THE LEARNING DIFFICULTIES EXPERIENCED BY INTRODUCTORY DATA SCIENCE STUDENTS

Sinem Demirci[1], Mine Dogucu[1,2], Andrew Zieffler[3] and Joshua M. Rosenberg[4]
[1]University College London, United Kingdom
[2]University of California, Irvine, USA
[3]University of Minnesota, USA
[4]University of Knoxville, USA
s.demirci@ucl.ac.uk

Introduction to Data Science (IDS) courses are being offered by many different departments either as a mandatory or an elective course. Because of the foundational nature of IDS courses to develop students' understanding of data science, it is important to be aware of students' potential learning difficulties. To that end, we conducted semi-structured interviews with 14 IDS instructors to study students' difficulties. Qualitative content analysis was used to analyze the data. IDS instructors reported that students without prior coding experience encountered more syntactic difficulties than their peers. In terms of conceptual and strategic knowledge, students experienced difficulties in understanding principles of data visualization, the basics of coding, joining data sets, debugging and data wrangling. These findings suggest that IDS courses could be improved by addressing student difficulties and invite conducting future research with students to understand the dimensionality of student learning to improve capacity in data science (education).

INTRODUCTION

Data science is a field that blends multiple areas of expertise and typically demands expertise in a range of skills and concepts spanning statistics, computer science, mathematics (De Veaux et al., 2017), and other domains (Mike & Hazzan, 2023). As Mike and Hazzan (2023) recently denoted, an agreement for a single definition for data science is a difficult task because of its multifaceted and interdisciplinary nature. This multifaceted nature of data science has also been discussed in the data science education community (e.g., Asamoah et al., 2020; Jiang et al., 2022) because it brings unique challenges to determine the scope and content of data science courses/majors (Yan & Davis, 2019). Although some initiatives such as providing curriculum guidelines for data science at tertiary levels (e.g., De Veaux et al., 2017; Danyluk & Leidig, 2021) and describing essential skills for a data scientist (e.g., De Veaux et al., 2017), more research on data science education is required to understand how to deliver this aspect efficiently and to cultivate proficient data scientists with a sound understanding of interdisciplinarity.

While data science education community strives to create curriculum guidelines and determine competencies for data scientists and data science students, we would like to direct attention to the importance of the learning difficulties of data science education students because these difficulties constrain them developing a sound understanding of data science and making progress (Qian & Lehman, 2017). We chose introductory data science courses (IDS) as a starting point because experiences in this course have a potential to attract students to pursue majors, minors, tracks, and certificates offered by institutions (National Academies of Sciences, Engineering and Medicine Consensus Report, 2018). Thus, IDS courses have an important role in developing a sound understanding of data science, influencing students' decisions and motivation to pursue additional data science courses —or a data science career. Lastly, it is evident that teachers' awareness of students' difficulties is an essential aspect of effective teaching (Sadler et al. 2013). Thus, determining students' difficulties might provide an insight into how to enhance IDS teaching and revise the scope of IDS courses as well as data science majors/programs. Accordingly, this study aims to determine students' difficulties in IDS courses reported by their instructors.

Theoretical Framework of the Study

In the context of our study, we adapted a framework that has been used in early computer science education (Bayman & Mayer, 1988; McGill & Volet, 1997; Qian & Lehman, 2017) to classify knowledge of programming into syntactic, conceptual, and strategic categories. Syntactic knowledge includes "knowledge of specific facts about a programming language and rules for its use" (McGill &

Volet, 1997, p. 277), while conceptual knowledge refers to an "understanding of computer programming constructs and principles" (p. 277). Strategic knowledge involves integrating syntactic and conceptual knowledge of programming to solve novel problems (Qian & Lehman, 2017). Although this framework is designed for programming, we expanded its scope to make it compatible with data science. We decided to retain the exact definitions of syntactic and strategic knowledge because they aligned well with the content covered in the IDS courses examined in our study. While we retained them, we extended the definition of conceptual knowledge to encompass understanding the constructs and principles of mathematics & statistics, computer science, subject-specific knowledge, and interdisciplinary knowledge which are included in defining data science (Mike & Hazzan, 2023) as a discipline.

METHOD

In this study, we chose *qualitative research* design (Merriam & Tisdell, 2016) because our aim was to understand how IDS instructors classify and interpret students' difficulties in IDS courses and "what meaning they attribute to their experiences" (Merriam, 2009, p. 23). This study is a subset of a larger study. For that larger study and the subset we report here, we defined the target population to consist of instructors who taught a course titled "Introduction to Data Science" at least twice at the undergraduate level. Often when an instructor teaches a course for the first time, they focus on multiple aspects of the course as a novice. We thought that instructors who have gone through the second iteration of the course would be able to reflect deeper about the course and the students. We recruited participants via mailing lists and online forums with large teacher-scholar communities.

We collected data through online semi-structured interviews with 14 participants. Each participant was compensated with a £ 50 gift card for their time. Most of the participants were teaching in North America at the time of the interview. The instructors had terminal degrees in varying subjects including statistics, mathematics, computer science, genetics, and economics. They all had been teaching an introductory data science course for a varying number of years, with a range from 1 to 10 years of experience. The students enrolled in these IDS classes were from a variety of majors. While majority of classes were composed of computer science, statistics, data science, mathematics students, there were also students from other majors such as cognitive science, engineering, social science, humanities, and those who had not yet declared their majors.

We designed specific questions to examine in which concepts/tasks IDS students experience difficulties. We asked participants 3 main questions: (1) With which concepts do your students have difficulties? (2) What are the difficulties, if any, that students have while performing DS tasks given in your course? And (3) What are the conceptual difficulties, if any, that students have in your course? We also had some follow-up questions depending on the responses to elaborate students' difficulties.

To qualitatively code these difficulties, we used qualitative content analysis (Merriam & Tisdell, 2016) for generating a comprehensive codebook for determining students' difficulties reported by IDS instructors. We adapted the framework of Qian and Lehman (2017) which covers introductory programming students' difficulties and extended this framework to introductory data science courses. We used this framework to determine initial themes for our codebook deductively, but we also extended our codebook inductively based on the findings of our study. In short, we benefited from both deductive and inductive coding in our content analysis. An example of qualitative codebook is given in Table 1.

Table 1. Sample codebook

Themes	Categories	Codes
Syntactic Knowledge	Markup Languages and Reproducibility Tools	e.g., R Markdown
Conceptual Knowledge	Mathematics	e.g., algorithms
	Statistics	e.g., types of variables, hypothesis testing
	Computer Science	e.g., how a loop works
	Domain-specific Knowledge	e.g., understanding technical writing
	Interdisciplinary Knowledge	e.g., ethics
Strategic Knowledge	-	e.g., debugging, data wrangling

To enhance the trustworthiness of the study, we collected indicators for *transferability, dependability,* and *credibility* (Merriam & Tisdell, 2016). Particularly, we provided a detailed description for our participants' profile, data collection and data analysis procedures. We also had different participants (e.g., differed in terms of year of experience, terminal degree etc.) based on our selection criteria which enabled maximum variation in our sample. Additionally, two researchers continuously compared and discussed to determine the extent of codebook based on the theoretical framework and qualitative data of the study.

RESULTS

In this section, we present the findings of our qualitative content analysis, which were categorized into three themes: (1) Syntactic Knowledge Difficulties; (2) Conceptual Knowledge Difficulties; and (3) Strategic Knowledge Difficulties.

Syntactic Knowledge Difficulties

Within the theme of Syntactic Knowledge Difficulties, we identified two categories based on the reports from IDS instructors. The first category related to students' difficulties with markup languages and reproducibility tools, while the second category related to difficulties with programming languages. The codebook for these difficulties is provided in Table 2.

Table 2. Syntactic knowledge difficulties

Markup Languages and Reproducibility Tools	HTML, R Markdown, Quarto Markdown, Jupyter Notebook, Linux, Git/GitHub
Programming Languages	Packages, Libraries, Misspelling, Adapting the Code, How to Read Data

Since IDS courses utilized various markup languages, reproducibility tools, and programming languages, IDS instructors reported distinct syntactic difficulties that were specific to their course. However, 11 out of 14 IDS instructors observed that students without prior coding experience encountered more syntactic difficulties. To support these students, some IDS instructors offer additional sessions and/or office hours. Here is an excerpt that provides an example related to this category:

> *Interviewer: And so, with which concepts do your students have difficulties? Do you have any observations for that?*
> *Participant-06: It varies by students [sic], right? Because they are computer science students [sic]in there, and they don't have problems with programming. But a lot of the other students have problems with programming where you know the typical are questions you get like you misspelled the data set name, or you forgot to change the data set name. We, we use the copy paste change model for teaching basic R here rather than the filling the blank which we use at the low level, elementary stat one. But, you know, did you, getting the coding to run is often very hard for the [sic] non-computer science students? The math students take to it okay, but the students outside of both of those tend to struggle a bit there.*

Conceptual Knowledge Difficulties

We categorized conceptual knowledge difficulties into five categories: (1) mathematics; (2) statistics; (3) computer science; (4) domain-specific knowledge; and (5) interdisciplinary knowledge. The codes that emerged from our data are presented in Table 3. These categories were derived from the description of data science available in the works of Mike and Hazzan (2023).

Table 3. Conceptual knowledge difficulties

Mathematics	Algorithms, Permutation Testing
Statistics	Types of Variables, Confidence Interval, Principles of Data Visualization, Hypothesis Testing, Correlation vs. Causality, Bootstrapping, Inductive Inference, Statistical Modelling, p-value, Sampling Distribution
Computer Science	I/O File Management, Working Mechanisms of Markup Languages, Basics of Coding, Filter Function, Basics of Web Scraping, Select Function, Joining Data Sets, Mapping Functions, Loops, Creating Functions
Domain-Specific Knowledge	Understanding Technical Writing, Understanding the Nature of Data
Interdisciplinary Knowledge	Ethics, Machine Learning

Of the IDS instructors, nine reported that students experienced difficulties in understanding statistical concepts, while six reported difficulties in understanding computer science concepts. Among the statistical concepts, the principles of data visualization were the most frequently mentioned. Regarding computer science concepts, understanding the basics of coding and joining data sets were two commonly reported difficulties. Additionally, five IDS instructors mentioned difficulties in understanding either the nature of data or technical writing in a specific domain.

As an example, the following excerpt from Participant-03 provides insight into conceptual knowledge difficulties in computer science: "*Yeah, so it's a picky little thing [web scraping], right? Finding the right code and the right thing that you're asking to pull from that from that web page and so the students struggle with that just a bit, putting together the right code to scrape something from a page, but then also creating functions and loops to do that on multiple pages. You know, and, and sort of automating that task. And so, I… I'd say that's probably one of the one of the topics they struggle the most with*."

Strategic Knowledge Difficulties

Except for 3 IDS instructors, 11 reported observing strategic knowledge difficulties in their IDS courses. The most frequently mentioned difficulties were debugging and data wrangling. Table 4 lists these and other difficulties reported by IDS instructors. Additionally, some of them denoted that students tend to oversimplify data science tasks given in IDS course and try to run a statistical analysis without thinking about the content and examining data set accordingly.

Table 4. Strategic knowledge difficulties

Strategic Knowledge Difficulties	Debugging, Communication, Data Wrangling, Appreciating the complexity of Interdisciplinary Research, Making Appropriate Data Visualization Decisions, Creative Thinking, Proper Use of Descriptive Statistics, Conducting a Good Research, Deciding Statistical Analysis Methods-Modelling, Working with Real and Messy Data, Handling Missing Data, Asking Good Questions, Web Scraping, Setting up Data Science Pipeline

A sample excerpt describing the difficulty of deciding statistical analysis methods/modelling were as following: *"…So certainly, so this so kind of so statistical analysis in so kind of correct statistical analysis in general is a problem. So, everyone is very tempted to just kind of throw any tool they can, they can at the problem and just like, look at the outputs to see if the if the p-value is significant. So, this so I try to instill this kind of skeptical mindset of like, you know, does that, does the model fit? Does the question make sense? … [conversation continues] So that, I would say, is kind of one of the more challenging things to teach."*

DISCUSSION

In this study, IDS instructors shared their observations regarding the difficulties that students encounter in their IDS courses. We categorized these data science knowledge difficulties into three areas: syntactic knowledge, conceptual knowledge, and strategic knowledge, using the framework developed by Qian and Lehman (2017). Most of the IDS instructors highlighted that students without prior programming knowledge tended to experience more syntactic difficulties and require additional support. Some of the commonly reported difficulties included difficulties in understanding statistical and computer science concepts as well as debugging and data wrangling.

We incorporated a framework utilized by various scholars (e.g., Bayman & Mayer, 1988; McGill & Volet, 1997; Qian & Lehman, 2017) to analyze the data. It is noteworthy that we adapted this framework from early computer science education to suit IDS education. While we found the framework useful in providing an initial understanding of IDS difficulties, we acknowledge the potential requirement for further modifications to actively classify knowledge in IDS classes.

Apart from students' difficulties, some IDS instructors in this study articulated that students have a tendency to oversimplify data science assignments in the IDS course, by attempting to run statistical analyses without adequately considering the content and carefully examining the dataset. While some students may oversimplify IDS tasks, we suggest that this oversimplification may also be partially attributed to the difficulties that students face in these courses, which are not yet fully understood. Therefore, further studies are needed to measure students' difficulties and identify the specific areas in which they struggle, to better understand the reasons for this "oversimplification". These studies would also provide an insight into not only for data science education in tertiary level but also lower levels because the integration of data science into secondary education, for instance, has started to be discussed (e.g., Heinemann et al., 2018; Pittard, 2018; Frischemeier et al., 2021).

It is noteworthy that while there were some commonalities among the IDS courses examined in this study, each course may have presented its own unique set of syntactic, conceptual, and strategic knowledge difficulties. Therefore, our findings may serve as informative rather than generalizable constructs that can inform IDS instructors and the wider data science education community about potential student difficulties and their types. What is more, these difficulties reported in this study were presented from the perspective of the instructors, not the IDS students. Thus, it is essential to conduct more systematic research to assess students' challenges in IDS courses incorporating both opinions and/or cognitive assessments of IDS students to be able to inform policymakers and educators on how to enhance meaningful learning experiences for IDS students.

The sample of this study consisted of North American IDS instructors, even though we did not have such a specific aim within the context of the study. The possible bias for sample selection might be related to the selection criteria (e.g., selecting participants based on similar course names). In other country settings, there might be similar courses with different names. Thus, further studies in other country settings might also provide an insight into other students' difficulties that we were not able to capture in this study.

REFERENCES

Asamoah, D. A., Doran, D., & Schiller, S. (2020). Interdisciplinarity in data science pedagogy: a foundational design. Journal of Computer Information Systems, 60(4), 370-377, https://doi.org/10.1080/08874417.2018.1496803

Bayman, P., & Mayer, R. E. (1988). Using conceptual models to teach BASIC computer programming. Journal of Educational Psychology, 80(3), 291, https://psycnet.apa.org/doi/10.1037/0022-0663.80.3.291

Danyluk, A., & Leidig, P. (2021). Computing competencies for undergraduate data science curricula: ACM data science task force. Peer-Reviewed Publications, 8, https://scholarworks.gvsu.edu/cispeerpubs/8

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

Frischemeier, D., Biehler, R., Podworny, S., & Budde, L. (2021). A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data. Teaching Statistics, 43(2021), S182– S189. https://doi.org/10.1111/test.12283

Jiang, S., Lee, V. R., & Rosenberg, J. M. (2022). Data science education across the disciplines: Underexamined opportunities for K-12 innovation. British Journal of Educational Technology, 53(5), 1073-1079.

Heinemann, B., Opel, S., Budde, L., Schulte, C., Frischemeier, D., Biehler, R., ... & Wassong, T. (2018, November). Drafting a data science curriculum for secondary schools. Proceedings of the 18th Koli Calling International Conference on Computing Education Research, 1-5, https://doi.org/10.1145/3279720.3279737

McGill, T. J., & Volet, S. E. (1997). A conceptual framework for analyzing students' knowledge of programming. Journal of Research on Computing in Education, 29(3), 276-297, https://doi.org/10.1080/08886504.1997.10782199

Merriam, S. B. (2009). Qualitative Research: A Guide to Design and Implementation. San Francisco: CA: Jossey-Bass.

Merriam, S. B., & Tisdell, E. J. (2016). Qualitative Research: A Guide to Design and Implementation (Fourth Edition). San Francisco.

Mike K. & Hazzan, O. (February 2023). What is data science? Communications of the ACM, 66(2), 12–13, https://doi.org/10.1145/3575663

National Academies of Sciences, Engineering and Medicine Consensus Report (2018). Data Science for Undergraduates: Opportunities and Options. Washington, https://nas.edu/envisioningds.

Pittard V. (2018). The Integration of Data Science in The Primary and Secondary Curriculum. Final Report: To the Royal Society Advisory Committee on Mathematics Education. https://royalsociety.org/-/media/policy/Publications/2018/2018-07-16-integration-of-data-science-primary-secondary-curriculum.pdf

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. ACM Transactions on Computing Education (TOCE), 18(1), 1-24, https://doi.org/10.3102/0002831213477680

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. American Educational Research Journal, 50(5), 1020-1049.

Yan, D., & Davis, G. E. (2019). A first course in data science. Journal of Statistics Education, 27(2), 99-109, https://doi.org/10.1080/10691898.2019.1623136