Social Sciences & Humanities Open 8 (2023) 100628

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular Article

Transdisciplinary teaching practices for data science education: A comprehensive framework for integrating disciplines

Nkosikhona Theoren Msweli [a,*], Tendani Mawela [a], Hossana Twinomurinzi [b]

[a] Faculty of Engineering, Built Environment and Information Technology, Department of Informatics, University of Pretoria, Lynnwood Road, Hatfield Pretoria, 0002, South Africa
[b] College of Business & Economics, University of Johannesburg, Auckland Park, Kingsway Campus, Johannesburg, 2092, South Africa

ARTICLE INFO

Keywords:
Data science
Education
Technological pedagogical content knowledge
Teaching
Educational technology

ABSTRACT

Teaching data science programmes poses challenges for instructors due to the transdisciplinarity of the field and the diverse backgrounds and skill levels of students. Effective data science education requires a comprehensive approach that incorporates theoretical knowledge, practical skills, and industry relevance. However, it is difficult to find appropriate teaching strategies and tools that successfully integrate all these elements into the classroom. Consequently, there is a need to identify and develop effective pedagogical methods, instructional resources, and technological solutions that enable instructors to deliver well-rounded data science education that caters to the diverse needs of students and prepares them for real-world data-driven challenges. Knowing which technology is appropriate to use in conjunction with a particular teaching pedagogy to deliver a particular piece of learning material to diverse students is crucial. Therefore, this study aimed to explore how the TPACK (technological pedagogical content knowledge) influences data science teaching practices. To achieve this, the study surveyed 26 data science instructors to assess their confidence in the seven TPACK constructs. The findings of the study showed a low representation of women in data science education. The findings also showed a balanced knowledge between pedagogy and technological content, indicating that instructors can contribute to a comprehensive and engaging learning environment that supports student success in data science education. Despite this positive finding being established, it was not clear which technological teaching and learning tools instructors are familiar with. To this end, future studies are recommended in this area. The results further showed that model evaluation is not taught at undergraduate level. Therefore, the study recommends continuous professional development for data science instructors to effectively contribute towards training current and future data scientists. This is necessary since technologies, data, and data science tools and techniques evolve. Furthermore, the study recommends research be conducted on the type of data science framework required to guide instructors in terms of curriculum design, pedagogies, and technological tools. Research that informs policy is also necessary to support efforts directed at data literacy, especially to support personnel involved in human capacity development in data science. Lastly, within the scope of data science, interdisciplinary collaboration at national and international levels is recommended so that instructors can stay updated with advancements in subject matter, technology, and pedagogy.

1. Introduction

Data science education (DSE) is an emerging transdisciplinary academic field that is gaining interest from researchers and practitioners. It integrates knowledge from computer science, mathematics and statistics, and other domains (Mike, 2020). Teaching data science requires transdisciplinary pedagogical approaches to deliver instructional programmes. Integration of technology is also important for how data science techniques are applied and resource sharing. Regarding data, literature shows that the use of technology advances data skills and improves knowledge relating to using real data sets as part of the DSE (Saddiqa et al., 2021). However, changes in organisational practices and new data science trends have out-paced DSE. This has created a knowledge gap, which may result in the misinterpretation of data, leading to misinformed policies and strategies (Schatsky et al., 2018).

While the number of data science programmes is growing (Loy et al.,

* Corresponding author.
E-mail address: u19401958@tuks.co.za (N.T. Msweli).

https://doi.org/10.1016/j.ssaho.2023.100628
Received 13 April 2023; Received in revised form 10 July 2023; Accepted 12 July 2023
Available online 24 July 2023

2019; Yu & Hu, 2019), recruiting trained and knowledgeable instructors in data science remains a challenge (Msweli, 2023; Paul & Aithal, 2018; Song & Zhu, 2016). Also, the foundational approaches applied in teaching data science are not flexible enough to adapt to changes and quickly fill the requisite skills gap (Donoghue et al., 2021). The other challenge is the limited information concerning how to teach students in this discipline. For instance, students usually struggle to understand real-time data, and advanced statistical models and algorithms (Liu & Wei, 2020). This poses constraints on democratising DSE because instructors cannot disseminate the knowledge at the expected level which cover various aspects of data science (Price & Ramaswamy, 2019), and produce data scientists who are experts in the field. The majority of these challenges can be addressed by adopting a TPACK (technological pedagogical content knowledge) to assess the level of competency among instructors (Koh & Chai, 2016). TPACK has been trending in the educational technology context (Tseng et al., 2020), however, it has not been studied within the field of data science (Mike, 2020). Improving the skills of academic staff will allow full data science participation, especially among higher institutions of learning, where students are looking forward to being part of the workforce or are already practicing in the field. Thus, it is the instructor's role to guide students and demonstrate to them how to navigate the existing data science technologies, and introduce them to the learning strategies they can use for learning future technologies (Donoghue et al., 2021).

Continued assessment of instructors' TPACK will support and improve teaching practices, and the development and continued review of data science programmes in line with the changing DSE landscape. DSE is technical and highly technologically focused. While students learn about technology, they must also learn how to use technology to improve the learning process. Several technologies can be adopted when teaching data science (Anderson et al., 2015; Beckman et al., 2021; Kim & Henke, 2021); however, it is important to adopt strategies that will suit the course content and teaching methods (Beckman et al., 2021; Msweli, 2023). Data science instructors with limited skills and knowledge concerning the subjects they teach may be unable to contribute to this academic field. For instance, without the requisite knowledge, these instructors are unable to teach data analytics (i.e. analysis of raw data with the aim of making specific predictions about or deriving conclusions from the data) (Chang et al., 2018; Daniel, 2019; Hernán et al., 2019). To teach algorithms adequately, instructors need specialized expertise, which is distinct from subject knowledge or generic pedagogy (Nijenhuis-Voogt et al., 2021). In addition, they need to be aware of technological tools that may afford students hands-on experience to gain a deeper understanding of skills requirements of the industry. Understanding the data science content and choosing the right pedagogy and technology to use in the classroom is a step toward closing the data skills gap.

For DSE instructors and students to master fundamental skills and knowledge, exposure to data science methods is necessary. Knowledge of data science teaching methodologies, subject matter, and tools, as well as how to combine them for effective learning can contribute towards such exposure. To encapsulate some of the crucial knowledge characteristics needed by facilitators for scholarly integration in their teaching methods and strategies, this study uses the TPACK paradigm. The purpose of this study is to investigate how the TPACK framework is currently integrated into data science curricula and instructional methods. With this objective in mind, the study examines TPACK integration in educational settings, identifying strategies, assessing effectiveness, and exploring challenges and best practices in DSE implementation. It is in this context that the study formulated the following research questions*:*

How does the TPACK framework influence teaching practices in data science?

By conducting this investigation, the study seeks to contribute to the understanding of how TPACK can enhance the quality of DSE and inform future curriculum development and instructional strategies in this field. Such a contribution has the potential to advance the field and ensure that students are well prepared for careers in data science. The remainder of the paper is structured as follows: after the introduction, the study distils prior work concerning the area of interest; thereafter, the study gives an account of the methods adopted to collect study data. Before venturing into a discussion of the data, the study showcases the results obtained using the above-mentioned methods. The paper rounds off with a section on the conclusion, implications, limitations, and areas for further research.

2. Literature review

2.1. Data science instructional content

Historically, DSE has focused attention on STEM (science, technology, engineering, and mathematics) programmes, thus leaving non-science programmes lagging behind (Yadav & Debello, 2019; Twinomurinzi et al., 2022). Despite previous research recommending that programmes should focus on cultivating skills that are less prone to automation such as business understanding and storytelling using data (Donoghue et al., 2021), content that focuses on these concepts is minimal (Dill-McFarland et al., 2021). This is especially true where there is specialisation or domain-specific training; it is still difficult to develop content that is customised for a particular industry (Garmire et al., 2017; Otero et al., 2014). The majority of the programmes lean towards computer science; others are, like (e.g. statistics), without specification of data (Wang et al., 2017). There is a need to understand how data science content from the various technology disciplines can be blended to create a pool of data science skills. Additionally, DSE students should be afforded opportunities to dissect data by understanding the problem, identifying patterns, and presenting evidence that proves the reliability and usefulness of the data patterns (Yu & Hu, 2019). Instructors armed with knowledge of the subject matter are better equipped to understand what the students should be taught and how it should be taught. Additionally, it is anticipated that instructors will also be able to identify specific concepts that make data science simple or complex and how technology can be used to improve learning (Mishra & Koehler, 2008). In a fast-changing and data-driven era, instructors must be prepared to grow with change where they are enabled to reimagine, adjust, improve, and adapt the relevant concepts (Niess, 2011).

2.2. CRoss Industry Standard Process for Data Mining

CRoss Industry Standard Process for Data Mining (CRISP-DM) is a popular framework for data science projects (Wirth & Hipp, 2000). It is also used in DSE programmes as a standard process for executing data science projects. The framework is also beneficial for teaching data science and analytics in many educational programmes, including universities, bootcamps, and online courses (Jaggia et al., 2020). CRISP-DM provides a structured approach to data science curricula that can help students to develop a more systematic and efficient way of working. It further provides a structured and standardized approach to data mining that can be applied to a wide range of data science problems.

2.3. Pedagogies in data science educational programmes

The nature of data science imposes a new kind of pedagogy that not only focuses on theoretical and practical training but also incorporates practices for inquiry and interpretation (Mike, 2020). A combination of pedagogies may be necessary to deliver data science instructional programmes, considering the transdisciplinary nature of concepts (Asamoah et al., 2020; Twinomurinzi et al., 2022). For instance, Yadav and Debello (2019) recommend various practices to teach Python as part of a data science course. Metcalf et al., (2016) also propose a set of pedagogical strategies to teach ethics in data science. The call for transdisciplinary pedagogy is a topic of growing interest (Asamoah et al., 2020). Based on the level of difficulty, data science instructors need to be familiar with the details of the content they teach and be able to apply proper teaching practices. Dubey and Gunasekaran (2015) argue that, if education is not backed by appropriate training, the learning outcomes may not translate into desired skills. However, topics on the application or adoption of pedagogical approaches in DSE remain minimal (Saltz & Heckman, 2015). Extant literature only focuses on teaching the technical part of data science, where project-based learning is framed as a suitable pedagogy for DSE (Kim et al., 2021; Saltz & Heckman, 2015; Donoghue et al., 2021; Schwab-McCoy et al., 2021), despite its challenges in various domains (Mike, 2020). Instructors already in STEM faculties, are at an advantage because they have experience in teaching complex concepts and some of the modules may already be part of the data science curriculum. However, it may be difficult to teach a student how the product domain concepts fit into the whole picture. It may also be a challenge to teach data science concepts to non-science students (Garcia-Algarra, 2020; Gil, 2014; Price & Ramaswamy, 2019; Sulmont et al., 2019).

2.4. Integration of technology into data science education

Integration of technology into teaching is not about technology, but relates to the content and effective instructional methods (Thomas et al., 2013). Game-based platforms (gamification or animation), cloud-based virtual labs, Github, and interactive learning platforms are some of the innovative tools that are used to teach data science (Beckman et al., 2021; Graux et al., 2021; Saddiqa et al., 2021). Technology provides an innovative environment for disseminating material and putting principles into action in more effective ways, however, focus is still placed on the curriculum, teaching, and learning. Incorporating technology for effective teaching of data science programmes has been reported (Saddiqa et al., 2021). Digital technologies are particularly useful for teaching and learning purposes in a sense that students can have a simulated data science project that portrays real-world scenarios, while working with fellow students (Anderson et al., 2015; Graux et al., 2021; Kim & Henke, 2021). Resource sharing, such as datasets and other learning material, including instructional videos, is also possible (BHEF, 2017; Saltz & Heckman, 2015; Van Dusen et al., 2019). Therefore, importance should be placed on how and why technology is used, rather than the type of technology. The challenge might be the adoption of technology among instructors or their beliefs and attitudes toward technology (Ertmer et al., 2012). For instance, instructors without the technological skills will find it challenging to adopt technology for teaching and learning purposes. Some strategies such as gamification have been proposed and adopted for a better understanding of data science (Garcia-Algarra, 2020; Hee et al., 2016). Interactive and integrative platforms are useful when introducing students to complex concepts and improving their success rate (Anderson et al., 2015; Hee et al., 2016).

2.5. Technological pedagogical content knowledge

TPACK results from the work of Mishra and Koehler (2008) expands on pedagogical content knowledge framework by incorporating the element of technological knowledge into the paradigm (Graham, 2011). The framework examines practices that can simplify learning to suit student needs and those of the course being taught. The framework effectively presents three domains, namely TK (technological knowledge), PK (pedagogical knowledge), and CK (content knowledge). Using the best pedagogy and technology available, this combination enables teachers to provide engaging course material to students (Mishra & Koehler, 2006). Table 1 summarises the knowledge components of TPACK.

The use of a framework has been recommended for planning, developing, and delivering student-centered educational activities and for creating virtual or simulated learning platforms (Salas-Rueda, 2020), in line with the current trends in DSE. To the best of our knowledge, there is limited research on the application of TPACK in DSE, especially in institutions of higher learning (Kim et al., 2021). Few studies have applied TPACK in STEM modules, such as maths (Salas-Rueda, 2020), artificial intelligence (AI) (Kim et al., 2021), and science (Jang & Chen, 2013). The use of TPACK can contribute to the establishment of interventions for accomplishing the needed knowledge, skills, and capabilities, and professional human capacity development (Scott & Nimon, 2021).

Table 1
Summary of TPACK from Mishra and Koehler (2008).

Knowledge Component	Description
TK	Knowledge about technologies and digital tools such as blackboard.
CK	Instructors must possess a deep understanding of the core concepts and techniques in DS, including data manipulation, statistical modeling, and Machine Learning (ML).
PK	Instructors must have a strong understanding of effective teaching strategies and how to design and deliver engaging and interactive lessons that cater for different learning styles.
PCK	In pedagogical content knowledge (PCK), pedagogical and content knowledge converge. The main issue when examining the link between content and pedagogy is how disciplines differ from one another and if disciplines can or should be taught using the same teaching methods.
TCK	The main goal of technological content knowledge (TCK) relies on understanding how content and technology impact and constrict one another. Apart from the topic matter they teach, facilitators must have a thorough awareness of how the use of technology might alter the subject matter.
TPK	Technological pedagogical knowledge (TPK) refers to knowledge of pedagogical affordances and constraints of a range of technological tools, which relate to disciplinarily and developmentally appropriate pedagogical designs and strategies.
TPACK	To improve teaching and learning, instructors must be able to use technology effectively. This entails utilizing interactive platforms and tools, producing interesting multimedia material, and making use of internet resources.

3. Methodology

3.1. Research instrument

This study adopted TPACK as a framework to understand the ability of instructors to teach data science with the aid of technology. The study adapted the research instrument developed by Elas et al. (2019). The instrument was updated to align the questionnaire with the CRISP-DM process model, a widely recognized and industry-standard framework for data science projects. By incorporating elements of CRISP-DM into the questionnaire, we aimed to capture relevant aspects and practices specific to DSE. The adaptation of the previously validated questionnaire was to enhance the validity and reliability of the study's findings. To establish whether it meets the study objectives, the new questionnaire was reviewed by two educational specialists in the field of data science. The questionnaire contained 31 items for measuring instructors' confidence in the seven TPACK constructs. After administering the questionnaire online, the collected data was statistically analysed using IBM SPSS AMOS version 21. The analysis encompassed various aspects, including demographic results of the participants and other relevant descriptive statistics. In addition to the descriptive statistics, cross-tabulation analysis was conducted on the collected data. These analyses allowed for the examination of the relationships between two or more categorical variables.

3.2. Participants

The study purposefully recruited 26 data science instructors as study's participants, through business networks, colleagues, and referrals. The recruitment of participants in this study was guided by the need for expertise, diversity, and active involvement in data science instruction, ensuring a comprehensive exploration of the challenges and strategies associated with teaching data science programmes. The researcher successfully applied for ethical clearance from the university research committee to collect data. All participants gave consent to participate in the study. Data was collected over a 2-month period during the 2022 academic year. Descriptive statistical analysis and cross-tabulations were conducted to analyse data.

3.3. Demography of participants

A brief demographic analysis of the participants covering the gender, highest qualification, level of data science qualification being taught by the instructor, and years of experience category is presented in Table 2.

Table 2
Demographic information of selected participants.

No	Gender	Highest qualification	Years of experience	DS programme taught
1	Male	Bachelor's degree	+10 years of experience	Undergraduate level
2	Male	Other	1–5 years of experience	SLP
3	Female	Vocational training	Less than 1 year	SLP
4	Female	Master's degree	1–5 years of experience	Postgraduate level
5	Male	Master's degree	1–5 years of experience	Undergraduate level
6	Male	Honours degree	+10 years of experience	Undergraduate level
7	Male	Honours degree	+10 years of experience	Postgraduate level
8	Male	Honours degree	+10 years of experience	Undergraduate level
9	Male	Master's degree	1–5 years of experience	Undergraduate level
10	Male	Bachelor's degree	Less than 1 year	SLP
11	Male	Bachelor's degree	+10 years of experience	SLP
12	Male	Other	+10 years of experience	SLP
13	Male	Honours degree	+10 years of Experience	Undergraduate level
14	Male	Honours degree	+10 years of Experience	Postgraduate level
15	Male	Honours degree	+10 years of Experience	Postgraduate level
16	Male	Master's degree	+10 years of Experience	Undergraduate level
17	Female	Master's degree	+5 years of Experience	Postgraduate level
18	Male	Bachelor's degree	+10 years of Experience	Undergraduate level
19	Male	Master's degree	Less than 1 year	Postgraduate level
20	Male	Honours degree	+5 years of Experience	Undergraduate level
21	Female	Honours degree	+5 years of Experience	SLP
22	Male	Bachelor's degree	Less than 1 year	Undergraduate level
23	Male	Master's degree	+10 years of Experience	SLP
24	Female	Master's degree	1–5 years of Experience	SLP
25	Male	Master's degree	+5 years of Experience	Undergraduate level
26	Male	Master's degree	1–5 years of Experience	Postgraduate level

Based on Table 2, the overwhelming majority (81%) of instructors are male, while 19% of the instructors are female (see Table 2). This indicates a gender disparity among instructors in the field of data science, with a significantly higher proportion of male instructors compared to female instructors. This gender disparity raises concerns about diversity and highlights the need for efforts to address this imbalance. The results indicate that a significant proportion of data science instructors have a master's degree (39%) as their highest qualification. This suggests that they have attained an advanced level of education in their respective fields, which can contribute to their expertise and knowledge in data science. Additionally, 31% of instructors have honours degrees, indicating a substantial portion with a strong foundational education. The presence of instructors with diverse qualifications, including master's degrees and honours degrees, can bring a variety of perspectives and expertise to the classroom. Different educational backgrounds can lead to varied approaches to teaching data science concepts, enriching the learning experience for students. The finding that about 4% of the instructors' qualifications could not be established raises the importance of proper qualification verification processes. It is crucial for educational institutions to ensure that instructors possess the necessary qualifications and expertise to teach data science effectively. Proper verification processes help maintain the credibility and quality of DSE. Based on the result, the significant percentage of instructors with more than 10 years of teaching experience in data science indicates a wealth of pedagogical expertise, subject matter mastery, adaptability, and potential for mentorship. The instructor's experience contributes to the quality of instruction, fosters innovation, and promotes continuity in the academic field of data science. According to the results, less than half (42%) of the instructors teach data science at an undergraduate level. It is worth noting that the distribution of instructors across different levels may vary based on factors such as institutional context, program offerings, and the specific goals of DSE initiatives. Understanding this distribution can help inform curriculum development, faculty hiring decisions, and the design of data science educational programmes.

4. Analysis and results

4.1. Results on central tendency

To determine how centered the distribution of the study's constructs is, central tendency measurements were used. A five-point Likert scale was used to measure the level of agreement of the participants with a particular statement. Numerical values ranging from 1 (denotes "Strongly disagree") and 5 (denotes "Strongly agree") were used to measure the attitude of the participants towards the constructs under investigation, namely: CK, PK, TK, PCK, TCK, TPK, and TPACK. The average mean for all constructs was 4.40. The instructors answered all the questions; the mean and standard deviation are reported for each subscale in Table 3.

4.1.1. Summary of descriptive statistics

CK (4.12) is high and very close to PK (4.13). This could mean that instructors of data science are familiar with the content that need to be

Table 3
Summary of descriptive statistics.

Construct	Number of items	Number of responses	Mean scores	Standard deviation
CK	6	26	4.12	.812
PK	7	26	4.35	.508
TK	4	26	4.49	.472
PCK	3	26	4.13	.811
TCK	2	26	4.35	.596
TPK	3	26	4.46	.574
TPACK	6	26	4.20	.769

featured in data science educational programmes, and by extension the teaching practices that are deemed appropriate to deliver lessons. However, these ratings slightly lower than those of other TPACK measured constructs (i.e., TK, TPK). The lower ratings suggest that instructors require additional support to effectively teach data science concepts to groups of students. It is important to note that data science is a complex and rapidly evolving field, and teaching it can be challenging due to its interdisciplinary nature and the need for a good understanding of various technical concepts being taught.

Based on the results presented herein, it appears that instructors have an equal level of understanding of both PCK and TCK; a rating score of 4.35 was achieved for both constructs. This score suggests that instructors have a solid grasp of data science concepts as well as a good understanding of how technological tools impact the subject matter. PCK refers to an understanding of how to effectively teach specific subject matter to students. In the context of data science, it involves knowing how to structure and deliver lessons, design learning activities, and assess student understanding in a way that promotes effective learning of data science concepts. TCK, on the other hand, refers to the understanding of the technological tools and resources relevant to the subject matter. In the case of data science, this involves knowledge of programming languages, statistical software, data visualization tools, and other technologies commonly used in the field. The equal rating scores achieved for both types of knowledge indicate that instructors possess a balanced understanding of the pedagogical aspects of teaching data science and the technological tools required to support the subject matter effectively. This is a positive finding as it suggests that, not only are instructors are equipped to teach the content, they also leverage on appropriate technologies to enhance the learning experience for their students.

Interestingly, the rating scores of TPK (4.46) and TK (4.49) are almost similar. This suggests that instructors have a good understanding of both constructs. This is a positive finding because it indicates that, not only are instructors familiar with the technological tools used in data science, they also know how to incorporate them effectively into their teaching practices. This way, instructors become well prepared to create dynamic and engaging learning environments that leverage technology to support student learning and mastery of data science concepts.

The overall TPACK mean score of 4.20 suggests that instructors, on average, have a solid understanding of how to integrate technology into their teaching practices while effectively conveying CK in the field of data science. This means that they are likely to be proficient in selecting and utilizing appropriate technological tools, designing engaging learning activities, and effectively integrating technology to support student learning and achievement. However, it is important to note that there might still be some variation among individual instructors. Some instructors may excel in specific aspects of TPACK while others may require further support or development in certain areas. Ongoing professional development opportunities, collaboration with colleagues, and staying updated with advancements in both technology and pedagogy can enhance instructors' TPACK further.

4.2. Cross-tabulation results

Cross-tabulation analysis was conducted to explore the relationship between the concepts of CRISP-DM and the level of qualification at which they are primarily taught. The objective was to gain insights into how the different phases or components of CRISP-DM are associated with specific levels of qualification in DSE.

4.2.1. Cross-tabulation between business understanding and the level of data science qualification

Table 4 reveals that 50% of instructors teach business understanding at the postgraduate level, as opposed to 80% of instructors teaching at undergraduate level. This indicates that the majority of instruction in business understanding is indeed taking place at undergraduate level.

Teaching business understanding at undergraduate level aligns with the typical educational pathway for students starting their studies in business or related fields. Undergraduate programmes often provide foundational knowledge and skills in business concepts, principles, and practices. Therefore, it is common for a larger number of instructors to be engaged in teaching business understanding at this level.

4.2.2. Cross-tabulation between data understanding and the level of DS qualification

The data presented in Table 5 shows that a very large proportion of instructors (83%) involved in teaching data understanding teach it at undergraduate level. This suggests that the majority of instruction in data understanding is taking place at the undergraduate level where students are typically pursuing their bachelor's degrees. In contrast, a very small percentage of instructors (39%) is engaged in teaching data understanding at the postgraduate level.

Looking at the percentage (39%) of instructors engaged in, this could mean that there is still a number of instructors teaching data understanding at other qualification levels. This distribution of instructors may reflect the educational structure of the institution or the specific focus and requirements of the data understanding curriculum. It is noteworthy that these percentages represent the sample of instructors in the study and may not be generalized to the entire population of data understanding instructors.

4.2.3. Cross-tabulation between data preparation and the level of data science qualification

The results presented in Table 6 indicate that 64% of the instructors that teach at the undergraduate level always teach data preparation. This leads to the belief a fair majority of instructors consistently include data preparation as part of their curriculum at undergraduate level.

Teaching data preparation at the undergraduate level equips

Table 4
Cross-tabulation of business understanding and data science qualification.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Business Understanding	Often	Count	1	3	4	8
		% within Business Understanding	12.5%	37.5%	50%	100%
		% within level of DS qualification	12.5%	27.3%	57.1%	30.8%
		% of Total	3.8%	11.5%	15.4%	30.8%
	Always	Count	1	4	0	5
		% within Business Understanding	20%	80%	0%	100%
		% within level of DS qualification	12.5%	36.4%	0%	19.2%
		% of Total	3.8%	15.4%	0%	19.2%
Total		Count	8	11	7	26
		% within Business Understanding	30.8%	42.3%	26.9%	100%
		% within level of DS qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

Table 5
Cross-tabulation between data understanding and data science qualification.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Data understanding	Often	Count	4	4	5	13
		% within Data Understanding	30.8%	30.8%	38.5%	100%
		% within level of DS qualification	50%	36.4%	71.4%	50%
		% of Total	15.4%	15.4%	19.2%	50%
	Always	Count	1	5	0	6
		% within Data Understanding	16.7%	83.3%	0%	100%
		% within level of DS qualification	12.5%	45.5%	0%	23.1%
		% of Total	3.8%	19.2%	0%	23.1%
Total		Count	8	11	7	26
		% within Data Understanding	30.8%	42.3%	26.9%	100%
		% within level of DS qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

Table 6
Cross-tabulation between data preparation and level of data science qualification.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Data Preparation	Often	Count	5	2	4	11
		% within Data Preparation	45.5%	18.2%	36.4%	100%
		% within level of DS qualification	62.5%	18.2%	57.1%	42.3%
		% of Total	19.2%	7.7%	15.4%	42.3%
	Always	Count	2	7	2	11
		% within Data Preparation	18.2%	63.6%	18.2%	100%
		% within level of DS qualification	25%	63.6%	28.6%	42.3%
		% of Total	7.7%	26.9%	7.7%	42.3%
Total		Count	8	11	7	26
		% within Data Preparation	30.8%	42.3%	26.9%	100%
		% within level of DS qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

students with the foundational knowledge and practical skills needed to work with real-world datasets. Moreover, it helps students understand the significance of data quality, data cleaning techniques, and the importance of preparing data for analysis to derive meaningful insights from the data. By including data preparation in the curriculum, instructors prepare students for the data-driven nature of various industries and thus help them to become proficient and more effective in working with data. This knowledge is vital for data analysts, data scientists, and professionals in related fields who need to navigate and manipulate data to extract valuable information and make informed decisions.

4.2.4. *Cross-tabulation between business understanding and the level of data science qualification*

Based on the results reported in Table 7, at least 46% of the instructors that teach at the undergraduate level also offer teach data modeling modules. It can therefore be deduced that a significant proportion of instructors at undergraduate level incorporate data modeling in their curriculum.

The percentage of instructors offering teaching data modeling indicates the recognition of data modeling as an essential component of undergraduate data science education. By emphasizing data modeling, instructors prepare students to become proficient in structuring and modeling data, thus enabling them to extract and derive meaningful insights from diverse datasets in their future careers.

4.2.5. *Cross-tabulation between model evaluation and the level of data science qualification*

As shown in Table 8, a percentage (50%) of instructors who teach SLPs rarely focus on explicit instruction and instead, place emphasis on model evaluation. The risk here is that instructors are predisposed to prioritizing other aspects of the short courses, such as introducing

Table 7
Cross-tabulation between business understanding and level of data science qualification.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Data Modeling	Often	Count	4	5	2	11
		% within Data Modeling	36.4%	45.5%	18.2%	100%
		% within level of data science qualification	50%	45.5%	28.6%	42.3%
		% of Total	15.4%	19.2%	7.7%	42.3%
	Always	Count	4	4	4	12
		% within Data Modeling	33.3%	33.3%	33.3%	100%
		% within level of data science qualification	50%	36.4%	57.1%	46.2%
		% of Total	15.4%	15.4%	15.4%	46.2%
Total		Count	8	11	7	26
		% within Data Modeling	30.8%	42.3%	26.9%	100%
		% within level of data science qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

Table 8
Cross-tabulation between model evaluation and level of data science qualification.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Model Evaluation	Rarely	Count	4	3	1	8
		% within Model Evaluation	50%	37.5%	12.5%	100%
		% within level of data science qualification	50%	27.3%	14.3%	30.8%
		% of Total	15.4%	11.5%	3.8%	30.8%
	Often	Count	3	4	3	10
		% within Model Evaluation	30%	40%	30%	100%
		% within level of data science qualification	37.5%	36.4%	42.9%	38.5%
		% of Total	11.5%	15.4%	11.5%	38.5%
Total		Count	8	11	7	26
		% within Model Evaluation	30.8%	42.3%	26.9%	100%
		% within level of data science qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

fundamental concepts, providing hands-on experience with data analysis tools, or focus on specific skills or applications within a limited timeframe. On the other hand, the data suggests that the low percentage of instructors teaching at the undergraduate level (40%) often teach model evaluation. This implies that within undergraduate programmes, greater emphasis is placed on introducing and instructing students on the importance and techniques of evaluating predictive models.

Model evaluation is a crucial aspect of statistical learning and ML. It helps data scientists to assess the performance and generalization capabilities of predictive models and aids in selecting the most suitable model for a given task. Understanding model evaluation is essential for students pursuing careers in data science or related fields because it equips them with the skills to critically assess the quality and reliability of their models. While results of this study indicate that model evaluation may not be frequently taught in SLPs or at the undergraduate level, it is still important to consider the overall curriculum and ensure that students receive exposure to the principles and techniques of model evaluation. Instructors and curriculum designers can explore ways to incorporate model evaluation in the curriculum, even if it is in a condensed or simplified form for SLPs. This could involve introducing key evaluation metrics, discussing best practices, or providing practical examples so that students can understand the concept better.

4.2.6. Cross-tabulation between deployment and the level of data science qualification they teach

The results in Table 9 indicate that 80% of the instructors that teach at the undergraduate always teach deployment. The high hign number of instructors teaching both deployment at undergraduate level testify to a strong focus on preparing students for the real-world implementation and utilization of their data science knowledge.

By emphasizing deployment, instructors aim to bridge the gap between theoretical understanding and practical application, equipping students with the skills needed to deploy their models effectively. Teaching deployment at the undergraduate level may help students understand the considerations, challenges, and best practices associated with deploying models in various settings. Curriculum on deployment topics may include model packaging, integration with software systems, scalability, performance optimization, and monitoring. By providing instruction on deployment, instructors enable students to understand the end-to-end process of taking a model from development to production.

5. Discussion

Assessing someone's teaching knowledge involves considering their qualifications, experience, expertise, and teaching approach (Niess, 2011). However, assessing an instructor's knowledge requires considering these elements in conjunction with each other. For instance, a qualified and experienced instructor with expertise in the subject matter might be ineffective if their teaching approach does not align with the needs and learning styles of their students (Jafar et al., 2016). Conversely, a teacher with a well-aligned approach but lacking in qualifications or expertise might struggle to deliver accurate and comprehensive instruction. Essentially this demands a probe into data science instructors' teaching knowledge. While the instructor TPACK aspects have been explored in other STEM areas (Başaran, 2020; Doukakis et al., 2021), research on data science has not investigated this area. This suggests a research gap in the field of DSE that needs to be attended to as the demand for data science skills continues to grow.

Instructors' gender is also noted as a factor that needs to be considered (Saeli et al., 2011; Spieler et al., 2019). Prior research has found men to be more technical and accustomed to technology, while women struggle with technology (Taopan et al., 2020). Even though the teaching profession often attracts women over men (Ambusaidi & Al-Maqbali, 2022), it is different when it comes to teaching technical modules. The underrepresentation of women in data science and other technical field is a well-recognized challenge in the field (Blake, 2019). The challenges start when few female students enrol for STEM programmes like data science (Rao et al., 2019), this continues to be an issue where organisation wants to appoint women but the number of women candidates are remarkably low (Spieler et al., 2019). This was noted in this study, where few data science instructors could be reached.

Table 9
Cross-tabulation between deployment and the level of data science education.

			Level of data science qualification			Total
			SLP	Undergraduate level	Postgraduate level	
Deployment	Always	Count	1	4	0	5
		% within Deployment	20%	80%	0%	100%
		% within level of data science qualification	12.5%	36.4%	0%	19.2%
		% of Total	3.8%	15.4%	0%	19.2%
Total		Count	8	11	7	26
		% within Deployment	30.8%	42.3%	26.9%	100%
		% within level of data science qualification	100%	100%	100%	100%
		% of Total	30.8%	42.3%	26.9%	100%

Despite efforts to promote gender diversity and inclusion such as Women in ML (WILM, 2023), and Women in Data Science (WIDS, 2023), there is still a gender gap in data science-related roles. This gender disparity can be attributed to various factors, including societal stereotypes, lack of representation and encouragement at early educational stages, and unconscious biases in hiring and promotion processes (Ambusaidi & Al-Maqbali, 2022; Taopan et al., 2020). Mbwilo et al. (2019) have also reported lack of STEM education in women, lack of mentorship programmes for women in data science, and policy that support gender balancing initiatives as potential contributors to the gender disparity. Efforts should be made to encourage and support women in pursuing careers in data science and related disciplines; this includes research and collaboration opportunities.

In the recent years, a number of studies on data science programme which places emphasis on their curriculum have been published, however, there has been little attention paid to the pedagogy of those programme (Mike, 2020). Data science is flagged as a challenging qualification to teach due to its interdisciplinary nature and its complexity (Sulmont et al., 2019). Given the challenges associated with teaching data science, it is essential for instructors to possess a strong foundation in the field, stay updated with the latest trends, and continually enhance their pedagogical skills to effectively teach this complex subject matter. Mikroyannidis et al. (2018) recommended agility in data science instructors to adjust learning material for suitability of diverse group of students as when data science landscape changes. This is in addition to applying technological tools that simplify learning around concepts that are deemed difficult. The challenge is that technology changes, therefore, those involved in teaching may need to learn these new technologies and be aware of the influence they have over a planned lesson. However, Ertmer and Ottenbreit-Leftwich (2013) advised that rather than the technology itself, emphasis should be placed on how it affects learning. Kim and Henke (2021) further advised that instructors should consider technologies that are easy for everyone to use.

Based on the assumption that many data science instructors have taken technology-related courses, it is expected that they will be knowledgeable and experienced in technological tools. This was noted in this study. However, academic departments have indicated that teaching a variety of technical concepts involved in data science is one of the biggest challenges they face, especially where there is diverse group of students from various disciplines (Schwab-McCoy et al., 2021). By experimenting with various teaching techniques and technological tools, instructors can adjust their teaching strategies to align with the diverse behaviours of their students. It has been suggested that integrated learning platforms be used, particularly when teaching machine learning for the model-training process (Kim et al., 2021; Yan & He, 2020). Despite integrated learning platforms being encouraged, literature is relatively silent on the direct impact of data science automation on data science skills. While this may be the case, it is important to recognize that automation is not a replacement for the skills and expertise of data scientists. Rather, automation can be seen as a complementary tool that enhances productivity and efficiency in certain aspects of data science workflows (Uzunalioglu et al., 2019). It is crucial for DSE to achieve a balance between imparting theoretical knowledge and practical skills while also acknowledging the place of automation in the field. Essentially, data science instructors need to be well-versed in the implications of automation in data science workflows. This includes adapting their teaching methods and materials to incorporate automation and staying abreast of the latest developments to prepare students for their data science careers.

The interdisciplinary nature of data science is one of the key factors contributing to its effectiveness in solving complex problems and extracting meaningful insights from data. However, it is difficult to design programmes that speak to these factors (Mike, 2020; Sulmont et al., 2019). Student interests together with their prior knowledge influence the way data science curriculum is built and taught (Hagen et al., 2019). The area is new and points of reference are limited concerning how to teach students data science to diverse students (Sulmont et al., 2019). For instance, non-science students may not see the relevance of certain concepts of programming and mathematics in data science (Garcia-Algarra, 2020). Addressing their specific challenges, making the subjects relatable, and providing support can help non-science students overcome the initial difficulties and develop a solid foundation for data science. Data science programmes such as CRISP-DM may provide a structured and standardized approach for data science programmes, and also determine the knowledge level that students should acquire at different levels.

With regards to data science content, the structural benefit that CRISP-DM provides in data science curricula has the potential to solve many inconsistencies that are currently experienced in the academic field of data science. Even though prior research indicates that it might be difficult to teach some of the data science concepts such as model deployment (Davenport & Malone, 2021; Song & Zhu, 2016), the use of collaborative tools or integrated technologies may be a solution (Garcia-Algarra, 2020). Only knowledgeable instructors will know how to adjust their teaching methods to incorporate concepts that are deemed difficult to teach especially in a natural setting.

Data science instructors may have realised that teaching and learning within data science disciplines can be more meaningful when technology is incorporated into the classroom. This realisation might include knowing how each technology improves the learning process and how it supports the content being taught, as well as the limitations it may have on specific pedagogical strategies. To support data science instructors on their journey, TPACK becomes a useful construct (Sulmont et al., 2019). However, there is a need to understand how it can be applied in transdisciplinary areas of learning (Mike, 2020). To illustrate this point, TPACK geared towards improving mathematical studies and ML (programming and algorithms) have been investigated as separate subjects (Doukakis et al., 2021; Salas-Rueda, 2020). However, computer science education has been studied as a whole discipline even though the main focus was on programming (Doukakis & Papalaskari, 2019). Based on this, one can assume that a need exists to identify concepts that are hard to teach within data science so that developmental training programmes for instructors can be identified. Prior studies have already indicated that having sufficient knowledge independently is not enough for technology integration into DSE (Ertmer & Ottenbreit-Leftwich, 2013).

Despite previous studies highlighting some challenges concerning teaching data science, for instance, access to and use of open data (Saddiqa et al., 2021), TPACK lacks the capability of determining which technologies are instructors competent in, that could support DSE. This study could not confirm which teaching technologies the instructors are familiar with. Literature has covered a few learning technologies that support data science programmes; however, their effectiveness remains unknown. The adoption and use of technology in teaching data science still need to be explored. Few of the instructors are aware of pedagogical techniques that use technology to teach data science content. This can be resolved by providing training on pedagogical approaches and how they can be used with technology.

Instructors have a role to play in preparing future data scientists for new and emerging technologies. One of the challenges to achieving this is that these instructors are not fully conversant with the content of these technologies, and some have no experience in using modern technologies. TPACK provides institutions with a framework that can be used as a self-assessment tool for data science instructors. This information is essential when preparing strategies for continuous professional development (Doukakis et al., 2021). However, discussions on what competencies instructors need for teaching data science are sparse (Kim et al., 2021).

The different elements of data science need to be fostered in DSE. Details of each element consist of concepts, techniques, competencies, and skills that are popular within the corporate world. Thus, the structuring and development of data science programmes should be both controlled and practical since these elements differ from one another. For instance, the study argued for a need to apply different or combinations of teaching pedagogies for each data science element where necessary. For instance, modeling algorithms cannot be taught using the same teaching pedagogies that are used to teach business understanding. Due to the transdisciplinary nature of data science, a variety of teaching approaches are employed to effectively teach different aspects of the field. What counts the most is the audience being taught. For instance, ML is difficult to teach to non-science students, but it is not impossible, as far as Sulmont et al. (2019) are concerned. Instructors must just understand the best ways to use technology to convey meaning (Beckman et al., 2021).

6. Conclusion, implications, limitations, and areas for further research

This study aimed to gain insights into instructors' skills and competencies when teaching data science through integration of technology. For this purpose, the data was collected from 26 instructors teaching data science at different learning institutions. The immediate findings of this study show that there are fewer female instructors in the data science educational field. By addressing the gender gap in DSE, we can work towards creating a more balanced and representative data science community. This will not only benefit individuals but also contribute to the advancement of the field as a whole, since diverse perspectives and experiences foster innovation and drive meaningful change.

Educational institutions have responded to the increase in demand for data science skills by providing data science programmes to fulfil the demands of both students and industry. Due to the field's infancy, research in DSE is necessary to address a number of issues and guarantee the efficient delivery of data science programmes. New instructors will therefore be better prepared when teaching data science, especially to non-science students, by defining what current instructors find easy and difficult to teach. This will allow new instructors to be more prepared for any support that may arise. Granting the fact that current instructors have an acceptable level of TPACK, it remains difficult to identify the technologies instructors are familiar given the ever evolving nature of technology. However, this denotes a good foundation for DSE where instructors can create meaningful learning experiences for students in data science. It further implies that learning institutions may effectively bridge the gap between theoretical concepts and practical applications, thus ensuring that students gain a comprehensive understanding of data science principles and develop the necessary skills to work with data.

Notwithstanding the fact that data science programmes are there, there is still a need to find ways to improve several aspects of DSE including curriculum development, instructional methods, assessment strategies, and the integration of data science across different disciplines. In addition, learning institutions need to invest adequate resources in initiatives that improve teaching complex data science concepts and those that are practical in nature such as model evaluation and deployment.

Despite the best efforts towards conducting this research study, few limitations are noted. Firstly, the study was conducted in a developing region, where the exploration of data science has not been fully explored. As a result, there are few learning institutions that offer data science programmes albeit with limited teaching staff. This has resulted in a smaller study sample. To illustrate, Twinomurinzi et al. (2022) found that only twelve universities offer data science programmes in South Africa, with only five programme offerings at undergraduate level. Therefore, the outcomes of this study may not be sufficiently generalizable, thus necessitating further investigation into data science instructors. Secondly, there are no items in the TPACK framework that refer to any particular data science subject topic, expertise, or practice. Data science is transdisciplinary; thus, the content knowledge and how it is measured might vary from one discipline to the next. Essentially, there is a need to conceptualise the TPACK framework, specifically for data science educational programmes, and understand the influence it has in each construct. The last limitation relates to the fact that only quantitative instruments were used in the current investigation. This emerging area of research can benefit from other research methods such as qualitative methods.

It is recommended that future studies focus on determining the impact technology has on data science pedagogical approaches, and what this means for those teaching within this discipline. Studies on professional development for data science instructors are also recommended. In doing so, it will be ensured that learning, developing, and experience-gaining will all be continual processes throughout the teaching of data science. It is further recommended that guidelines be formed to guide the initiating and development of DSE and thus increase support resources that are needed such as but not limited to teaching practice and technological tools. Professional bodies are therefore necessary to advance and govern knowledge and practices in teaching data science. It is also advisable to examine the policies that support such efforts to ensure that adequate resources are allocated.

CRediT authorship contribution statement

Nkosikhona Theoren Msweli: Conceptualization, Conducted data collection and, Formal analysis. **Tendani Mawela:** Reviewed and approved the data collection instrument, Supervision. **Hossana Twinomurinzi:** Reviewed and approved the data collection instrument, Supervision, All authors discussed the results and contributed to the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ssaho.2023.100628.

References

Ambusaidi, A. K., & Al-Maqbali, F. Y. (2022). Exploring pedagogical decision making from the lens of science teachers in response to different pedagogical issues. *Social Sciences & Humanities Open, 5*(1), Article 100236. https://doi.org/10.1016/J.SSAHO.2021.100236

Anderson, P. E., Turner, C., Dierksheide, J., & McCauley, R. (2015). An extensible online environment for teaching data science concepts through gamification. In *Proceedings - frontiers in education conference, FIE*. https://doi.org/10.1109/FIE.2014.7044205

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

Başaran, B. (2020). Investigating science and mathematics teacher candidate's perceptions of TPACK-21 based on 21st century skills. *Elementary Education Online, 19*(4), 2212–2226. https://doi.org/10.17051/ilkonline.2020.763851

Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with git and GitHub as a learning objective in statistics and data science courses. *Journal of Statistics and Data Science Education, 29*(1), 132–144. https://doi.org/10.1080/10691898.2020.1848485

BHEF. (2017). *Investing in America's data science and analytics talent (Issue April)*.

Blake, A. (2019). *Dynamics of data science skills: How can all sectors benefit from data science talent?*.

Chang, C.-L., McAleer, M., & Wong, W.-K. (2018). Big data , computational science , economics , finance , marketing , management , and psychology : Connections. *Journal of Risk and Financial Management Review, 11*(15). https://doi.org/10.3390/jrfm11010015

Daniel, B. K. (2019). Big data and data science: A critical review of issues for educational research. *British Journal of Educational Technology, 50*(1), 101–113. https://doi.org/10.1111/bjet.12595

Davenport, T., & Malone, K. (2021). Deployment as a critical business data science discipline. *Harvard Data Science Review, 3*(1). https://doi.org/10.1162/99608f92.90814c32

Dill-McFarland, K. A., Konig, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y., & Hallam, S. J. (2021). An integrated, modular approach to data science education in microbiology. *PLoS Computational Biology, 17*(2). https://doi.org/10.1371/JOURNAL.PCBI.1008661

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Taylor & Francis, 29*(S1), 27–39. https://doi.org/10.1080/10691898.2020.1860725

Doukakis, S., & Papalaskari, M. A. (2019). *Scaffolding technological pedagogical content knowledge (TPACK) in computer science education through learning activity creation. 2019 4th South-East Europe design automation, computer engineering, computer Networks and social media conference*. https://doi.org/10.1109/SEEDA-CECNSM.2019.8908467. *SEEDA-CECNSM 2019, June*.

Doukakis, S., Psaltidou, A., Stavraki, A., Adamopoulos, N., Tsiotakis, P., & Stergou, S. (2021). *Measuring the technological pedagogical content knowledge (TPACK) of in-service teachers of computer science who teach algorithms and programming in upper secondary education.* arXiv preprint arXiv:2105.09252.

Dubey, R., & Gunasekaran, A. (2015). Education and training for successful career in big data and business analytics. *Industrial & Commercial Training, 47*(4), 174–181. https://doi.org/10.1108/ICT-08-2014-0059

van Dusen, E., Suen, A., Liang, A., & Bhatnagar, A. (2019). Accelerating the advancement of data science education. *Proceedings of the Association for Information Science and Technology, 56*(1), 601–603. https://doi.org/10.1002/PRA2.103

Elas, N. I. B., Majid, F. B. A., & Narasuman, S. Al (2019). Development of technological pedagogical content knowledge (TPACK) for English teachers: The validity and reliability. *International Journal of Emerging Technologies in Learning, 14*(20), 18–33. https://doi.org/10.3991/ijet.v14i20.11456

Ertmer, P. A., & Ottenbreit-Leftwich, A. (2013). Removing obstacles to the pedagogical changes required by Jonassen's vision of authentic technology-enabled learning. *Computers in Education, 64*, 175–182. https://doi.org/10.1016/j.compedu.2012.10.008. March 2018.

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers in Education, 59*(2), 423–435. https://doi.org/10.1016/j.compedu.2012.02.001

Garcia-Algarra, J. (2020). *Introductory machine learning for non STEM students.* Proceedings of the European Conference on Machine Learning.

Garmire, L. X., Gliske, S., Nguyen, Q. C., Chen, J. H., Nemati, S., Van Horn, J. D., Moore, J. H., Shreffler, C., & Dunn, M. (2017). *The training of next generation data scientists in biomedicine. Pacific Symposium on biocomputing 2017*. www.worldscientific.com.

Gil, Y. (2014). Teaching parallelism without programming: A data science curriculum for non-CS students. *2014 Workshop on Education for High Performance Computing*, 42–48. http://www.journalism.columbia.edu/page/1058-the-lede-program-an-introduction-.

Graham, C. R. (2011). Theoretical considerations for understanding technological pedagogical content knowledge (TPACK). *Computers & Education, 57*, 1953–1960. https://doi.org/10.1016/j.compedu.2011.04.010

Graux, D., Janev, V., Jabeen, H., & Sallinger, E. (2021). A big data learning platform for the west balkans and beyond. In *Annual conference on innovation and technology in computer science education, ITiCSE, April* (pp. 617–618). https://doi.org/10.1145/3456565.3460026

Hagen, L., Andrews, J., Federer, L., & Benoit, G. (2019). Data science education in library and information science schools. *Proceedings of the Association for Information Science and Technology, 56*(1), 536–537. https://doi.org/10.1002/PRA2.84

Hee, K., Zicari, R. V., Tolle, K., & Manieri, A. (2016). Tailored data science education using gamification. *Proceedings of the International Conference on Cloud Computing Technology and Science, 0*, 627–632. https://doi.org/10.1109/CloudCom.2016.0108

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance, 32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578

Jafar, M. J., Babb, J., & Abdullat, A. (2016). Emergence of data analytics in the information systems curriculum. In *Proceedings of the EDSIG conference*. http://iscap.info.

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM framework for teaching business analytics. *Decision Sciences Journal of Innovative Education, 18*(4), 612–634. https://doi.org/10.1111/dsji.12222

Jang, S.-J., & Chen, K.-C. (2013). Development of an instrument to assess university students' perceptions of their science instructors' TPACK. *Journal of Modern Education Review, 3*(10), 771–783. http://www.academicstar.us.

Kim, B., & Henke, G. (2021). Easy-to-Use cloud computing for teaching data science. *Journal of Statistics and Data Science Education, 29*(S1), 103–111. https://doi.org/10.1080/10691898.2020.1860726

Kim, S., Jang, Y., Choi, S., Kim, W., Jung, H., Kim, S., & Kim, H. (2021). Analyzing teacher competency with TPACK for K-12 AI education. *KI - Kunstliche Intelligenz, 35* (2), 139–151. https://doi.org/10.1007/s13218-021-00731-9

Koh, J. H. L., & Chai, C. S. (2016). Seven design frames that teachers use when considering technological pedagogical content knowledge (TPACK). *Computers in Education, 102*, 244–257. https://doi.org/10.1016/j.compedu.2016.09.003

Liu, Y., & Wei, X. (2020). How to use stock data for data science education: A simulated trading platform in classroom. In *Proceedings of 2nd international conference on computer science and educational informatization, CSEI 2020* (pp. 5–8). https://doi.org/10.1109/CSEI50228.2020.9142534

Loy, A., Kuiper, S., & Chihara, L. (2019). Supporting data science in the statistics curriculum. *Journal of Statistics Education, 27*(1), 2–11. https://doi.org/10.1080/10691898.2018.1564638

Mbwilo, B., Kimaro, H., Justo, G., & Godfrey, J. (2019). Data science postgraduate education at university of dar es salaam in Tanzania: Current demands and opportunities. In *15th international conference on social implications of computers in developing countries (ICT4D)* (pp. 349–360). https://doi.org/10.1007/978-3-030-19115-3_29ï

Mike, K. (2020). Data science education: Curriculum and pedagogy. In *ICER 2020 - proceedings of the 2020 ACM conference on international computing education research* (pp. 324–325). https://doi.org/10.1145/3372782.3407110

Mikroyannidis, A., Domingue, J., Phethean, C., Beeston, G., & Simperl, E. (2018). Designing and delivering a curriculum for data science education across Europe. *Teaching and Learning in a Digital World: Proceedings of the 20th International Conference on Interactive Collaborative Learning –, 2*(2), 540–550. https://doi.org/10.1007/978-3-319-73204-6

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054.

Mishra, P., & Koehler, M. J. (2008). *Introducing technological pedagogical content knowledge.* New: Annual Meeting of the American Educational Research Association.

Msweli, N. (2023). Instructors ' perception of the competencies required to teach DS in HEI. In *Proceedings of NEMISA digital skills conference 2023: Scaling data skills for multidisciplinary impact education* (Vol. 5, pp. 89–103).

Niess, M. L. (2011). Investigation TPACK: Knowledge growth in teaching with technology. *Journal of Educational Computing Research, 44*(3), 299–317. https://doi.org/10.2190/EC.44.3.c

Otero, P., Hersh, W., & Ganesh, A. U. J. (2014). *Big data: Are biomedical and health informatics training programs ready? In IMIA yearbook of medical informatics*. https://doi.org/10.15265/IY-2014-0007

Paul, P. K., & Aithal, P. S. (2018). Computing academics into new age programs and fields: Big data analytics & data sciences in Indian academics-an academic investigation of private universities. *IRA-International Journal of Management & Social Sciences, 10*(3), 107–118. https://doi.org/10.21013/jmss.v10.n3.p3

Price, R., & Ramaswamy, L. (2019). Challenges and approaches to teaching data science technologies in an information technology program with non-traditional students. In *Proceedings - 2019 IEEE 5th international conference on collaboration and internet computing, CIC 2019* (pp. 49–56). https://doi.org/10.1109/CIC48465.2019.00015. *Cic*.

Rao, A. R., Desai, Y., & Mishra, K. (2019). Data science education through education data: An end-to-end perspective. In *2019 9th IEEE integrated STEM education conference, ISEC 2019* (pp. 300–307). https://doi.org/10.1109/ISECon.2019.8881970

Saddiqa, M., Magnussen, R., Larsen, B., & Pedersen, J. M. (2021). Open Data Interface (ODI) for secondary school education. *Computers in Education, 174*, Article 104294. https://doi.org/10.1016/J.COMPEDU.2021.104294

Saeli, M., Perrenet, J., Jochems, W. M. G., & Zwaneveld, B. (2011). Teaching programming in secondary school: A pedagogical content knowledge perspective. *Informatics in Education, 10*(1), 73–88.

Salas-Rueda, R.-A. (2020). TPACK: Technological, pedagogical and content model necessary to improve the educational process on mathematics through a web application? *International Electronic Journal of Mathematics Education, 15*(1), 551. https://doi.org/10.29333/iejme/5887

Saltz, J., & Heckman, R. (2015). Big data science education: A case study of a project-focused introductory course. *Themes in Science & Technology Education, 8*(2), 85–94. https://www.learntechlib.org/p/171521/.

Schatsky, D., Chauhan, R., & Muraskin, C. (2018). *Democratizing data science to bridge the talent gap*.

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education, 29*(S1), S40–S50. https://doi.org/10.1080/10691898.2020.1851159

Scott, K. C., & Nimon, K. (2021). Construct validity of data from a TPACK self-assessment instrument in 2-year public college faculty in the United States. *Journal of Research on Technology in Education, 53*(4), 427–445. https://doi.org/10.1080/15391523.2020.1790444

Song, I. Y., & Zhu, Y. (2016). Big data and data science: What should we teach? *Expert Systems, 33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Spieler, B., Grandl, M., Ebner, M., & Slany, W. (2019). "Computer Science for all": Concepts to engage teenagers and non-CS students in technology. https://bwinf.de/biber/.

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education, 19*(4), 1–16. https://doi.org/10.1145/3336124

Taopan, L. L., Drajati, N. A., & Sumardi. (2020). TPACK framework: Challenges and opportunities in EFL classrooms. *Research and Innovation in Language Learning, 3*(1), 1–22. https://doi.org/10.33603/RILL.V3I1.2763

Thomas, T., Herring, M., Redmond, P., & Smaldino, S. (2013). Leading change and innovation in teacher preparation: A blueprint for developing TPACK ready teacher candidates. *TechTrends, 57*(5), 55–63. https://doi.org/10.1007/s11528-013-0692-7

Tseng, J.-J., Chai, C. S., Tan, L., & Park, M. (2020). *A critical review of research on technological pedagogical and content knowledge (TPACK) in language teaching*. https://doi.org/10.1080/09588221.2020.1868531

Twinomurinzi, H., Mhlongo, S., Bwalya, K. J., Bokaba, T., & Mbeya, S. (2022). *Multidisciplinarity in data science curricula. African Conference on information Systems and technology*.

Uzunalioglu, H., Cao, J., Phadke, C., Lehmann, G., Akyamac, A., He, R., Lee, J., & Able, M. (2019). *Augmented data science: Towards industrialization and democratization of data science*. https://doi.org/10.1145/nnnnnnn.nnnnnnn. *AutoML*.

Wang, H., Gao, H., Yin, S., & Zhu, J. (2017). The design of course architecture for big data. *ACM International Conference Proceeding Series, Part, F1277*, 1–6. https://doi.org/10.1145/3063955.3063968

WIDS. (2023). *Women in data science (WiDS)*. https://www.widsconference.org/.

WILM. (2023). *Women in machine learning*. https://wimlworkshop.org/.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1). http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

Yadav, N., & Debello, J. E. (2019). Recommended practices for Python pedagogy in graduate data science courses. In *Proceedings - frontiers in education conference, FIE, 2019-octob*. https://doi.org/10.1109/FIE43999.2019.9028449

Yan, C., & He, Y. (2020). Auto-Suggest: Learning-to-Recommend data preparation steps using data science notebooks. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1539–1554. https://doi.org/10.1145/3318464.3389738

Yu, B., & Hu, X. (2019). Toward training and assessing reproducible data analysis in data science education. *Creative Commons Attribution 4.0 International, 1*, 381–392. https://doi.org/10.1162/dint_a_00053 *(CC BY 4.0)*.