Exploring Data Science Education: From Tutorials to Assessment

Duke Statistical Science | Graduation with Distinction

Evan Dragich supervised by Dr. Mine Çetinkaya-Rundel, PhD.

February 16, 2023

Table of contents

Abstract

Presentation Title: *

Creating a standardized assessment to measure learning in introductory data science courses

Brief Abstract:

As data science (DS) continues to grow in popularity among university course offerings, it is becoming crucial to successfully measure students' learning outcomes in introductory courses. To do this requires an assessment designed to which could additionally be used to evaluate pedagogical techniques or curriculum interventions in data science curriculum

To develop a blueprint for the assessment, a multi-institutional team of statistics and data science education researchers identified common DS content (e.g., data wrangling, interpreting visualizations), drawing from published guidelines/recommendations and introductory DS syllabi. A draft of the assessment was written and used to conduct three think-aloud interviews with field-relevant faculty members.The interviews consisted of both open-ended brainstorming on the assessment's scope as well as individual examinations of each item for relevance, clarity, and efficacy in measuring the desired learning objective. Think-aloud interviews were also conducted with introductory DS students to gauge item clarity and gain insight into the reasoning for their responses.

This poster includes the blueprint developed, as well as example items, and results from the faculty and student think aloud interviews. We also present next steps for the project including plans for larger scale piloting and further analyses.

Goals

By sharing the work, we hope that participants will become familiar with an assessment they may use for designing intro data science curriculum or researching classroom innovations. We also hope that this instrument can serve as an inspiration or a starting point to be tailored by future researchers more specifically to their courses or to another discipline (eg. by adding more programming concepts instead of data visualization to better serve a computing-focused introductory data science class, etc.)

Introduction

next section to write after development: just like about the motivation for the project

Assessment Background

Background

An essential component of any educational research is a validated, relevant instrument to measure students' learning outcomes. Whether to award college credit like the College Board's AP and CLEP exams, to measure students' previously-held misconceptions, or just as a tool upon which to evaluate educational interventions, such assessments have been developed and statistically analyzed for a wide range of subjects such as Spanish, Psychology, Chemistry, and Calculus (Godfrey and Jagesic 2016; Solomon et al. 2021; Mulford 1996; Epstein 2013).

In the field of statistics, previous work on measuring students' reasoning skills led to the development of the Comprehensive Assessment of Outcomes in Statistics (CAOS). The revised CAOS 4, comprising 40 multiple-choice items on a variety of commonly-taught first-semester introductory concepts, was first administered in 2005 and allowed instructors to measure whether their courses were successfully resulting in their desired learning outcomes. However, many instructors noted that their findings reflected a much lower understanding than expected, specifically regarding the topics of data visualization and data collection (Delmas et al. 2007). However, a key feature of the CAOS was the lack of hard computation nor need to recall specific formulas or definitions, allowing greater accessibility for a variety of statistics-adjacent uses.

In 2022, as more universities begin to support the emerging field of data science via specific courses, concentrations, or even majors, there is a need to measure students' learning outcomes in these introductory classes analogously to the subjects named above (Swanstrom, n.d.). Lacking a clearly-defined scope, empirical studies of so called "data science" curricula suggest that the field can be thought of an augmentation of traditional statistical modeling concepts, with emphases on computing, data visualization and manipulation, as well as a consideration of ethics and the role data plays in society (Zhang and Zhang 2021).

good places to start: read DelMas 2007 paper, look at what theyve done to get it beginnning to end, at some point acknowledge that were not finishing with just my theiss (analyzing student data, subscale stuff that wont be part of my thesis)

refer to this for background: https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1804497 in on eo fthe earlier sections, summary of data science courses and what they cover (and compare and contrast to our assessment). developed collectively based on the faculty and what they teach/what ive taken, but also capturing most of what's "out there"

are there any papers on: teaching data science with or without code, kind of justify our choice for language agnosticism

Just found this paper and thought you all would find it interesting: Emmanuel Schanzer, Nancy Pfenning, Flannery Denny, Sam Dooman, Joe Gibbs Politz, Benjamin S. Lerner, Kathi Fisler, and Shriram Krishnamurthi. 2022. Integrated Data Science for Secondary Schools: Design and Assessment of a Curriculum. In Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2022). Association for Computing Machinery, New York, NY, USA, 22–28. https://doi.org/10.1145/3478431.3499311 add one sentence on like: its in high schools too!

ask andy/chelsey–do they have other good references how to run think aloud interviews

Assessment Development

Phase 0: Initial Cleaning and Feedback

In January 2022, I inherited a repo with several documents: many were background information on what topics would be included, with one containing all currently-written questions. These questions were not yet organized into specific groups of stem and items, many sections were commented out or overwritten, and it was clearly something that had been written piecewise by a group of people. My first true task was to run through the current questions myself, answer them how I would, and provide feedback on clarity, wording, and reflect on the topic covered.

From this initial feedback, I created three pull requests attempting to solve some of these issues. These first fixes were minorly substantial edits: changing "most" to "the majority of" to make a question less ambiguous, adding a "fill-in-the-blank" slot to make even more clear what the question is asking, and background information on confidence intervals to an item that referenced them. This was my first exposure to what I would call "advanced" GitHub usage, or more than just typical cloning, pushing, and pulling from ordinary group projects. I used the very helpful `pr_*()` series of functions from the `usethis` package, which allowed me to easily create the pull requests for review.

While the team discussed my pull request proposals, I worked in parallel to clean up any "obvious" fixes to all items. These "no-brainer" edits were mainly typos, distorted or obscured plots and items that don't actually pose a question. It was also time to clean up the current document by converting it to a Quarto book, which would allow for easy webpage-like navigation. This marked my first exposure to Quarto, and I quickly grew to love its improvements to the RMarkdown workflow, like ease of rendering to different formats and intuitive structure of the index and YAML files. This was also when I received my first moment of creative liberty with the project, as I was tasked with dividing the ungrouped set of questions into discrete passages consisting of a stem and one or more corresponding items. Each of these passages–26 at the time–was given a title to identify it in the Quarto book sidebar.

At this point, the team came to a verdict on my initial pull request proposals: they approved and merged the slight text modification to "the majority" and the "fill-in-the-blank" title, but rejected the confidence interval language, asserting that the stem already provided sufficient motivation. The final step before presenting the assessment externally was to improve reproducibility, as many figures were not being rendered with each update, but rather embedded as images. In a throwback to my former STA 313 days, I used the current images as guides to recreate a map of the US colored by region, movie-themed data tables, and pseudocode chunks. I had some slight HTML and CSS exposure in previous classes, but this new challenge of matching an existing format allowed me to expand my knowledge about web layout and styles.

By April 2022, concluding my first semester working on the project, we had a polished, reproducible, website-hosted prototype of the assessment ready to present externally and gather feedback.

Phase 1: Faculty Interviews

The remainder of the time spent working on the assessment–April 2022 to February 2023–was spent gathering feedback via interviews, iteratively updating items, and removing weaker questions based on group discussions. We conducted a series of three think-aloud informational interviews with various faculty nationwide who teach or have taught introductory data science courses. Participants were recruited from a shortlist of Dr. Çetinkaya-Rundel's data science assessment contacts, and were specifically chosen to represent a breadth of data science curricula. The three interviewees chosen were, in order of interview:

- Dr. Nicholas Horton (Professor of Statistics and Data Science at Amherst College)
- Dr. Kristin Stephens-Martinez (Assistant Professor of the Practice of Computer Science at Duke University)
- Dr. Amelia McNamara (Assistant Professor of Computer and Information Science at the University of St. Thomas)

Each interview was scheduled for two hours long and consisted of three sections: open-ended introductory and concluding discussions, sandwiching an item-by-item run through of the assessment. In the initial discussion, we asked participants two big picture questions: What topics *must* be in an introductory data science course? And, what topics are *nice to have* in an introductory data science course? For each item in the assessment run-through, we asked interviewees to narrate out loud their thinking process from start to finish: any initial reactions, their process to arrive at an answer, and what said answer would be. We then asked for additional comments or suggestions, ranging from small- (formatting changes) to large-scale (removing the question entirely). We then concluded with three big-picture questions: What are the strengths of the current assessment? What topics are missing from the current assessment? And, what is in the current assessment, but doesn't belong?

While I scheduled and directed the flow of the interviews myself, I was joined by Dr. Çetinkaya-Rundel for all three, Dr. Legacy for the second two, and Dr. Beckman for the third. Silently observing, these team members took notes in real-time while I was conversing with the interviewee as a supplement to the transcript.

0.0.0.0.1 * Discipline-Specific Perspectives

After conducting each of three interviews, we met to discuss the new feedback and, when appropriate, made modifications and deletions to the current assessment. While each interviewee came from distinct backgrounds, there were some salient themes about their perspectives on introductory data science that emerged through patterns in their responses.

Dr. Horton seemed to have a somewhat similar perspective on "what is data science" as the research team–chiefly, visualization and wrangling. He noted positively when items featured multivariate data, particularly when interpreting visualizations. He also notably brought a wealth of pedagogical experience with real-life data to our consideration: such as that movie budgets and revenue display less compelling of a relationship than one would think, that movie-themed data at all is less relevant to younger generations, and that county data is generally too heterogenous to be useful in most contexts. He also grounded our conceptions of what a pre-test student would know by claiming that residuals are now part of the Common Core in K-12 education. However, he thought residual analysis and other related modeling topics like supervised learning didn't align with his experience teaching introductory data science. But, as in the case of Realty Tree, he acknowledged that some topics may fall outside the scope of an introductory class, but could be reasoned through and provide motivation to learn more.

Dr. McNamara held similar opinions on the scope of data science as did Dr. Horton–visualization and wrangling. The main takeaway from her feedback was that we needed to consider carefully the scenarios posed in question stems. Hurricane data, like in Hurricane Andy, she claimed, may provoke discomfort in students from hurricane-prone areas. She noted that the former wording of Austen Gender, in which the verbs on the plot itself were in the past tense while those in the items were in the present, may present a burden to non-native English speakers. The SAT, referenced in Banana Conclusions, may need more explanation for international students who are less familiar with the US admissions process. She also, like Dr. Horton, thought that some of our more technical questions breached the scope of introductory data science, such as residual analysis and the distinction between training and validation sets. She was similarly a fan of Realty Tree. However, while Dr. Horton was in favor of summary statistic questions and mapping them to graphs, Dr. McNamara called out language like "margin of error" as too statistical, and not as data science related.

As a computer science education representative, we knew Dr. Stephens-Martinez's interview would offer a unique perspective. This was evident to the start, when her first response to my question of "what is essential" was to ask me whether this course assumed prior coding experience or not. After I clarified that we are assuming no coding experience, she answered that learning coding "in the order that matters for data science", e.g. functions first, is a primary topic, as is some unerstanding of transforming and cleaning data. But, she qualified, not too much as much cleaning can be done for them in an introductory class. It wasn't until the "what is nice to have" question that she mentioned data visualization. This CS-focused perspective continued when she pointed out our pseudocode didn't specify which type of join would be performed, that basic English vocabulary like "filter" and "select" might not be known in their data wrangling context, and generally dissuaded us from using pseudocode on something that would be a pre-test. Another notable pattern was that some of her suggested edits went against data visualization best practices: she suggested we change from a rainbow color scheme to a gradient one for categorical data, and to reorder the y-axis in Austen Gender to be alphabetical rather than allowing for easy comparison. Finally, she led us to remove the logarithmic transformation present in all Movie Budgets questions, explaining that since the transformation itself isn't a learning outcome, students seeing those words continually repeated may start losing sight of the item's objective.

Amelia: change hurricane context to paths of animals wandering; log-transform cognitive overload; doesn't think margin of error belongs in a "data science course"; thought conflicting sources was "too inference", doesn't like judging r2 from scatterplot; tense change in austen gender; liked regression tree for being "different than what they might have done before" (acknowledges could answer without knowing); park wrangling pseudocode too R-specific, related to that she completely missed the "load in the data wrong" question; SAT contextual knowledge; COVID Map contextual knowledge; highest on data viz and wrangling/cleaning

Nick: loved the statty stuff (even referenced connecting summary stats to graphical display as building on K-12); loves when questions were multivariate; didn't like movie dataset; wanted to define r2 and asserts residuals are part of K-12 now, "We don't do any regression in our data science course; also issues with county data; loved regression tree as pretest motivation; does teach unsupervised learning though; PR as a state; thought banana conclusions was more data acumen/data literacy; repeated Time series in covid map; really liked ethics and visualization

Kristin: first question was coding background or not–already off to a differnet start. coding in the order that matters for DS (ie. function first), some understanding of cleaning (thinks cleaning can be done for them in an intro class), probability > p-values or t-test, visualizatino only a "nice tohave", cognitive load of log transform; some tweaks kinda went against data viz best practices (said we should change to gradient colors for categories, another cognitive load to sort Austen Gender); didnt like pseudocode for complete intro class, especially terms like group by, filter, select (and especially for this questioin that doesnt specify join type), planning a solution vs writing code.

highlight a few things: one thing that hadn't really thought about that Kristin pointed out, something we hadnt thoguht about that Amelia/Nick pointed out. one paragraph each, nothing crazy but just explaining what its like to get feedback

https://docs.google.com/document/d/1edz9Erh2aB_RA9WsHPEqxm_dvOMvNkqj1BjmL73b-3A/edit

0.0.0.0.2 * Regrouping to synthesize feedback

how to write questions that arent too obvious and get at specific things. this ended up being a lot of quetions that we cut–amelia and nick (kristin didnt get to it) missed the R-specific data load in question, the options of Data Confidentiality went through more than a handful of iterations, they (Amelia especially) had a lot of "hmm i see what concept youre getting at but this question doesnt really.... okay yeah, this is something that's hard to write to be autogradable but also actually measure _____". also was just soooooo hard to cut questions. we had ~50 at this point, and every week would cut ~2 or so. it's not that people on the research team had diff schools of thought like the faculty did, or that people were protective of what they wrote–we just genuinely saw value in all questions at that point, and decided we would let student interviews guide the necessary culling.

Phase 2: Student Interviews

with students with exposure: goal is to see if these topics are in your wheelhouse and if wording makes sense as a student balancing size and pacing. we've hammered out a lot of the larger-scale scope type issues, now time to start drilling down into the weeds and seeing if they land how we think they do. if not, or way too easy/hard/confusing, take it out!

• Are the pacing and length appropriate? • Based on what you remember learning in intro data science, what topics are missing from the current assessment? • Based on what you remember learning in intro data science, what is in the current assessment, but doesn't belong?

- Student A (2nd year Masters of Statistical Science student)
- Student B (4th year undergraduate Statistical Science minor)
- Student C (1st year Masters of Statistical Science student)

highlights were–we originally had items grouped by section, but we had a clear case of Student A being primed by earlier ethics question (Image Recognition) to be aware of in Application Screening. Student B missed Austen Gender bear trap and just generally had a very air of uncertainty throughout. Student C attended undergrad at a python, computing-heavy program so feedback was very similar to that of Kristin's. Mention or not–Student A seemed to be a non-native English speaker, and nothing clocked them on that (at least i dont think?). Oh other takeaway i guess is that all of the students couldn't read `theme_minimal()`, which i think is funny because I "grew up on that" since i never did 199 and did 313 as part of my first semester of R classes.

0.0.0.0.1 * Regrouping to synthesize feedback/final pilot assessment

summarising feedback from live sessions, working with the rest of the group to incorporate feedback. really didnt make many changes at this stage, just axed a large chunk of quetions after being too wishywashy during the fall 2022 gap. oh changed movie budgets 2 from individual items-> single matrix

0.0.0.0.2 * Final assessment themes

somethign good would be good to have content: use 15 passages and assign them to rough buckets

our previous reviews are in here (I guess leave item # blank for now?)

Passage	Item Number	Learning Objective(s)
Storm Paths		
Movie Budgets 1		
Movie Budgets 2		
Application Screening		
Banana Conclusions		
COVID Map		
Disease Screening		
Austen Gender		
Recreate Unemployment		
Realty Tree		
Website Testing		
Image Recognition		
Data Confidentiality		
Activity Journal		
Movie Wrangling		

Phase 3: Large-scale Student Pilot

future direction: students in 199 this spring would complete the assessment (either in full or a randomized subset of items) as an extra credit component of class. moving beyond individual item tweaking and refinement, this phase of the project will allow for real-world observations of pacing and length, feasibility for intro data science students. have to convert Quarto book to a qualtrics, very high-maintenance IRB, etc.

Package Development

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).

Discussion

also about future directions with 199 stuff

talk about how much quarto, html, css, yaml, qualtrics, github actions, CRAN etc ive learned

Bibliography

Çetinkaya-Rundel, Mine, and Victoria Ellison. 2021. "A Fresh Look at Introductory Data Science." *Journal of Statistics and Data Science Education* 29 (sup1): S16–26.

Delmas, Robert, Joan Garfield, Ann Ooms, and Beth Chance. 2007. "ASSESSING STUDENTS' CONCEPTUAL UNDERSTANDING AFTER A FIRST COURSE IN STATISTICS." *STATISTICS EDUCATION RESEARCH JOURNAL* 6 (2): 28–58.

Epstein, Jerome. 2013. "The Calculus Concept Inventory-Measurement of the Effect of Teaching Methodology in Mathematics." *Notices of the American Mathematical Society* 60 (8): 1018–27.

Godfrey, Kelly E., and Sanja Jagesic. 2016. *Validating College Course Placement Decisions Based on CLEP Exam Scores: CLEP Placement Validity Study Results. Statistical Report.* College Board.

Mulford, DouglasRobert. 1996. "An Inventory for Measuring College Students' Level of Misconceptions in First Semester Chemistry." *Unpublished Master's Thesis, Purdue University, IN.*

Solomon, Erin D., Julie M. Bugg, Shaina F. Rowell, Mark A. McDaniel, Regina F. Frey, and Paul S. Mattson. 2021. "Development and Validation of an Introductory Psychology Knowledge Inventory." *Scholarship of Teaching and Learning in Psychology* 7: 123–39.

Swanstrom, Ryan. n.d. "Data Science Colleges and Universities." http://datascience.community/colleges.

Zhang, Zhiyong, and Danyang Zhang. 2021. "What Is Data Science? An Operational Definition Based on Text Mining of Data Science Curricula." *Journal of Behavioral Data Science* 1 (1): 1–16.

Appendix A: Current Assessment Prototype

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).

Appendix B: Item Graveyard

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).

Appendix C: dsbox in action

Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).