An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment

Benjamin Xie, Matthew J. Davidson,
Min Li, Andrew J. Ko

Information School
UNIVERSITY of WASHINGTON
COLLEGE OF EDUCATION
UNIVERSITY of WASHINGTON

- narrative around main takeaway for 80% of audience (CS educators who make their own items)
- hook: discuss how you make sense of exam scores and how you know you can trust them

Which **wrong** answer does a **high-performing** student choose?

A **low-performing** student?

Why?

Given the function definition for computing the surface area of an object:

```
DEFINE compute (x, y)
    answer = x * x + 2 * x * y
RETURN answer
ENDDEF
```

And the following code statements:

```
a = 3
b = 7

answer = compute(a, b)
```

Which of the following statements is true after the call to `compute(a, b)` has completed execution?

A. The order of function inputs is not important; how the inputs are used is declared by the function definition. e.g., `compute(a, b) == compute(b, a)`

B. ~~`x` and `y` are undefined.~~ **(correct answer)**

C. `x = 3, y = 7`

D. `compute` is called with up to 2 variables.

E. If the value of `x` (inside the compute function) changes, the value of `a` also changes.

tricky question

- mock from SCS1 (Miranda Parker)
- assess function scope
- function which calculates surface area
- code statement
- “which of the following statements is true after the call to compute(a,b) has completed execution"

How to validate our interpretation of test scores?

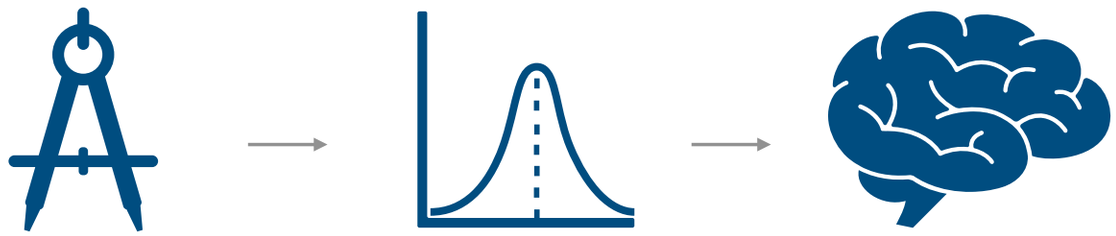
Validity refers to the degree to which <u>evidence and theory support</u> the <u>interpretations of test scores</u> entailed by <u>proposed uses</u> of test scores.

–AERA, APA, and NCME 1999 (from Kane 2009)

standards may be old?

Be skeptical of how you interpret test scores.
Design, evaluate, and iterate towards better assessments.

a psychometric process:
design + data + domain-expertise



understand
test **design**

analyze
response **data**

follow-up w/
domain-experts

how to be skeptical at each step:

- skeptical of how test scores interpreted, used
- skeptical of response patterns (finding anomalies)
- skeptical of causes of problems

understand ARGUMENT for how interpret test scores
assume less and develop evidence-centered design

Argument to justify score interpretation.

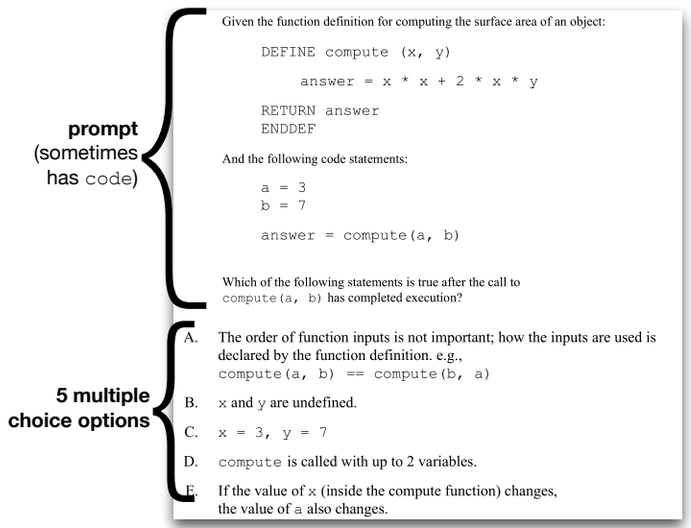
- What does a score of ___ mean?
- Who is the test for?
- How will the scores be used?

Kane, Michael T. 2013. "Validating the Interpretations and Uses of Test Scores."

developing argument for why test score can be interpreted in such a way
READ PAPER
TALK TO US AFTER

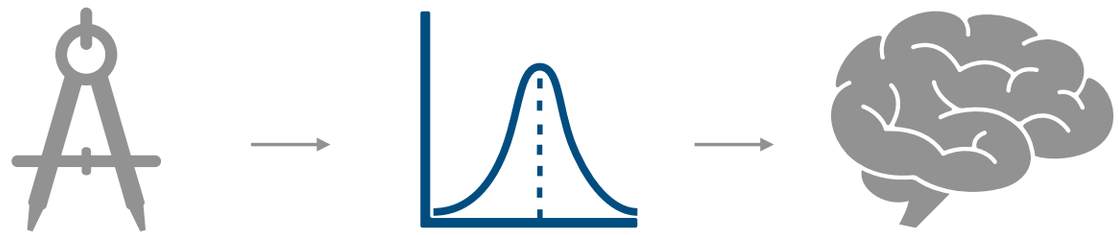
SCS1: Test to measure CS1 knowledge

- measure novice's working knowledge of CS1
- 27 multiple-choice questions
- low-stakes



modified from Parker, Guzdial, & Engleman 2016

- standardized, multiple choice assessment to identify misconceptions, investigate learning, measure student understanding of core concepts
- used by researchers



understand
test **design**

analyze
response **data**

follow-up w/
domain-experts

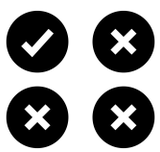
Which response patterns are ***potentially*** problematic?

in essence, we're trying to find if some of the test questions don't work the way we would expect them to based on the test design.

Question properties of interest

difficulty: harder to get question correct

discrimination: distinguish high, low learners



distractors: patterns in wrong answer selection

difficulty: for easier questions, we expect fewer learners to answer correctly
discrimination: how well does a question distinguish between high performing learners and low performing learners? in other words, does a learner's answer to this question tell us a lot or a little about their knowledge?
distractors: which wrong answers are chosen most often?

Data analysis overview

1. classical test theory (**CTT**): first-pass for "descriptive" statistics
2. item response theory (**IRT**): modeling learner, test properties

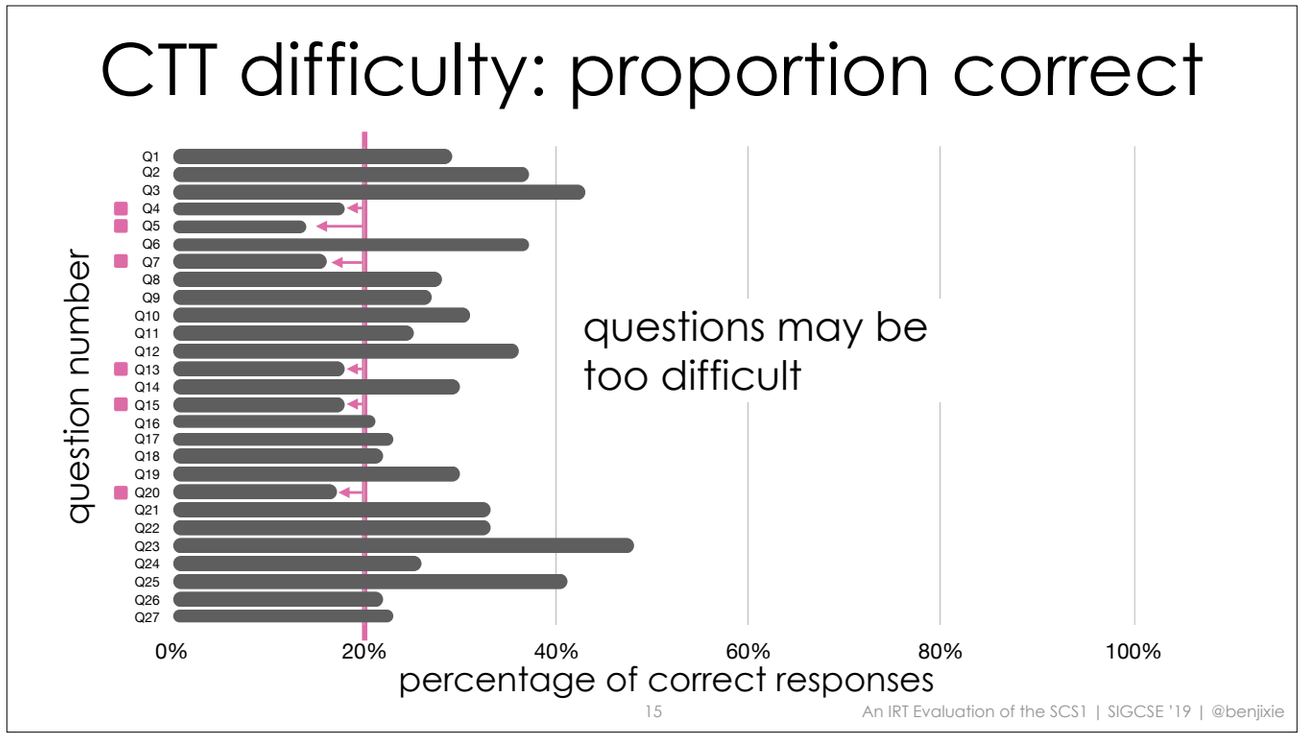
CTT: sample-specific properties
IRT: model to estimate and predict properties about learners and the test

CTT: "*within this sample…*"

- considering total test score
- limitations:
  - sample-dependent statistics
  - score confounds learner, test properties
  - unfalsifiable model (*read paper!*)

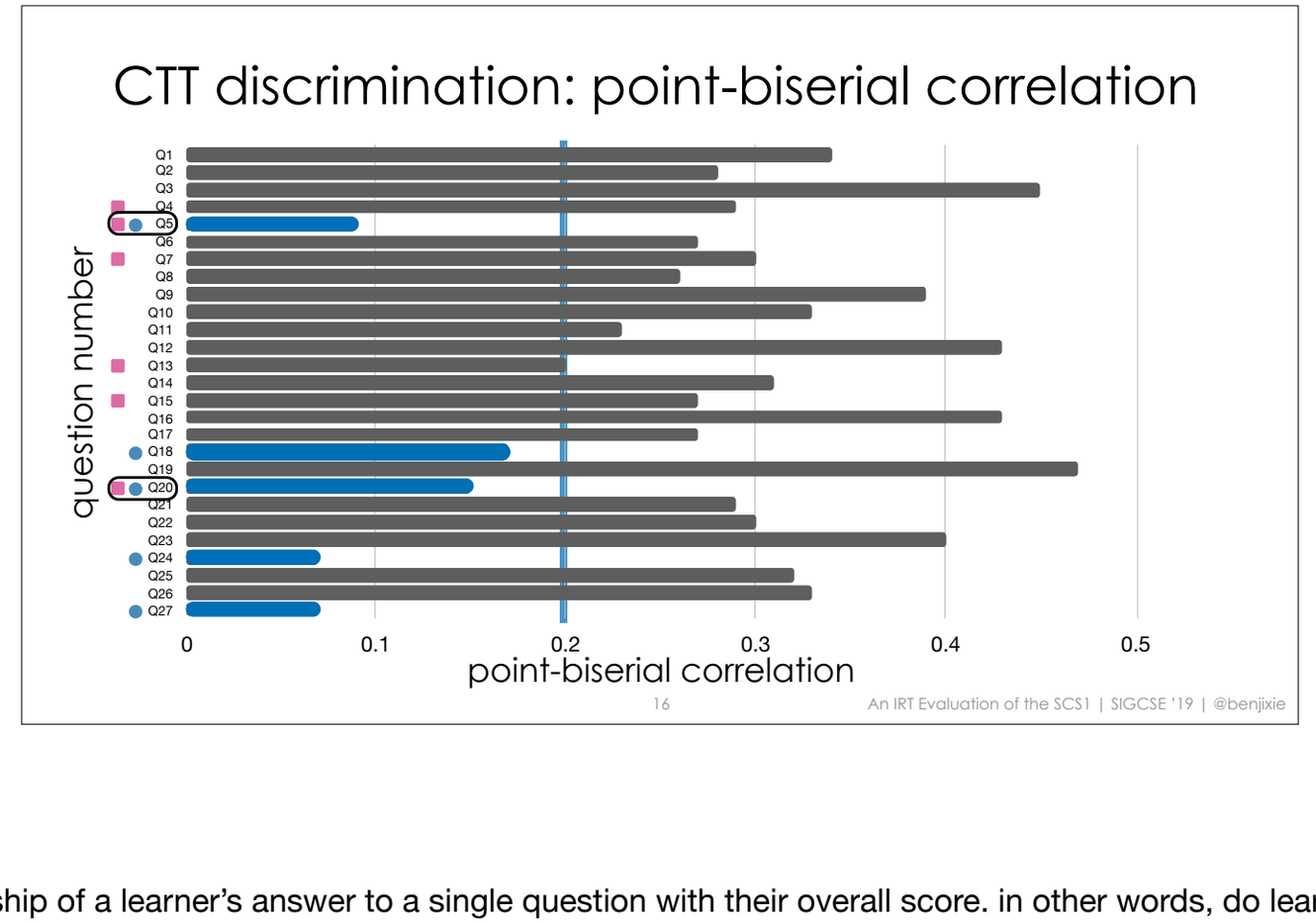
Allen & Yen 2001

grading a test is CTT
looking at aggregate statistics is CTT



questions difficult, but are they *too* difficult for learners?

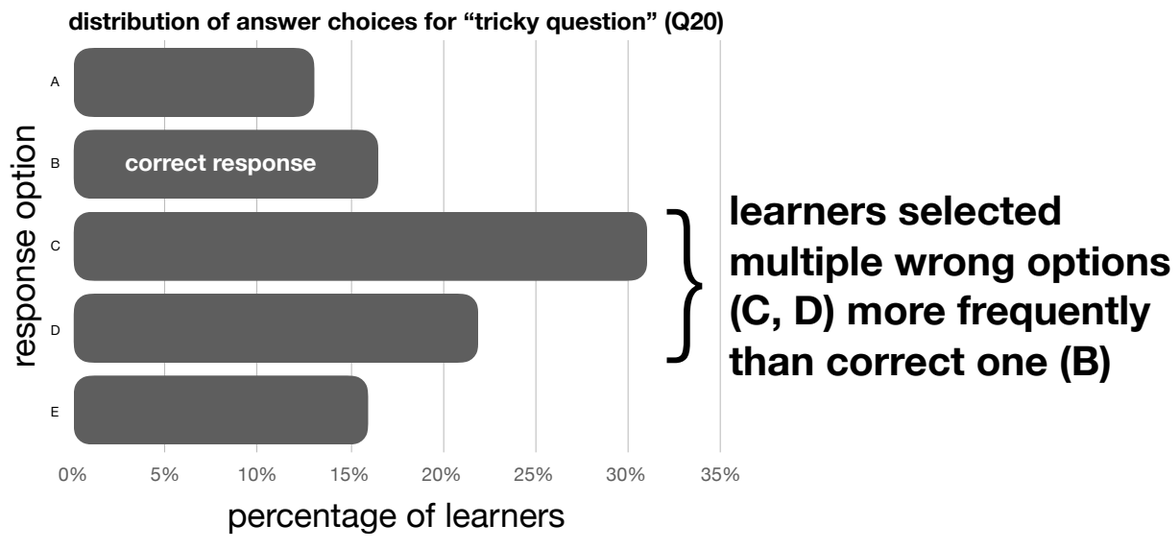
multiple choice test with 5 options, so we could assume guessing would get a 20% correct response rate, so maybe those below 20% are too hard?



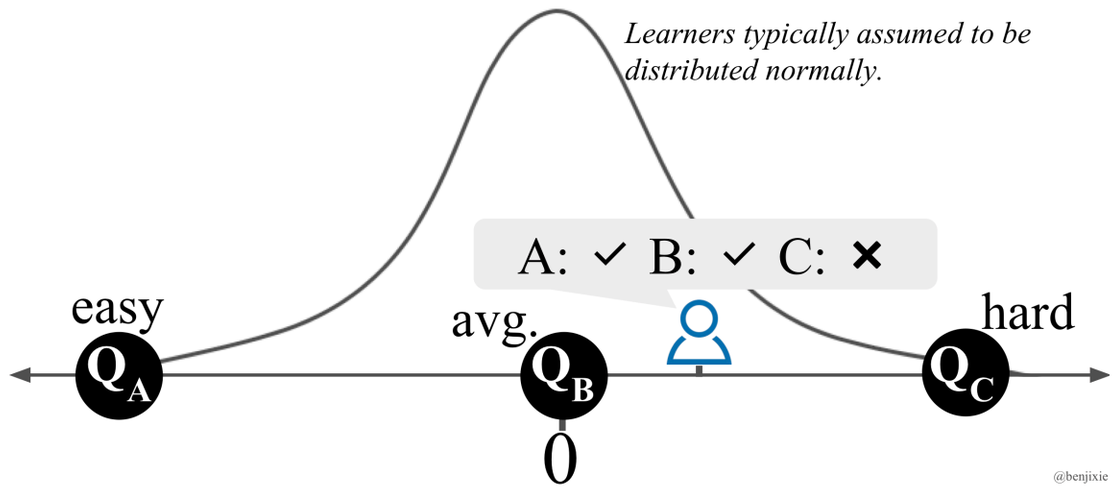
point-biserial correlation is the relationship of a learner's answer to a single question with their overall score. in other words, do learners with high test scores generally answer that question correctly? do learners with low scores generally answer it incorrectly?
higher point-biserial correlation means there was a stronger relationship between the answer and the overall score

CHANGE: LINE, COLORED BARS

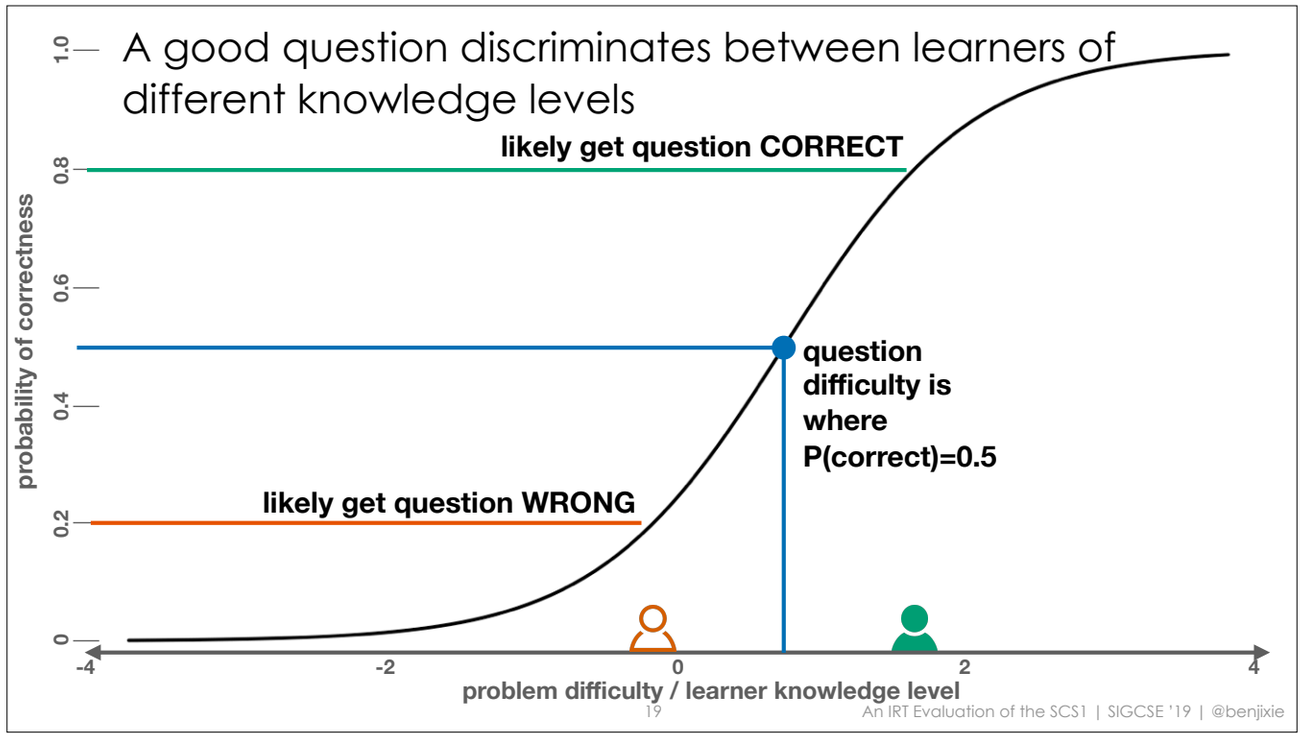
CTT distractors: distribution of answer selection



IRT overview

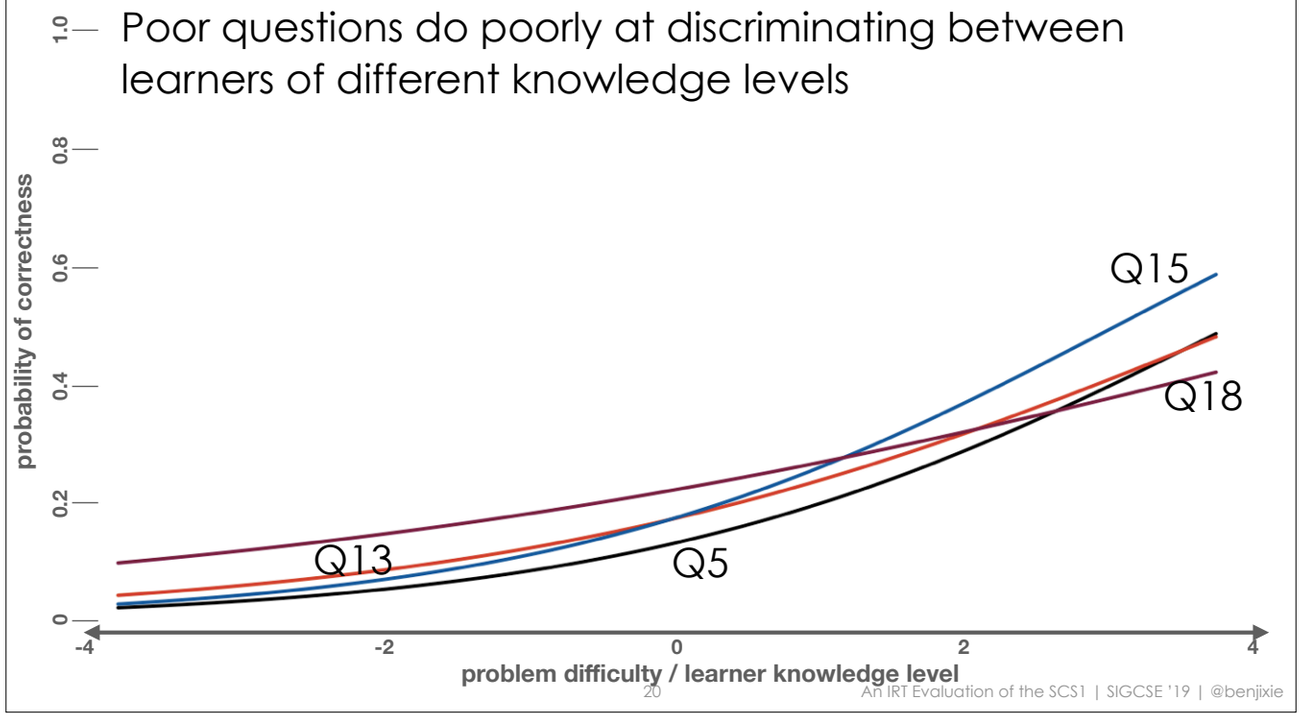


IRT places questions and learners on the same scale. that means we can both estimate a learner's knowledge level AS WELL AS properties of the item that are independent of the sample.

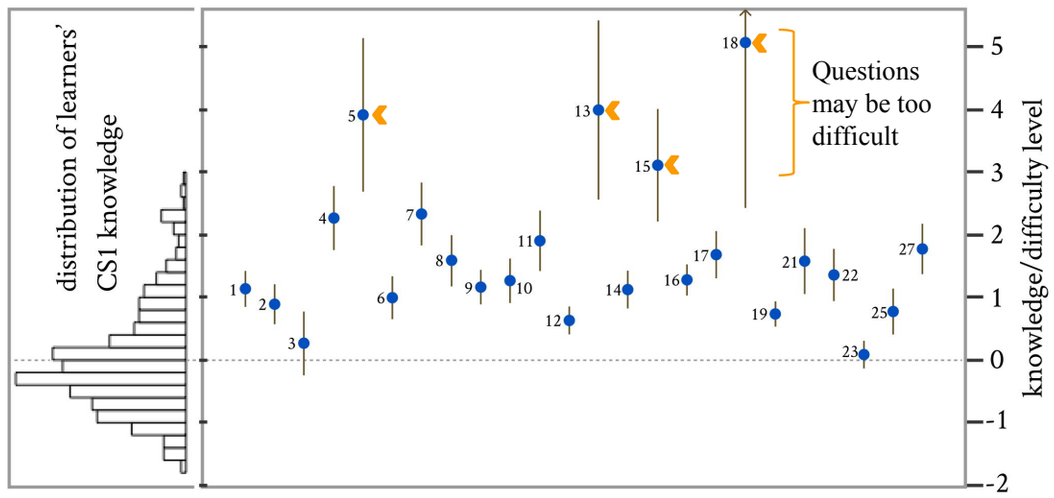


DOUBLE ENCODE LEARNER ICONS

Poor questions do poorly at discriminating between learners of different knowledge levels

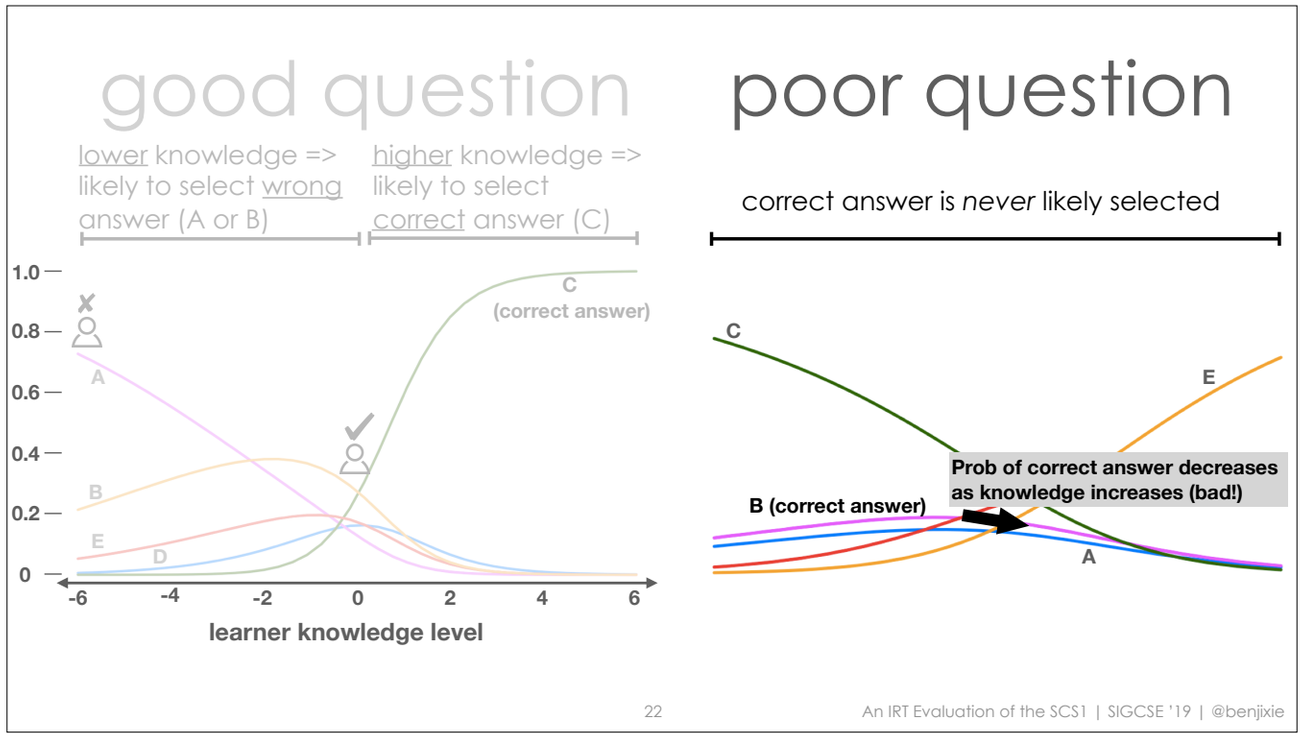


test/learner (mis)alignment



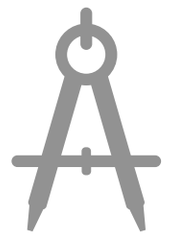
IRT can help us know if questions are too hard. the left is the distribution of learners knowledge—essentially normal, but shifted a little negative. Then we can see how this matches up with the question difficulty, which is on the same scale as learner knowledge. this suggests that a lot of the test questions were too difficult for the students in our sample!

TODO: build this slide

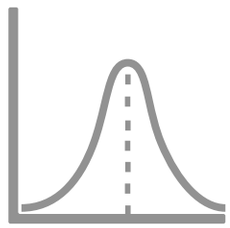


we can see a learner with low knowledge is more likely to select the wrong answer choice A
for a learner with slightly below average knowledge, they are likely to select the wrong answer B
however, for learners with greater than average knowledge, they choose C
for the poor question, (click) we see that learners are less likely to select the correct answer if they have higher knowledge

TRANSITION: this is about as far as we can go with the response data. now, Benji will explain some of the follow-up you can do if you find a problematic item like this.



understand
test **design**



analyze
response **data**

follow-up w/
domain-experts

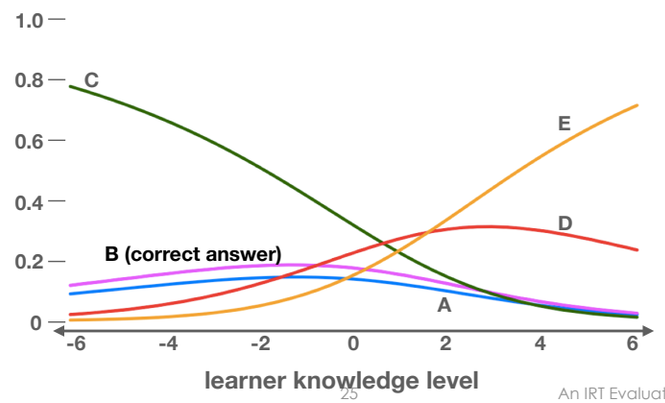
Understanding *potential* problems

follow-up w/
domain-experts

- Why are questions problematic?
- Should questions be revised, removed?

(Why) is “tricky question” (Q20) bad?

- CTT: question may be too difficult, discrimination is low
- IRT: low-performers: C. high-performers: E



Why low-performers selected C?

1. Misconception about scope?
2. Confusing prompt?

Given the function definition for computing the surface area of an object:

```
DEFINE compute (x, y)
    answer = x * x + 2 * x * y
RETURN answer
ENDDEF
```

And the following code statements:

```
a = 3
b = 7

answer = compute(a, b)
```

Which of the following statements is true after the call to `compute(a, b)` has completed execution?

A. The order of function inputs is not important; how the inputs are used is declared by the function definition. e.g., `compute(a, b) == compute(b, a)`

B. `x` and `y` are undefined. ***(correct answer)***

C. `x = 3, y = 7`

D. `compute` is called with up to 2 variables.

E. If the value of `x` (inside the compute function) changes, the value of `a` also changes.

Why high-performers selected E?

1. Misconception about about scope?
2. Prompt was confusing?
3. Confusion in wording of answer?

Given the function definition for computing the surface area of an object:

```
DEFINE compute (x, y)
    answer = x * x + 2 * x * y
RETURN answer
ENDDEF
```

And the following code statements:

```
a = 3
b = 7

answer = compute(a, b)
```

Which of the following statements is true after the call to `compute(a, b)` has completed execution?

A. The order of function inputs is not important; how the inputs are used is declared by the function definition. e.g., `compute(a, b) == compute(b, a)`

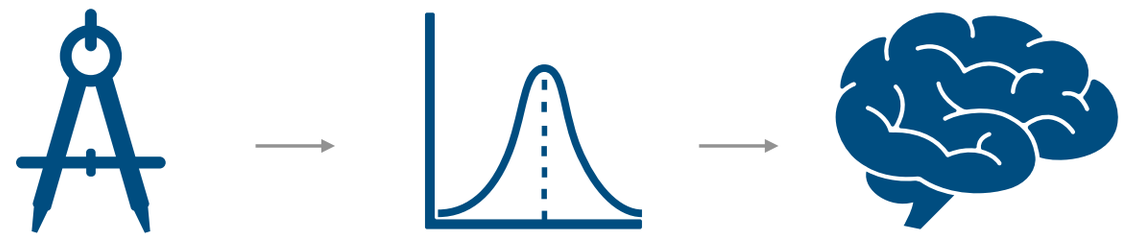
B. `x` and `y` are undefined. ***(correct answer)***

C. `x = 3, y = 7`

D. `compute` is called with up to 2 variables.

E. If the value of `x` (inside the compute function) changes, the value of `a` also changes.

Be skeptical of how your tests measure learning.
Design, evaluate, and iterate towards better assessments.



understand test **design**

analyze response **data**

follow-up w/ **domain**-experts

IRT is part of a process of finding identifying potential problems, and domain-specific follow-up analysis can help you improve your assessments.

Why be skeptical: SCS1 explicitly created to be “good” and there is still room for improvement
How: our way
what should they do: more than “be skeptical” (a bit nihilist). test design, validation is iterative process

An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment

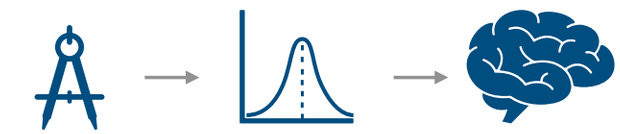
Benjamin Xie
bxie@uw.edu
@benjixie
benjixie.com

Matthew Davidson
mattjd@uw.edu

Min Li
minli@uw.edu

Andrew J. Ko
ajko@uw.edu

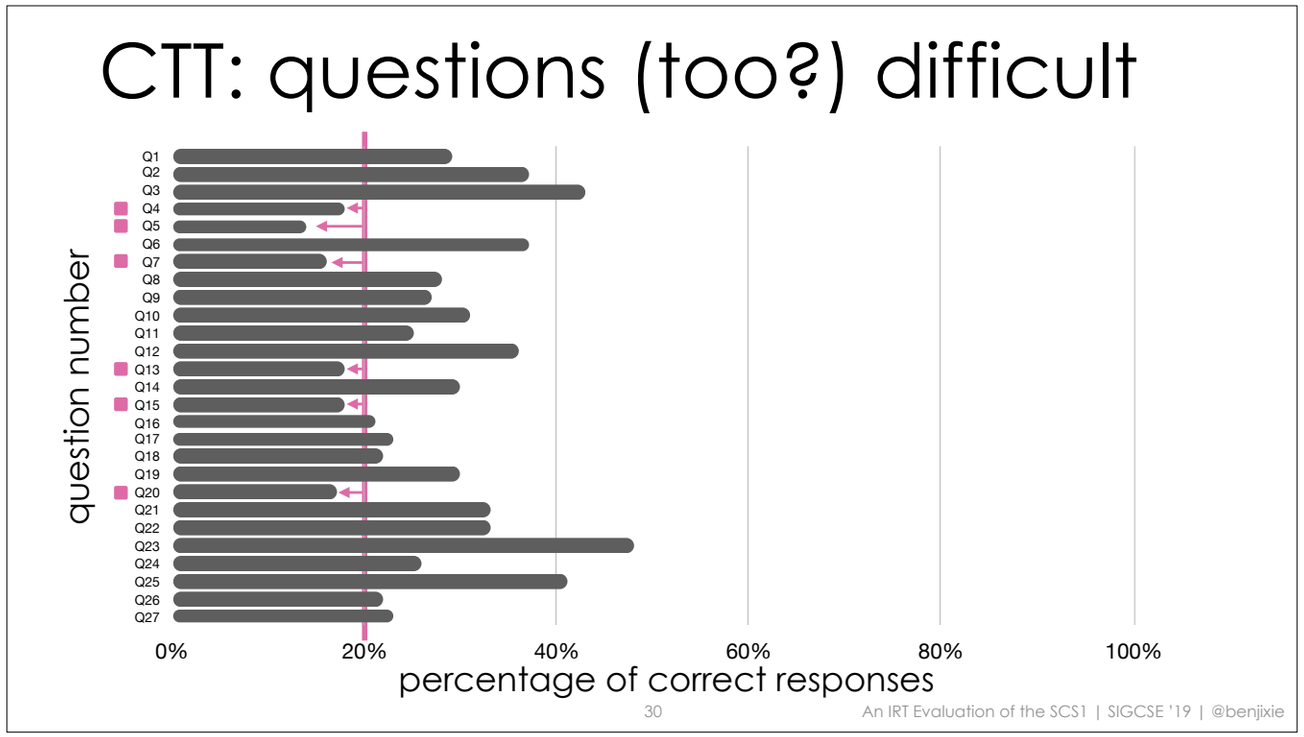
Main takeaway: Be skeptical of how your tests measure learning. Design, evaluate, and iterate towards better assessments.



understand test **design** → analyze response **data** → follow-up w/ **domain**-experts

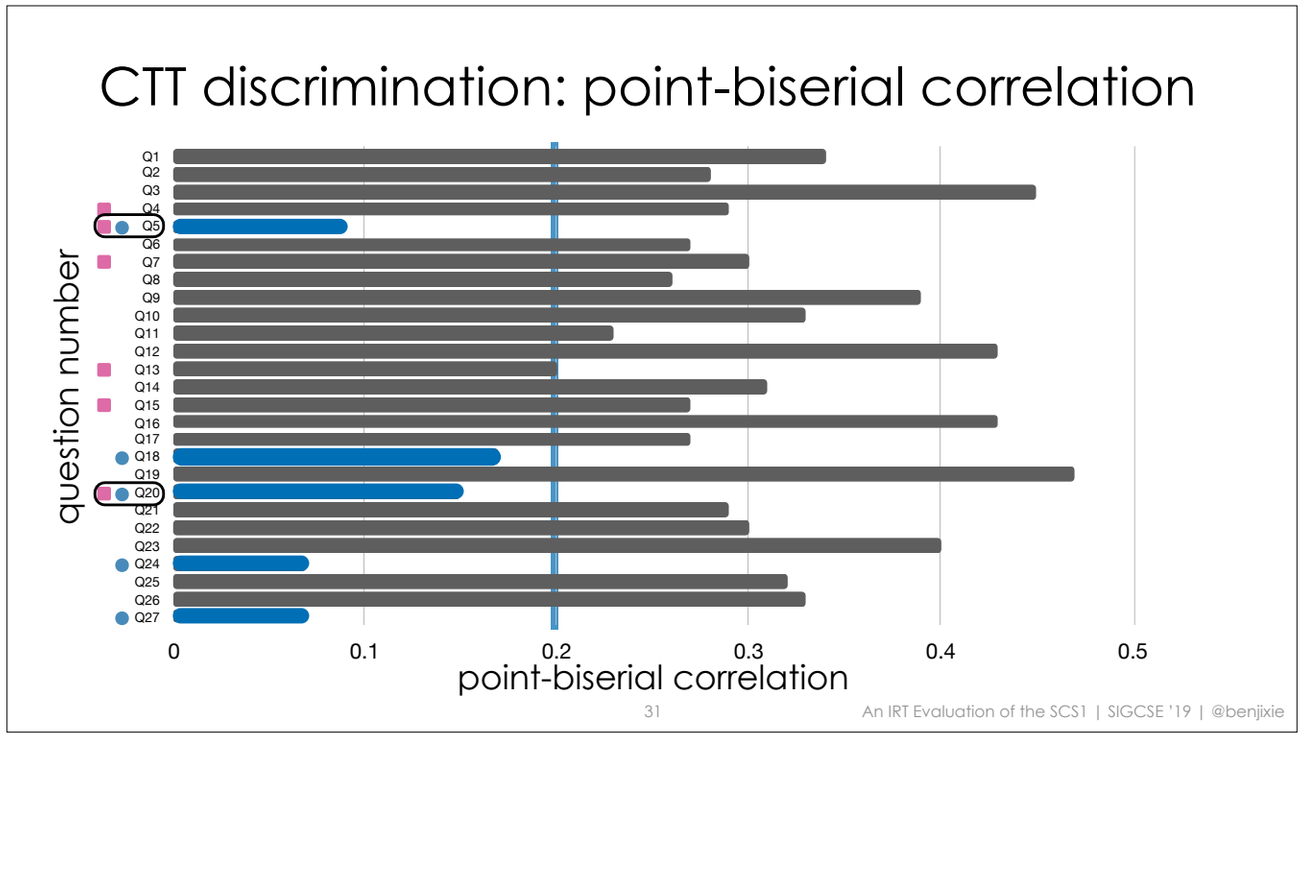
benjixie.com/sigcse2019
blog post w/ more details, slides, paper, supplementary materials

"Dedicated to my dad, who was strong enough to begin his fight with cancer while I attended SIGCSE." -Benji



questions difficult, but are they *too* difficult for learners?

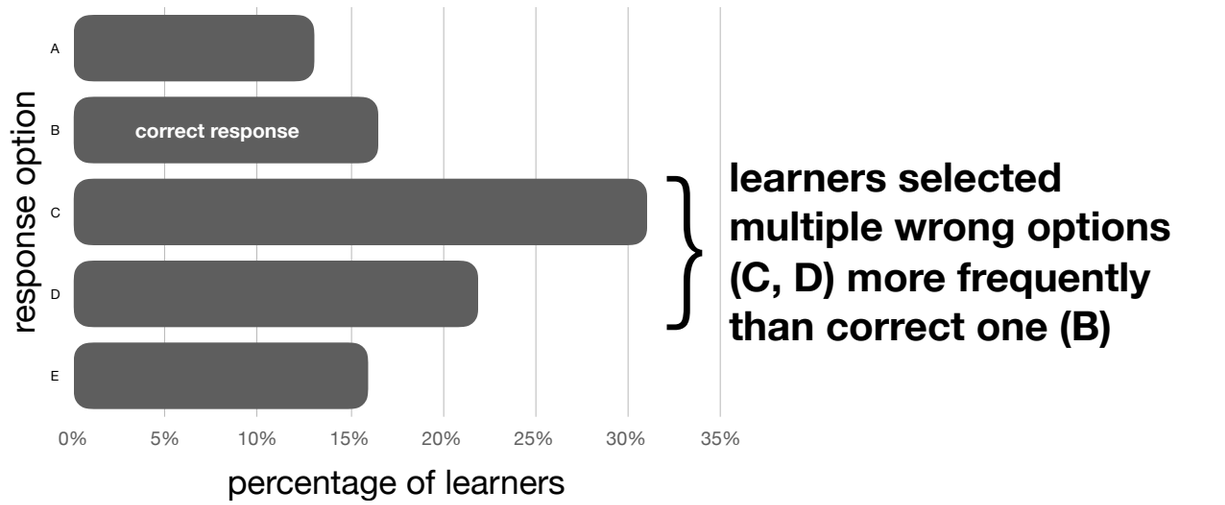
multiple choice test with 5 options, so we could assume guessing would get a 20% correct response rate, so maybe those below 20% are too hard?



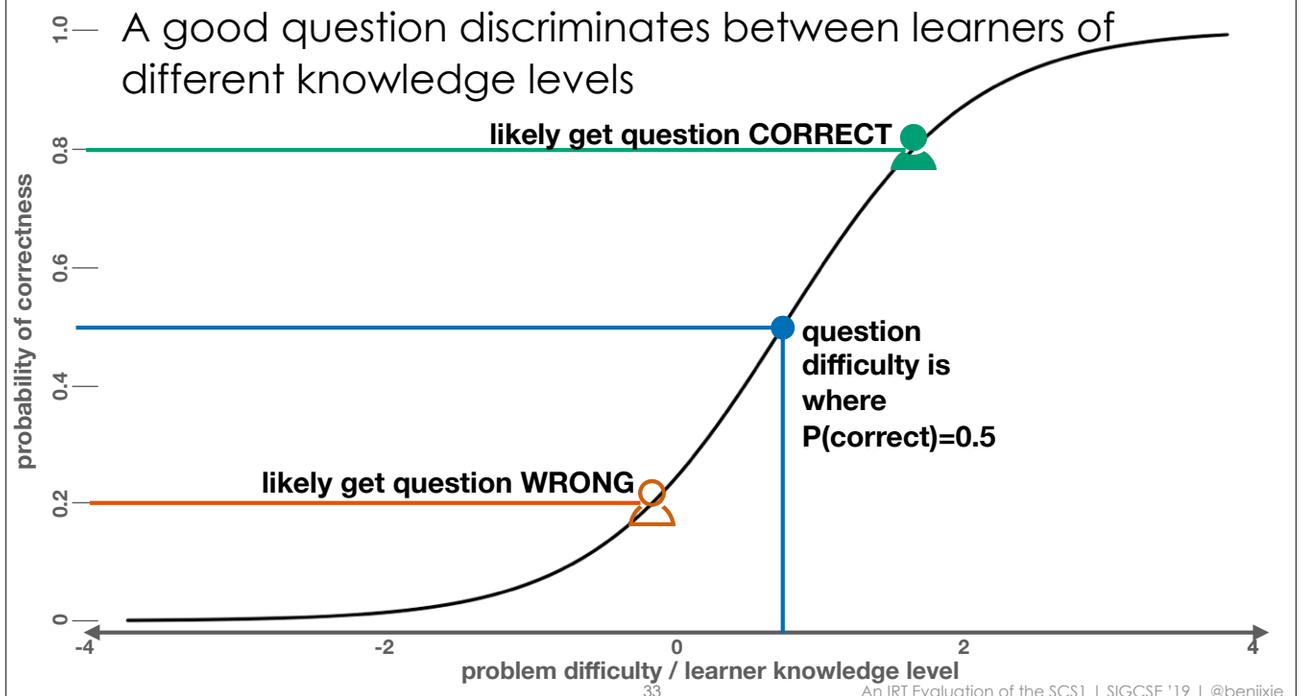
point-biserial correlation is the relationship of a learner's answer to a single question with their overall score. in other words, do learners with high test scores generally answer that question correctly? do learners with low scores generally answer it incorrectly?
higher point-biserial correlation means there was a stronger relationship between the answer and the overall score

CHANGE: LINE, COLORED BARS

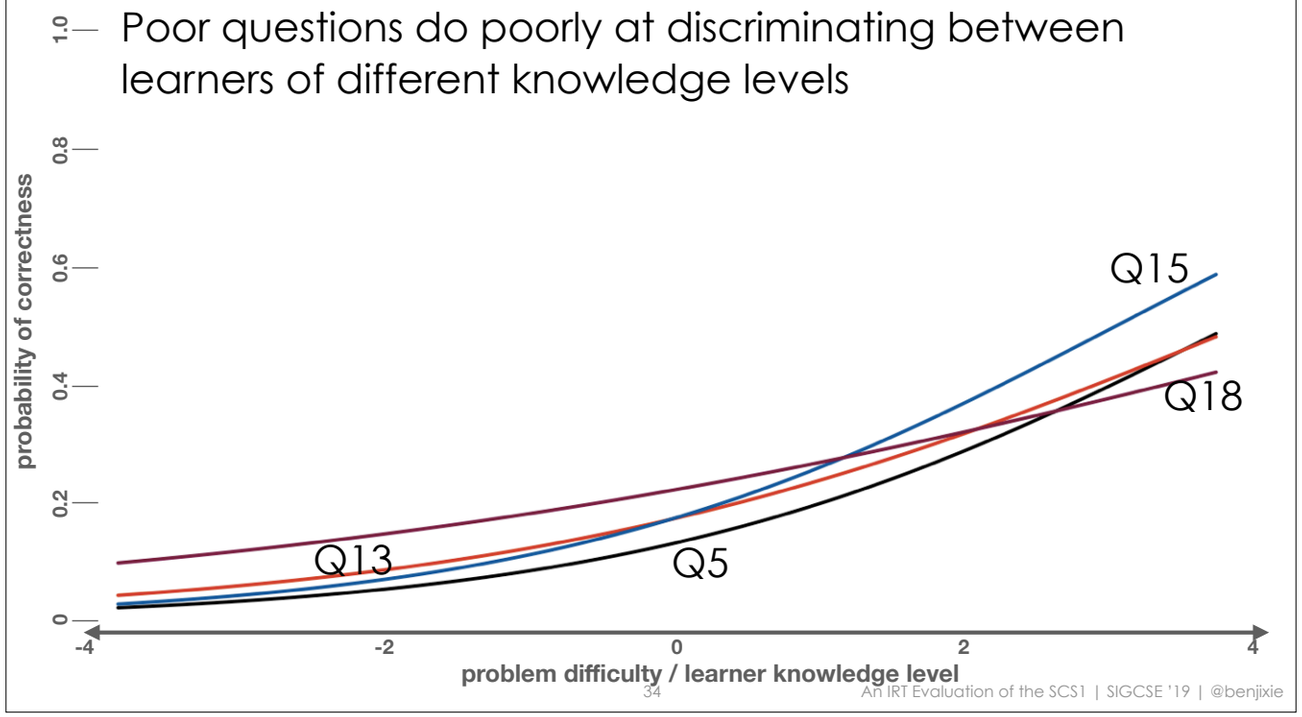
Distribution of answer selection



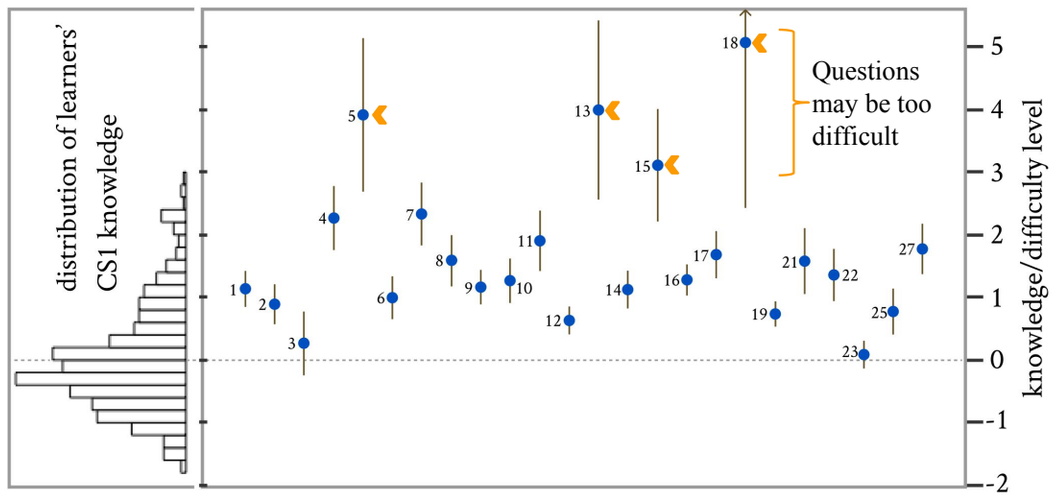
A good question discriminates between learners of different knowledge levels



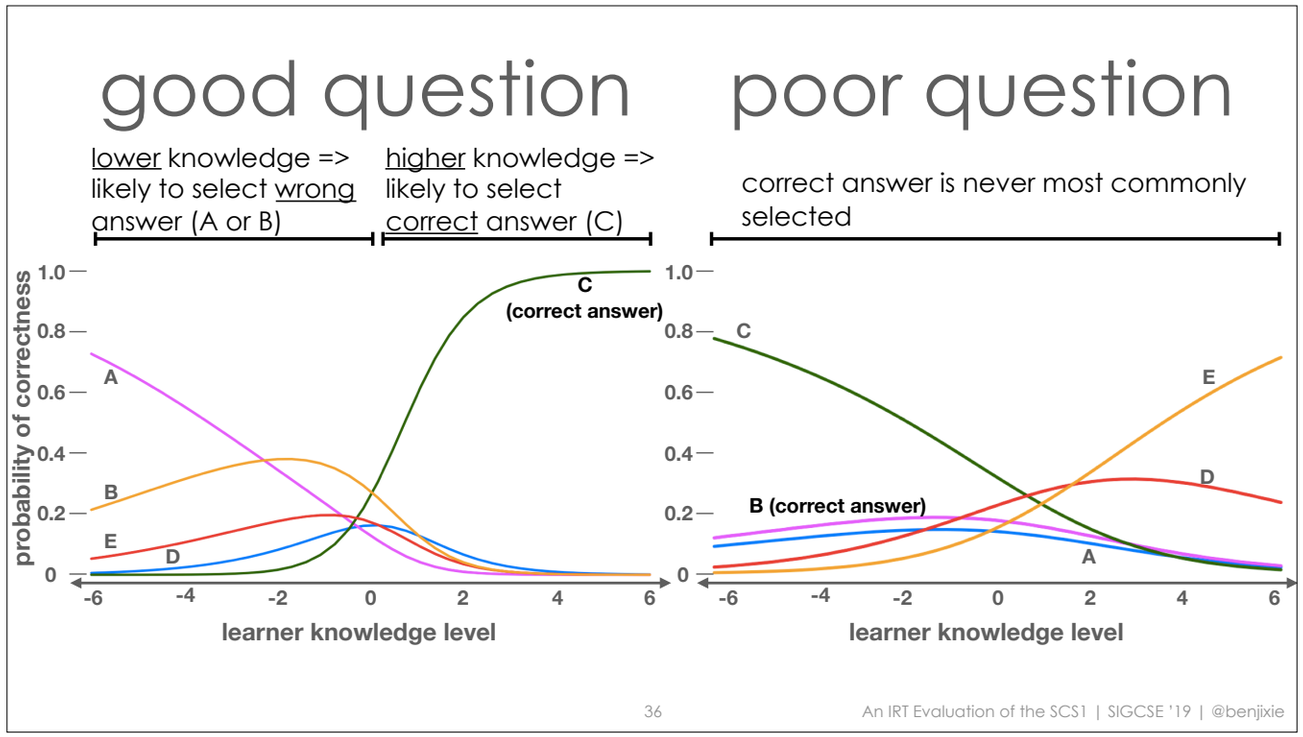
Poor questions do poorly at discriminating between learners of different knowledge levels



test/learner (mis)alignment



IRT can help us know if questions are too hard. the left is the distribution of learners knowledge—essentially normal, but shifted a little negative. Then we can see how this matches up with the question difficulty, which is on the same scale as learner knowledge. this suggests that a lot of the test questions were too difficult for the students in our sample!



this shows us the relationship between a learner's CS1 knowledge and which answer choice they selected. lets us diagnose that something is going wrong with Q20, because the correct answer, B, is less likely to be chosen as ability increases.
By placing learners and question difficulty on the same scale, IRT can help us disentangle specific performance on test questions from properties of the questions themselves.

Potential explanations for odd response patterns:

1. Misconception about about scope?
2. Prompt was confusing?
3. Confusion in wording of answer?

Given the function definition for computing the surface area of an object:

```
DEFINE compute (x, y)
    answer = x * x + 2 * x * y
RETURN answer
ENDDEF
```

And the following code statements:

```
a = 3
b = 7

answer = compute(a, b)
```

Which of the following statements is true after the call to `compute(a, b)` has completed execution?

A. The order of function inputs is not important; how the inputs are used is declared by the function definition. e.g., `compute(a, b) == compute(b, a)`

B. `x` and `y` are undefined. ***(correct answer)***

C. `x = 3, y = 7`

D. `compute` is called with up to 2 variables.

E. If the value of `x` (inside the compute function) changes, the value of `a` also changes.

often selected by low-performers → C

often selected by high-performers → E