cut score impact cut score classification?

Classification Consistency

		Harder Difficulty	
		No	Yes
Different Classification	No	83.7%	2%
	Yes	14%	.5%

Discussion

- It doesn’t take long for a random form to be equivalent on average
- Only a small percent of people are subjected to a “more difficult” form
- An even smaller percent of people had a classification change based on receiving a harder difficulty exam (0.5%)
- There is a tradeoff

Conclusions

Conclusions

Science: SmartItems are a legitimate way to enhance item formats in order to drastically reduce security risks.

Keeping traditional item formats is very risky today and the level of risk will increase.

Worth Repeating…

Psychometric analysis of items shows similar item statistics for smart items and their MC counterpart

Thank you!

- Contact us with questions at:
  - Chris.Foster@caveon.com
- Check out the Lockbox, which often contains articles about SmartItems:
  - https://www.caveon.com/resources/the-lockbox-test-security-e-zine/