Identifying misconceptions of introductory data science students using a think-aloud protocol

Philipp Burckhardt, Peter Elliott, Ciaran Evans, Sangwon Hyun, Kevin Lin, Amanda Luby, Christopher Peter Makris, Josue Orellana, Alex Reinhart, Jerzy Wieczorek, Ronald Yurko, Gordon Weinberg & Rebecca Nugent

Department of Statistics & Data Science
Carnegie Mellon University
Background and motivation

**Goal:** Develop conceptual assessment for intro stats/data science students

**Method:** Think-aloud interviews

- Provide insight into *why* students answered the way they did
- Check whether question tests what we think it tests
- 37(?) interviews conducted during the Spring 2018 semester, in three rounds corresponding to topics covered in class
- Interviews proved invaluable to revise and validate questions
- Think-aloud protocol reviewed by CMU IRB

(For more on think-alouds, see Adams and Wieman, 2011)
Prior concept inventories

- There are several prior concept inventories in statistics, like the Comprehensive Assessment of Outcomes in Statistics (CAOS)
- CAOS was revised with instructor input and validated by a group of 18 expert raters
- All raters agreed that “CAOS measures basic outcomes in statistical literacy and reasoning that are appropriate for a first course in statistics”
- But think-aloud interviews were not used to ensure the questions really elicited the intended student reasoning

(delMas, R., Garfield, J., Ooms, A., & Chance, B., 2007)
Example 2: Sampling distributions

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 $\bar{X}$, one for sample size $n = 5$, and one for sample size $n = 50$.

Prompt: circle the most likely distribution for each description.

(a) Population distribution
A B C

(b) Sampling distribution for $n = 5$
A B C

(c) Sampling distribution for $n = 50$
A B C

All 9 students got this wrong, for a variety of reasons...
Example 2: Sampling distribution question, redux

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.

The three final histograms are shown below, in scrambled order.

Match each person to the histogram they made.
One of the rows below shows her three histograms. Using the shape of the histograms, choose the correct row.

Intended to test a misconception uncovered in think-alouds: students believe all large populations, or large samples, are normally distributed.

“More sample size -- there’s a less chance for data to vary”
What’s next?

- We recommend think-alouds as a way to understand student reasoning and develop assessments, interventions, and activities
- We plan to:
  - Develop a set of questions to assess understanding of introductory concepts
  - Administer test to a large sample of students
  - Compare pre-test and post-test scores to measure learning
- Look out for our survey!
- [www.stat.cmu.edu/teachstat](http://www.stat.cmu.edu/teachstat)
References


Many thanks to our numerous student participants.
REJECT SLIDES
Example 1: Histogram matching

**Prompt:** Match the description to the appropriate histogram:

A. A distribution for a set of quiz scores where the quiz was very easy

B. A distribution for a set of wrist circumferences taken from the right wrist of a random sample of newborn female infants

C. A distribution for the last digit of phone numbers sampled from a phone book

“There are 10 digits, and IV was the only one with 10 bins”

(Question from CAOS: delMas, R., Garfield, J., Ooms, A., & Chance, B., 2007)
Quotes from think-aloud

“There are 10 digits, and (iv) was the only one with 10 bins(!!)”

“Every digit, so it is iv”

“it's pretty random, it's from a phone book, there should be a lot of data”

“I Chose this answer because there were 10 bins.”
Notes from think-alouds

Student did not know what a sampling distribution was.

Initially thought population must be normal. Later realized population should have some people studying 8 hours in it, so student picked B

“Population should be normally distributed.”

“Small n means few bars.”

“I wasn't thinking about the average, more about the distribution of the data.”

“True sample size for population should approach normal.”