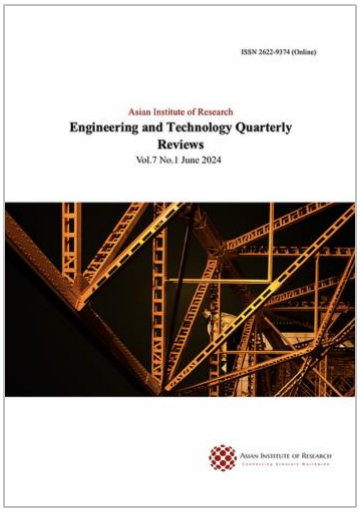

Engineering and Technology Quarterly Reviews

Malallah, S., Osiobe, E., & Shamir, L. (2024), Data Science Self-Efficacy Assessment Tools: A Foundational Guide to Evaluating Progress. In: *Engineering and Technology Quarterly Reviews*, Vol.7, No.1, 78-92.

ISSN 2622-9374

Published by:
The Asian Institute of Research

The Asian Institute of Research
Engineering and Technology Quarterly Reviews
Vol.7, No.1, 2024: 78-92
ISSN 2622-9374

Data Science Self-Efficacy Assessment Tools: A Foundational Guide to Evaluating Progress

Safia Malallah[1], Ejiro Osiobe[2], Lior Shamir[3]

[1] Kansas State University, safia@ksu.edu
[2] Baker University, Jiji.osiobe@bakerU.edu
[3] Kansas State University, lshamir@ksu.edu

Abstract
Data science education research faces a notable gap in assessment methodologies, leading to uncertainty and unexplored avenues for enhancing learning experiences. Practical assessment is crucial for educators to tailor teaching strategies and support student confidence in data science skills. We address this gap by developing a data science self-efficacy survey to empower educators by identifying areas where students lack confidence, enabling the design of targeted plans to bolster data science education. Collaboration among computer science, business, and statistics experts was instrumental in crafting a comprehensive survey that caters to the interdisciplinary nature of data science education. The survey evaluates 13 essential skills and knowledge areas, synthesized from literature reviews and industry demands, to provide a holistic assessment framework for educators in the field. Rigorous reliability and validity tests were conducted to ensure the survey's robustness and efficacy in accurately assessing student proficiency.

Keywords: Data Science, Self-Efficacy, Assessment Tools

1. Introduction

Data science has experienced remarkable global demand, solidifying its position as one of the fastest-growing professions worldwide. However, this demand is met with a shortage of freshly graduated, qualified data scientists, raising concerns for academia and industries (Davenport & Patil, 2022; Haben & Hinton). Additionally, research on data science education assessments is lacking, leaving many uncertainties surrounding students' pre-graduation skills. This paper addresses this limitation and develops a data science self-efficacy survey to evaluate and quantify individuals' confidence levels in applying data science skills to build data-driven solutions, intending to enhance the learning experience within data science education. Also, remedial activities were proposed to boost students' confidence based on individual confidence levels. Survey development followed a modified Vinay approach, which guided the construction of customized assessments for data science aligned with organizational needs (Vinay, 2024). This was carried out by collaborating with computer science, business, and statistics experts, crafting a comprehensive lens that caters to the interdisciplinary nature. The survey evaluated 13 items representing applying data science life cycle steps and using related interdisciplinary skills to fulfill step requirements identified from literature reviews. The survey comprises 48 questions organized into eight sections, answered with a 5-point Likert scale from strongly disagree to agree strongly. The survey was distributed to students and researchers in six educational institutions in Kansas (KS), the United States of America (USA), and Kuwait (KT). Pilot results showed that the survey has high reliability, stability, and suitability. The final analysis indicates that 11.56% of students report low confidence, 11.54% record high confidence, and the majority express moderate confidence. Lower confidence levels were found around "model development" and "model evaluation," which can be tied to "analysis and calculation skills," "optimization skills," and "technical and computing skills." To boost students' confidence using the remedial suggestions, individualized support sessions should be used to discuss student concerns, address any questions or misunderstandings they may have, and offer personalized guidance and encouragement. Additionally, peer support groups can show students that they are not alone and provide opportunities to encourage one another during regular check-ins. Highly confident students need opportunities for advanced learning through independent research, creative projects, or leadership roles within the learning environment, thus encouraging confident participants to share their knowledge and expertise with their peers.

2. Limitations

A primary limitation of this study is the biases or inaccuracies that self-efficacy assessments carry. Self-efficacy often focuses on specific tasks or domains, which may not fully capture an individual's overall sense of efficacy across different situations. Moreover, self-efficacy is inherently subjective and self-reported, lacking objective measurement and increasing the prevalence of bias or inaccuracies. Our small size and distributed populations can present significant limitations in research papers by compromising generalizability, statistical power, comparability, external validity, and replicability.

3. Background

3.1. Confidence and Learning

Confidence plays a pivotal role in students' academic success and overall well-being. Social Cognitive Theory suggests that self-efficacy, or belief in one's ability to succeed, significantly influences behavior and performance. Students with low confidence often exhibit hesitancy, self-doubt, and reluctance to engage in academic tasks. Interventions targeting low-confidence students should build self-efficacy through incremental successes, constructive feedback, and role modeling (Locke, 1987). Additionally, fostering a supportive classroom environment that encourages risk-taking and emphasizes growth mindset principles can empower students to develop resilience and confidence in their abilities (Works, 2017). Self-determination theory posits that autonomy, competence, and relatedness are fundamental psychological needs that drive motivation and well-being. To support moderate-confidence students, educators can provide opportunities for autonomy by offering choices and promoting student agencies in their learning process.

Furthermore, scaffolding instruction and targeted interventions tailored to individual learning needs can enhance students' sense of competence and foster a positive learning experience (Deci & Ryan, 2012). High-confidence students typically believe in their abilities and may seek challenges or leadership roles. However, excessive confidence without corresponding competence can lead to overestimating skills and performance (Hornstra et al., 2023). The Zone of Proximal Development suggested that learning occurs most effectively within the "zone" where tasks are challenging yet achievable with appropriate support. Educators can support high-confidence students by providing opportunities for intellectual challenge and promoting metacognitive skills, such as self-reflection and self-regulation. Encouraging collaboration and peer feedback can also help high-confidence students better understand their strengths and areas for improvement (Training, 2017).

3.2. Data Science Assessment Pathway

Vinay proposed a nine-step assessment pathway to create a customized data science assessment aligned with organizational goals using these competencies. These steps include identifying critical competencies, categorizing and prioritizing, defining competency levels, developing assessment tools, scoring and evaluation rubrics, integrating organizational goals, feedback mechanisms, implementation and training, and iterative refinement. We incorporated the first five steps to develop our survey, which were relevant to our goal of creating an assessment process for academia (Vinay, 2024).

In our study, developing an instrument for assessing self-efficacy confidence in data science requires carefully considering the specific skills, tasks, and challenges relevant to the field. Here's a step-by-step foundational guide to help you create such an instrument:

1. **Define the Scope:** Determine the specific areas within data science that you want to assess. This could include programming skills, statistical knowledge, machine learning expertise, data visualization proficiency, problem-solving abilities, etc.
2. **Review Existing Literature:** Look for existing self-efficacy scales or instruments related to data science or similar fields. This can provide insights into relevant constructs and items that you might include in your instrument. Adaptation of existing scales can be a time-saving approach.
3. **Item Generation:** Generate a pool of items/questions that reflect the skills and tasks you want to measure in data science. These items should be clear, specific, and cover a range of difficulty levels. Consider consulting with data science experts to ensure the items' relevance and validity.
4. **Pilot Testing:** Administer the initial set of items to a small sample of individuals representing your target population (e.g., students and professionals in data science). Collect feedback on the clarity, relevance, and difficulty of the items. Use this feedback to refine the items and eliminate any ambiguities.
5. **Validity and Reliability:** Assess the validity and reliability of your instrument. Validity ensures that the instrument measures what it intends to measure, while reliability ensures consistency in measurement. You can use techniques such as factor analysis to assess construct validity and Cronbach's alpha to measure internal consistency reliability.
6. **Finalize the Instrument:** Based on the results of pilot testing and validity/reliability analyses, finalize the items for your instrument. Ensure that the instrument is comprehensive yet concise enough to be administered efficiently.
7. **Scoring:** Determine the scoring mechanism for your instrument. This could involve assigning numerical values to responses (e.g., Likert scale) or using a categorical scoring system. Consider whether reverse scoring is necessary for specific items to prevent response bias.
8. **Administration:** Decide on the method of administration for your instrument. It could be administered online or in person, depending on your resources and the preferences of your target population. Ensure to work with the Institutional Review Board (IRB) within your institution, county, state, and country.
9. **Data Collection:** Administer the instrument to your target population and collect the responses. Ensure that participants understand the instructions and have enough time to complete the instrument accurately.
10. **Analysis and Interpretation:** Analyze the collected data to assess self-efficacy and confidence in data science. Calculate descriptive statistics (e.g., mean, standard deviation) to understand the distribution of scores. Compare scores across different groups (e.g., novices vs. experts) to identify patterns and trends.
11. **Validation:** Validate the instrument by comparing scores with external criteria (e.g., performance on data science tasks and academic achievement). This can provide evidence for the validity of your instrument.
12. **Iterative Improvement:** Continuously evaluate and refine your instrument based on feedback and further research findings. This iterative process helps ensure the instrument's effectiveness and relevance over time.

By following these steps, you can develop a robust instrument for assessing self-efficacy confidence in data science that can be used for research, education, or professional development purposes.

4. Method

4.1. Design

This study employed a quantitative approach to develop a self-efficacy survey to assess students' confidence levels in utilizing data science skills and knowledge. The experiment consisted of two phases: survey development and survey implementation. In the development phase, a framework inspired by Vinay's data science assessment pathway guided the process through four key stages (Vinay, 2024) (Malallah, Weese, & Alsalmi, 2023; Malallah, Weese, Shamer, et al., 2023). First, a comprehensive literature review was conducted to understand the current landscape of data science assessment. No scientific research directly addressing data science assessment was found, prompting the creation of a foundational framework for survey development. Second, a thorough literature review was conducted to identify the requisite knowledge and skills for a data scientist, guided by educator and industry recommendations. Data saturation determined the depth of the review. The third stage aimed to establish a coherent sequence of data science concepts within the survey, satisfying interdisciplinary needs. This involved identifying the appropriate data science cycle to guide the arrangement of concepts. Finally, the survey questions were crafted in stage four, drawing from the intersection of the data science cycle steps and the necessary knowledge to fulfill them. The research implementation phase spanned eight weeks. Initially, the survey underwent review and modification based on feedback from experts in statistics, computer science, and business analytics. Subsequently, the survey was distributed online to 163 participants enrolled in data science and data analytics courses across collaborating universities in the USA, Kuwait, and KSA. A pilot study involving 33 randomly selected students from the same population, not included in the analysis, was conducted. Participants were required to complete an online consent form before beginning the survey, with an expected survey completion time ranging between 25 minutes and 40 minutes.

4.2. Sample

The sample encompassed a diverse population of 163 individuals engaged in various data science disciplines, comprising 64.7% males and 32.4% females. Participants represented fields such as computer science, statistics, mathematics, and business; they were drawn from six educational institutions, including four universities and two community colleges. Geographically, 32% of participants hailed from the USA, 38% from Kuwait, and 29% from Saudi Arabia. Among the participants, 25% were researchers. The remainder were students (46.4% seniors, 21.4% juniors, and 7.1% first-year students). A notable portion of the sample, 42.4%, possessed prior working experience, albeit only 21% had worked within the technology sector. Regarding educational background, 26% of participants had never taken research courses before, 3% had never taken statistics classes, 8.8% had never taken coding classes, and 44% had never taken machine learning/artificial intelligence (AI) courses. Additionally, 32% had never enrolled in business analytics courses. The remaining participants had varying degrees of exposure to these subjects as part of their curriculum through one or multiple courses (see Figure 1).

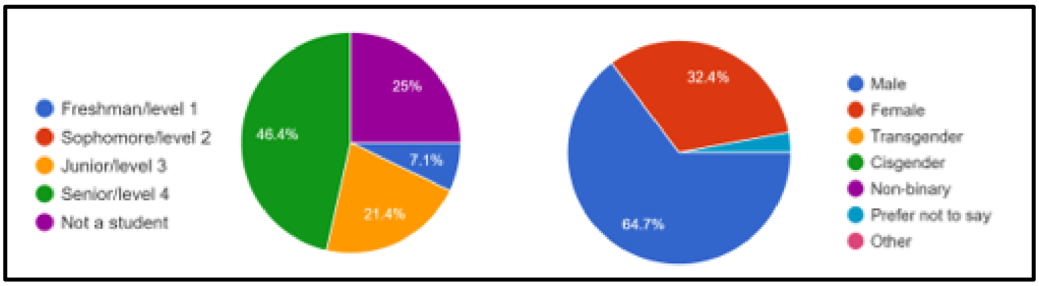

Figure 1: Sample Population
Source: Authors' creation

4.3. Keywords, Database, and Criteria

The literature reviews were conducted using specific keywords tailored to each investigation area. The first literature review searched "assessment||self-efficacy" + "data science." The second literature review used the keywords "knowledge ||skills" + "literature review" + "data science ||data science education ||teaching ||learning ||teaching and learning." The third literature review utilized the keywords "data science||statistic|| mathematics ||computer Science ||business" + "life cycle." Searches were conducted in Google, Google Scholar, and ScienceDirect. Various source types were considered, including conference papers, journals, and blogs. The results were meticulously filtered by isolating abstracts and titles aligned with the search criteria. Studies that did not primarily focus on data science were excluded from the analysis. The search was refined to include only results from 2020 to 2024, except in cases concerning the data science life cycle. Furthermore, research on specific medical fields (e.g., medicine, dentistry, nursing, health professions, neuroscience, pharmacology, toxicology, pharmaceutical science, cancer, effect, and psychological studies) was excluded.

4.4. Instruments

The survey was carefully developed based on thorough analyses from literature reviews (see the Results section). Table 1 presents the investigation's final findings, outlining the 13 elements assessed. Column 2 categorizes these elements as data science life cycle steps and interdisciplinary skills utilized within those steps. The last column specifies the questions targeting each skill. Table 2 contains the survey questions—48 items that evaluate the 13 distinct aspects identified in Table 1. Responses are assessed using a 5-point Likert scale ranging from strongly disagree to agree strongly.

Table 1: The Data Science (DS) Skills and Knowledge of DS Life Cycles

#	Concept	Description	Examples	Questions
1	DS life cycle step 1	Domain knowledge and research design		Q1-Q7
2	DS life cycle step 2	Data planning and data collection		Q8-Q12
3	DS life cycle step 3	Data cleaning, wrangling, and feature engineering		Q13-Q19
4	DS life cycle step 4	Feature selection		Q20-Q28
5	DS life cycle step 5	Model design		Q29-Q35
6	DS life cycle step 6	Model evaluation		Q36-Q40
7	DS life cycle step 7	Communicate and propose action		Q41-Q48
8	Researching and Planning Skill	The ability to formulate well-defined questions, creating a road map for successful project execution while incorporating critical thinking, strategic reasoning, and the ability to navigate, follow, and evaluate both the process and the outcome	Domain Knowledge - Scientific Research Knowledge & Ethic Knowledge.	Q1-Q6, Q19-Q20, Q34, Q42, Q47
9	Analysis & Calculation Skill	The capability to comprehend and utilize statistical concepts and mathematical operations for analysis	Statistical Proficiency Mathematics Proficiency.	Q16, Q18, Q20-Q23, Q26, Q28, Q32, Q34, Q38-Q40.
10	Optimization Skill	The capacity to pinpoint weaknesses within a problem and devise solutions to bolster and enhance it, thereby optimizing efficiency and effectiveness while also facilitating growth to meet or surpass specified requirements and expectations	Optimization – Scalability – Quality - Continuous learning and adaptability - Analytical thinking and problem-solving	Q10, Q19, Q24-Q27, Q31, Q34, Q36-37
11	Technical & Computing Skills	The ability to utilize computing skills, including general computing, advanced machine learning, and AI, along with technical knowledge, to effectively leverage technology for developing innovative solutions	General computing, Machine Learning, AI proficiency, technical knowledge	Q7, Q16, Q18, Q20-Q24, Q27-Q29, Q30-Q34, Q37-Q40
12	Data Management & Handling Skill	The ability to comprehend data structures and the language of data manipulation technology to harness technology effectively for managing and manipulating both small and big data sets to explore and prepare data, ensuring its accuracy and usability.	Data handling, Management, and Database proficiency - Big Data, Data Preparation, and Exploration proficiency.	Q8-Q15, Q17-Q18, Q25, Q27, Q31, Q33, Q36, Q41-Q42, Q48.
13	Business & Communication Skill	Proficiency in translating and aligning business strategies into actionable technical findings, effectively communicating them to stakeholders		Q4, Q16-Q17, Q20, Q23, Q28-Q29, Q34, Q43-Q48

Source: Authors' Creation

4.5. Instruments Rubric

The instrument's rubric outlines thresholds for confidence levels using a 5-point Likert scale by categorizing responses. Self-efficacy confidence scores obtained from the survey were divided into three levels: 1–2.9 (low confidence), 3–3.6 (moderate confidence), and 3.7–5 (high confidence). This categorization applies specifically to the sample analyzed in this paper and may not be generalized to all populations. Future studies aiming to replicate this research should categorize results into three quartiles to determine an appropriate threshold for the data.

Table 2: Data Science Self-Efficacy Survey

#	Questions
1	Creating a plan and designing an effective strategy to develop necessary solutions in a data science project.
2	Establishing realistic timelines and defining achievable milestones using the data science life cycle.
3	Exploring a domain to acquire the necessary knowledge for a specific data science project.
4	Exploring trends and preparing reviewed literature and other scholarly justification from the data science project
5	My ability to formulate investigative questions that align with the nature of the problem.
6	My ability to consider ethical implications related to data privacy, bias, and fairness throughout the process.
7	Creating clear documentation for code, models, and any essential insights made during the project.
8	Articulating the investigated problem and identifying suitable and trustworthy data sources to help derive insights.
9	My ability to design an efficient data collection method while identifying challenges that might arise in the collection process.
10	My ability to [iteratively] adapt modifications to the data collection and cleaning process in response to new findings.
11	My ability to identify and use suitable tools for data collection.
12	My ability to effectively handle the collection of both big and small, structured, unstructured, numerical, quantitative, and qualitative data.
13	Understanding the structure and characteristics of diverse datasets.
14	Merging or joining datasets from different sources to create a unified dataset.
15	Using appropriate tools to visualize data distributions of missing values, duplicate values, inconsistency types, and outliers.
16	My ability to inform decisions to standardize or normalize values as needed, depending on project requirements.
17	In making informed decisions on handling invalid data. Based on the visualized data distributions and stakeholders ' requirements.
18	My ability to validate and ensure data quality after cleaning to determine whether the data is cleaned, structured, and ready for feature extraction.
19	My ability to identify when there is a need to create subsets based on project requirements.
20	My ability to understand the meaning of each feature and the relationships between features by communicating with domain experts, ensuring a comprehensive understanding of the features.
21	In applying exploratory data analysis to understand the dataset better using basic statistics (Central Tendency Descriptive Summary), Principal Component Analysis (PCA), or Self-Organizing Map (SOM)
22	Use descriptive statistics and machine learning measures to rank features based on their relationship with the target variable.
23	My ability to filter the features using selection techniques like Forward Selection, Backward Elimination, Recursive Feature Elimination, or Akaike information criterion (AIC), Schwarz or Bayesian Information Criterion (SIC), and Likelihood results selection.
24	My ability to experiment with multiple techniques to find the most effective approach for a specific model.
25	Creating new features by transforming existing ones to enhance the model outcome.
26	(If necessary), in applying transformations to variables, such as transforming values from categorical to numerical data, to strengthen model efficiency.
27	In removing redundancies and selecting features to improve model efficiency.
28	My ability to identify trends and patterns detecting anomalies or novel patterns in the data.
29	My ability to develop a model-building and validation plan.
30	Choosing the appropriate tools suited for model development.
31	Evaluating trade-offs between model complexity, interpretability, and performance.
32	Determining when to use statistical inference, simulation, classification, regression, or clustering methods.
33	Customizing my dataset to match the suitable learning algorithm (supervised, unsupervised).
34	My ability to choose suitable machine learning or statistical models based on the nature of the problem that can minimize the loss function.
35	Identifying when sampling is needed and selecting appropriate sampling methods.
36	That I can scale the model to handle larger datasets.
37	Performing hyperparameter tuning and addressing potential biases or imbalances during model building.
38	Performing validation techniques (e.g., cross-validation) to assess the model's generalization ability.
39	Defining metrics for evaluating model performance, such as accuracy, precision, and recall metrics.
40	Performing diverse analyses on the developed model and its outcome, such as hypothesis testing, estimation, prediction intervals, and determining the significance of relationships.
41	Generating appropriate data visualizations for model outcomes.
42	Using the model's outcomes to inform insight.
43	My ability to provide explanations for model outcomes.
44	Interpreting my result to the lowest denomination so that non-academic readers understand it.
45	Connecting my results to exciting trends and literature to draw inferences when applicable.
46	Combining complex visualized structures, encompassing multidimensional and hierarchical data, to create a non-complex, meaningful, and insightful representation of our results through data storytelling.
47	My ability to tailor visualizations to the specific needs and understanding of different audiences, including non-technical stakeholders.
48	My ability to follow best practices for data visualization, including appropriate chart selection, color usage, and labeling

Source: Authors' Creation

5. Results

This study analyzed students' confidence in building data-driven solutions in a data science education environment to deliver a coherent assessment. The following research questions were considered, and the responses were analyzed through repeated measures (analysis of variance [ANOVA] and descriptive statistics) using Statistical Package for Social Science (SPSS) software and Excel.

5.1. Research Questions

RQ1: What specific data science skills and knowledge are essential for students to acquire to align with the demands of the industry?
RQ2: What are the key steps involved in the process of constructing data science solutions?
RQ3: How can insights from industry needs and solution-building methodologies inform the creation of a tailored survey?
RQ4: How reliable is the survey? (Instrument reliability and validity)
RQ5: Which skills and steps do students feel less confident about, as identified through the survey? (Instrument analysis)
RQ6: How can interventions be designed to address these areas?

RQ1 - What specific data science skills and knowledge are essential for students to acquire to align with the demands of the industry?

The literature reviews below were used to design and set the survey content. Table 3 lists the 136 created data science skills, knowledge, and tool ability. The first 39 were taken from Vinay's work (Vinay, 2024), the next 50 items from Usama Fayyad's and Hamit Hamock's work (Fayyad & Hamutcu, 2020), and the remaining from Guoyan's work (Li et al., 2021). The list was clustered and filtered to generate the final list, which has eight categories presented in Table 1, Skills 8–13.

Table 3: The Identified Items from the Literature Reviews

1. Programming Languages
2. Data Processing Frameworks
3. Machine Learning Libraries
4. Data Visualization Tools
5. Database Management Systems (DBMS)
6. Version Control Systems
7. Big Data Technologies
8. Cloud Platforms
9. Integrated Development Environments
10. Automation and Workflow Management
11. Problem Formulation
12. Hypothesis Generation
13. Data Exploration
14. Statistical Analysis
15. Machine Learning Application
16. Iterative Refinement
17. Critical Thinking
18. Optimization Strategies
19. Interdisciplinary Collaboration
20. Continuous Learning
21. Industry Contextualization
22. Relevant Data Variables
23. Customized Modeling Approaches
24. Understanding Business Objectives
25. Data Privacy and Compliance
26. Effective Communication with Stakeholders
27. Identification of Key Performance Indicators
28. Adaptability to Industry Trends
29. Problem-Solving Relevance
30. Strategic Decision Support
31. Rapid Technological Advancements:
32. Expanding Methodological Landscape
33. Lifecycle of Data Science Projects
34. Adapting to Diverse Data Types:
35. Embracing Interdisciplinary Knowledge
36. Professional Development:
37. Adoption of New Tools and Frameworks:
38. Peer Collaboration and Knowledge Sharing
39. Proactive Problem-Solving:

40. Basics of the scientific method, research methods, hypothesis formulation
41. Problem identification.
42. Basic math
43. Calculus
44. linear algebra
45. Data structures and Algorithms
46. Databases and Data Processing Systems
47. Software Engineering and Development
48. Operating Systems
49. Deep Learning
50. Descriptive Statistics
51. Inferential Statistics
52. Bayesian Statistics
53. Stochastic Processes, Time Series, Survival Analysis
54. Statistical Sampling
55. Linear programming,
56. Nonlinear optimization
57. Probability basics, descriptive, inferential, and Bayesian statistics, stochastic processes and time series, causality, sampling
58. Data Preparation and Transformation
59. Data Cleaning
60. Data Exploration and Visualization
61. Unsupervised Learning
62. Supervised Learning
63. Reinfforcement Learning
64. Parallel and Distributed Computing
65. Text Mining and Natural Language Processing
66. Statistical Sampling
67. Linear programming,
68. Nonlinear optimization
69. Data Preparation and Transformation
70. Data Cleaning
71. Data Exploration and Visualization
72. General-Purpose Programming Languages
73. Computing Fundamentals
74. Stochastic Processes, Time Series, Survival Analysis
75. Virtualization/ Containerization
76. Cloud Platforms
77. Statistical
78. Mathematical/Numeric
80. ML Libraries
81. Development Environments
82. Visualization
83. RDBMS and SQL
84. NoSQL and NewSQL
85. Data Warehousing
86. Querying and Presentation
87. Infrastructure
88. Processing and Execution
89. Access
90. Integration

91. Data Mining
92. Big Data
93. Statistics
94. Algorithms
95. Data Engineering
96. Agile Methodology
97. Extract Transform Load
98. Data Modeling
99. Data Warehousing
100. Data Visualization
101. Database Administration
102. Relational Databases
103. Business Intelligence
104. Scalability
105. Mathematical Optimization
106. Data Architecture
107. Automation
108. Artificial Intelligence
109. Data Management
110. Operations Research
111. Deep Learning
112. Data Quality
113. Machine Learning
114. SQL
115. Python
116. R
117. Apache Hadoop
118. Java
119. Tableau
120. Apache Spark
121. Scripting
122. SAS
123. Microsoft SQL Servers
124. Apache Hive
125. Amazon Web Services
126. C++, C
127. MATLAB
128. Scala
129. NoSQL
130. Power BI
131. Object-Oriented Programming
132. Apache Kafka
133. Microsoft Azure
134. PostgreSQL
135. Apache Cassandra
136. PyTorch

Source: Authors' Creation

Google Scholar shows seven results, and ScienceDirect shows 73. All were excluded except one. Twenty-five results were found from Google Scholar. Two were chosen as they included extensive literature reviews with new information, and data saturation was satisfied. Vinay (2024) introduced a comprehensive framework to assess and categorize the essential competencies of proficient data scientists. This framework—which stemmed from a literature review exploring technical proficiency, analytical thinking and problem-solving, domain-specific knowledge, continuous learning, and adaptability in data science—provides valuable insights into the field. Vinay defined critical skills for proficient data scientists. The 39 competencies he identified were: Technical proficiency (1–10); analytical thinking and problem-solving (11–20); domain-specific knowledge (21–30); and continuous learning and adaptability (31–39). Although we did not use all his competencies directly, we cross-referenced them with other resources in the following steps (Vinay, 2024).

Fayyad and Hamock, in their study, introduced a comprehensive Data Science Knowledge Framework to foster industry standardization and the creation of measurement and assessment methodologies. Emphasizing data science's dynamic and multidisciplinary nature, the authors constructed the framework through an extensive literature review, identifying pivotal topics and technologies crucial for analytics and data science professionals. The findings were systematically organized into a hierarchical knowledge structure (Fayyad & Hamutcu, 2020).

Guoyan Li et al. analyzed the data science and analytics skills gap in the Industry 4.0 reports to identify the critical technical skills and domain knowledge required for data science in today's manufacturing industry. The authors used Emsi job posting and profile data to gain insights into the trends in manufacturing jobs, leveraging data science (Li et al., 2021).

The process of clustering 136 items was extensive. The list contained various categories, making it challenging to perform definitive clustering without specifying a purpose or desired level of granularity. Several options were available for clustering: domain, function, level of expertise, and tool/technology. We clustered the terms by skill, as it is our objective. We clustered the groups several times, and with every iteration, we merged groups until 14 categories remained: domain knowledge, scientific research method, statistical proficiency, mathematics proficiency, optimization/continuous learning and adaptability, data preparation and exploration, machine learning, general computing, technical proficiency, data management handling and database proficiency, business proficiency and communication, big data, analytical thinking and problem-solving and ethic. The categories have been reduced to eight after being validated by the experts.

RQ2- What are the critical steps involved in the process of constructing data science solutions?
A data science life cycle embodies an iterative series of steps crucial for project or analysis delivery, tailored to each project's unique needs. Although no standardized workflow exists for data science, selecting appropriate steps is essential for survey coherence and suitability. Four models were identified and compared for common factors to address this, ultimately revealing eight key steps presented in Table 1.

Table 4 and Figure 2 showcase the identified data science models, where each row represents a model with its associated steps. Model (a) emphasized a data science education lens, encompassed the holistic data life cycle, and integrated workflow with environmental and social considerations such as regulations and ethics (Stodden, 2020). Model (b), viewed statistically, identified seven crucial steps in the data investigation process, including framing the problem, data gathering and processing, exploration, and visualization, model consideration, and communication of findings (LEE et al., 2022). Model (c), from a business and computer science perspective, leveraged Microsoft's Team Data Science Process (TDSP) framework for collaborative learning, and aimed to convert data into actionable insights (Saltz & Hotz, 2020). Model (d), which adopted a computer science and statistic lens, relied on CRISP-DM, guided data mining projects through six phases, from understanding business objectives to deploying models into operational systems (Gupta, 2022).

All models began with problem understanding, progressed through data acquisition and comprehension, and concluded with communication, either as a standalone step or integrated within the evaluation, depending on the model. While tasks such as feature engineering were categorized differently in various models, expert feedback determined the sequence, and the last row served to structure the survey flow and cluster competencies.

Table 4: Identified Data Science Life cycle models

Model	Sequence								
a [11]	Acquire	Clean			Use/ reuse				Publish
b [12]	Frame problem	Consider and gathering	Process data	Explore & visualize	Consider models				Communicate & propose action
c [13]	Business understanding	Data acquisition and understanding		Deployment	Modeling Feature engineer		Modeling training	Modeling evaluation	
d [14]	Business understanding	Data understanding	Data preparation Data cleaning: Data integration Data transformation: Data reduction: Data discretization: Feature engineering				Modeling	Evaluation	
	Domain knowledge and research design		Data collection	Data wrangling	Feature engineering	Feature selection	Model design	Model evaluation	Communicate and propose action

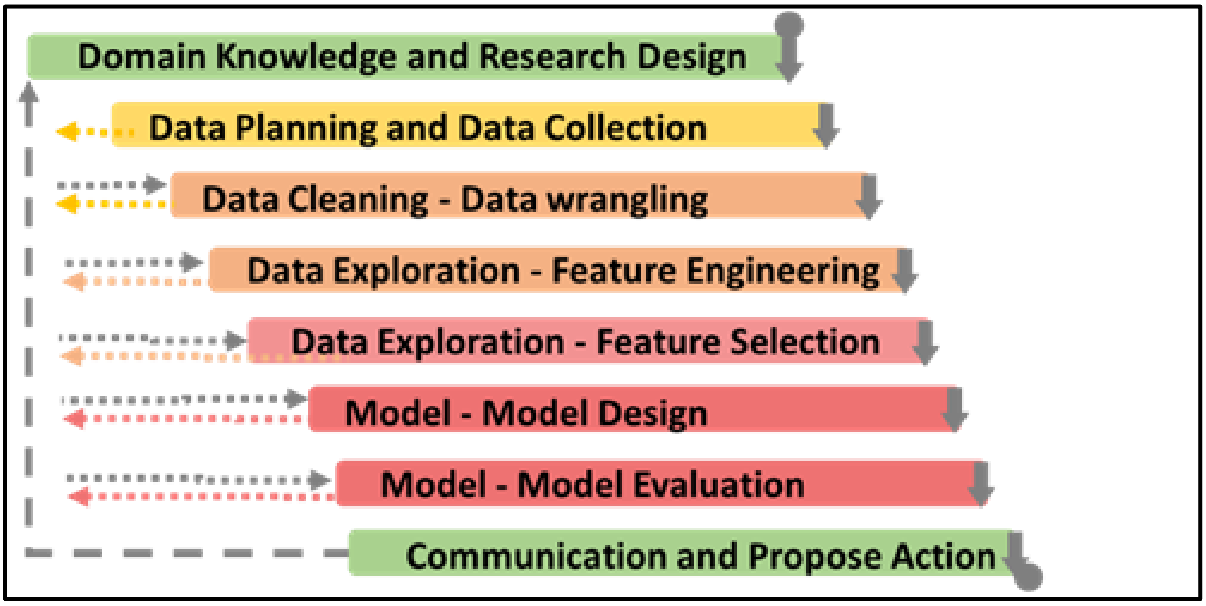

Figure 2: Identified Data Science Life cycle models

Source: Authors' Creation

RQ3 - How can insights from industry needs and solution-building methodologies inform the creation of a tailored survey?

Table 5 presents the fundamental elements necessary for crafting pertinent questions. It aligned the identified skills with the data science steps to create a question flow that effectively fulfills dual purposes. Based on this approach, the final formulated questions are presented in Table 2.

Table 5: The Used Skills and Data Science Steps to Construct the Survey Questions

DB Cycle\Skills	Researching & Planning	Analysis & Calculation	Optimization Skill	Technical & Computing Skills	Data Management & handling	Business & Communication
Domain Knowledge & ...	x			x		x
Data Planning & Data ..	x		x		x	
Data cleaning, wrangling.		x	x	x	x	x
Feature Selection		x	x	x	x	x
Model design		x	x	x	x	x
Model evaluation	x	x		x		x
Communicate & propose,	x				x	x

Source: Authors Creation

RQ 4 - How reliable is the survey? (Instrument reliability and validity)

The pilot stage was subjected to validation through Cronbach's alpha testing to evaluate the reliability of survey statements; the validity was assessed using the Pearson correlation coefficient, presented in Tables 6 and 7. The calculated Cronbach's α coefficient resulted in a value of 0.915, indicating a high level of internal consistency among the survey items. This implied strong reliability, with the items collectively measuring the intended construct effectively, surpassing the widely accepted threshold of 0.7. Furthermore, the Cronbach's α coefficient was separately computed for the 13 sections, revealing internal consistency validity within the range of .6–.8. All scales exhibited convergent validity, with correlations among items exceeding 0.3, indicating robust convergent validity statistically, except for the correlation between Q28 and Q21, which was not statistically significant ($p = 0.45$). Assessment of internal consistency validity using the Pearson correlation coefficient showed correlations ranging from .57 to 0.90 for the survey statements. All correlation coefficients were statistically significant at the 0.01 level, highlighting the high internal consistency and validity of the questionnaire.

Table 6: Person Correlations of All the Questions

	S1Q1	S1Q2	S1Q3	S1Q4	S1Q5	S1Q6	S1Q7	S2Q8	S2Q9	S2Q10	S2Q11	S2Q12	S3Q13	S3Q14	S3Q15	S3Q16	S3Q17	S3Q18	S3Q19	S4Q20
Pearson Correlation	.901**	.759**	.832**	.814**	.679**	.705**	.713**	.820**	.789**	.836**	.808**	.812**	.793**	.751**	.733**	.792**	.834**	.827**	.704**	.693**
	S4Q21	S4Q22	S4Q23	S4Q24	S4Q25	S4Q26	S4Q27	S4Q28	S5Q29	S5Q30	S5Q31	S5Q32	S5Q33	S5Q34	S5Q35	S6Q36	S6Q37	S6Q38	S6Q39	S6Q40
	.577**	.692**	.590**	.816**	.751**	.687**	.740**	.861**	.818**	.855**	.792**	.660**	.719**	.696**	.825**	.759**	.704**	.723**	.831**	.873**
	S7Q41	S7Q42	S7Q43	S7Q44	S7Q45	S7Q46	S7Q47	S7Q48												
	.822**	.854**	.790**	.823**	.879**	.839**	.716**	.789**												
	**. Correlation is significant at the 0.01 level (2-tailed).																			
	*. Correlation is significant at the 0.05 level (2-tailed).																			

Source: Authors' Creation

Table 7: Cronbach Alpha for the 13 Sections

	Domain Knowledge .	Data Collection	Data Wrangling..	Feature Selection	Model design	Model evaluation	Communicate & propose ..
Items	7	5	7	9	7	5	8
Cronbach's α	.821	.639	.701	.741	.787	.707	.791
	Researching & ..	Analysis & Calc ..	Optimization Skill	Technical & Computing	Data Management..	Business & ..	All
Items	11	13	10	19	18	14	48
Cronbach's α	.617	.656	.721	.671	.787	.707	.915

Source: Authors' Creation

RQ 5 - Which skills and steps do students feel less confident about, as identified through the survey? (Instrument analysis)

Four of the 130 participants did not complete the survey and were excluded. Table 8 results were scrutinized based on gender (male, female), major (computer science, statistics, business, math, non-STEM), and the 13 identified skills/steps (see Table 1). Significant findings corresponding to associated p-values were highlighted. The effect size, denoted by eta-squared (η^2 = SS_effect / SS_total), was classified as small, moderate, or large. Notably, bold font indicated a large effect ($\eta^2 = .14$), underlined results indicated a moderate effect ($\eta^2 = .06$), and no markings denoted a small effect ($\eta^2 = .01$). The abbreviation "M" represented the mean, and "SD" represented the standard deviation. The analysis revealed a significant difference in scores ($F(4,152) = .549$, $p = .00$, partial-eta-squared = .086). All main interactions reached statistical significance at the .05 level—except for the data planning, feature selection, and model evaluation scores. The effect size was small for data planning and feature selection and moderate for domain knowledge, data cleaning, model design, and communication. Confidence levels exhibited similar means for data planning (M = 3.5, SD = .9) and data cleaning (M = 3.5, SD = .8), followed by a lower but comparable trend between domain knowledge (M = 3.4, SD = .8) and communication (M = 3.4, SD = 1).

Group interactions did not show any significant differences. Descriptive analysis of group interactions revealed that the highest domain knowledge scores were observed near male statistics majors and female business majors (M = 3.4). The lowest was found among non-STEM females (M = 2.7, SD = .0). For data planning, the highest scores were attributed to male computer science majors and female statistics majors (M = 3.8). The lowest scores were observed among non-STEM females (M = 2.3, SD = .0). Regarding data cleaning, male business majors scored the highest (M = 3.08, SD = .4), while the lowest scores were among non-STEM females (M = 2.9, SD = .0). Female statistics groups attained the highest scores in feature selection (M = 3.4, SD = .7). In model design, statistics majors consistently achieved the highest scores, followed by computer science and business majors, with similar scores, and then math, and finally non-STEM. Female statistics students displayed almost the highest confidence levels compared to males across all skills and steps. Notably, computer science was intermediate, with business majors scoring higher than females in the same major. Female math and non-STEM students displayed the lowest scores in all areas. Research skills were most confidently identified with math (73%) and least with math again (61%), along with non-STEM. As expected from non-STEM students, analysis skills were highest among statistics and business majors and lowest among math students. Research skills were most confidently identified with math (73%) and least with math again (61%), along with non-STEM. As expected from non-STEM students, analysis skills were highest among statistics and business majors and lowest among math students. Lastly, business and statistics majors achieved the highest scores for business knowledge skills, with a confidence level of 72%, while computer science scored the lowest at 67%. The results indicate that 11.56% identified themselves with.

Table 8: Mean of Participants Confidence level Over the 13 Sections

		Domain Knowledge & Research design		Data Planning & Data collection		Data cleaning, wrangling & Feature Engineering		Feature Selection		Model design		Model evaluation		Communicate & propose action		Researching & Planning Skill		Analysis & Calculation Skill		Optimization Skill		Technical & Computing Skill		Data Management & handling Skill		Business & Communication Skill	
Gender		M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Computer Science	M	3.1	3.0	3.8	3.4	3.6	3.4	3.3	3.2	3.2	3.0	3.1	3.0	3.4	3.3	68.0%	68.5%	62.0%	63%	70%	65%	72%	63%	74%	67%	67%	66%
	SD	0.7	0.9	0.9	1.0	0.7	0.9	0.7	0.9	0.8	0.8	0.9	1.1	0.9	1.0	9.6	11.0	11.0	12.5	8.1	9.8	15.5	17.3	14.5	17.2	11.5	13.6
Statistics	M	3.4	3.2	3.4	3.8	3.4	3.7	3.0	3.4	3.0	3.3	3.1	3.3	3.4	3.6	68.2%	75.3%	61.1%	68%	63%	68%	61%	69%	67%	73%	65%	72%
	SD	0.9	0.6	0.9	0.4	0.9	0.6	0.8	0.7	0.7	0.8	1.0	0.9	1.1	0.8	12.0	7.4	11.0	9.4	9.5	7.9	15.4	13.1	17.0	11.3	13.2	8.9
Buisness	M	3.2	3.4	3.6	3.4	3.8	3.6	3.3	3.2	2.9	3.0	3.2	3.2	3.7	3.4	73.1%	68.9%	67.4%	66%	66%	64%	67%	65%	73%	69%	72%	66%
	SD	0.4	0.7	0.5	0.8	0.4	0.7	0.5	0.8	0.6	0.9	0.8	0.9	0.6	1.0	6.6	9.8	8.1	10.7	6.1	7.7	11.2	15.2	8.7	14.0	7.8	12.7
Math	M	3.3	2.9	3.5	3.1	3.5	3.0	3.3	2.8	3.4	2.7	3.3	2.8	3.4	3.1	73.5%	61.4%	65.2%	55%	68%	59%	65%	56%	71%	61%	68%	59%
	SD	1.0	0.9	1.0	1.1	0.9	1.0	0.9	0.8	0.8	0.7	1.1	1.0	1.2	1.1	12.2	12.3	12.2	11.2	10.3	9.5	17.0	16.0	18.2	18.5	14.7	13.3
NoneSTEM	M	3.1	2.7	3.4	2.3	3.3	2.9	3.2	2.5	2.9	2.0	3.0	2.0	3.1	2.5	61.1%	63.6%	63.7%	51%	62%	48%	63%	51%	64%	59%	62%	56%
	SD	0.9	0.0	0.8	0.0	0.8	0.0	0.9	0.0	1.3	0.0	1.2	0.0	1.1	0.0	9.1	0.0	14.1	0.0	11.0	0.0	20.7	0.0	15.9	0.0	14.9	0.0
Total	M	3.4	3.3	3.5	3.5	3.6	3.4	3.3	3.2	3.2	3.0	3.3	3.1	3.4	3.3	69%	68%	64%	61%	66%	61%	66%	61%	70%	66%	67%	64%
	SD	0.8	0.9	0.9	1.0	0.8	0.9	0.8	0.8	0.9	0.9	0.9	1.0	0.9	1.0	9.89	8.10	11.28	8.75	8.99	6.98	16.0	12.3	14.9	12.2	12.4	9.69
	M	3.4		3.5		3.5		3.2		3.1		3.2		3.4													
	SD	0.8		0.9		0.8		0.8		0.9		1.0		1.0													

Source: Authors' Creation

Figure 3 illustrates that 11.56% of cases fall within the low confidence range; moderate confidence accounts for 11.54%, and high confidence is 76.92%. Lower confidence levels were observed, particularly in model design, followed by feature selection and model evaluation, which can be attributed to deficiencies in analysis and calculation, optimization, and technical and computing skills. Conversely, higher confidence levels were associated with research design, data management, and data cleaning, possibly indicating stronger proficiency in these areas.

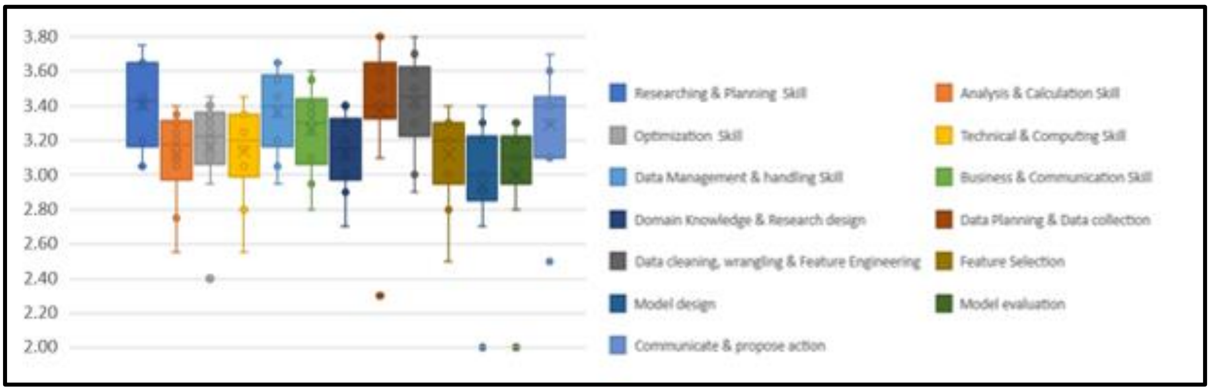

Figure 3: Students' Confidence Level in Using Data Science Skills for Building Data-driven Solutions

Source: Authors' Creation

Note style: Note. M = mean; SD = standard deviation

5.2. A Suggested Intermediate Plan to Support Confidence in Data Science Education

An intermediate plan was derived from the background section to bolster confidence in using data science skills across various proficiency levels. Following the application of survey data, educators in data science can pinpoint specific skills or steps in the data science life cycle that require particular attention during instruction. Educators can select activities tailored to their classes upon identifying the skills/knowledge and the corresponding confidence levels.

Low confidence: (1) Individualized Support Sessions: Schedule one-on-one meetings with participants to discuss their concerns and address any questions or misunderstandings they may have confidently. Offer personalized guidance and encouragement to help boost their confidence. **(2)** Additional Learning Resources: Provide supplementary materials—articles, videos, or tutorials—to reinforce key concepts and provide alternative explanations. Recommend books or online courses that align with participants' learning needs and preferences. **(3)** Peer Support Groups: Facilitate peer support groups where participants can collaborate, share experiences, and provide encouragement to one another. Encourage group members to discuss challenges openly and offer constructive feedback and support. **(4)** Regular Check-Ins: Conduct regular check-ins with participants to monitor progress, address new concerns, and provide ongoing support and encouragement. Use these opportunities to celebrate small victories and acknowledge participants' efforts and improvements.

Moderate confidence: (1) Clarification Sessions: Organize group or question-and-answer sessions where participants can ask questions, seek clarification, and discuss areas of uncertainty. Provide clear explanations and examples to reinforce understanding and address common misconceptions. **(2)** Practice Opportunities: Offer practice exercises, quizzes, or problem-solving tasks to allow participants to apply their knowledge and skills in a supportive environment. Please provide feedback and guidance to help participants identify areas for improvement and build confidence in their abilities. **(3)** Mentorship Program: Pair participants with mentors or more experienced peers who can offer guidance, advice, and encouragement. Encourage mentors to provide personalized support and share their own experiences and strategies for success. **(4)** Self-Reflection Activities: Encourage participants to reflect on their learning journey, identify strengths and growth areas, and set achievable goals for themselves. Provide prompts or reflection questions to guide their self-assessment and encourage deeper engagement with the material.

High confidence: (1) Advanced Learning Opportunities: Offer advanced workshops, seminars, or projects for participants confident in their abilities and eager to challenge themselves further. Provide opportunities for independent research, creative projects, or leadership roles within the learning community. **(2)** Peer Teaching Sessions: Encourage confident participants to share their knowledge and expertise with their peers through peer teaching sessions or mini-workshops. Facilitate opportunities for participants to develop their presentation and communication skills while helping others learn. **(3)** Professional Development Resources: Provide access to professional development resources such as webinars, conferences, or networking events to help participants further their skills and expertise. Offer guidance on career pathways, industry trends, and continued growth and advancement opportunities. **(4)** Recognition and Rewards: Acknowledge and celebrate participants' achievements and contributions within the learning community. Offer certificates of achievement, badges, or other forms of recognition to acknowledge their dedication and accomplishments.

6. Conclusion

The field of data science is experiencing rapid global growth, yet there is a notable shortage of qualified data scientists, posing concerns for academia and industries alike. Moreover, the lack of research in data science education assessments leaves uncertainties about students' skills before graduation. This paper addresses these gaps by developing a data science self-efficacy survey to gauge individuals' confidence levels in applying data science skills and proposing activities to boost confidence based on their levels. The survey—developed with input from computer science, business, and statistics experts—evaluates 13 items representing data science life cycle steps and related interdisciplinary skills. Distributed to students and researchers across six educational institutions, pilot results indicated high reliability and stability. Analysis revealed varying confidence levels among participants, with the majority exhibiting moderate confidence. Remedial suggestions include individualized support sessions and peer support groups for those with low confidence. High-confidence individuals are encouraged to pursue advanced learning opportunities and share their expertise with peers.

7. Future Work

The survey will compare results across a broader sample from various continents, enabling a more comprehensive understanding of trends and variations in data science proficiency across diverse geographical regions. Further investigation will be conducted regarding the threshold scale.

Acknowledgment: The opinions expressed in this work are solely those of the authors and do not necessarily reflect the views of the NSF. We sincerely thank King AbdulAziz University and Mariam Alsalmi for their invaluable support in conducting this research.

Author Contributions: All authors contributed to this research.

Funding: The National Science Foundation (NSF) funded this work under Grant No. DRL GEGI008182.

Conflict of Interest: The authors declare no conflict of interest.

Informed Consent Statement/Ethics Approval: Not applicable.

Acknowledgment: The opinions expressed in this work are solely those of the authors and do not necessarily reflect the views of the NSF. We sincerely thank King AbdulAziz University and Mariam Alsalmi for their invaluable support in conducting this research.

References

Davenport, T. H., & Patil, D. (2022). Is data scientist still the sexiest job of the 21st century. *Harvard Business Review*, *15*.

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology*, *1*(20), 416-436.

Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. *Harvard Data Science Review*, *2*(2).

Gupta, S. (2022). *Data Science Process: A Beginner's Guide in Plain English*. https://www.springboard.com/blog/data-science/data-science-process/

Haben, S., & Hinton, S. DATA SCIENCE: FROM ACADEMIA TO INDUSTRY.

Hornstra, L., Stroet, K., Rubie-Davies, C., & Flint, A. (2023). Teacher expectations and self-determination theory: Considering convergence and divergence of theories. *Educational Psychology Review*, *35*(3), 76.

LEE, H., MOJICA, G., THRASHER, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal*, *21*(2), 3-3.

Li, G., Yuan, C., Kamarthi, S., Moghaddam, M., & Jin, X. (2021). Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis. *Journal of Manufacturing Systems*, *60*, 692-706.

Locke, E. A. (1987). Social foundations of thought and action: A social-cognitive view. In: Academy of Management Briarcliff Manor, NY 10510.

Malallah, S., Weese, J., & Alsalmi, K. (2023). *The "besTech" Technology Practice Framework for Early Childhood Education* ASEE Annual Conference,

Malallah, S., Weese, J., Shamer, L., & Hsu, W. (2023). *Data Science (Dataying) for Early Childhood* ASEE Annual Conference,

Saltz, J. S., & Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. 2020 IEEE International Conference on Big Data (Big Data),

Stodden, V. (2020). The data science life cycle: a disciplined approach to advancing data science as a science. *Communications of the ACM*, *63*(7), 58-66.

Training, D. o. E. a. (2017). High impact teaching strategies: Excellence in teaching and learning. In: Department of Education and Training, State of Victoria Melbourne.

Vinay, S. (2024). Data Scientist Competencies and Skill Assessment: A Comprehensive Framework. *Journal ID*, *1660*, 1544.

Works, M. (2017). Decades of scientific research that started a growth mindset revolution. *Retrieved from the World Wide Web https://www. mindsetworks. com/science/on October*, *7*, 2018.