Exploring Data Science Education: From Tutorials to Assessment

Duke Statistical Science | Graduation with Distinction

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

February 21, 2023

Table of contents

Abstract

Download PDF here

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:

Name	Department Affiliation	Institution Type	Field of Ph.D. Dissertation
Professor X	Statistics	Liberal arts college	Biostatistics
Professor Y	Computer science	R1 research university	Computer science education
Professor Z	Computer and information science	Liberal arts school within university	Statistics

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.

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.

Professor X 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.

Professor Z held similar opinions on the scope of data science as did Professor X–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 Professor X, 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 Professor X was in favor of summary statistic questions and mapping them to graphs, Professor Z called out language like "margin of error" as too statistical, and not as data science related.

As a computer science education representative, we knew Professor Y'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.

Regrouping to synthesize feedback

While the team met once bewteen each faculty interview to make any minor changes before the next one, the bulk of the revision took place after all were completed, in the Fall 2022 semester. A common theme from all interviews was that our questions designed to test tricky, nuanced concepts (e.g. reading in data to statistical software) in a multiple choice format weren't landing as we hoped. While we thought we had written a former item, [Data Reading], with enough language agnosticity, it ended up being too R-specific and baffling all faculty. To this item, and several others, Professor Z summed up the crux of our dilemma well, paraphrasing: "hmm.. I see what concept you're getting at, but this question doesn't really get there… Okay, well, this is something that's hard to write to be autogradable but also actually measure. I don't know how I would fix this but I really like the underlying idea."

Encouraged to keep these concepts on the assessment with modifications, we found it difficult to cut out any of the ~50 pilot questions we had at this point. One notable large-scale fix was combining the concepts tested on three former wrangling passages into a single context, hoping to further reduce students' cognitive load. We knew we needed to cut items down to the ~30 range, but found it very difficult to identify those items that were bringing the least to the current assessment. Interestingly, this involved almost no large disagreements between team members–there were no particularly polarizing items, like we observed during the faculty interviews–rather, we all genuinely couldn't identify any items we wanted to cut.

We did eventually make rounds of sacrificial cuts, choosing to prioritize culling questions that even slightly overlapped and those that were most straightforward. We acknowledged the tradeoff that this may lead to students taking the assessment in a pre-test context unable to confidently answer a majority of the questions. However, we were constrained by the need for an instrument that could reasonably be given in an hour, and cutting out several entire passages was the only way forward. Nevertheless, there were a handful of similar items that we decided to keep for the student interviews, such as two pseudocode chunks (in Movie

Wrangling) that varied by a single statement. Here, we were ambivalent on which would be more effective, and hoped to let student feedback dictate whether one "stuck" better.

Phase 2: Student Interviews

The chief purpose of Phase 2 was to see if the topics covered would be in the wheelhouse of students with DS exposure, as well as continuing to refine wording and pacing. With the big picture questions around "what is data science" previously discussed, a similar series of interviews with Duke Statistical Science students allowed us to start drilling down and seeing if items landed the way we thought they would.

To reflect this different type of feedback being solicited, we omitted the initial "big picture" questions while interviewing TAs. Combined with the fact that that the assessment was now about half the original length, these interviews were only one hour long. The final portion of big picture questions was kept, though, with slightly modified prompts: 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?

We first reached out in November 2022 to all TAs at that time for STA 199 or STA 198, its health-themed analog. However, we were unable to recruit any students so close to finals, and ended up reaching out to the same group in January. The interviews were conducted in early February, thus consisting of students who were at least TAs for STA199 in the previous Fall semester. The three students recruited were, in order of interview:

Name	Degree Year and Level	Program
Student A	2nd year Masters	Statistical Science
Student B	4th year Undergraduate	Economics major, Statistical Science minor
Student C	1st year Masters	Statistical Science

In general, while still engaging well and demonstrating a strong command of DS skills, the students tended to be much less verbose in their feedback. Across all interviews, a former item on [Realty Tree] that asked students to trace a non-trivial regression tree four times stood out as an interruption to the otherwise smooth flow of the assessment. Also, on a more cosmetic note, all students suggested that plots which had used `theme_minimal()` be modified to more clearly show which titles corresponded to which faceted subplots–this stood out to me, as I am someone who also has taken many Duke Statistical Science courses, which are often `theme_minimal()` heavy.

Other issues revealed during interviews was the question order. We originally arranged items near others testing a similar topic to facilitate our iterative editing. However, there was a clear case of cognitive priming in Student A's interview. Having just correctly answered [Image

Recognition], he then immediately jumped to examining ethical implications in the following item, [Application Screening], while the faculty members who had answered the questions in a different order required more consideration. As well, Student B was the first interviewee to answer incorrectly to the item I think is the single trickiest (#3 in [Austen Gender]), and best written in terms of getting students to think critically about what exactly plots represent. This stood out to me, as she was the only undergraduate interviewed; we observed a stratification of data science comfortability even among TAs. Finally, Student C was the quickest interviewee to run through the questions, answering all correctly and citing the logic we were looking for. As the most relative newcomer to the Duke Statistical Science program and having graduated from an undergraduate data science program that used Python, we interpreted his notable ease in getting through all questions in much less than an hour as a good sign that the current length might be ideal for students, who would be less experienced with the material but would also not have to think aloud or give feedback.

Regrouping to synthesize feedback/final pilot assessment

Not many changes were made following the student interviews in Phase B: a sign that we were converging on a viable prototype for distribution. To remedy the priming issue encountered in Student A's interview, we shuffled the question order between the student B and C interviews and landed on a solid layout: distributing topics well throughout and gradually ramping up our intended difficulty to prevent less-knowledgeable students from getting frustrated and giving up early on. As part of this pacing reform, we improved the flow of the assessment by reformatting [Movie Budgets 2] from five nearly-identical items into a single matrix.

Final assessment themes

The following table summarizes the learning objectives for each item in the most current version of the assessment.

Passage	Learning Objective(s)
Storm Paths	simulated data; interpreting uncertainty
Movie Budgets 1	comparing summary statistics visually
Movie Budgets 2	R^2; comparing trends visually
Application Screening	ethics; proxy variable
Banana Conclusions	causation; statistical communication
COVID Map	interpreting complex visualization; spatial data; time series data interpreting complex visualization; sophisticated scales
Disease Screening	comparing classification diagnostics visually comparing classification diagnostics visually

Passage	Learning Objective(s)
	comparing classification diagnostics visually
	comparing classification diagnostics visually
Austen Gender	interpreting basic visualization
	interpreting basic visualization
	interpreting basic visualization; sophisticated scales
Recreate Unemployment	data to visualization process
Realty Tree	modeling; regression tree
	modeling; regression tree; variable selection
Website Testing	interpret trends visually; visualize uncertainty; time series data
	interpret trends visually; visualize uncertainty; time series data; extrapolation
	interpret trends visually; visualize uncertainty; time series data; extrapolation
Image Recognition	ethics; representativeness of training data
Data Confidentiality	ethics; data deidentification
Activity Journal	structuring data; storing data
Movie Wrangling	data cleaning; column-wise operations; string operations
	data cleaning; column-wise operations; string operations
	data cleaning; extrapolation
	data wrangling; pseudocode, joins
	data wrangling; pseudocode, joins
	data wrangling; pseudocode, joins

Phase 3: Large-scale Student Pilot

Looking ahead, the next step after individual item tweaking and refinement will be real-world measurements of pacing and length and the feasibility for introductory data science students. Collaborating with the professors of Duke's Introductory Data Science Course STA199, ____ students across three sections will be invited to take the assessment, on their own time, for a chance to earn extra credit in the course. We will not solicit general feedback like in the previous interviews, instead asking just a single question at the end to get a rough estimate of how long it took to complete, since they will be doing it on their own time. The extra credit points students receive will not reflect their performance on the assessment, but rather will be simply awarded proportionally to how much was completed. To ensure students are actively engaging with the questions, there will be ____ attention checks scattered throughout that must also be passed to earn credit.

In order to distribute the assessment at large and record students' responses, the Quarto book has been converted into a Qualtrics survey. Once complete, large-scale analysis will be conducted, both to ensure that all questions were indeed appropriate for introductory level students and to explore potential instrument subscales.

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).