Using think-aloud interviews to assess student understanding of statistics concepts

http://www.stat.cmu.edu/teachstat/

New curriculum calls for new assessments: New courses taken by new students

Carnegie Mellon University

~~Statistical Reasoning and Practice~~
→ Reasoning with Data

Now taken by humanities & social sciences, fine arts & creative writing, and statistics & data science students

Colby

~~Introduction to Statistical Methods~~
→ Introduction to Statistics and Data Science

Now taken by math, global studies, biology, and environmental science & policy students

We want to build on prior assessments to evaluate our new curriculum’s success

- Prior assessments include **CAOS**, **SCI**, **AIRS**, **LOCUS**, and **REALI**
- These have been widely and successfully used in classrooms and for educational experiments
- Validated psychometrically for internal consistency, difficulty, …
- We want to build on these to
  - Adapt to our curriculum
  - Avoid expert blind spot
  - Find new misconceptions

Today’s goal: Demonstrate **think-aloud interviews** as a tool to develop assessment questions and **learn how your students learn**.

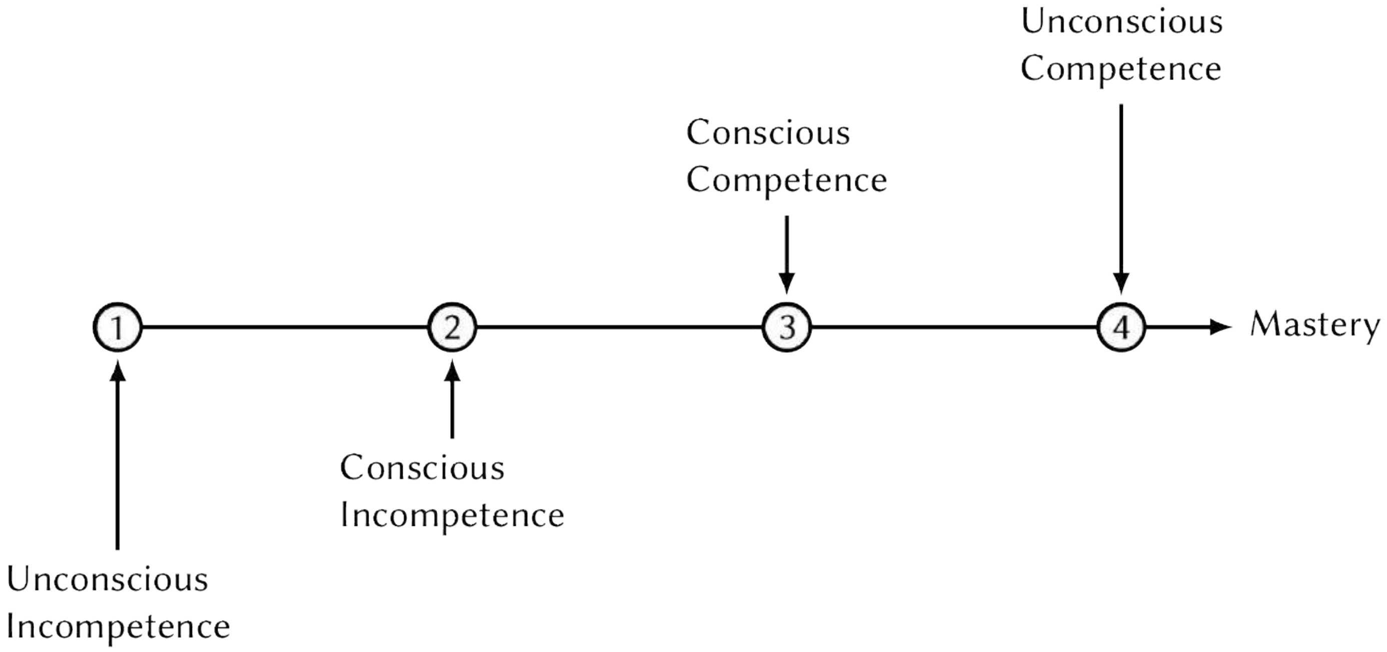
Why are good assessments hard to write?

Draft: Some potentially lucrative, but very uncertain, investments can be made independently. Each has the probability of 0.1 of being a success. As an investment program, a firm invests in 10 of these. Find the probability that the firm gets **at least** one success.

A. 1
B. 0.9
C. $1 - 0.9^{10}$
D. 0.9^{10}
E. 0

(Mosteller, Fienberg, and Rourke, *Beginning Statistics with Data Analysis*, 1983)

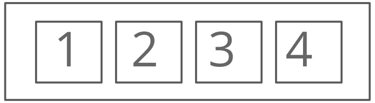
Experts assess student learning, but they don't think like students



(Sprague & Stuart, *The Speaker's Handbook*, 2002)
(Ambrose et al., *How Learning Works: 7 Research-Based Principles for Smart Teaching*, 2010)

An unexpectedly confusing question:

Two draws are made at random from the box containing



1 2 3 4

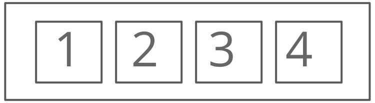
After taking out the first draw, you lose it, and nobody knows what was written on it. You draw a second time. Are the two draws independent?

A. The draws are independent
B. The draws are dependent
C. Not enough information to tell

(Freedman, Pisani, and Purves, *Statistics*, 1978)

A revision:

Two draws are made at random from the box containing



1 2 3 4

After taking out the first draw, a **duck eats it**, and nobody knows what was written on it. You draw a second time. Are the two draws independent?

A. The draws are independent
B. The draws are dependent
C. Not enough information to tell

Think-aloud interviews

- Developed in cognitive science by Ericsson and Simon
- Interviewees read the question aloud and narrate their thinking
- No feedback from the interviewer (verbal reactions, facial expressions, ...)
- Benefits:
  - Better understand clarity of questions
  - Observe why students are getting questions right or wrong and find unexpected misconceptions
  - Hear how confident they are in their responses
- We tested draft questions with think-aloud interviews, then drafted revisions and tested those too

Think-aloud warm-up

“How do you make your favorite kind of toast?”

Think-aloud activity

In pairs, think-aloud the following questions; for each question, take turns playing the role of interviewer and interviewee:

- How many windows are there in your current home?

- Gambler Gary is rolling two dice, hoping for snake-eyes: a roll where both dice show a 1. Gary keeps rolling the dice until he's successful, and on average, it would take 36 rolls to win. After 8 unsuccessful rolls, Gary's friend Fiona joins him to watch him roll the dice. How many rolls should Fiona expect to watch before Gary is successful?
  A. 36
  B. 28
  C. 44
  D. She totally jinxed him and he'll never roll snake-eyes

Some things to reflect on...

As the interviewer, what was your experience like?

Awkward? Tough to avoid feedback?

As the interviewee, what was your experience like?

Also awkward? Tough to explain your thoughts when the answer feels obvious? Could you identify every step you used to reach the answer?

Testing draft questions

In small groups, pick a couple draft questions on the handout to think-aloud in pairs. After completing the interviews, discuss your experience with the group. Is there anything you would change in the questions? How do you think students would respond?

Interviewee: Please remember to think-aloud; don’t pause if you can avoid it, feel free to ask any questions you have about the question (though the interviewer can’t respond), ...

Interviewer: If you feel your interviewee isn’t thinking aloud properly, feel free to say: “Please remember to think aloud,” ...

What we heard in interviews about those questions:

`horse-races` (pg. 5):

- thought that fact that scale broke meant the 5th duck must be very heavy

`study-time` (pg. 2):

- thought the population should be “most normal”
- thought *n* = 5 means histogram contains 5 samples
- thought that more samples should produce a “smoother” curve

`vitamin-c` (pg. 4):

- correlation does not equal causation
- picked the correct answer even though they thought correlation is not causation, because it “made sense”

Lessons from think-alouds

1. Some questions have small issues that cause irrelevant misunderstandings

   Solution: Simple edits

2. Students have multiple misconceptions about a single question

   Solution: Split into multiple parts/write new questions

3. Students sometimes get right answer for wrong reason

   Solution: depends on situation

Irrelevant misunderstanding: `horse-races` (p. 5)

You have 25 ducks. You want to know which duck is the lightest and which is the heaviest, so you begin to weigh the ducks in a random order. However, after you have weighed five ducks, ~~the scale breaks~~.

all the ducks fly away

Multiple misconceptions: `study-time` (pg. 2)

To estimate the average number of daily hours that students study at a large public college, a researcher randomly samples some students, then calculates the average number of daily study hours for the sample.

Pictured below (in scrambled order) are three histograms: One of them represents the population distribution of study hours; the other two are sampling distributions of the mean, one for sample size n = 5, and one for sample size n = 50.

Revised version of `study-time`

Jeri, Steve, and Cosma are conducting surveys of how many hours students study per day at a large public university.

Jeri talks to two hundred students, **one at a time**, and adds each student's answer to her histogram.

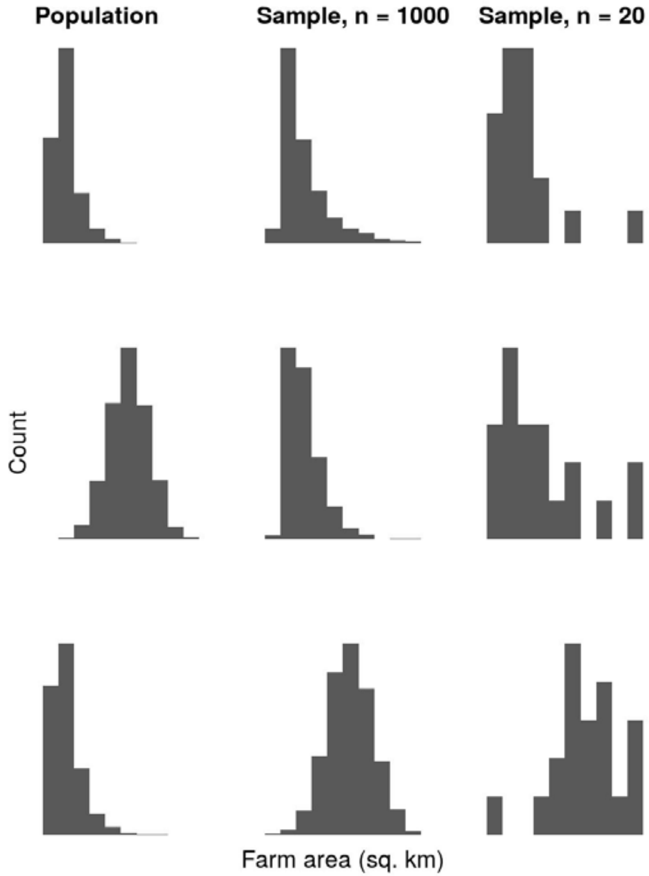
Steve talks to two hundred **groups of 5 students**. After asking each group of 5 students how much they study, Steve takes the **group's average** and adds it to his histogram.

Cosma talks to two hundred **groups of 50 students**. After asking each group of 50 students how much they study, Cosma takes the **group's average** and adds it to his histogram.

New question: `farm-areas`

Farmer Brown collects data on the land area of farms in the US (in square kilometers). By surveying her farming friends, she collects the area of every farm in the US, and she makes a histogram of the population distribution of US farm areas. She then takes two random samples from the population, of sizes *n* = 1000 and *n* = 20, and plots histograms of the values in each sample.

One of the rows below shows her three histograms. Using the **shape** of the histograms, choose the correct row.



Right answer for wrong reasons: `vitamin-c` (p. 4)

A clinical trial randomly assigned subjects to receive either vitamin C or a placebo as a treatment for a cold. The trial found a statistically significant negative correlation between vitamin C dose and the duration of cold symptoms.

Revised version of `vitamin-c`

A clinical trial randomly assigned subjects to either *perform mindfulness meditation* or a placebo relaxation exercise as a treatment for a cold. The trial found a statistically significant negative relationship between *mindfulness meditation* and the duration of cold symptoms.

We used think-alouds to build a question bank

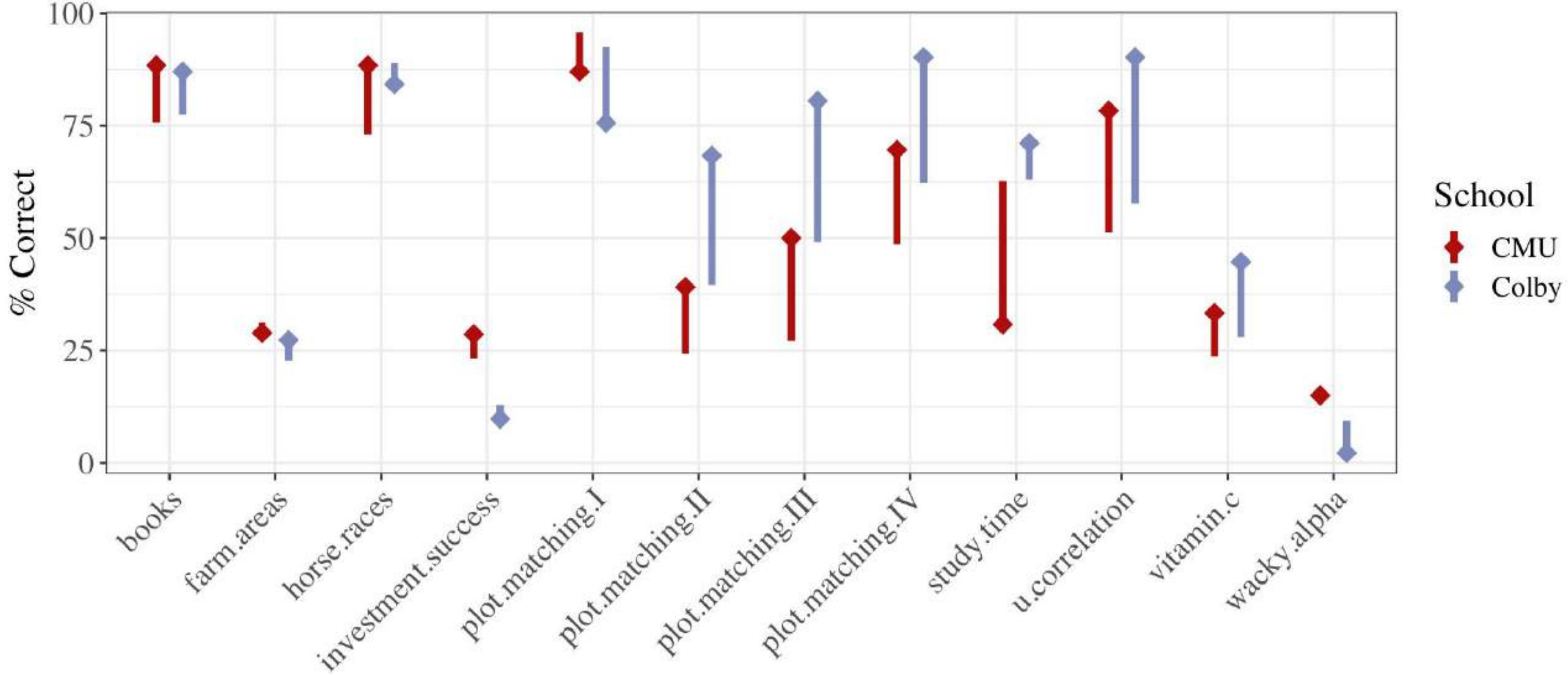
- ~50 original and adapted questions (and growing!), validated through 36 think-aloud interviews
- 18+ questions required revision after think-aloud interviews

EDA (~15)	**Probability (~20)**	**Inference (~15)**
Visual Interpretation	Independence	Correlation/causation
Variation	Conditional Probability	Null Distributions
Estimate Relationship	Probability Rules (Bayes, etc.)	Confidence Intervals
⋮	⋮	⋮

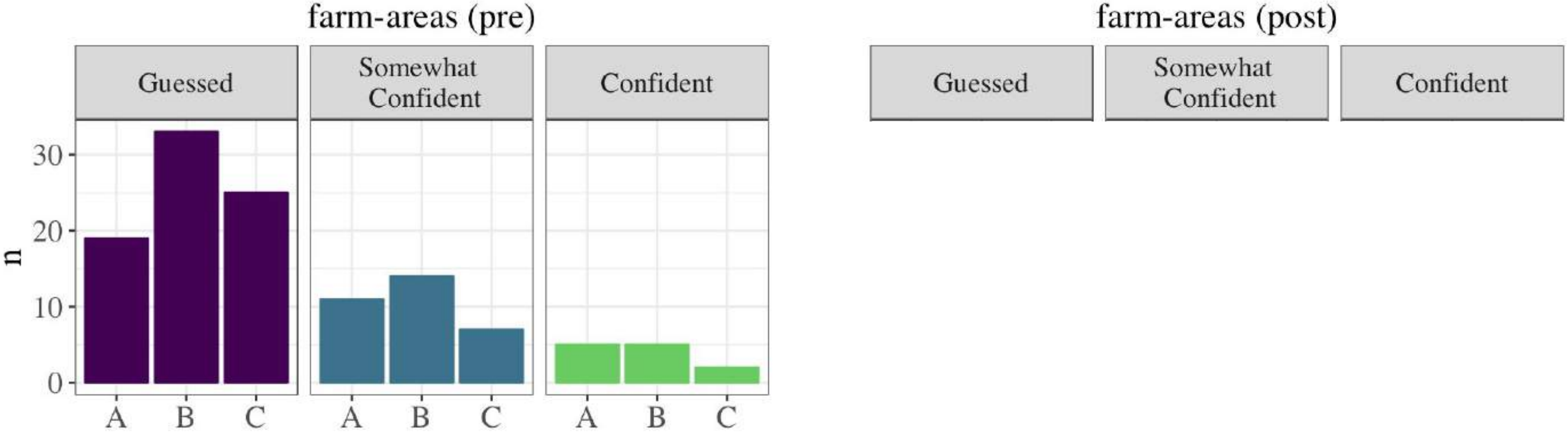
After giving our questions to the entire class:

1. Better sense of what the entire class has learned, but no longer have reasoning behind student answers
2. Confirm or refute that misconceptions found in think-alouds are widespread (`farm-areas`)
3. We can use assessment results to identify additional issues with questions that weren't obvious from think-alouds (`investment-success`)
4. Results also inspire additional questions, which will require additional think-alouds (`vitamin-c and books`)

For the most part, scores for each question **improve** after taking an intro course

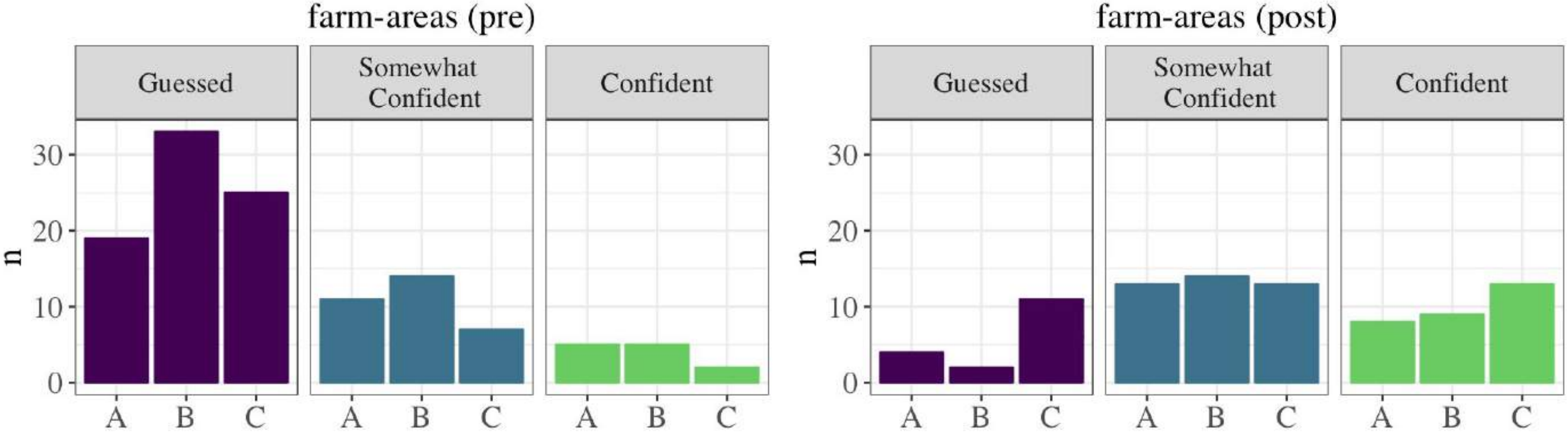


Misconceptions from think-alouds are prevalent



A: Skewed population and samples (correct)
B: Normal population, skewed samples
C: Skewed population, normal samples

Misconceptions from think-alouds are prevalent



A: Skewed population and samples (correct)
B: Normal population, skewed samples
C: Skewed population, normal samples

Using assessment results to improve questions

Draft: Some potentially lucrative, but very uncertain, investments can be made independently. Each has the probability of 0.1 of being a success. As an investment program, a firm invests in 10 of these. Find the probability that the firm gets **at least** one success.

A. 1
B. 0.9
C. $1 - 0.9^{10}$
D. 0.9^{10}
E. 0

(Mosteller, Fienberg, and Rourke, *Beginning Statistics with Data Analysis*, 1983)

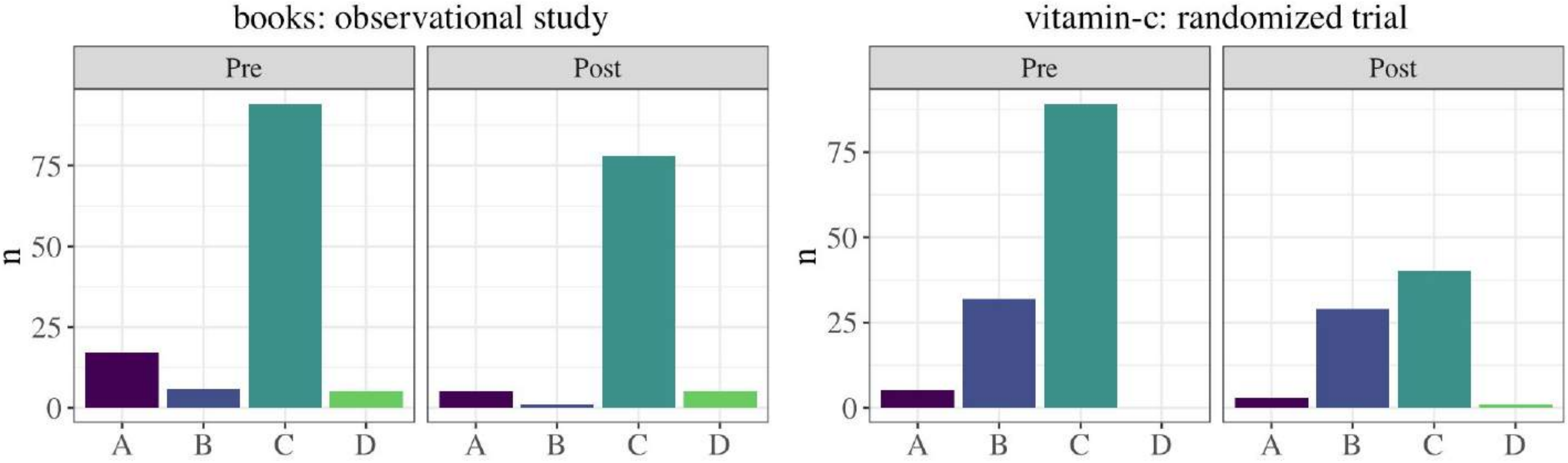
Using assessment results to improve questions

Revision: Some potentially lucrative, but very uncertain, investments can be made independently. Each has the probability of 0.1 of being a success. As an investment program, a firm invests in 8 of these. Find the probability that the firm gets **at least** one success.

A. 8×0.1
B. 0.1^8
C. 0.9^8
D. $1 - 0.9^8$
E. $1 - 0.1^8$

Assessment results inspire new questions

Both `vitamin-c` and `books` (pg. 6) deal with causation; results suggest that students are over-generalizing “correlation ≠ causation” to randomized trials



books: observational study

vitamin-c: randomized trial

C: Can’t draw causal conclusions because correlation ≠ causation

Think-alouds are challenging to implement, but worth the trouble

- Time constraints, interviewer training, getting students to participate, etc.
- Currently preparing a paper on our assessment and its results
- Questions will be available to anyone who asks
  - Looking for research partners to collect data and participate in future studies
- Future directions: using think-alouds to transform an open-ended question into a multiple choice question; sophomore mathematical statistics course
- Think-alouds are useful tools to learn about learning
  - They’re fascinating even if you’re not designing an assessment
  - Think-aloud results can help guide instruction

But wait, there’s more!

We are gathering instructor feedback on our question pool, and would love your input!

https://isle.heinz.cmu.edu/surveys/instructor/

> "Best survey I've ever taken"
>
> — *Alex Reinhart*
>
> "Such a rewarding way to spend 20 minutes"
>
> — *Anonymous*
>
> "This looks fine."
>
> — *Rebecca Nugent*

Backup

Possible assessment names

- KILTS@CMU: Knowledge Investigation for Learning and Teaching Statistics
- QUACKS: Questions for Understanding and Assessing Conceptual Knowledge of Statistics
- SABRE: Statistics Assessment Built for Research in Education
- TREBUCHET: Test to Reliably Establish Baseline Understanding of Concepts like Hypotheses, EDA, and Testing
- IS.TRUE: Introductory Statistics Test for Research in Undergraduate Education
- ...and many more.

Pilot testing was performed in Fall 2018, and pre/post testing was implemented Spring 2019

	Fall 2018	**Spring 2019**

CMU 200	N = 136 (post)	N = 145 (pre)
		N = 82 (post)

CMU 202	N = 95*	N = 121*

Colby 212	N = 115
	(mid-semester)
		N = 89 (pre)
		N = 67 (post)

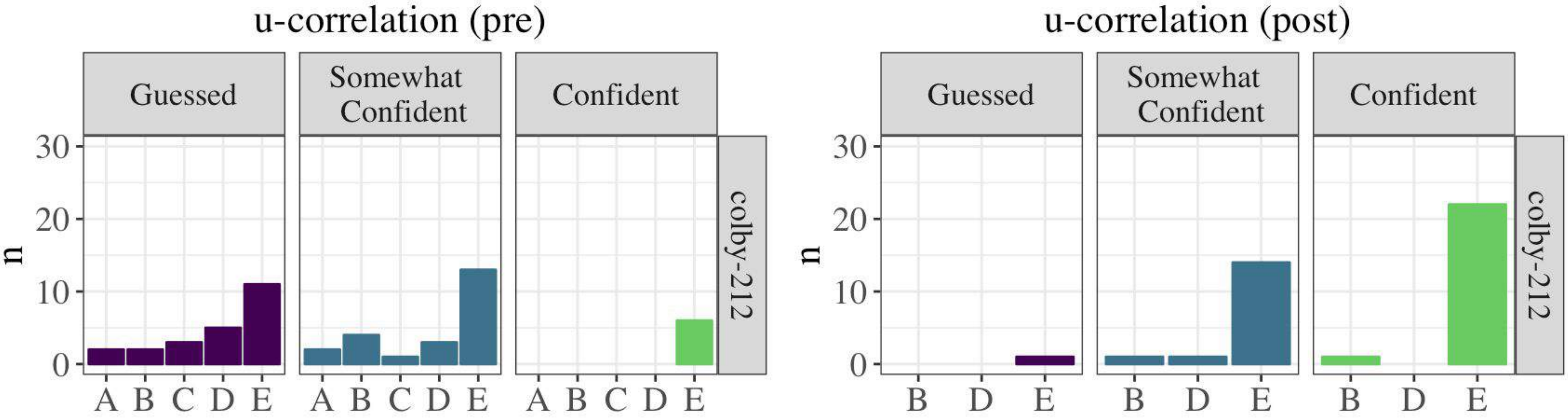
We built an assessment

- ~50 original and adapted questions (and growing!), validated through 36 think-aloud interviews
- Administered online through *ISLE** system
- Students see 30 questions (or 30 minutes, whichever comes first)

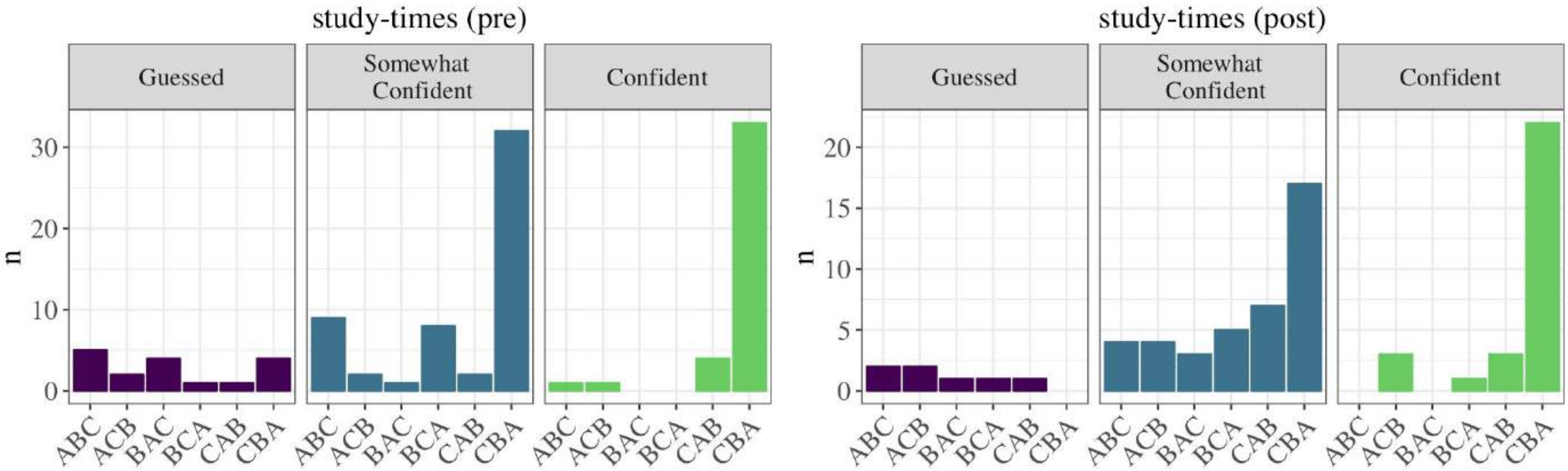
EDA (~15)	Probability (~20)	Inference (~15)
Visual Interpretation	Independence	Correlation/causation
Variation	Conditional Probability	Null Distributions
Estimate Relationship	Probability Rules (Bayes, etc.)	Confidence Intervals
⋮	⋮	⋮

**Integrated Statistics Learning Environment:* http://www.stat.cmu.edu/isle/

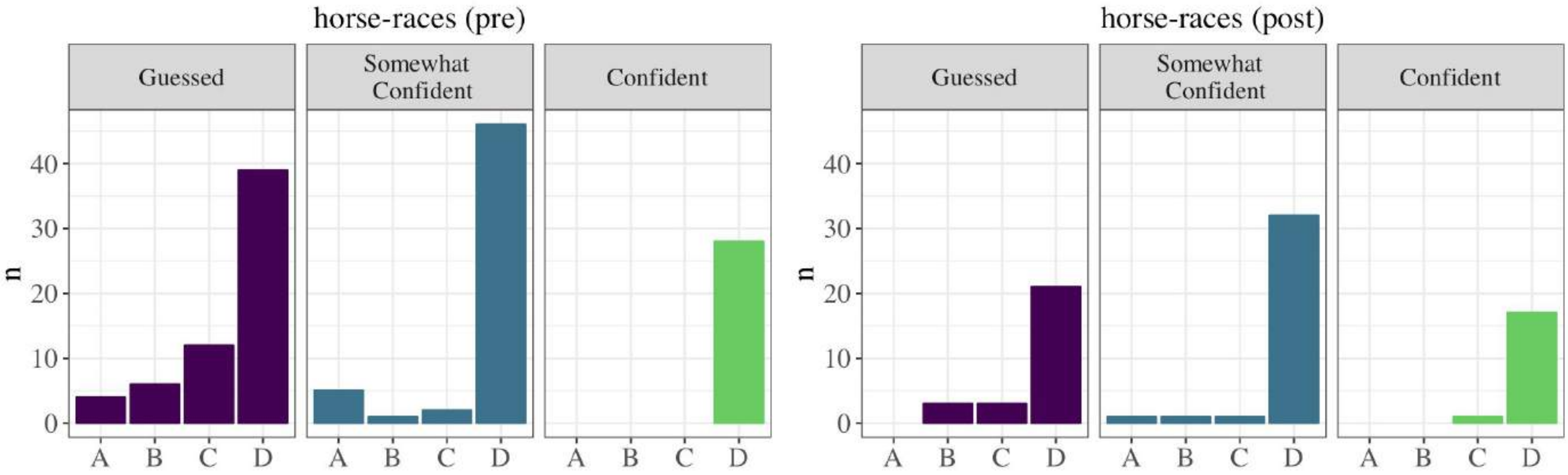
u-correlation



But for some questions, students get **more** confident in **wrong** answers



Students are **not confident** about finding the **correct** answer, even though few get it wrong



Think-alouds bring limitations and challenges

- Time-consuming to gather good data
  - Time to conduct interviews, transcribe detailed notes, revise questions
- Some students are reluctant to think aloud
  - Students are tempted to justify conclusions rather than narrate thought process
- Challenge of standardizing across interviewers
  - Some interviewers ask more probing questions post-think-aloud, details in notes vary…
- Still won’t catch every issue with questions
  - Selection bias in students who volunteer for think-aloud interviews
  - Can further improve questions by analyzing results from administered assessment
- Good way of writing open-ended questions, but these are hard to administer to a large class
  - Might be some misconceptions not captured by focus on multiple-choice questions