**Developing an Instrument for Assessing Self-Efficacy Confidence in Data Science**

Safia Malallah, Kansas State University, safia@ksu.edu

Ejiro Osiobe, Baker University's, Jiji.osiobe@bakerU.edu

Zahraa Marafie, Kuwait University, Zahraa.Marafie@ku.edu.kw

Patricia Coronel, ULEAM, patricia.henriquez@uleam.edu.ec

Lior Shamir, Kansas State University, lshamir@ksu.edu

Ella Carlson, Kansas State University, ellacarlson23@ksu.edu

Joshua Levi Weese, Kansas State University, weeser@ksu.edu

**Abstract**

The field of data science education research faces a notable gap in assessment methodologies, leading to uncertainty and unexplored avenues for enhancing learning experiences. Effective 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 aimed to empower educators by identifying areas where students lack confidence, enabling the design of targeted plans to bolster data science education. Collaboration among experts from the fields of computer science, business, and statistics 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.

# 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 both academia and industries [1, 2]. Additionally, research on data science education assessments lacks, 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, with the goal 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 construction of customized assessments for data science aligned with organizational needs [3]. This was carried out by a collaboration among experts from computer science, business, and statistics, 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 strongly agree. The survey was distributed to students and researchers in six educational institutions in KSA, the United States of America (USA), and Kuwait. 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 confidence 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.

# Background

*Confidence and Learning*

Confidence plays a pivotal role in students’ academic success and overall well-being. Social Cognitive Theory suggested that self-efficacy, or one’s 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 focus on building self-efficacy through incremental successes, constructive feedback, and role modeling [4]. 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 [5]. 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 [6]. High-confidence students typically demonstrate a strong belief in their abilities and may seek out challenges or leadership roles. However, excessive confidence without corresponding competence can lead to overestimation of skills and performance [7]. 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 develop a more accurate understanding of their strengths and areas for improvement [8].

*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 identification of key competencies; categorization and prioritization; definition of competency levels; development of assessment tools; scoring and evaluation rubrics; integration with organizational goals; feedback mechanisms; implementation and training; and iterative refinement. We incorporated steps first five steps to develop our survey, as they were relevant to our goal of creating an assessment process for academia [3].

# Method

## *Design*

This study employed a quantitative approach to develop a self-efficacy survey aimed at assessing 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 [3]. 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 8 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.

*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% freshmen). 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 courses in machine learning/artificial intelligence (AI). 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).

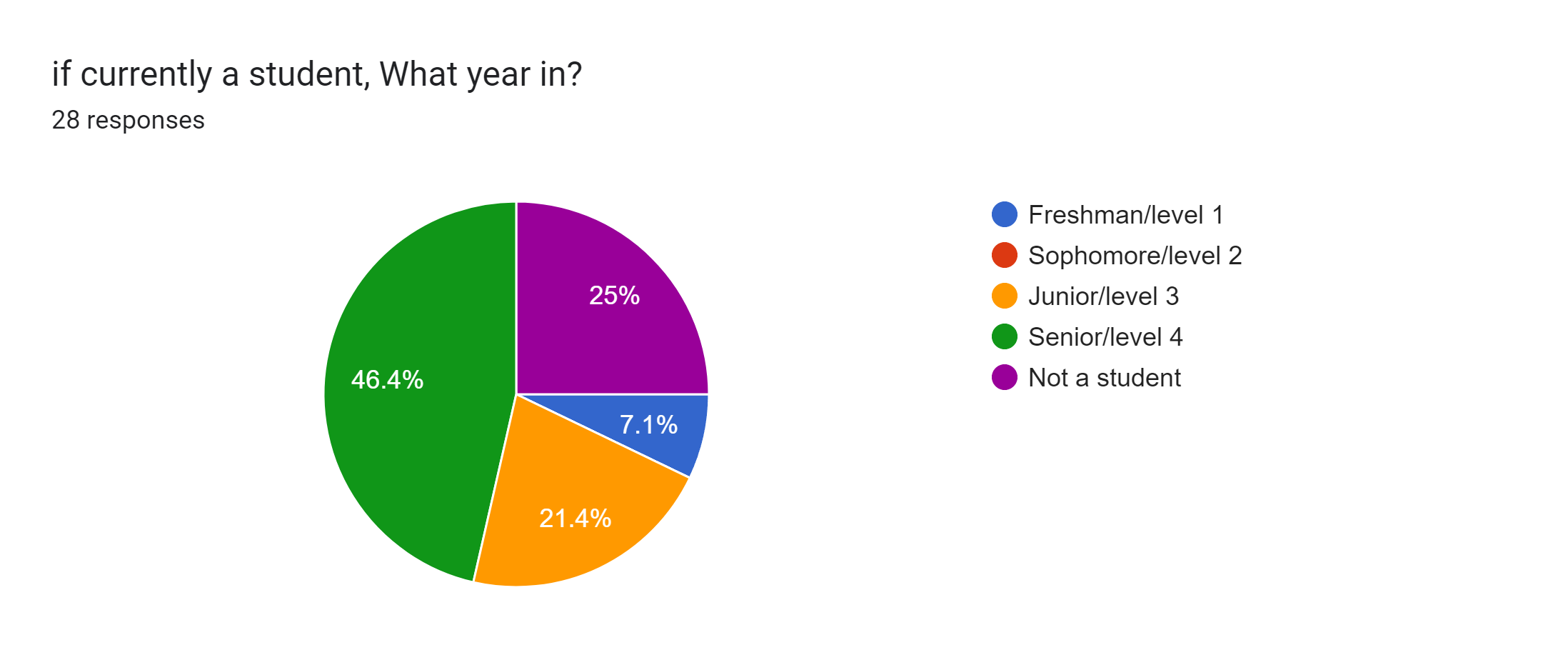
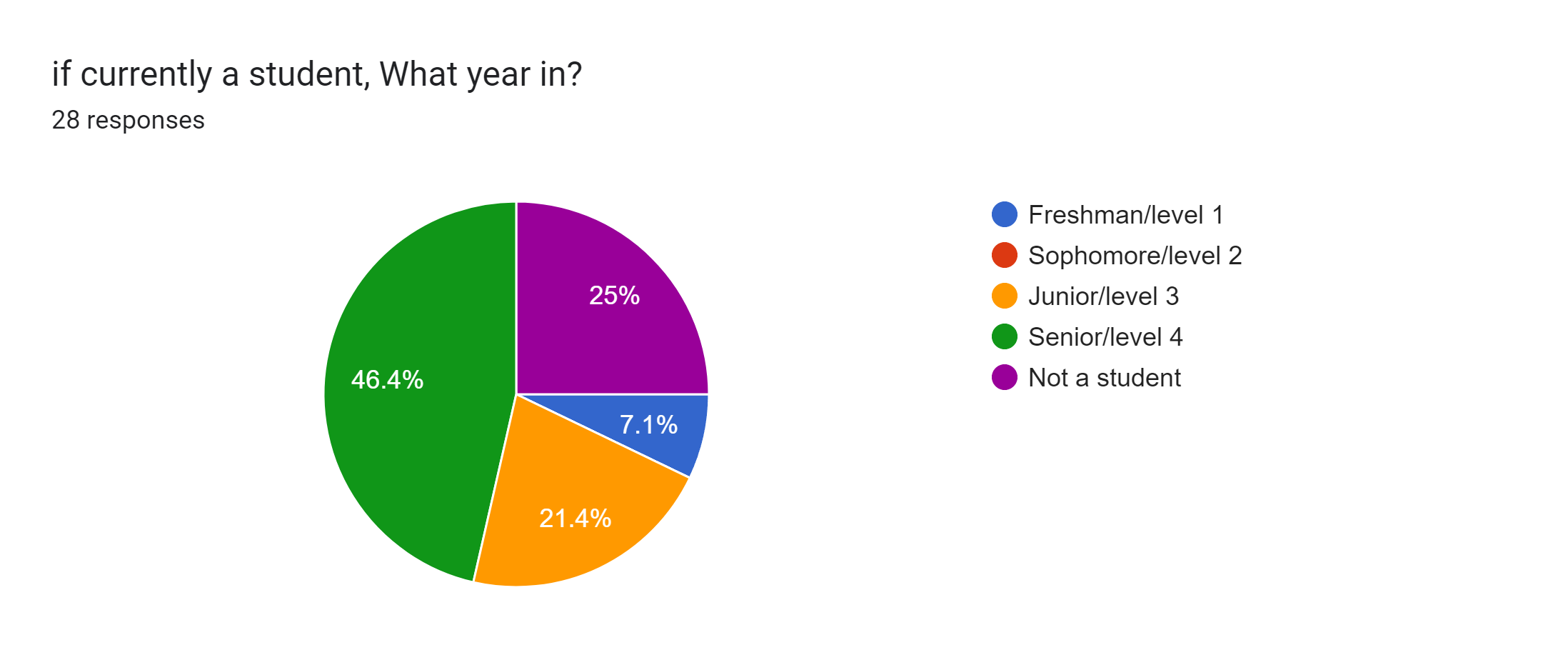
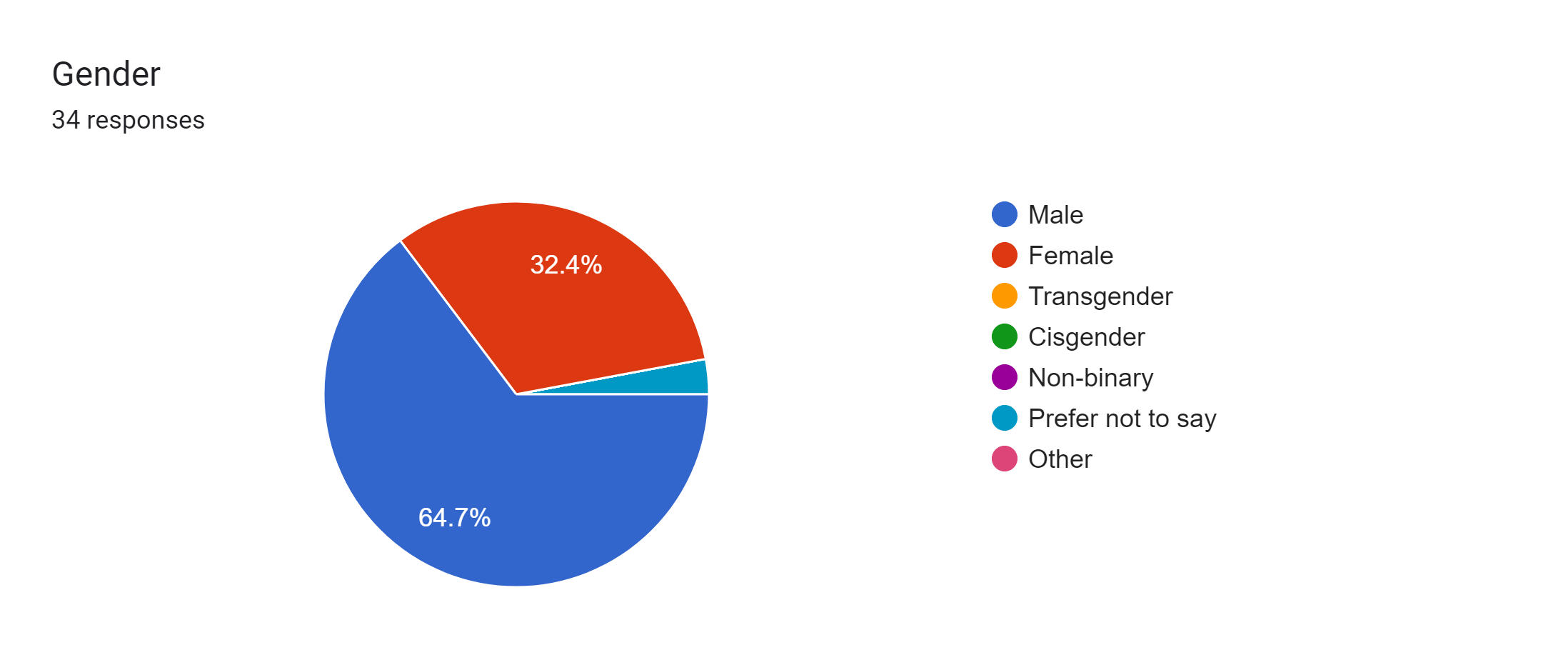
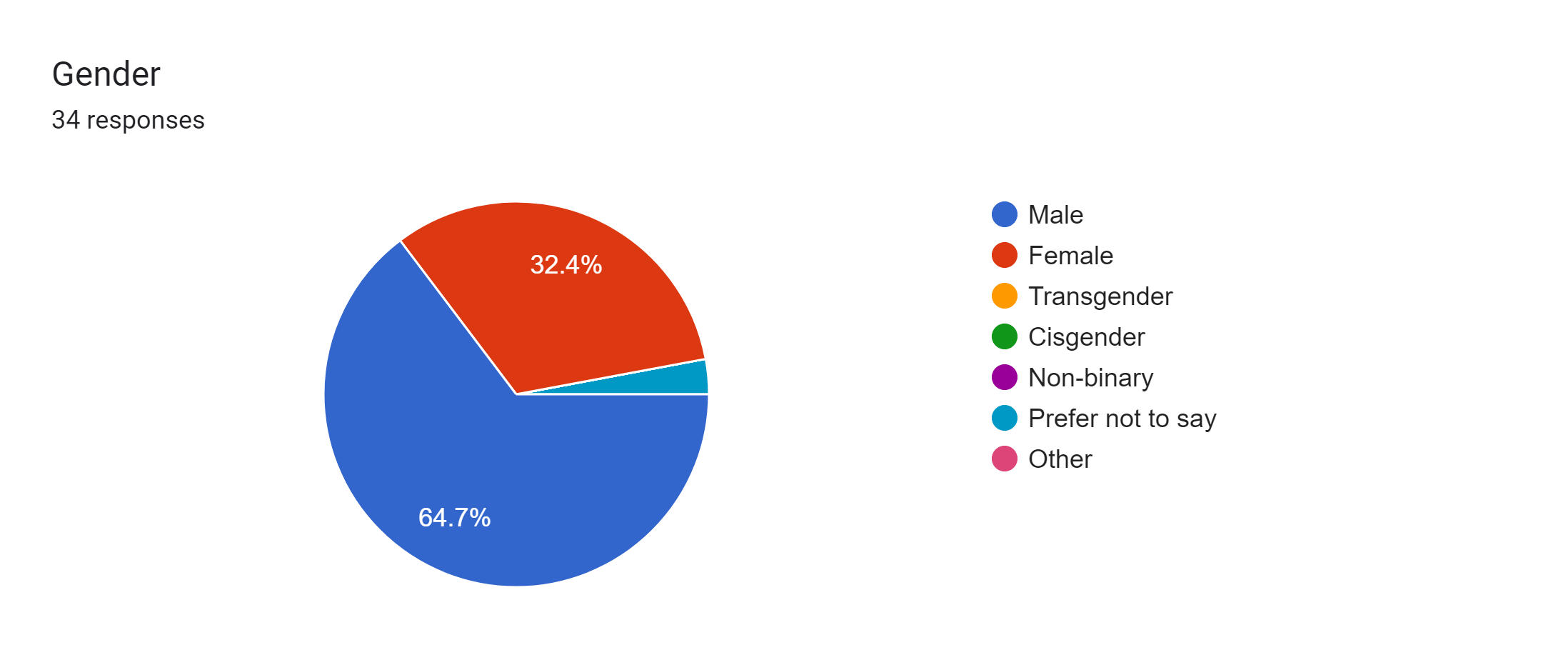
 

Figure . Sample Population

## *Keywords, Database, and Criteria*

The literature reviews were conducted using specific keywords tailored to each investigation area. The first literature review searched the keywords “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—including conference papers, journals, and blogs—were considered. The results were meticulously filtered by isolating abstracts and titles that aligned with the search criteria. Studies that did not primarily focus on data science were excluded from analysis. The search was further refined to only include results from 2020 to 2024, except in cases concerning the data science life cycle. Furthermore, research pertaining to specific medical fields (e.g., medicine, dentistry, nursing, health professions, neuroscience, pharmacology, toxicology, pharmaceutical science, cancer, effect, and psychological studies) were excluded.

*Instruments*

The survey was carefully developed based on thorough analyses from literature reviews (see the Results section). Table 1 presents the final findings of the investigation, 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 strongly agree.

Table . 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 Skill | 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 | The proficiency in translating and aligning business strategies into actionable technical findings, effectively communicating them to stakeholders—both ways |  | Q4, Q16-Q17, Q20, Q23, Q28-Q29, Q34, Q43-Q48 |

*Instruments Rubric*

The instruments 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 . Data Science Self-Efficacy Survey



**Results**

This study analyzed students’ confidence level 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 analysis through repeated measures (analysis of variance [ANOVA] and descriptive statistics) using Statistical Package for Social Science (SPSS) software and Excel.

**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’s ability. The first 39 were taken from Vinay’s work [3], the next 50 items from Usama Fayyad’s and Hamit Hamock’s work [9], and the remaining from Guoyan’s work [10]. The list was clustered and filtered to generate the final list that has eight categories presented in Table 1, skills 8–13.

Table . The Identified Items from the Literature Reviews



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 aimed at assessing and categorizing 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 directly use all his competencies, we cross-referenced them with other resources in the next steps [3].

Fayyad and Hamock (2020) introduced a comprehensive Data Science Knowledge Framework to foster industry standardization and the creation of measurement and assessment methodologies. Emphasizing the dynamic and multidisciplinary nature of data science, the authors constructed the framework through extensive literature review, identifying pivotal topics and technologies crucial for professionals in analytics and data science. The findings were systematically organized into a hierarchical knowledge structure [9].

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 [10].

The process of clustering 136 items was extensive. The list contained various categories, making it difficult 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 together 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 validating them with the experts.

*RQ2- What are the key 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. To address this, four models were identified and compared for common factors, 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 [11]. 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 [12]. 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 [13]. 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 [14].

All models began with problem understanding, progressed through data acquisition and comprehension, and concluded with communication, either as a standalone step or integrated within 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 . Identified Data Science Life Cycles 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 |

A diagram of a model

Description automatically generated with medium confidence

Figure . Identified Data Science Life Cycles Models

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

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

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| DB Cycle\Skills | Researching & Planning | Analysis & Calculation | Optimization Skill | Technical & Computing Skill | 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 |

*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 level of internal consistency and validity of the questionnaire.

Table . Person Correlations of all the Questions



Table . Cronbach Alpha for the 13 Sections



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

Of the 130 participants, four 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 were highlighted, corresponding to associated p-values. The effect size, denoted by eta-squared (η² = SS\_effect / SS\_total) was classified as small, moderate, or large. Notably, bold font indicated a large effect (η² = .14), underlined results indicated a moderate effect (η² = .06), and no markings denoted a small effect (η² = .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 were 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. Analysis skills were highest among statistics and business majors and lowest among math students, as expected from non-STEM students. Research skills were most confidently identified with math (73%) and least with math again (61%), along with non-STEM. Analysis skills were highest among statistics and business majors and lowest among math students, as expected from non-STEM students. Lastly, for business knowledge skills, business and statistics majors achieved the highest scores with a confidence level of 72%, while computer science scored the lowest at 67%. The results indicate that 11.56% identified themselves with

Table . Mean of Participants Confidence level Over the 13 Sections



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

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 skills; optimization skills; 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.

A graph of data on a white background

Description automatically generated with medium confidenceA graph of data on a white background

Description automatically generated with medium confidenceA graph of data on a white background

Description automatically generated with medium confidence

Figure . Students’ Confidence Level in Using Data Science Skills for Building Data-driven Solutions

**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. Upon identifying the skills/knowledge and the corresponding confidence levels, educators can select activities tailored to their classes.

**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 sessions 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 give participants opportunities to apply their knowledge and skills in a supportive environment. 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 who are 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 opportunities for continued growth and advancement. **(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.

**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 experts in computer science, business, and statistics—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.

# 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.

**Future Work**

The survey will be used to 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.

# Acknowledgement

The National Science Foundation (NSF) provided funding for this work under Grant No. DRL GEGI008182. The opinions expressed in this work are solely those of the authors and do not necessarily reflect the views of the NSF. We extend our sincere gratitude to King AbdulAziz University, and Mariam Alsalmi for their invaluable support in conducting this research.

**References**

[1] T. H. Davenport and D. Patil, "Is data scientist still the sexiest job of the 21st century," *Harvard Business Review,* vol. 15, 2022.

[2] S. Haben and S. Hinton, "DATA SCIENCE: FROM ACADEMIA TO INDUSTRY."

[3] S. Vinay, "Data Scientist Competencies and Skill Assessment: A Comprehensive Framework," *Journal ID,* vol. 1660, p. 1544, 2024.

[4] E. A. Locke, "Social foundations of thought and action: A social-cognitive view," ed: Academy of Management Briarcliff Manor, NY 10510, 1987.

[5] M. Works, "Decades of scientific research that started a growth mindset revolution," *Retrieved from the World Wide Web. mindsetworks. com/science/on October,* vol. 7, p. 2018, 2017.

[6] E. L. Deci and R. M. Ryan, "Self-determination theory," *Handbook of theories of social psychology,* vol. 1, no. 20, pp. 416-436, 2012.

[7] L. Hornstra, K. Stroet, C. Rubie-Davies, and A. Flint, "Teacher expectations and self-determination theory: Considering convergence and divergence of theories," *Educational Psychology Review,* vol. 35, no. 3, p. 76, 2023.

[8] D. o. E. a. Training, "High impact teaching strategies: Excellence in teaching and learning," ed: Department of Education and Training, State of Victoria Melbourne, 2017.

[9] U. Fayyad and H. Hamutcu, "Toward foundations for data science and analytics: A knowledge framework for professional standards," *Harvard Data Science Review,* vol. 2, no. 2, 2020.

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

[11] V. Stodden, "The data science life cycle: a disciplined approach to advancing data science as a science," *Communications of the ACM,* vol. 63, no. 7, pp. 58-66, 2020.

[12] H. LEE, G. MOJICA, E. THRASHER, and P. Baumgartner, "Investigating data like a data scientist: Key practices and processes," *Statistics Education Research Journal,* vol. 21, no. 2, pp. 3-3, 2022.

[13] J. S. Saltz and N. Hotz, "Identifying the most common frameworks data science teams use to structure and coordinate their projects," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020: IEEE, pp. 2038-2042.

[14] S. Gupta. "Data Science Process: A Beginner’s Guide in Plain English." https://www.springboard.com/blog/data-science/data-science-process/ (accessed 2024).